FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

# Discovery of food exposure markers in urine and evaluation of dietary compliance by untargeted LC-MS metabolomics

PhD thesis 2014

MAJ-BRITT SCHMIDT ANDERSEN

# Discovery of food exposure markers in urine and evaluation of dietary compliance by untargeted LC-MS metabolomics

PhD thesis 2014

Maj-Britt Schmidt Andersen

**Title:**
Discovery of food exposure markers in urine and evaluation of dietary compliance by untargeted LC-MS metabolomics

**Author:**
Maj-Britt Schmidt Andersen
Department of Nutrition, Exercise and Sports
Faculty of Science, University of Copenhagen, Denmark

**Date of submisssion:**
31/10-2013

**Academic advisors:**
Professor Lars Ove Dragsted
Department of Nutrition, Exercise and Sports
Faculty of Science, University of Copenhagen, Denmark

Associate professor Åsmund Rinnan
Department of Food Science
Faculty of Science, University of Copenhagen, Denmark

**Assessment committee:**
Associate professor Cristina Andrés-Lacueva
Department of Nutrition and Food Science
Faculty of Farmacy, University of Barcelona, Spain

Professor Knud Erik Bach Knudsen
Department of Animal Science -Molecular nutrition and cell biology
Aarhus University, Denmark

Associate professor Francesco Savorani
Department of Food Science -Quality and Technology
Faculty of Science, University of Copenhagen, Denmark



ISBN 978-87-7611-686-6

Printed by SL grafik, Frederiksberg C, Denmark (www.slgrafik.dk)

# Preface

The work presented in this thesis has been conducted at the Department of Nutrition Exercise and Sports, Faculty of Science, University of Copenhagen from November 2010 to October 2013 with a 2.5 months external stay at the Human Nutrition Unit, INRA, Clermont-Ferrand-Theix in France in the Autumn of 2012. The analyzed data is from a meal study and an intervention study that were both carried out as part of the large research project OPUS which is aiming to develop a healthy palatable New Nordic Diet and investigate its health effects. OPUS is a Danish acronym for 'optimal well-being, development and health for Danish children through a healthy New Nordic Diet' and is supported by a grant from the Nordea Foundation. In addition to the OPUS project, my PhD study was supported by a scholarship from the University of Copenhagen.

Maj-Britt Schmidt Andersen, Frederiksberg, October 2013

# Acknowledgements

When I began my PhD studies three years ago, I was very excited to work on what I thought (and still think) would be an important new methodology for studying nutrition. I knew the concept of untargeted metabolomics of course and its wide perspectives but at the same time, I did not know the details. I knew very little on chemometrics and mass spectrometry. I had never thought about the need for preprocessing in data generation, the need for structure elucidation to understand the findings, and basic programming skills to handle large datasets. Actually, I had never even heard about the program Matlab which turned out to be one of the most central tools in my work, including a large number of small celebrations, when it worked, and a larger number of obscure lines of red text and annoyance, when it did not. Three years on, I have become so much wiser! I guess this is exactly how a PhD process is supposed to work -a continuous facing of new challenges, new ideas, walls, ups and downs. I would like to thank my supervisor Lars Ove Dragsted for inviting me into this and for his support when it was most needed. It is interesting to have been supervised by someone who is a curious researcher by heart and whose optimism and generation of ideas are without limits. Many thanks also to my co-supervisor Åsmund Rinnan, who was willing to make a special chemometrics course for me, so I could get into this new world of chemometrics as fast as possible. Åsmund has been excellent in his supervision of the multivariate modeling.

I am sure that I would have gotten stuck very early into my PhD project, if it was not for all the people who have helped me to move on, work on, understand and cheer up. First of all, I have had the great pleasure of sharing office with two great personalities, Gözde and Daniela. I have learned a whole lot about metabolomics from them and they have always been open for questions which in most cases turned into very good discussions. There is always a reason to smile and get a good laugh with G&D around and that is invaluable. Another hard core metabolomics person, who I have received great help from, is Jan. Especially, when it comes to everything related in some way or another to computers.

For the two main datasets I have worked on, I have been dependent on others for information on the study details, as I have not been involved in designing and conducting the studies. Helene has helped me out with the meal study, always with a fast answer and interest in the results. Sanne has been the expert on the intervention study. I do not know how many times I have asked her for data and about data. Despite this, she has always helped me with whatever question I was interested in and with a very positive attitude. It has been a great joy to work with her.

For marker identifications, I had an instructive external stay in France and I would like to thank Blandine Comte for sharing her lunch time with me to discuss the possibilities of this stay at a conference and Claudine Manach for taking good care of me, while I was at INRA.

Finally, I would like to thank Hanne and Ümmühan for their help on the laboratory work, Jeanette, Julie and Anna who have been occasional but very good office mates (also outside the office), Suzanne, for always helping me with administrative issues, as well as the OPUS secretariat with Eva and Anne-Marie for nice “indianergruppe møder” and good support in everything related to the

OPUS project. I have had many other nice and helpful colleagues, not least in the kitchen and the corridor.

Outside work, I am very grateful for the support I have received from my family and friends. I have not always been very easy to deal with during the past three years and I really appreciate that you have shown so much understanding. A final and very special thank you goes to Teunis for helping me back on track whenever needed and for his outstanding ability to decomplicate things.

# Summary

Accurate measurement of dietary intake in nutrition studies is crucial to investigate relationships between diet and health. Dietary effects on disease risk are often small and may in some cases be distorted due to errors in dietary assessment or lack of compliance in dietary studies. The common tools used for assessment of dietary exposure in humans rely almost solely on self-reporting, which is associated with a range of random and systematic errors such as under- and overreporting of certain foods and inter-individual differences in reporting behavior. Biomarkers measured in biological samples, most often urine or plasma, provide a promising supplement to self-reporting, as they are objective measures. However, the few currently available biomarkers cover the diet poorly and more markers, in particular for intake of individual foods, are needed. A relatively new approach to discover dietary exposure markers is by untargeted metabolomics, also known as metabolomic fingerprinting. With this method, a large number of compounds are measured in a biological sample and the resulting so-called 'fingerprints' are explored to discover patterns of metabolites related to certain dietary exposures.

In this thesis, untargeted metabolomics has been applied to: 1) discover new urinary exposure markers of individual foods (Paper I and II) and 2) develop a model based on measurements of urine samples to estimate compliance to two dietary patterns (Paper III). Based on the results in Paper I-III, it is discussed how metabolomic fingerprinting can contribute to the development of new exposure and compliance measures for use in future nutrition studies.

Data from three studies have been analyzed. The first study is a controlled cross-over meal study with three meals, each prepared with three different protein sources. This study has been used to find markers for the individual meals and protein sources. The second study is a meal study with single foods, which was used to confirm the food sources of the markers found in the first study. The last study is a parallel intervention study with two dietary patterns; A New Nordic Diet (NND) and an Average Danish Diet (ADD). Three analyses were conducted on this study. One to discover exposure markers of individual foods, one to validate the markers found in the first study, and one to develop a compliance model to estimate compliance to ADD and NND based on 24 h urine samples.

It was possible to find exposure markers for several individual foods and food groups such as cruciferous vegetables, citrus fruits, fish, walnuts and chocolate. Some markers from the meal study were also markers in the intervention study. Other markers were related to the meal matrix, or could not be validated in the other study because the foods had not been reported often enough. When validating the identified markers according to a range of common biological validation criteria, using other literature, several of the markers found were very promising exposure marker candidates.

The two dietary patterns ADD and NND were clearly reflected in urine samples and a multivariate model, including 52 metabolites, proved to be a potentially suitable compliance measure for the diets. The model was able to classify 81 percent of 139 validation samples to the correct dietary pattern. The metabolites in the model were from several characteristic foods in ADD, while the

NND diet was characterized by more general dietary traits, such as a high fruit and vegetable intake. An analysis of other compliance measures, in which subjects with misclassified and correctly classified samples in the model were compared, supported the findings in the model.

Overall, the results in this thesis substantiate metabolomic fingerprinting as a promising tool to develop new dietary compliance measures and exposure markers. However, for all analyses performed, a large proportion of the markers found could not be identified which is a prerequisite to better understand how dietary compliance is represented in the multivariate model and to validate and further investigate new exposure markers. For the future, the identification of unknown markers needs to be improved to gain more knowledge from studies applying metabolomic fingerprinting. In addition, identified as well as unidentified marker candidates from previous studies should be taken into account in the analyses of new studies. Finally, quantitative analyses of the best marker candidates should be conducted to understand the full value of the markers found.

# Resume

Præcis måling af kostindtag i ernæringsforskning er essentielt for at kunne bestemme sammenhænge mellem kost og sundhed. Kostens effekt på risikoen for at udvikle forskellige sygdomme er ofte lille og kan i nogle tilfælde blive overset som følge af uprækis bestemmelse af kostindtag eller mangel på komplians (en persons efterlevelse af retningslinjerne i et studie) i ernæringsstudier. De gængse metoder der anvendes til at bestemme kostindtag afhænger næsten udelukkende af selvrapportering og er associeret med en række tilfældige og systematiske fejl såsom under- og overrapportering af bestemte fødevarer og inter-individuelle forskelle i måden at rapportere på. Biomarkører der måles i biologiske prøver, oftest i urin eller plasma, udgør et lovende supplement til selvrapportering, da de er objektive mål. Dog dækker de få eksisterende biomarkører ikke en fuld kost særlig godt og der er et behov for nye markører, især for indtag af individuelle fødevarer. Untargeted metabolomics, også kendt som metabolomic fingerprinting, er en relativt ny metodisk tilgang til at finde kostmarkører. Med denne metode er det muligt at måle en lang række stoffer i en biologisk prøve på samme tid og de resulterende såkaldte "fingerprints" (fingeraftryk) udforskes derefter for at bestemme hvilke mønstre af metabolitter, der er associeret med indtag af bestemte fødevarer eller kosttyper.

I denne afhandling er untargeted metabolomics blevet anvendt til at: 1) finde nye markører i urin for indtag af forskellige fødevarer (Artikel I og II) og 2) udvikle en model baseret på målinger af urinprøver til at estimere komplians til to kostmønstre (Artikel III). Baseret på resultaterne i Artikel I-III, diskuteres det, hvorvidt metabolomic fingerprinting kan bidrage til udviklingen af nye eksponerings- og komplians mål, der kan anvendes i fremtidige ernæringsstudier.

Data fra tre studier er blevet analyseret. Det første studie er et kontrolleret måltidsstudie med overkrydsningsdesign, hvor der blev indtaget tre måltider, der hver blev tilberedt med tre forskellige proteinkilder. Dette studie er blevet brugt til at finde markører for de enkelte måltider og proteinkilder. Det andet studie er et måltidsstudie med enkelte fødevarer, som blev brugt til at bekræfte hvilke fødevarer markørerne i det første studie stammede fra. Det sidste studie er et parallelt interventionsstudie med to kostmønstre: En ny nordisk hverdagsmad (NNH) og en gennemsnitlig dansk kost (GDK). Tre analyser blev udført på dette studie. En for at finde nye markører for indtag af enkelte fødevarer, en for at validere de markører, der blev fundet i det første studie og en for at udvikle en komplians model, der kan bruges til at estimere komplansen til NNH og GDK baseret på 24 timers urinprøver.

Det var muligt at finde eksponeringsmarkører for en række enkelte fødevarer og fødevaregrupper som for eksempel korsblomstrede grøntsager, citrusfrugter, fisk, valnødder og chokolade. Nogle markører, der blev fundet i måltidsstudiet, var også markører i interventionsstudiet. Andre markører var relateret til måltidet som helhed eller kunne ikke valideres, fordi fødevaren ikke var rapporteret hyppigt nok i interventionsstudiet. En validering af markørerne i forhold til en række biologiske valideringskriterier ved brug af anden litteratur viste, at flere af de fundne markører er meget lovende kandidater som eksponeringsmarkører.

De to kostmønstre NNH og GDK var tydeligt afspejlet i urinprøverne og en multivariat model, hvori der indgik 52 metabolitter, var velegnet som et muligt kompliansmål for kostmønstrene. Modellen kunne klassificere 81 procent af 139 valideringsprøver til det rigtige kostmønster. Metabolitterne i modellen var fra flere karakteristiske fødevarer i GDK, mens de repræsenterede mere generelle fødevaregrupper i NNH. For eksempel højt indtag af frugt og grønt. En analyse af andre kompliansmarkører, hvor personer med misklassificerede prøver i modellen blev sammenlignet med personer med korrekt klassificerede prøver, understøttede resultaterne i modellen.

Alt i alt underbygger resultaterne i denne afhandling metabolomic fingerprinting som et lovende værktøj til at udvikle nye eksponerings- og komplians markører for kostindtag. Dog er der i alle analyser en stor andel af de fundne markører, der ikke kunne identificeres, hvilket er en forudsætning for bedre at forstå, hvordan komplians er repræsenteret i den multivariate model og for at validere og undersøge nye eksponeringsmarkører nærmere. For fremtiden skal identifikationen af nye markører forbedres for at få mere viden ud af studier, der anvender metabolomic fingerprinting. Desuden skal viden om identificerede, såvel som ikke identificerede markører, fundet i tidligere studier inddrages i analysen af nye studier. Endelig skal der laves kvantitative analyser af de bedste markørkandidater for at forstå den fulde værdi af de fundne markører.

# List of abbreviations

| | |
|---|---|
| ADD | Average Danish diet |
| BMI | Body mass index |
| EE | Energy expenditure |
| EI | Energy intake |
| FFQ | Food frequency questionnaires |
| FOOD | Short-term single food studies (information on the studies is given in section 3.2) |
| INTER | Parallel intervention study comparing two dietary patterns (information on the study is given in section 3.3) |
| LC | Liquid chromatography |
| LV | Latent variables |
| MS | Mass spectrometry |
| MEAL | Cross-over meal study with *Brassica*-containing meals (information on the study is given in section 3.1) |
| NMR | Nuclear magnetic resonance |
| MUFA | Monounsaturated fatty acids |
| NND | New Nordic diet |
| m/z | Mass-to-charge ratio |
| OPUS | Danish acronym for 'Optimal well-being and health for Danish children through a healthy New Nordic Diet' |
| PABA | Para-aminobenzoic acid |
| PC | Principal component |
| PCA | Principal component analysis |
| PEM | Potential exposure marker |
| PLS-DA | Partial least squares discriminant analysis |
| RT | Retention time |
| SFA | Saturated fatty acids |
| TIC | Total ion current |
| UPLC | Ultra-performance liquid chromatography |
| qTOF | Quadrupole time of flight |
| VIP | Variable importance in projection |
| WDR | Weighed dietary records |

# Table of Contents

# 1. Introduction

'In an ideal world, nutrition scientists would like to control precisely what a person eats or monitor everything that a person has consumed, much the same way that we can control or monitor the diets of caged lab animals'
(Wishart, 2008)

## 1.1 Impact of exposure and compliance markers in nutrition research

It is well-known that poor nutrition increases the risk of developing lifestyle related diseases, such as cardiovascular diseases, some forms of cancer and type two diabetes (Key et al., 2002; Mann, 2002; Mente et al., 2009; Ajala et al., 2013). Nutrition research is addressing this issue, aiming to unravel the optimal nutrition for good health and disease prevention. In order to succeed, a key is to establish causal relationships between diet and disease through well-designed studies and one of the major obstacles in this regard is the need for accurate dietary assessment (Bingham, 2002; Fairweather-Tait, 2003; Penn et al., 2010). If dietary exposure is not estimated correctly, the association found between dietary intakes and markers of disease risk will be biased (Livingstone & Black, 2003). Measurement of dietary intake is most often done by self-reporting and a large number of methods have been developed to try to capture the habitual diet of individuals in this way. The most commonly applied self-reporting methods are food frequency questionnaires (FFQ), where individuals are asked about the average intake of a predefined number of food groups within a defined period of time, and methods to estimate the diet on individual days, for example from weighed dietary records (WDR) or 24 h recalls (Bingham, 2002). These methods have major drawbacks due to inherent systematic and random errors, such as misreporting, definition of appropriate food categories and errors when converting food based records to individual nutrients using information from food databases. In order to overcome some of the problems with dietary reporting, biomarkers have been introduced as a means for data validation (Jenab et al., 2009). Biomarkers are measured in biological samples, most often urine or plasma, and are therefore a more objective way to investigate dietary exposure. Two common biomarkers for validation of dietary records are estimation of total energy expenditure (EE) from double labeled water and estimation of protein intake from total urinary nitrogen excretion (Potischman & Freudenheim, 2003). By applying those measures, it has been demonstrated that underreporting is a major problem in dietary assessment, in particular among overweight subjects. Underreporters tend to overreport healthy foods such as fruit, vegetables and fish and underreport unhealthy foods such as sugar, confectionary and cakes (Livingstone & Black, 2003; Bingham, 2003). Unfortunately, only few biomarkers exist for validation of reported dietary intakes and the existing markers only cover limited aspects of the diet (Bingham, 2002; Jenab et al., 2009). Availability of exposure markers for intake of individual foods would be a way to broaden validation of data from dietary records but very few markers of individual foods are known and even fewer have been validated. Discovery of new food exposure markers therefore have a great potential to supplement the current dietary assessment tools. In some cases exposure markers may also provide a better measure than what can be achieved from dietary self-reporting.

This is true in particular for estimation of dietary exposure to individual nutrients or food components for which the content varies highly within and between the common food sources (Potischman & Freudenheim, 2003; Wishart, 2008).

Another application of dietary exposure markers is for estimating compliance to a nutritional intervention study (Llorach et al., 2012). The purpose of intervention studies is to investigate the effects of well-defined dietary exposures that may consist of individual foods, compounds or whole diets. An unexposed control group is compared to an intervention group and a good compliance to the study is crucial for the study outcome and for the interpretation of the results (Vitolins et al., 2000). Objective measures of dietary exposure in these studies would enable the investigator to identify non-compliant subjects and remove data from those subjects prior to the statistical analysis in order to reduce bias caused by non-compliance.

## 1.2 What defines a good exposure marker?

Dietary exposure markers should ideally reflect the intake of a food or food component accurately and be applicable in many populations (Jenab et al., 2009). However, because foods undergo digestion which is a complex process, sometimes with substantial inter-individual variation, ideal exposure markers do not exist. Rather, the ideal exposure marker is a marker that is well understood and thereby applicable because its limitations and advantages are known. The most important validation criteria for dietary exposure markers are listed in Table 1.

**Table 1** Biological and analytical validation criteria for dietary exposure markers

| | | |
|---|---|---|
| **Biolgical validation** | **Known relation to the exposure** | It must always be ascertained that the association between food exposure and the exposure marker is causal. For example the marker may be a metabolite of a compound known to be present in a food. |
| | **Sensitivity and specificity** | The measured marker should be as unique to the exposure as possible. This implies that the percentage of true positive measures (sensitivity), for exposure, and true negative measures (specificity), for no exposure, should be as high as possible. |
| | **Dose-response** | There should be a positive association between the level of exposure and the measured level of the marker. The range of intake with a dose-response should be known. |
| | **Inter-individual variation** | The main potential sources of inter-individual variation should be investigated, such as genotypes, gender, age, smoking, gut microbiota etc. |
| | **Time of exposure** | An exposure marker may be acute, reflecting only recent intake of a food, or long term, reflecting intake from days, months or even years. What type of exposure is reflected by a marker should be determined. |
| | **Population** | The population where the exposure marker is applicable should be known. |
| **Analytical validation** | **Sample** | Timing and type of sample should be defined as well a storage conditions and sample preparation. |
| | **Defined measure** | The marker should be quantifiable by a defined method and the analytical error should be known. |

The information in the table is based on the following references: Spencer et al., 2008, Jenab et al., 2009, Manach et al., 2009.

It is obvious from Table 1 that it takes a lot of work to validate an exposure marker and this is probably the main reason why so few markers are routinely used. In addition, one marker may not be sufficient to describe an exposure. Rather, a combined measure including several markers may be necessary and provide a stronger measure in some cases (Wishart, 2008), which further complicates the validation procedure. All validation criteria can rarely be completely assessed. There will almost always be more aspects to investigate which can further consolidate a marker for a given study setting. Validation is therefore an ongoing process that can contribute to understanding which factors are most important for the metabolism of a food or food compound and what a given dietary exposure marker reflects.

## 1.3 Metabolomic fingerprinting as a tool for marker discovery

Metabolomic fingerprinting, also known as untargeted metabolomics, is a holistic approach for discovery of metabolic changes following some sort of perturbation such as a change in dietary habits or development of a disease. As many metabolites as possible are measured simultaneously in a

biological sample and together, these metabolites consist a so-called 'fingerprint' which can be compared between subjects or within subjects to find discriminant metabolites of the perturbation studied (Dettmer et al., 2007). A metabolite is defined as any small molecule (typically <1500 Da) that can be found in an organism (Wishart, 2008) and metabolites can be subdivided into several categories depending on their origin or function. The group of exogenous metabolites is of interest as dietary exposure markers, as it includes all metabolites derived from extrinsic sources such as the diet (Scalbert et al., 2009; Dunn et al., 2011). The term nutrimetabolomics has been introduced by Zhang et al. (2008) for metabolomics studies concerning nutrition. Studies conducted within nutrimetabolomics are mainly aiming to find new markers of dietary exposure and effect as well as exploring potential interactions between dietary characteristics and phenotypes (Llorach et al., 2012). The ultimate goal of nutrimetablomics together with other 'omics' disciplines is to make possible a so-called personal nutrition where dietary recommendations can be provided at an individual level (Zhang et al., 2008).

Our diet contributes a vast number of compounds that needs to be handled in one way or the other by the body after intake (Gibney et al., 2005). Plant foods are rich in secondary metabolites. Considering polyphenols alone, more than 500 different compounds have been identified in plant foods so far (Neveu et al., 2010). Processed foods can contain various additives and as soon as food is prepared, cooked, mixed, heat-treated etc., even more compounds are formed. After intake, the dietary compounds are digested during which they may undergo several transformations before excretion. The gut microbiota of the colon comprises more than 400 species and contributes largely to digestion and biotransformation of non-nutritive compounds (Gibney et al., 2005). Microbial and non-microbial products absorbed from the gut can be further modified by the endogenous metabolism. As an example, phase I and II reactions are common in which compounds are typically glycinated, sulfated or glucuronidated to increase the polarity which makes them more soluble in urine and thereby easier to excrete (Gibney et al., 2005; Spencer et al., 2008). Altogether, our diet can introduce an almost innumerable number of exogenous metabolites to the body and these are to a large extent excreted into urine (Scalbert et al., 2009). In the study of dietary exposure, urine is therefore often the biofluid of choice to identify markers of food intake.

With such a vast diversity of compounds originating from the diet, it is extremely difficult to predict which metabolites would be the optimal exposure markers of choice to estimate intake of a given food. This is why metabolomic fingerprinting is a very promising approach to identify new dietary markers. Performing metabolomic fingerprinting on dietary intervention and cohort studies can reveal how individual foods and complex diets are reflected in urine and lead to discovery of new promising exposure marker candidates (Wishart, 2008; Llorach et al., 2012).

The most commonly applied methods for metabolomics analysis are mass spectrometry (MS), in most cases coupled with chromatographic separation, and nuclear magnetic resonance (NMR). However, due to the chemical diversity and large variation in concentration ranges of metabolites, no single analytical method can cover them all (Dunn et al., 2011). The highest number of metabolites, several thousands, can be detected with MS but because the metabolites have to be ionized, and ionization is matrix dependent, the method is semi-quantitative (Dettmer et al., 2007). Any

finding in untargeted metabolomics must therefore be validated further in a quantitative analysis. Before this can be done, however, the metabolites must be identified. Compounds detected by ultra-performance liquid chromatography quadrupole time of flight mass spectrometry (UPLC-qTOF-MS), which was applied in the present work, are characterized by a retention time (RT), a mass-to-charge ratio (m/z) and maybe a fragmentation pattern. The identification of a metabolite, keeping in mind all the possible exogenous metabolites that may be formed from foods, is therefore a major task. Even though several metabolite and chemical databases are available as well as a number of automated tools to aid the identification of new compounds, this step is generally recognized as the main bottleneck for discovery of new biomarkers by metabolomics (Dunn et al., 2013).

## 1.4 Nutrimetabolomic studies on exposure and compliance

The application of metabolic fingerprinting in nutrition research is still a relatively new field and a large part of the research conducted has been focused on methodical aspects of the metabolomics analysis in order to ensure sample stability and analytical reproducibility (Maher et al., 2007; Gika et al., 2008a; Gika et al. 2008b; Guy et al., 2008). Main sources of biological and inter-individual variation have also been investigated such as the effect of dietary standardization, gender, Body mass index (BMI) and age (Wang et al., 2005; Slupsky et al., 2007; van Velzen et al., 2008; Rasmussen et al., 2010; Favé et al., 2011). Together, these studies have demonstrated that it is possible to conduct a robust untargeted metabolomics analysis and discover clear patterns in the metabolomic fingerprints for different physiological characteristics and dietary standardizations. Within the last ten years, an increasing number of studies have started to apply untargeted metabolomics in human nutrition. For the area of dietary exposure and compliance markers in urine, which are most relevant to this thesis, the majority of published metabolomics studies performed in humans are short intervention studies and meal studies. Meal studies have been conducted on chocolate, salmon, raspberries, biscuits, broccoli and an almond skin extract (Stella et al., 2006; Llorach et al., 2009; Llorach et al., 2010; Lloyd et al., 2011b). The short term metabolic changes following intake of a meal are generally very clear and a large number of discriminant markers are found in these studies, especially if the diet is controlled. In addition, time-changes following intake of a meal have been traced in some meal studies where urine has been collected at various time-points after the meal (Llorach et al., 2009; Pujos-Guillot et al., 2013). Intervention studies have been performed to investigate exposure and effect markers of capsules containing polyphenol rich extracts of red wine, grapes and black tea (van Velzen et al., 2009; van Dorsten et al., 2010), chocolate (Martin et al., 2009; Llorach et al., 2009; Llorach et al., 2013), nuts (Tulipani et al., 2011), soy products (Solanky et al., 2005), an apple, strawberry and carrot drink (Walsh et al., 2007), different teas (Wang et al., 2005; van Dorsten et al., 2006), milk and meat (Bertram et al., 2007) and whole grains as compared to refined grains (Bondia-pons et al., 2013; Ross et al., 2013). A few studies have investigated dietary characteristics instead of individual foods such as high versus low protein diets (Stella et al., 2006; Rasmussen et al., 2012b), meals with different fatty acid composition (Legido-Quigley et al., 2010), high and low dietary glycemic index, diets differing in fiber contents (Rasmussen et al., 2012a) and three dietary patterns defined from self-reported intakes of 33 food groups

(O'Sullivan et al., 2011). Other observational studies have been conducted to find markers of habitual food exposures, most often based on FFQ. The advantage of observational studies is the mixed dietary background and the uncontrolled setting, which increases the chance of finding food markers with a high specificity. However, not all markers may be discovered in such studies due to many sources of biological variation, errors associated with use of FFQ and the semi-quantitative nature of MS data. In a study based on FFQ dietary records and 24 h urine samples collected by a standardized procedure, models were developed to discover markers of 38 individual foods and food groups (Lloyd et al., 2013). A few known food markers from previous studies were identified among the discriminant metabolites to confirm that the models gave biologically meaningful discriminants. In general, frequently consumed foods and food categories consisting of distinct foods resulted in the best models, whereas reliable models could not be developed for a range of less frequently consumed and more diverse food groups. In two recent studies, data from cohort studies and controlled dietary studies have been analyzed and compared to find markers of citrus (Pujos-Guillot et al., 2013) and a diet high in fruit and vegetables as defined by consumption of cruciferous vegetables, soya foods and citrus (May et al., 2013). These studies are interesting because they highlight the interdependency of the study design for the markers discovered. Pujos-Guillot et al. (2013) compared the markers found by LC-MS based metabolomics from a meal study, an intervention study and a cohort study on citrus. This study demonstrated that the citrus markers found were not the same in all studies and that the same marker could be ranked high in one study setting and low in another. The number of discriminant features found decreased from 605 in the meal study to 19 in the cohort study. In May et al. (2013), the results from a dietary intervention study on high and low fruit and vegetable diets were compared to results from a cohort study based on the third and first tertiles of self-reported fruit and vegetable intake from FFQ and from three day WDR. While 2857 discriminant ions were found in the intervention study, only around 50 ions were significantly different also in the cohort study in which the differences between high and low fruit and vegetable consumers were much less pronounced.

Overall, the studies applying metabolomic fingerprinting to discover urinary exposure and effect markers of foods and diets have proved that dietary exposures can be distinguished in the urine metabolome. The dietary exposure markers found by metabolomics so far has been reviewed by (Llorach et al., 2012). According to this review food exposure markers can be divided into two main classes: A class of compounds which is metabolised by the gut microbiota before absorption and another class that is not. Common examples of microbially derived markers are hippuric acid, 3- and 4- hydroxyhippuric acid and hydroxyphenylacetic acids. Microbial markers are commonly found for polyphenol rich diets and are normally not specific to individual foods. Creatine, creatinine and carnitine, which are not produced by the gut microbiota, have been found several times as markers of high meat diets and proline betaine, which is thought to be an inert metabolite, has consistently been found as a marker of citrus consumption. Except from these common examples, the studies conducted until now are generally too few and diverse to point out the most promising markers. A great advantage of the metabolomics approach is that identified markers which occur consistently across different studies and study designs already at this discovery stage can be evaluated according to most of the biological validation criteria listed in Table 1, such as sensitivity and

specificity, known relation to the exposure and time of exposure. This provides a short-cut to pick out the best marker candidates before performing an analytical validation and further investigation of a marker.

## 1.5 The New Nordic Diet

All studies in this thesis are related, directly or indirectly, to a dietary pattern which has been named The New Nordic diet (NND). The term NND was first put forward by Bere & Brug (2009) as a regional healthy diet that could be a Nordic equivalent to the Mediterranean diet, which has long been famous for its health potential. The advantage of developing regional healthy diets is that such diets are based on familiar foods to the region and therefore probably easier to adopt. A diet consisting of foods that can be grown within the region is also more environmentally friendly (Bere & Brug, 2009). Much in line with this, the NND applied in the studies in the present thesis, was developed based on four main criteria: Health, palatability, local produce and sustainability (Mithril et al., 2012). The NND is defined by intakes of fifteen food groups which were selected based on their health potential and the potential for a large scale local production. The macro- and micronutrient compositions of the diet are largely based on the Nordic Nutrition Recommendations (Nordic Council of Ministers, 2004) and only deviates for a few nutrients (Mithril et al., 2013). To make the diet more sustainable, it is seasonal, based on local products and mainly organic. The palatability of the diet has been taken into account by including foods which develop good organoleptic properties when grown in a Nordic climate.

In Figure 1 and 2, the macronutrient composition and intake levels of the fifteen food groups in NND are given in comparison with an Average Danish diet (ADD). In addition to the listed criteria for macronutrients and food groups, the NND diet contains at least 50 % organic foods and more than 95 % foods produced in the Nordic region. The corresponding numbers for ADD are less than 10 % organic foods and less than 50 % foods of Nordic origin (Poulsen et al., 2014).

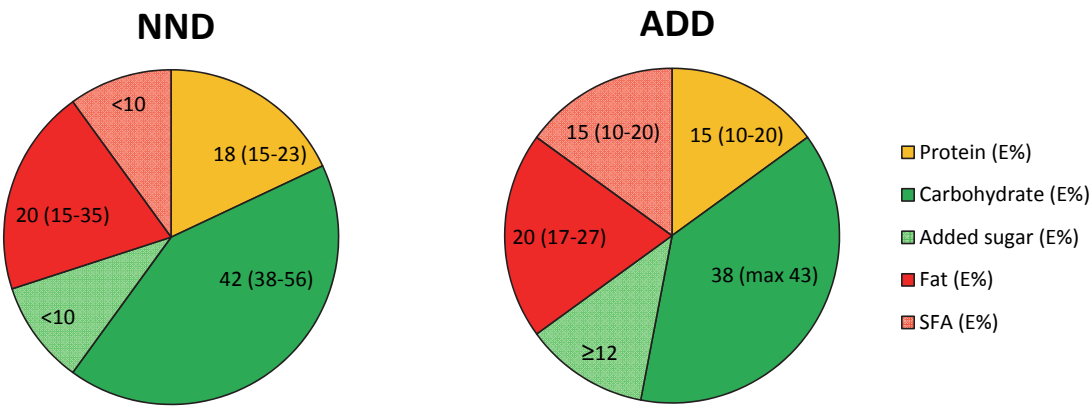



**Figure 1** Macronutrient composition of NND and ADD. Target ranges for energy percentages (E%) of protein, added sugar, carbohydrate (including fibre and excluding added sugar), saturated fatty acids (SFA) and fat (excluding SFA). Data from Poulsen et al. (2014).

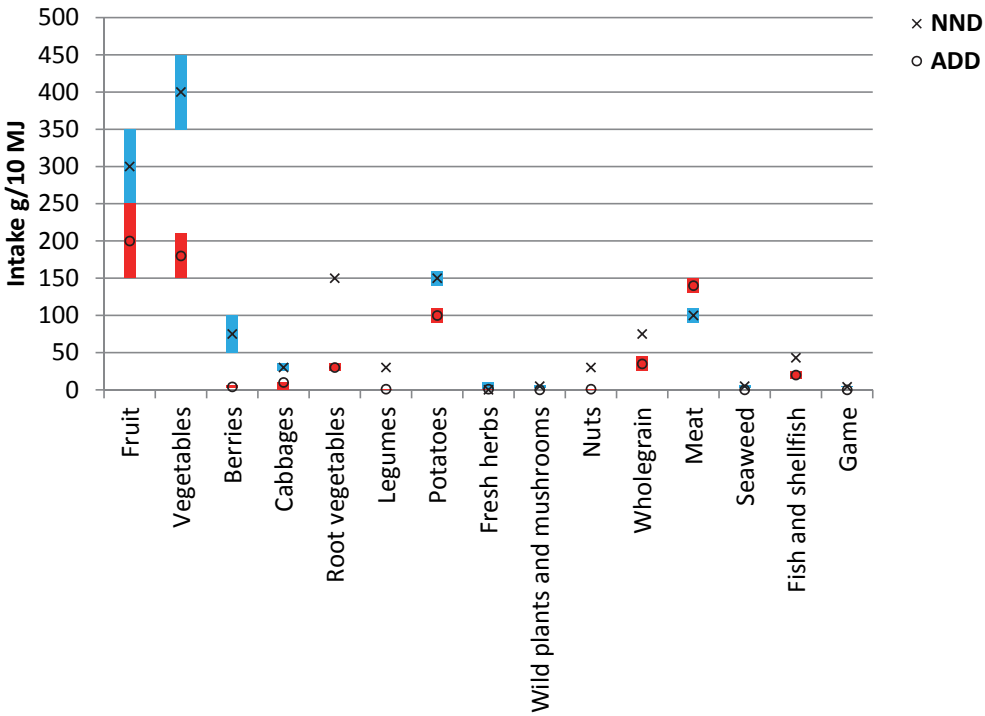



**Figure 2** Target ranges for intake of food groups in NND as compared to ADD. The target intakes of NND and ADD are marked as crosses and circles, respectively. The bars represent the target ranges of intake as applied in the intervention study (section 3.3). For NND data points without bars there is no upper limit of intake. Data from Poulsen et al. (2014).

It is evident from Figure 1 and 2 that large contrasts can be found for several food groups between the diets, especially vegetables, berries and root vegetables, while the macronutrient composition is very similar. This makes NND a very interesting diet to study from a metabolomics perspective. With metabolomic fingerprinting, individual characteristic NND foods as well as the whole NND dietary pattern in comparison to ADD can be assessed and contribute new exposure markers of foods and the dietary pattern as a whole. Several of the food components in NND have never been investigated with metabolomics before such as lingonberry, sea buckthorn, rhubarb, beetroot, celeriac and parsley root. Working with NND therefore opens up for new discoveries in the area of food exposure markers. As part of the OPUS project (see preface), large intervention studies have been conducted with NND in school children (Damsgaard et al., 2012) and in overweight adults (Poulsen et al., 2014) in order to investigate the potential health benefits of the diet. The metabolomics conducted in this PhD concerns mainly the adult intervention study which is presented in section 3.3. In addition, a meal study with nine *Brassica*-containing NND meals has been analyzed (section 3.1).

# 2. Objectives

Two main objectives are investigated in this thesis:

1. How untargeted metabolomics can be applied in the search for new food exposure markers.

   Paper I and II deal with exposure marker discovery in three different nutritional study settings. The results from the studies are summarized and the potential food exposure markers (PEMs) are validated based on other published findings. Based on the validation, it is discussed how metabolomic fingerprinting can contribute to exposure marker discovery. What are the main strengths and limitations and which promising aspects are still largely unexploited.

2. How untargeted metabolomics can contribute to development of compliance measures in nutritional intervention studies.

   To my knowledge, the compliance model in Paper III is the first attempt to determine dietary compliance to a dietary intervention study from characteristic urinary metabolomic fingerprints. Due to the limited literature available on the subject, the second objective of the thesis will be discussed by elaborating on the findings in Paper III. A range of parameters related to the urine samples, the subjects and compliance will be compared between misclassified and correctly classified samples in the developed compliance model, with the aim of obtaining a better understanding of the potential of such a model.

The two study objectives are addressed separately in section five and six, respectively. Section subheadings are organized by subject to provide an easy overview of the issues considered. Results and discussions are not separated in the individual sections. Overall conclusions for the study objectives are provided in section seven.

## 2.1 Focus areas and limitations

Metabolomics is a comprehensive method that involves multiple steps from sample preparation to marker identification (van den Berg et al., 2011). The choice of analytical strategy for samples, preprocessing and the statistical analysis will unavoidably influence marker discovery (Scalbert et al., 2009; Rasmussen et al., 2010; Gürdeniz et al., 2012) and the findings in a metabolomics study are therefore very closely related to the whole analytical pipeline. It is not possible to carry out and compare every possible procedure for sample and data analysis in order to find the best one. In untargeted metabolomics you win and you lose. The method opens up for finding new biomarkers that would be difficult to discover by other means but you can never be sure that you find all and that the ones you find are true markers and not artifacts introduced somewhere in the experiment (Theodoridis et al., 2008). The explorative nature of metabolomics is exciting because you never know where data will lead you to. However, at the same time, it is innate in the method that a dataset can

neither be explored fully nor fully understood. There is always a large proportion of unknown features and a high level of complexity in the data that cannot be completely elucidated. This thesis will not go into detail with the more technical aspects of the metabolomics methodology, such as the choice of analytical method and the common statistical tools for data analysis. Only the strategies applied for data analysis in the studies in Paper I-III will be discussed.

All work in the present thesis has been done on urine and urinary markers will therefore be the focus even though other biological samples can of course be used for biomarker discovery as well.

When reviewing the literature, metabolomics studies and, preferably, LC-MS metabolomics studies are emphasized. Dietary exposure markers can also be found by measuring known metabolites of foods directly in a targeted approach (Kuhnle, 2012). However, such markers would not necessarily be found in a metabolomics study due to the limited metabolite coverage and the dependency on all other detected compounds during marker discovery.

# 3. Studies analyzed

Data from three studies have been investigated: A cross-over meal study with *Brassica*-containing meals (MEAL), a range of short-term single food studies (FOOD) and a six month parallel intervention study comparing two dietary patterns (INTER).

## 3.1 Design of MEAL

The MEAL study was a cross-over study with nine test meals performed in 17 healthy normal weight males and females. On each test day, the diet was controlled from 9 am to 5 pm, including a breakfast at 9 am, a snack between 9 and 11 am, a test meal between 12 and1 pm and a dinner between 4 and 5 pm. The breakfast, dinner and snack were standardized meals and the test meal consisted of three types of *Brassica*-containing meals (a Soup, a Pie and a Barleyotto), each served in three versions with three different protein sources (vegetarian, meat and fish). On each test day, one meal was served and subjects were randomized to one of the three protein sources. Each meal was served once per week in a random order and the study was completed within three weeks.

Urine samples were collected from 9 am until intake of the test meal (Urine 1), from intake of the test meal until 2 h after the test meal (Urine 2), from 2 h after the test meal until dinner (Urine 3) and from intake of the dinner until 9 am the following morning (Urine 4). After removing 1.5 mL from each of Urine 1-4, the urine samples were pooled into a 24 h sample.

Data used for the metabolomics study:

- List of ingredients from the test meals
- UPLC-qTOF-MS analysis of Urines 1-4 and the pooled 24 h urines

More detailed information on the study is given in Paper I.

## 3.2 Design of FOOD

In the FOOD studies, experiments were performed with eight single foods (white cabbage, Brussels sprouts, carrot, parsley root, kale, chicory salad, brown beech mushroom and fava beans) prepared in the same way as they were served in the MEAL study (raw, fried or boiled) except that all other ingredients were omitted. The food was served ad libitum between 12 and 1 pm. Urine was collected from 11am until intake of the food (Urine 1) and from intake of the food until 3 pm (Urine 2). Except from the test food, participants were only allowed to drink water from 9 am until 3 pm. Three to four subjects participated in each study. The participant characteristics were comparable to the MEAL study but no subject participated in both studies and the participants were not the same in all FOOD studies.

Data used for the metabolomics study:

- UPLC-qTOF-MS analysis of Urines 1-2

More detailed information on the study can be found in Paper I.

## 3.3 Design of INTER

The INTER study had the two dietary patterns, ADD and NND, as intervention arms. Three day WDR were collected three times (week 0, 12 and 26) and 24 h urine samples were collected five times (week 0, 4, 12, 20 and 26) during the study. At baseline (week 0), the diet was standardized. The two intervention diets were defined as described previously in section 1.5 (Figure 1 and 2). Participants collected all their foods free of charge at a small supermarket at the University of Copenhagen and the dietary intake was monitored by registering all foods brought home from the shop buy each household (couples who both participated in the study or singles). During monitoring, participants could select their own foods in the supermarket as long as the foods chosen overall were in accordance with their assigned dietary pattern. Para-aminobenzoic acid (PABA) tablets were given to the subjects on days of urine collections to measure the completeness of urine samples.

The subjects enrolled were men and women aged 18-65 with an increased waist circumference and preferably one or more additional risk factors of the metabolic syndrome. Out of 181 subjects, who were randomized to the two diets, 107 study completers provided urine samples and dietary records at week 0,12 and 26.

For the metabolomics studies in Paper II and III, the following data from the study were used:

- One day WDR made on the same day as the urine collections
- Total amount of foods registered over the full study period in the supermarket for participants with a household size of one
- UPLC-qTOF-MS analysis of 24 h urine samples

More details on the study and the subsets of data applied in the metabolomics analyses is described in Paper II and III and in the main paper from the study (Poulsen et al., 2014).

# 4. Analytical strategies applied

The metabolomics data analyses carried out for discovery of potential exposure markers (PEMs) and for investigation of compliance are illustrated in Figure 3 and Figure 4, respectively. The different analytical steps (data preprocessing and –treatment, statistical analyses and feature validation) are described briefly in section 4.1-4.3.

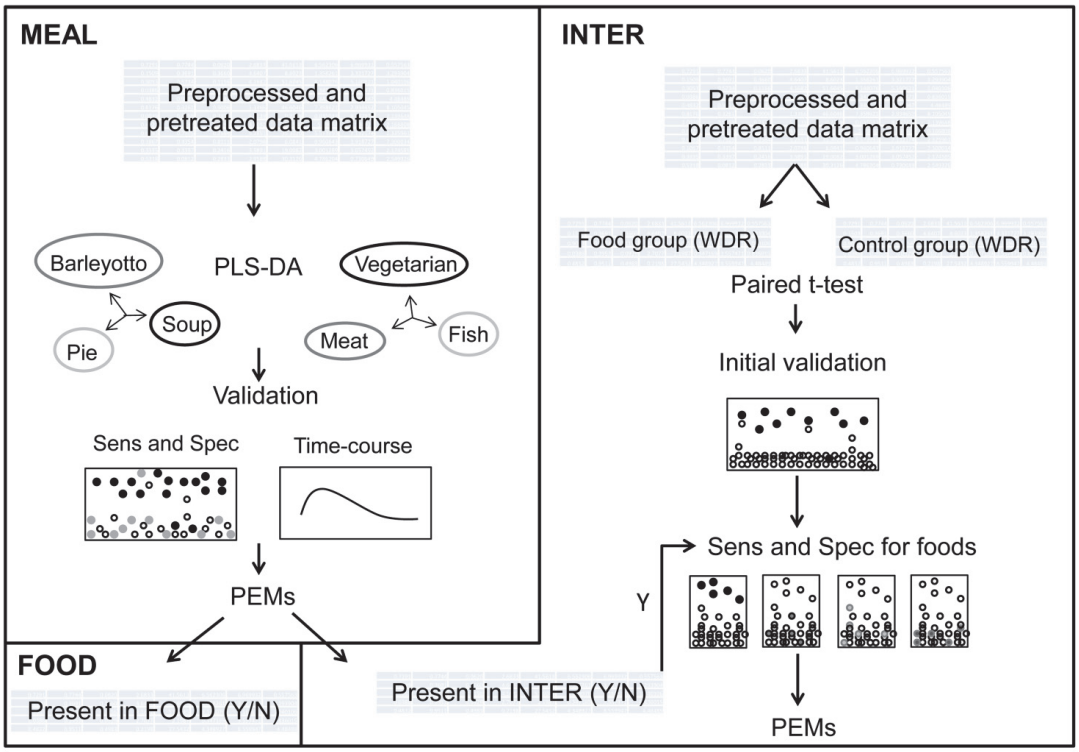



**Figure 3** Analytical strategy applied for discovery of PEMs. Separate analyses were carried out for the MEAL and INTER datasets to find PEMs. The PEMs in MEAL were further validated using data from FOOD and INTER. Sens: sensitivity; Spec: Specificity; Y/N: yes/no; PLS-DA: Partial least squares discriminant analysis.

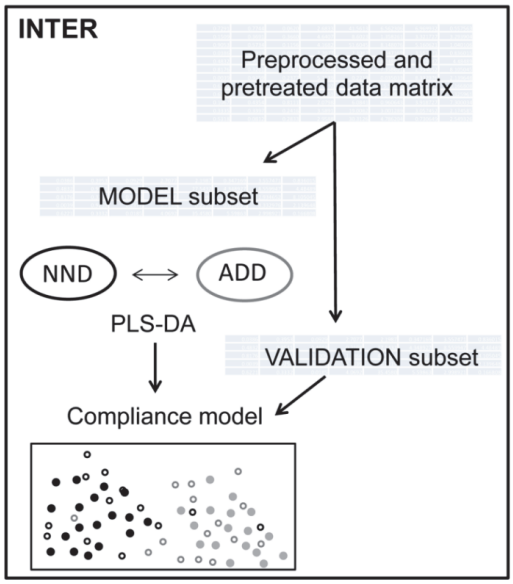



**Figure 4** Analytical strategy for development of a compliance measure in INTER. A partial least squares discriminant analysis (PLS-DA) model was developed as a compliance measure to distinguish the dietary patterns NND and ADD. The compliance model was built from a subset of urine samples (MODEL) and another set of samples from the same study was used to evaluate model performance (VALIDATION).

## 4.1 Data preprocessing and -treatment

### 4.1.1 Preprocessing

Preprocessing is in many ways the foundation to performing a good metabolomics study. The purpose of preprocessing for LC-MS based metabolomics is to convert raw data files, in which RT, m/z and ion intensities are measured continuously throughout a sample run, into a two-dimensional data matrix that is used throughout the following data analysis. The two-dimensional data consist of paired RT and m/z values that together define a detected feature in one direction and samples in the other direction. For each combination of a feature and a sample, the ion intensity, represented by the integrated peak area of the feature, is calculated. The key in preprocessing is to determine the biologically relevant signals in raw data as accurate as possible when detecting and integrating peaks and at the same time minimize peaks from sources of analytical noise (Katajamaa et al., 2007).
Two main critical steps in data preprocessing are peak detection within samples and peak alignment between samples. Various methods can be applied for these steps and the choices of software for preprocessing as well as the parameter settings are crucial for the result (Gürdeniz et al., 2012). In Paper I-III, mzMine2 (Pluskal et al., 2010) has been applied for preprocessing of raw data. This software has the advantage of a gap filling algorithm after peak alignment which decreases the risk of missing peaks in a sample. Also, mzMine2 has good visualization tools for the conversion and alignment steps (Gürdeniz et al., 2012). For the MEAL and the INTER datasets, a subset of 10-20

representative samples were used to optimize the parameter settings. Due to a large number of samples in MEAL (close to 1000), a high noise level of 40-50 for peak detection was applied, as the software otherwise could not handle the data. In INTER, on the other hand, the noise level was set to 15. This is the main explanation why 6933 features were detected in total in MEAL compared to 13327 in INTER. Features with low intensities in all samples may have been missed in MEAL during preprocessing, while more noise will be present in INTER that should be reduced before the statistical analysis (see section 4.1.3).

### 4.1.2 Normalization

For nutrition studies, the metabolic changes of interest are often subtle and smaller than inter-individual differences (Rezzi et al., 2007; Scalbert et al., 2009; Heinzmann et al., 2011). Nutritional effects may therefore be overshadowed, if random and systematic errors caused by sample collection, as well as biological and analytical variation, are not minimized. A robust analytical method and study design are the basis for performing a good metabolomics study but the choice of additional normalization procedures and data pretreatment are also influencing the study outcome (van den Berg et al., 2011). For urine samples, an important issue is the biological variation in urine concentrations (Ryan et al., 2011). Concentration differences should ideally be corrected by determination of the renal elimination rate but such a measure is not feasible to obtain. Common approaches suggested for normalization of urine are to adjust the samples according to creatinine level, osmolality or ion intensity across samples (Warrack et al., 2009). In the literature, creatinine, volume and normalization according to total ion current (TIC) has been applied for nutrition studies analyzed by MS based untargeted metabolomics (Legido-Quigley et al., 2010; van Dorsten et al., 2010; Lloyd et al., 2011b; Pujos-Guillot et al., 2013). To my knowledge, only one study has compared the different approaches in which it was concluded from the separation in a principal component analysis (PCA) that an ion intensity based measure and normalization to the same osmolality performed best (Warrack et al., 2009). However, this study was an in vivo toxicology study on rats and the conclusion may therefore not be the same for human nutrition studies. More work would need to be done to understand how the variations in urine concentration influence data analysis and how to best solve this issue. In the MEAL and INTER studies, TIC was used for normalization across samples.
Besides normalization across samples, normalization can also be performed in the feature direction (Goodacre et al., 2007). In the MEAL and INTER studies, all samples from the same subject were placed within the same plate to minimize analytical intra-individual variation. This allows a normalization of data for either analytical variation between plates or inter-individual differences. However, as individual and plate variation will always be mixed, normalizing for one of these will always be biased by the other.

In the MEAL dataset, a large number of samples (40-45) were available from each subject due to the cross-over design. For this reason, few subjects were represented on a plate and inter-individual differences within plates were evident (Figure 5). In the statistical analysis applied for the study, intra-individual differences were not considered in the multivariate model and it was therefore important to make this correction before the statistical analysis to make full use of the cross-over study

design (van Velzen et al., 2008). Normalization in this study was performed feature-wise by adjusting the mean peak area of a feature across all samples from a subject to the same mean value for all subjects. In the INTER dataset, the study had a parallel design. Adjusting such data for individual differences would interfere with the nutritional intervention. Instead, plate correction was performed for this dataset by a feature-wise normalization to obtain the same mean value of each feature on each plate. When a large number of subjects are represented on a plate, the biological variation is assumed to be randomized across the plates and plate differences can therefore be corrected with minimal influence on the biological variation.

In Figure 5, the effect of the different normalization steps is illustrated with data from MEAL (Paper I). From the left plot in Figure 5A, it is clear that several systematic sources of variation are present in the data. Subjects are clearly separated from each other and this is also true for the subjects, whose samples have been run on the same plate (person 3 and 4). Within each sampling point, there is a tendency for many subjects that TIC is decreasing with increased urine volume, demonstrating the influence of different urine concentrations. The unwanted systematic variation on the left plot in Figure 5A is removed, following normalization and person correction, as illustrated in the plot to the right. In Figure 5B, it is exemplified how the different normalization steps affect an individual feature that was found as a PEM. Feature intensities are becoming more similar when comparing data before (on the top) and after normalization to TIC (middle). In addition, the differences between the meals are becoming more and more pronounced when moving from the top plot to the bottom plot on the right side of Figure 5B.

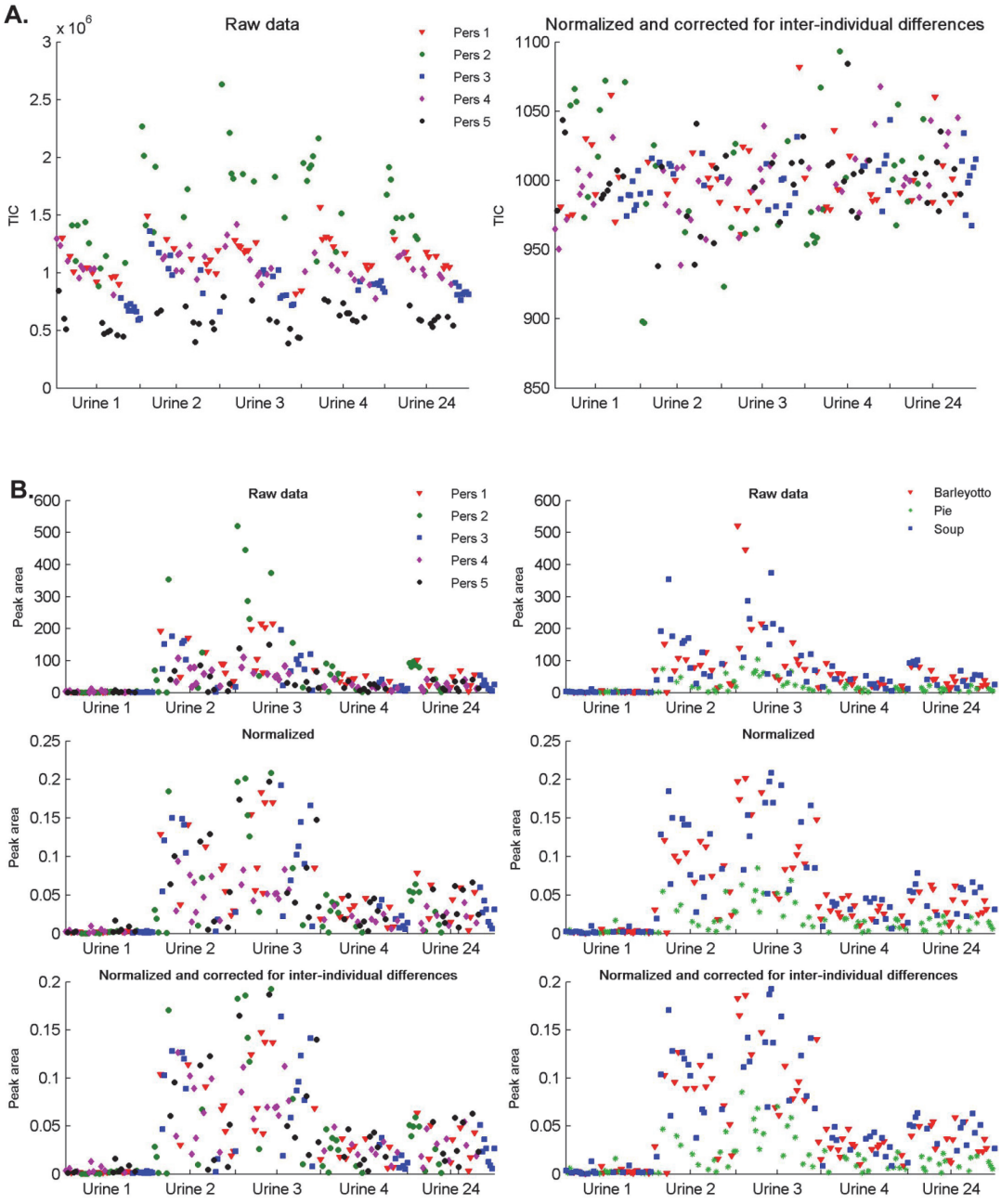



**Figure 5** Effect of normalization to TIC and correction for inter-individual differences in MEAL. Data from five representative subjects are shown for each time point (Urine 1-4 and 24 h urines). Within each time-point, data has been sorted according to urine volume starting from the lowest on the left to the highest on the right. Person 3 and 4 were run on the same plate. A. TIC for raw data (left) and for samples after normalization to TIC and inter-individual correction (right). B. Example of raw data (top), data after normalization to TIC (middle) and normalization to TIC and correction for inter-individual differences (bottom) from a feature that was found as a PEM in the study (m/z 263.054, RT 3.54). The plots on the left side are colored according to person and on the right side according to meal.

### 4.1.3 Data reduction

Metabolomics data is characterized by a low number of samples and a high number of variables of which a large proportion is noise introduced particularly during the preprocessing steps for peak detection and alignment. Preferably, the number of features should be reduced to include only the ones that are related to the nutritional effects of interest before the statistical analysis (Kjeldahl et al., 2010). One way to eliminate noisy features and features unrelated to the treatment is the 80 % rule introduced by Bijlsma et al. (2006). According to this rule, data is divided into different treatment groups, depending on the study design, and features not present in at least 80 % of the samples in any of the groups are excluded. The rule was originally made for preprocessing methods without a gap filling algorithm which generates a large number of zeros but the rule can easily be adapted, if a threshold defining the noise level is used instead of zero. An iteration of thresholds was performed with the INTER and MEAL studies to find the optimum cut-off value for defining the noise level and afterwards this value was applied to exclude features that were noise or not related to the nutritional effect of interest. In the compliance study, a feature only had to be present in 5 % of the samples within any group instead of the usual 80 % due to the expected diversity within a dietary pattern (Paper III). For the other studies, 80 % were used. Between 29 % (the MEAL study) and 94 % (analysis of PEMs for individual foods in the INTER study) of all detected features were excluded after applying the 80 % rule, demonstrating the impact of this pretreatment step. The very high percentage of features excluded in the INTER study is partly caused by strict criteria for data analysis of PEMs for individual foods, where 80 % was used as cut-off despite a sample size down to ten for the food groups (Paper II). However, the main reason of the divergence is different preprocessing parameter settings in the two studies as described in section 4.1.1.

## 4.2 Statistical analyses

### 4.2.1 Multivariate analysis (PCA and PLS-DA)

Multivariate analysis is commonly applied for metabolomics data because it, as opposed to univariate approaches, can handle a large number of variables and does not require that variables are independent (Wold et al., 2001). The features detected in LC-MS based metabolomics are often strongly correlated due to adduct formation and fragmentation during analysis as well as biological associations such as shared metabolic pathways. Another advantage of a multivariate approach is that biological patterns can be explored which may include features that would not be significant in a univariate statistical analysis.

In Paper I and III, two common multivariate methods were applied; PCA and partial least squares discriminant analysis (PLS-DA). PCA is an unsupervised method that is useful to explore the main sources of variation in a dataset. It decomposes a data matrix into a new orthogonal space consisting of principal components (PCs) that describe the main variation in the data (Boccard et al., 2010). For the work presented in this thesis, PCA has been applied continuously during data analysis to explore how data is affected by different mathematical operations. It has also been used to explore

the food patterns in the INTER study (Paper III). In situations where the main sources of variation are not explained by the nutritional intervention, as is normally the case for metabolomics data, PLS-DA is necessary to investigate the effect of interest (Barker & Rayens, 2003). PLS-DA is a supervised classification method. A dummy class matrix with zeros and ones is made corresponding to defined groups of samples in the data, for example different meals, protein sources or dietary patterns (Figure 3 and 4). The PLS-DA model is built in such a way that the new orthogonal space of latent variables (LVs) maximizes the correlation between the data matrix and the class matrix (Wold et al., 2001). In this way, features that best predict the class vector can be discovered. Application of PLS-DA is convenient for metabolomics data but two main issues need to be addressed in order to obtain meaningful results: Validation of the model and feature selection (Westerhuis et al., 2008; Rajalahti et al., 2009). With a large number of variables there is always something that correlates to the class matrix and it is therefore crucial to validate the PLS-DA model to make sure that it can actually predict which class a new sample belongs to. The other issue, feature selection, is challenging because it is difficult to know how many features should be included in a PLS-DA model to not lose any relevant biological information. The same strategy for developing a PLS-DA model applying double cross-validation and feature selection based on Variable importance in projection (VIP) scores was applied in Paper I and III. In all multivariate models, data was mean-centered and autoscaled prior to analysis. The purpose mean-centering is to move the between sample variation to the lower PCs and LVs by subtracting the mean of each feature from all the individual measurements of the feature (Boccard et al., 2010). Autoscaling gives all variables equal weight in the analysis, ensuring that metabolites with low abundance are also taken into account (Wold et al., 2001). The drawback of autoscaling is a higher risk of chance findings when noise is given equal weight to biological compounds measured in the samples. To reduce the extent of chance findings as much as possible extra feature validation steps were applied for PEM discovery in Paper I-II, by taking advantage of other aspects of the study designs (see section 4.3).

### 4.2.2 Univariate analysis

The fact that metabolomics data contains correlating features, different variable distributions and a large number of variables compared to samples violates the underlying assumptions in a univariate statistical approach (Broadhurst et al., 2006). Despite this, univariate statistics can still be applied in metabolomics as long as the limitations are kept in mind.

In Paper II, a paired t-test was performed. This univariate measure was preferable over multivariate models for two reasons: 1) Food intake in the study was not controlled on individual days implying that only few subjects consumed each food group of interest on the days with urine collection. 2) Inter-individual differences could not be adjusted as explained in section 4.1.1. With few observations and large inter-individual differences, it is not possible to develop a valid multivariate model without a high risk of over-fitting. With a paired t-test, the inter-individual variation was taken into account which increases the chance of finding markers. In addition, iteration was performed with the available control samples from the included subjects to take into account all observations from the selected subjects and thereby strengthen the analysis. The higher rate of false positive results

(type 1 errors), due to multiple testing was corrected by the positive false discovery rate as defined in Storey (2002). A limitation of the applied method for univariate analysis is the assumption that each feature follows a normal distribution which may not be fulfilled. There may be cases where a discriminant feature would be significant only in a non-parametric test and such a feature would not be considered in the data analysis applied. The paired t-test in Paper II was performed with different, often small, subgroups of participants and different subsets of features. It would be a major task to test each feature separately for normality and make a reasonable conclusion as to which test to apply in each case with so many small datasets. It has been demonstrated in a clinical study that a t-test was robust even for small sample sizes from a distorted distribution with a large proportion of zeros (Sullivan & D'Agostino, 1992). However, it cannot be excluded that some discriminant features are missed in the t-test. To ensure that the findings in the statistical analysis were valid for the purpose of the study, emphasis was placed on feature validation to minimize false positive findings, rather than possible false negatives.

## 4.3 Feature validation

The purpose of the feature validation step in the metabolomics analysis is to reduce the number of artifacts and chance findings among the discriminant features in the statistical analysis to a minimum, in order to focus on the features with the best potential as exposure markers. The term validation when applied as part of the metabolomics analysis therefore should not be confused the general validation criteria for exposure markers listed in Table 1. A feature validation procedure was applied for all features remaining after performing the statistical analyses in Paper I and II by taking advantage of other sources of information in the study designs. In the MEAL study, the different time-points (Urine 1-4) were available and it was investigated if the most discriminant features in the multivariate analysis had a meaningful time-course of excretion after intake of the test meals. In addition, the sensitivity and specificity of the features as markers of one or more meals were investigated. In the INTER study, the sensitivity and specificity of the significant features from the paired t-test were calculated for the full dataset, including all subjects who never reported intake of the foods on the days of urine collection. By including more individuals, information from a larger percentage of the study population was included in the analysis compared to the often small subsets of consumers in the food group on which the statistical analysis was performed. It was also investigated if other reported foods had higher sensitivity and specificity for the feature to take into account possible confounding foods.

The PEMs from the MEAL study were further validated in the FOOD and INTER studies. By investigating if the PEMs were present after intake of different individual foods, it was possible to establish the likely food sources of the PEMs (Paper I). Data from the INTER study was used to investigate if the MEAL PEMs were also related to individual foods in a study setting with a more mixed dietary background and with another study population (Paper II). All feature validation steps for PEMs are illustrated in Figure 3.

By performing the feature validation steps in the MEAL study, the number of discriminant features was reduced by 93 % from 775 to 57 (in this calculation, features present in the PLS-DA model for meals and in the PLS-DA model for protein sources are only included once in the total number before and once in the total number after validation). For the INTER study, the number of features was reduced by 94 % from 568 to 35 (features that came out as significant for more foods are only included once in the calculation of the total number before and once in the total number after validation). When the PEMs from the MEAL study were validated in the INTER study, seven out of 30 unique PEMs were also PEMs in the INTER study, corresponding to a reduction of 77 %. These numbers demonstrate clearly that feature validation procedures can be used to reduce the number of PEMs considerably. Especially for identification, which is very time-consuming, it is of major importance that time is spent well on the features with the best potential as future dietary exposure markers, instead of trying to cover all 1343 features that are of potential interest based solely on the statistical analyses.

# 5. Application of untargeted metabolomics for discovery of exposure markers

## 5.1 Summary and discussion of results from Paper I and II on exposure markers

The PEMs found in the INTER and MEAL studies can be divided into three categories: 1) Identified or tentatively identified PEMs in both studies, 2) identified or tentatively identified PEMs in one study and 3) unidentified PEMs. If the results on exposure markers from this thesis were considered independently, the markers with the best potential are in category one, because PEMs in this category have been found in one study and validated in a complementary study which consolidates the findings. However, it was only possible to validate the PEMs from the MEAL study in the INTER study and not the other way around and other markers from the INTER study may of course be good PEMs as well. The same is true for markers in the MEAL study of foods that were very rarely reported in the INTER study and therefore could not be validated.

To investigate the potential of the markers found as exposure markers, other studies should be taken into account in order to evaluate the validation criteria listed in Table 1. In Table 2, all category one and two PEMs from the MEAL and INTER studies are listed. In the same table, an overview of other relevant research on the same markers is provided. Based on the summary in Table 2, the biological validation criteria in Table 1 will be discussed separately in section 5.1.1-5.1.6 with a focus on how metabolomics studies can contribute to the validation of dietary exposure markers. Analytical validation of the markers will not be considered, as this is not something metabolomic fingerprinting can contribute to.

**Table 2** Identified and tentatively identified PEMs found in the INTER study and/or in the MEAL study

| PEM | Food | Cat | IN/ME/FO | Precursor | Other findings |
|---|---|---|---|---|---|
| ***N*-acetyl-S-(*N*-3-methylthiopropyl)-cysteine (III)** | white cabbage, Brussels sprouts, red cabbage, pointed cabbage | 1 | Y/Y/Y | Glucoibervirin[1,2] | - |
| **AITC-NAC (III)** | white cabbage, Brussels sprouts, red cabbage, horseradish | 1 | Y/Y/Y | Sinigrin[1,2] | **Studies:** Targeted meal[3, 4,5]<br>**Urine samples:** Spot[3, 4], 48 h[5]<br>Time-course of excretion (24 h)[3, 4, 5]<br>Matrix effects[3, 4]<br>Content in foods[3, 4, 5] |
| **IB-NAC (III)** | white cabbage, Brussels sprouts, red cabbage, pointed cabbage | 1 | Y/Y/Y | Glucoiberin[1,2] | Content in foods[4] |
| **SFN-NAC (I)** | white cabbage, Brussels sprouts, red cabbage, | 1 | Y/Y/Y | Glucoraphanin[1,2] | **Studies:** Targeted meal[1,4]<br>**Urine samples:** spot[1,4], 24 h[1]<br>Time-course of excretion (24 h)[1,4]<br>Matrix effects[4]<br>Content in foods[4] |
| **4-iminopentyl-isothiocyanate (III)** | Brussels sprouts | 1 | Y/Y/NF | 4-iminopentyl-glucosinulate (?) | - |
| **TMAO (I)** | cod, smoked mackerel, monkfish, Pollock, halibut | 1 | Y/Y/NI | TMAO[6], TMA[6] Choline[6], Phosphatidylcholine[7] Carnitine[6] | **Studies:** Untargeted cohort[8], Untargeted intervention[9, 10], Untargeted meal[11,12], Targeted meal[6], Targeted observational[13]<br>**Urine samples:** spot[6, 8, 11, 12, 13], 24 h[9, 10]<br>**Analytical method:** NMR[8, 9, 10, 12], MS[11]<br>Content in foods[14, 15, 16] |
| **Proline betaine (I)** | citrus fruits | 2 | Y/NI/NI | Proline betaine[17] | **Studies:** Targeted meal[18], Untargeted meal[12, 17, 19, 27], Untargeted intervention[19, 20], Untargeted cohort[17, 19, 20, 21]<br>**Urine samples:** spot[27, 12, 18, 19, 20, 21, 17], 24 h[17, 19]<br>Time-course of excretion (24 h)[17, 18]<br>**Analytical method:** NMR[12,17], MS[27, 19,20, 21]<br>Content in foods[17, 18, 22] |

**Table 2** (continued)

| PEM | Food | Cat | IN/ME/FO | Precursor | Other findings |
|---|---|---|---|---|---|
| **Hesperetin glucuronide (I)** | citrus fruits | 2 | Y/NI/NI | Hesperidin[23] | **Studies:** Targeted meal[24,25], Targeted intervention[23], Targeted cohort[26] , Untargeted meal[19, 27], Untargeted intervention study[19], Untargeted cohort[19]<br>**Urine samples:** spot[19, 23, 24, 25, 27], 24 h[19, 24, 25, 26]<br>Time-course of excretion 24 h[24], 48 h[28, 29]<br>**Analytical method:** MS[19, 27]<br>Content in foods [30, 31] |
| **4-ethyl-5-amino-pyrocatechol sulfate (III)** | beetroot | 2 | Y/NI/NI | Betanin (?) | - |
| **3-hydroxyhippuric acid sulphate (II)** | red cabbage | 2 | Y/NI/NI | Same as 3-hydroxyhippuric acid (?) | - |
| **3-hydroxyhippuric acid (II)** | red cabbage | 2 | Y/NI/NI | Various polyphenols[32] | **Studies:** Targeted meal[32, 33], Targeted intervention[34, 35], Untargeted intervention[36]<br>**Urine samples:** spot[32, 33, 34, 35], 24 h[36, 35]<br>Time-course of excretion (48 h)[35]<br>**Analytical method:** MS[36] |
| **5-hydroxyindole-3-acetic acid (I)** | walnut | 2 | Y/NI/NI | Serotonin, Tryptophan[37] | **Studies:** Targeted meal[37,38], Untargeted intervention study[39]<br>**Urine samples:** spot[37], 24 h[38, 39]<br>Time-course of excretion (40 h)[37]<br>Content in foods[38]<br>**Analytical method:** MS[39] |
| **6-AMMU (II)** | chocolate | 2 | Y/NI/NI | Theobromine[40] | **Studies:** Untargeted meal[41], Untargeted intervention[42]<br>**Urine samples:** spot[41], 24 h[41, 42]<br>**Analytical method:** MS[41, 42] |
| **Theobromine (I)** | chocolate | 2 | Y/NI/NI | Theobromine[40]<br>Caffeine[43] | **Studies:** Untargeted meal[41], Untargeted intervention[42, 44], Targeted meal[45]<br>**Urine samples:** spot[41, 44, 45], 24 h[41, 42]<br>**Analytical method:** MS[41, 42], NMR[44] |
| **7-methyluric acid (I)** | chocolate | 2 | Y/NI/NI | Theobromine[40], | **Studies:** Untargeted meal[41], Untargeted intervention[42]<br>**Urine samples:** spot[41], 24 h[41, 42]<br>**Analytical method:** MS[41, 42] |
| **N-acetyl-cysteine conjugate (III)** | white cabbage, Brussels sprouts | 2 | N/Y/Y | Glucosinulate (?) | - |

**Table 2** (continued)

| PEM | Food | Cat | IN/ME/FO | Precursor | Other findings |
|---|---|---|---|---|---|
| **ERN-NAC (III)** | white cabbage, Brussels sprouts | 2 | NF/Y/Y | Glucoerucin, Glucoraphanin[1, 2, 4] | **Studies:** Targeted meal[4]<br>**Urine samples:** spot[4]<br>Time-course of excretion (24 h)[4]<br>Matrix effects[4]<br>Content in foods[4] |
| **BITC-NAC (I)** | white cabbage, Brussels sprouts | 2 | N/Y/Y | Glucotropaeolin[1,2] | **Studies:** Targeted meal[4, 46]<br>**Urine samples:** spot[4,46], 24 h[46]<br>Time-course of excretion (24 h)[4, 46]<br>Content in foods[4] |
| **SFN-CYS (III)** | white cabbage, Brussels sprouts | 2 | N/Y/Y | Glucoraphanin[1,2] | **Studies:** Targeted meal[1]<br>**Urine samples:** spot[1], 24 h[1]<br>Time-course of excretion (24 h)[1]<br>Content in foods[4] |

**PEM**, **Food**, **Cat** and **IN/ME/FO** summarize the results from the INTER, MEAL and FOOD studies. PEM is the potential exposure marker (roman numbers I-III designates the level of the identification as defined by Sumner et al. (2007). Food is a list of foods for which the PEM was found as a marker. Cat is the PEM category as defined in section 5.1, and IN/ME/FO designates if the PEM was found in the INTER, MEAL and FOOD studies, respectively by the following letters: Y (yes), N (no), NF (not found), NI (Not investigated).
**Precursor** and **other findings** summarize relevant knowledge from literature. Precursor is the precursor found in food to the PEM. Other findings are divided into six subcategories: **Studies** provides an overview of the type of studies carried out where the PEM has been found. It is divided into untargeted metabolomics studies and targeted studies (none of the targeted studies were metabolomics studies in this case) and the study types: meal, intervention and cohort. **Urine samples** is the type of urine samples analyzed in the studies. Spot is a urine sample from one void or one or more short time intervals. 24 h refers to 24 h urine collections. **Time-course of excretion** indicates if the time-course of excretion has been investigated for the PEM. Only time-courses of excretion after food intake (not the pure compounds) are included. **Analytical method** is only given for metabolomics studies. **Matrix effects** refers to studies where the PEM has been investigated and compared between different food matrices (in this case raw and boiled food). **Content in foods** designates if the content of the PEM (or where relevant, the precursor) has been measured in different foods.
**Abbreviations:** AITC-NAC: *N*-acetyl-S-(*N*-allylthiocarbamoyl)cysteine; BITC-NAC: N-Acetyl-(*N*-benzylthiocarbamoyl)-cysteine ; IB-NAC: Iberin *N*-acetyl-cysteine; SFN-CYS: Sulforaphane N-cysteine; SFN-NAC: Sulforaphane *N*-acetyl-cysteine; ERN-NAC: Erucin N-acetyl-cysteine ; TMAO: Trimethylamine *N*-oxide; TMA: Trimethylamine; 6-AMMU : 6-amino-5-[N-methylformylamino]-1- methyluracil.

**References:** The numbers given in the column **'other findings'** corresponds to the following references: [1]Janobi et al., 2006, [2]Clarke, 2010, [3]Rouzaud et al., 2004, [4]Vermeulen et al., 2006, [5]Jiao et al., 1994, [6]Zhang et al., 1999, [7]Wang et al., 2011, [8]Xu et al., 2010, [9]Stella et al., 2006, [10]Rasmussen et al., 2012b, [11]Lloyd et al., 2011b, [12]Heinzmann et al., 2011, [13]Svensson et al., 1994, [14]Chung & Chan, 2009, [15]Zeisel et al., 2003, [16]Seline &Johein et al., 2007, [17]Heinzmann et al., 2010, [18]Atkinson et al., 2007, [19]Pujos-Guillot et al., 2013, [20]May et al., 2013, [21]Lloyd et al., 2011a, [22] de Zwart et al., 2003, [23]Franke et al., 2005, [24]Erlund et al., 2001, [25]Ito et al., 2005, [26]Brantsæter et al., 2007, [27]Favé et al., 2011, [28]Manach et al., 2003, [29]Krogholm et al., 2010, [30]Harnly et al., 2006, [31]Aturki et al., 2004, [32]Rechner et al., 2001, [33]Rechner et al., 2002b, [34]Rechner et al., 2002a, [35]Hollands et al., 2008, [36]van Dorsten et al., 2010, [37]Helander et al., 1992, [38]Feldman & Lee, 1985, [39]Tulipani et al., 2011, [40]Rodopoulos et al., 1996, [41]Llorach et al., 2009, [42]Llorach et al., 2013, [43]Rodopoulos & Norman, 1996, [44]Martin et al., 2012, [45]Ptolemy et al., 2010, [46]Mennicke et al., 1988.

### 5.1.1 Known relation to the exposure

For most of the PEMs in Table 2, the precursor or precursors of the compounds are known. This is a prerequisite to establish a causal association between food consumption and urinary excretion of the exposure marker. Only for a few tentatively identified PEMs (4-iminopentylisothiocyanate, 4-ethyl-5-amino-pyrocatechol sulfate, 3-hydroxy-hippuric acid sulphate and N-acetyl-cysteine conjugate) that were not reported in other studies, it was not possible to confirm the suggested precursor. It is important to keep in mind that Table 2 only contains identified and tentatively identified compounds found in the INTER and MEAL studies. The fact that the route of excretion for the majority of PEMs is already known is therefore mainly because previously reported compounds are easier to identify. For unidentified compounds, it is difficult to causally link the PEM to intake of any food, which is crucial in the validation process.

Exposure markers with several precursors (3-hydroxyhippuric acid, 5-hydroxyindole-3-acetic acid, TMAO and theobromine, in Table 2) are difficult to interpret, as the excretion of those depends on exposure to all the precursors which may originate from various food sources. A classic example of this is hippuric acid and 3- and 4-hydroxyhippuric acids. These compounds are microbial products of various polyphenols (Rechner et al., 2002a) and therefore not suitable to discriminate between intake of individual foods. For example 3-hydroxyhippuric acid has been found to increase after intake of coffee (Rechner et al., 2001), blackcurrant products (Rechner et al., 2002b; Hollands et al., 2008) and extracts of wine and grape juice (van Dorsten et al., 2010), while it was found as a marker of red cabbage in the INTER study. For other compounds, such as 5-hydroxyindole-3-acetic acid and theobromine, one precursor, in this case serotonin (Helander et al., 1992) and theobromine (Rodopoulos & Norman, 1996), respectively, is by far the dominating one. Despite this, the content of these precursors in other food sources must be known to investigate, if they can be neglected when the markers reach certain concentrations. For TMAO, the compound itself and TMA are the main precursors but the contribution from carnitine and choline-containing compounds which are converted to TMAO in the gastrointestinal tract, may be considerable (Svensson et al., 1994; Zhang et al., 1999). With the exception of ERN-NAC, which may be produced endogenously from sulforaphane (Vermeulen et al., 2006), the isothiocyanate mercapturic acids (*N*-acetyl-S-(*N*-3-methylthiopropyl)-cysteine, AITC-NAC, IB-NAC, SFN-NAC, BITC-NAC and SFN-CYS) have one precursor each.

Only proline betaine and theobromine of the PEMs in Table 2 are excreted unmetabolized. Unmetabolized compounds are typically very good candidates as exposure markers because they are less dependent on inter-individual variation in the metabolic pathway. While proline betaine is considered to be an inert metabolite (Heinzmann et al., 2010), a large proportion of the ingested theobromine is metabolized to other products (Rodopoulos et al., 1996) and the recovery of theobromine in urine may therefore exhibit considerable inter-individual variation (see section 6.3.4. for a further discussion of this).

### 5.1.2 Sensitivity and specificity

The sensitivity and specificity of a PEM is mainly determined by the distribution of the compound or precursor(s) in different foods, the normal range of consumption of these foods and the metabolism of the compound. Which foods the subjects are exposed to during a study depends on the dietary restrictions applied, the dietary habits of the subjects and the intervention diet. All these factors vary between nutrition studies and a food exposure marker found in several untargeted metabolomics studies obviously has a stronger potential than a marker found only in individual highly controlled untargeted studies or in targeted studies. Presence of a marker in several untargeted metabolomics studies, especially cohort studies, is per se an indication of a marker with a high sensitivity and specificity. Despite this, it is not straightforward to compare marker findings between metabolomics studies as the metabolite coverage of the analytical methods applied and the procedures for data analysis are not uniform. How important methodological issues are for the results obtained from different metabolomics studies is not easy to assess at this stage since there is still not a large number of studies published dealing with exposure marker discovery. However, Table 2 clearly demonstrates the potential of untargeted metabolomics to validate food exposure marker findings.

*Proline betaine and hesperetin glucuronide*

The best documented marker discovered by untargeted metabolomics is proline betaine. This compound has been found consistently as a marker of citrus consumption in all the listed untargeted meal-, intervention- and cohort studies in Table 2, except from one study by May et al. (2013), where it was found as a marker of a fruit and vegetable diet including citrus, soya foods and cruciferous vegetables. However, the finding of proline betaine in this study is most likely solely caused by citrus being part of the intervention diet. In two untargeted cohort studies, the sensitivity and specificity of proline betaine as a marker of citrus consumption have been calculated and in both studies, the sensitivity and specificity were above 80 % for classification of low/high citrus consumers or non-consumers/consumers (Heinzmann et al., 2010; Lloyd et al., 2011a). A sensitivity and specificity above 80 % was also obtained in the INTER study as this was one of the validation criteria for PEMs (section 4.3). Hesperetin glucuronide is another marker that has been found to be related to citrus consumption in untargeted meal, intervention and cohort studies (Table 2). Interestingly, hesperetin has already been investigated in a targeted cohort study on pregnant women (Brantsaeter et al., 2007). In this study, a significant correlation was found between reported intake of citrus in WDR and FFQ, and hesperetin (with and without) glucuronide measured in 24 h urine, confirming the findings by untargeted metabolomics. For hesperidin and proline betaine, the content of the compounds in other commonly consumed foods is negligible (de Zwart et al., 2003; Harnly et al., 2006). Orange is the citrus fruit with the highest content of both compounds.

*Theobromine, 7-methyluric acid, 6-AMMU and 5-hydroxyindole-3-acetic acid*

Other promising PEMs in Table 2 that seem to be found consistently by untargeted metabolomics studies are theobromine, 7-methyluric acid and 6-AMMU as markers of chocolate intake and 5-hydroxyindole-3-acetic acid as a marker of nuts or walnuts (Llorach et al., 2009; Tulipani et al., 2011; Llorach et al., 2013). In one study, however, only theobromine (Martin et al., 2009) was found as a marker of chocolate and not 7-methyluric acid and 6-AMMU which may be caused by this study being the only one applying NMR and not MS for the metabolomic fingerprinting.

Chocolate is the dominant food source of theobromine (Shively & Tarka, 1984) but theobromine can also be produced endogenously as a metabolite of caffeine. Serotonin is found in a range of commonly consumed foods (Feldman & Lee, 1985). Even though walnuts contain the highest amounts of serotonin, it remains to be investigated, if consumption of other serotonin rich foods, for example bananas (Feldman & Lee, 1985; Helander et al., 1992) in absence of walnuts may give rise to 5-hydroxyindole-3-acetic acid as an exposure marker as well.

*TMAO*

TMAO is a marker that is commonly found in untargeted metabolomics studies but for which there is not complete consistency in the findings. One meal study confirm the findings in the MEAL and INTER study of TMAO as a fish marker (Lloyd et al., 2011b), while TMAO has been found as a marker of non-vegetarian diets, animal protein and diets high in protein in other studies (Stella et al., 2006; Xu et al., 2010; Heinzmann et al., 2011; Rasmussen et al., 2012b). In the study by Rasmussen et al. (2012b), TMAO in 24 h urine was correlated to 24 h urinary nitrogen which supports the finding of TMAO as a marker of protein intake. Unfortunately, only the macronutrient composition of the diet was monitored in this study and it is not possible to assess the fish intake, which would have been interesting. In the study by Stella et al. (2006), fish intake was not part of the experimental diet, demonstrating that TMAO is also related to high meat consumption. This is explained by the other precursors of TMAO. Choline, phosphatidylcholine and especially carnitine are mainly found in meat products (Zeisel et al., 2003; Seline & Johein, 2007), while TMAO is present only in fish. The content of TMAO in fish varies considerably from ten to more than 1000 mg/kg (Chung & Chan, 2009). TMAO is virtually absent from freshwater fish and may reach very high levels in marine fish. This difference between fish species was also observed in the results from Paper II. In a targeted meal study with 46 foods, the urinary excretion of TMAO (0-8 h after the meal) was much higher following fish intake compared to meat intake (Zhang et al., 1999) and in a targeted observational study, TMAO levels in overnight urine correlated with the self-reported habitual weekly fish intake (Svensson et al., 1994). However, the concentration of TMAO in urine was not significantly higher in moderate and high consumers compared to non-consumers of fish in this study. Overall, TMAO is probably a good fish- or meat protein marker in some study settings but because of the various dietary sources and precursors of TMAO, more studies are needed to establish under which circumstances TMAO can be used as a marker.

*Isothiocyanate mercapturic acids*

Few metabolomics studies have been conducted on cruciferous vegetables and none of them have confirmed the finding of isothiocyanate mercapturic acids as PEMs (Edmands et al., 2011; May et al., 2013). In Edmands et al. (2013), NMR was applied which may explain why the same markers are not found and in May et al. (2013), identification was only performed by searches in the Human Metabolome Database (Wishart et al., 2009), where isothiocyanate mercapturic acids are not included yet. While isothiocyanate mercapturic acids have known and specific precursors, that are almost only present in foods from the *Brassica* species, the distribution of the precursors vary between *Brassica* vegetables (Vermeulen et al., 2006). These markers are therefore probably not specific to individual *Brassica* species, which is also the case for the findings in the MEAL and INTER study. In addition, it has been demonstrated that the excretion of isothiocyanate mercapturic acids depends on the food matrix. A higher concentration of isothiocyanate mercapturic acids is found following intake of raw compared to boiled cruciferous vegetables (Vermeulen et al., 2006).

### 5.1.3 Dose-response

Dose-response is rarely taken into account in metabolomics studies. Meal studies are generally performed with one dose and for cohort studies, low or non-consumers are generally compared to high consumers to obtain a high contrast in the model. In Lloyd et al. (2011a), a tendency for a dose-response relationship between proline betaine in urine and citrus intake reported in FFQ was demonstrated for three intake levels. It is questionable, however, if semi-quantitative measurements are appropriate for investigating dose-response. As illustrated in Paper II, there may be large inter-individual variation between subjects for the same reported food intakes and, unless a large dataset is available, it is not a good criterion for evaluation of the potential of an exposure marker. For this reason, dose-response was also not included in the validation of PEMs as markers in Paper II. Dose-response should be investigated in a targeted approach and a controlled study setting. The only PEM in Table 2 for which this has been done is hesperetin (with and without glucuronide), where a dose-response relationship has been demonstrated in two studies, which together cover an intake range of zero to one liter orange juice (Manach et al., 2003; Brevik et al., 2004). For discovery of food exposure markers, it may be valuable to include different doses in the intervention diets of controlled cross-over studies. For example a mixed meal could be prepared with varying contents of each ingredient in turn. This would enable an investigation of dose-response as well as possible confounding effects on a marker of other ingredients.

### 5.1.4 Inter-individual variation

Large inter-individual differences have been measured for several of the PEMs in Table 2. For example, 3-hydroxyhippuric acid (Rechner et al., 2002a) and TMAO (Zhang et al., 1999) vary considerably between individuals, probably due to the dependency of gut microbiota for production of these compounds. For TMAO the conversion from TMA to TMAO in the liver is impaired in some

persons, a condition known as trimethylaminuria or the fish odor syndrome, due the fishy odor of TMA. Presence of trimethylurea in some subjects may therefore also influence the findings for this exposure marker, even though the prevalence of the condition is low (Rehman, 1999). At present, inter-individual differences are rarely considered in the analysis of dietary studies by metabolomic fingerprinting even though there may be a great value of including person characteristics such as sex, genotype or phenotype in the analysis, as some exposure markers may be valid only in well-defined subpopulations. Heinzmann et al. (2011) demonstrated the potential of metabolomic fingerprinting for this purpose and found that inter-individual variation in the choline degradation pathway affected the response to a dietary meal challenge. In another study, genetic and environmental determinants linking dietary characteristics to metabolite concentrations in fasting serum samples were separated by studying monozygotic twins (Menni et al., 2013). This approach revealed that only about 25 % of the metabolites found to be associated to the diet were not genetically determined.

### 5.1.5 Time of exposure

The time-course of excretion has been investigated for most of the markers in Table 2 but, unfortunately, the majority of the studies performed only cover up to 24 h after consumption, which is not enough in all cases to return to baseline levels. Hesperetin glucuronide, proline betaine and 5-hydroxyindole-3-acetic acid are almost completely excreted within 24 h (Helander et al., 1992; Erlund et al., 2001; Heinzmann et al., 2010). The same is true for isothiocyanate mercapturic acids, even though the excretion patterns over 24 h for these compounds depends on the preparation method with a later peak in excretion for cooked cabbage compared to raw cabbage (Rouzaud et al., 2004). For TMAO, the peak in excretion is probably around 2-4 hours after consumption for fish, as shown in Paper I, but the excretion may not be complete after 24 h and there may be differences in the rate of excretion following fish and meat consumption. If that is true, it could explain why TMAO is a marker of fish in acute studies (Lloyd et al., 2011b) and a marker of protein in other intervention and cohort studies (Stella et al., 2006; Rasmussen et al., 2012b). In both Paper I and II, acute markers are favored in the data analysis because urine was collected on the same day as the food was consumed.

It is not possible to distinguish if PEMs are short- or long term markers, unless the time-course of excretion is investigated. In cohort studies, slowly excreted markers may be favored because the timing of sampling and food intake is not controlled. However, slowly excreted markers cannot be distinguished from acute food markers of frequently consumed foods in an observational study. The other way around, acute markers found in the INTER study may also be slowly excreted, if a food is not eaten frequently. Even though I have not been able to find data on the time-course of excretion for the chocolate markers in Table 2, a study of theobromine ingestion has demonstrated that theobromine and the theobromine metabolites are not completely excreted within 24 h (Rodopoulos et al., 1996). In a study with proline betaine, differences were found in the excretion pattern and the amount of proline betaine excreted between orange juice and a comparable dose of proline betaine

added to apple juice (Atkinson et al., 2007). The excretion of theobromine taken as the pure compound and theobromine as a constituent of chocolate may therefore not be the same. If the excretion of theobromine from chocolate also extends beyond 24 h, however, it indicates that studies favoring acute markers like the INTER study, can also provide information on markers that are excreted more slowly. The finding of proline betaine and hesperetin as markers in cohort studies (Heinzmann et al., 2010; Pujos-Guillot et al., 2013), even though they are excreted fast in urine, is probably because citrus is consumed frequently in the populations studied. It would have been interesting to carry out an analysis based on reported foods that represents dietary habits better than a single day food record in the INTER study in order to explore if the PEMs found would then be supplemented with more slowly excreted markers.

### 5.1.6 Population

The majority of the metabolomics studies referred to in Table 2 have been conducted on healthy normal or overweight adult males and females. It is important that an exposure marker is only applied in a population where it has been validated but this is less relevant in the discovery phase even though it should be taken into account when comparing results between metabolomics studies.

### 5.1.7 Summary of the validity of exposure marker findings in Paper I and II

The short review on the knowledge regarding PEMs in Table 2 in relation to biological validation criteria for food exposure markers presented in section 5.1.1-5.1.6 demonstrates that untargeted metabolomics is a valuable tool for exposure marker discovery. Except from a few PEMs for cabbage and beetroot, that have not been reported previously, all PEMs in Table 2, are supported by other studies of which a large part is untargeted metabolomics studies. This demonstrates consistency in findings by metabolomic fingerprinting despite applications of a range of analytical parameters and strategies for data pretreatment and analysis. Several PEMs are found in various study designs including cohort studies, indicating that a metabolomics approach can provide a shortcut to markers with high sensitivity and specificity for individual foods. Based on the biological validation criteria in Table 1, proline betaine and hesperetin-glucuronide are supported by most evidence as exposure markers at present among the PEMs found in the INTER and MEAL studies. Both of these are almost unique to citrus fruits and are found as markers in a large number of studies. Interestingly, a lot of work had been done on hesperetin as an exposure marker of fruits and citrus in targeted studies before introduction of metabolomic fingerprinting. The finding of hesperetin-glucurinide by untargeted metabolomics therefore consolidates that the method works for exposure marker discovery. On the other hand, proline betaine is an example of a marker, where the potential of untargeted metabolomics has been fully utilized and where the marker has been put forward as an exposure marker almost solely based on evidence from untargeted metabolomics studies. Targeted studies on dose-response and inter-individual differences are needed as the last step before proline betaine can be applied in nutrition studies. Theobromine and theobromine metabolites as markers of chocolate and 5-hydroxy-indole-3-acetic acid as a marker of walnuts also seem promising but for

those, the contribution of other food sources needs to be further elucidated. There may well be markers that can only be found in high concentration after intake of one food because the intake required to reach the same levels for other foods exceeds habitual intakes. TMAO is a controversial marker for which much more work needs to be done in order to find out how and if it can be applied as an exposure marker. The large inter-individual variation and the many precursors of this metabolite complicate the validation procedure considerably.

It is important to keep in mind that the presented markers in Table 2, only represent the markers that were identified or tentatively identified in the INTER and MEAL studies and that these markers were found, following very strict criteria for marker validation within the dataset (section 4.3). The findings therefore reflect very strong markers and such markers are of course more likely confirmed in other studies. For several of the foods in Table 2, other metabolomics studies have found a much higher number of markers and it would have been interesting to investigate how these markers perform in a study like the INTER study. Are they not found in the statistical analysis? Are they eliminated during validation? Or are they simply not detected? This way of targeting the metabolomics analysis based on previous findings could support a faster validation of previous findings as well as guide the criteria for marker validation in a dataset. The border distinguishing markers from non-markers in metabolomics is often fluent and markers may be lost that might have been acceptable if the criteria for marker selection had been modified slightly. The fact that highly controlled studies generate more markers in a metabolomics analysis than cohort studies is probably not only caused by markers from intervention studies being less specific and dependent on a high dose but also the fact that cohort studies rely on dietary records and semi-quantitative measurements which implies large sources of errors in the data. Probably a higher number of markers would be valid in observational studies if a set of markers was investigated in a targeted way.

Overall, metabolomic fingerprinting offers a short cut to evaluating some of the most essential biological validation criteria defining a good exposure marker: Sensitivity and specificity, time-course of excretion and validity across populations. Another very important criterion, where the potential of untargeted metabolomics is enormous, is for investigation of inter-individual variation. However, inter-individual variation is at present most often not considered and rather used as a limitation for marker selection than a way to identify subgroups who respond differently. There may be a potential for exploring dose-response relationships by untargeted metabolomics, especially in controlled dietary intervention studies but this is another area that is not studied yet.

Other biological validation criteria call for a targeted approach. Establishing a causal link between exposure and a PEM relies on knowledge of metabolic pathways from targeted studies. Also, investigations of dose-response and possible matrix effects are more feasible to determine by performing quantitative measurements. Targeted studies should be conducted for the final validation of the exposure markers found by metabolomics as soon as the markers have a proven potential. So far, the work of validating outcomes of untargeted metabolomics studies in targeted analyses is lacking behind. I have not been able to find a single published study where this has been done.

## 5.2 Unidentified PEMs

Category three (section 5.1) of unknown or unidentified PEMs consists the largest group of PEMs found in the INTER and MEAL studies. In total, 67 % and 41 % of PEMs in the MEAL and INTER studies, respectively, were unknowns. Unknown compounds are not necessarily artifacts or poor markers but they are extremely difficult to validate. When a PEM is identified, it is often possible to link it directly to intake of a food and this is a prerequisite for any exposure marker as discussed in section 5.1.1. Knowing the identity of a compound also makes it easier to understand which other foods may give rise to the same marker and a known marker can be validated further in a quantitative analysis as opposed to unknown compounds. It is hard, though not impossible, to confirm if an unknown compound has been found in other metabolomics studies. Only the m/z of an unknown marker can be compared, unless the analytical method is the same and m/z is not unique for any compound and therefore not sufficient to conclude from. Since the same analytical method was applied in the MEAL and INTER studies, findings of unidentified PEMs could also be validated by comparing m/z and RT. In this way, one unknown PEM from the MEAL study was found to be a marker also in the INTER study (Paper II). It was also possible to compare two unidentified markers of citrus in INTER to unknown citrus markers published in another metabolomics study (Paper II). This successful comparison of unknowns between studies demonstrates that there may be valuable information to share, even if the identity of a compound is not known yet. Unfortunately, unknown compounds are not always reported in metabolomics studies and a large proportion of the knowledge gained from these studies is therefore unavailable for other research groups at present. Especially for urinary exposure markers, it is expected that hitherto undescribed compounds will be found by untargeted metabolomics. The possible metabolites that can be produced from intake of foods are vast and largely unknown because studies conducted before it was possible to use metabolomic fingerprinting were targeted. Metabolomics can and should contribute with new findings but as long as the process of identification is not given much more attention, a large proportion of perfectly qualified findings is lost. This is a major issue that I will return to in the perspectives, section 8.

# 6. Application of untargeted metabolomics for development of compliance measures

The aim of applying untargeted metabolomics data as a compliance measure is to investigate if it is possible to identify non-compliant subjects to a dietary pattern by an explorative multivariate approach. There is a large potential in identifying non-compliant subjects as inclusion of those in the data analysis may obscure the findings in a nutrition study. In a similar manner as for the discovery of exposure markers of individual foods described in section 5, the idea behind the development of a compliance measure is to find out which metabolites characterize a diet as a whole. A few metabolomics studies have been conducted to identify markers of habitual dietary patterns (Peré-Trepat et al., 2010; Altmaier et al., 2011; O'Sullivan et al., 2011; Floegel et al., 2013; Menni et al., 2013) but the studies are too diverse in design and analytical approach to compare if there is any consistency in the dietary patterns found and the associated metabolites. Even though the results from the studies demonstrate that dietary habits are reflected in urine, serum and plasma, it is unknown how robust a dietary pattern as determined by metabolomic fingerprinting is over time for an individual subject and how generalizable dietary patterns are across study populations. It may well be that a compliance measure which reflects a complex interplay between characteristic foods in a diet over time may not simply be the sum of exposure markers of individual foods. In support of this, the markers identified in a study conducted by O'Sullivan et al. (2011) on urine, reflected general dietary traits such as diets high in meat and vegetables rather than individual foods. A compliance measure based on urinary metabolic fingerprints may therefore be superior to use of individual food exposure markers for identifying subjects that have been non-compliant to a complex diet or subjects who respond differently to a dietary pattern.

## 6.1 Summary of the compliance model developed in Paper III

In Paper III, a PLS-DA model has been developed to distinguish the two dietary patterns NND and ADD in 24 h urine samples. Urine samples from week 12 and 26 of one hundred and seven subjects in the INTER study, for which dietary records were available, were used to develop the model. Afterwards, the model was validated with 139 other samples, representing all sample points (week 4, 12, 20 and 26) in INTER. The final PLS-DA model consisted of four latent variables and included 67 features from 52 metabolites. The cross validation error of the model was 0.10 and the misclassification error for the validation set was 0.19, which corresponds to 26 samples being misclassified. Twenty-one of the metabolites used in the model were identified or tentatively identified. A higher number of metabolites in the model were characteristic of an ADD compared to the NND and these ADD metabolites were also ranked higher in the model based on VIP scores. Metabolites related to intake of citrus, chocolate, heat treated foods and probably animal protein and foods containing medium chain fatty acids characterized the ADD diet in urine. The metabolites in the model characterizing the NND diet were related to intake of fish as well as high intake of fruit and vegetables. More details on the PLS-DA model and the identified features can be found in Paper III.

## 6.2 Limitations of the PLS-DA model for compliance

Paper III represents a first step in the development of a compliance measure to distinguish between the NND and the ADD. In the paper, it was demonstrated that the identified metabolites characterizing consumption of the two diets mainly reflect food items and food groups that are either unique for one diet or that are consumed more frequently in one diet compared to the other. This, together with the low misclassification rate for the validation set, is a good indication that the model contains relevant information to estimate compliance to the diets. However, there are still a large number of unidentified metabolites in the model and therefore it is not fully transparent how the diets are reflected by the selected features. In addition, the samples used to develop the model do not necessarily represent compliant subjects. The subjects whose samples were included in the model had provided urine samples as well as dietary records at three sampling points in the study. Even though there may be some relation between dietary compliance and the likelihood of a subject providing all requested samples and dietary records, it is not an optimal selection procedure. If samples from non-compliant subjects are included in the model, it implies that information from samples that are not representative of the correct dietary pattern will be taken into account during model development and this is of course problematic. Ideally, compliant subjects to the dietary intervention should be identified to select the best samples for model development. Some common tools for measuring compliance were applied in the study. For example, 24 h nitrogen excretion and PABA recovery were measured in all urine samples and energy intakes (EI) as estimated by the foods registered in the supermarket at the department and by WDR, were also available. However, the recommended number of urine collections needed to use nitrogen excretions as a compliance measure of protein intake at an individual level is eight (Bingham, 2003), and only four were available in the study. Also, calculations of EE are associated with uncertainty when the doubly-labeled water method is not used, especially when applied at an individual level (Livingstone & Black, 2003; Hall et al., 2011). Even though it may have been possible to exclude samples from some subjects before developing the model based on the ratio EI/EE and nitrogen measures, there would not be much evidence that these subjects actually represent the least compliant subjects. For a study with a main interest on the food composition of a diet, compliance measures of food intake would be more relevant than macronutrient and energy based measures but, as already described, such exposure markers are still lacking. It is therefore questionable if it would have been possible to make a better selection of samples for model development.

A main problem with the compliance model is that it is not possible to distinguish if misclassified samples in the model are from truly non-compliant subjects or simply represent other subject or sample characteristics that are not directly related to dietary compliance. In order to better understand what characterizes the misclassified samples in the model, an analysis of the misclassified samples is carried out in section 6.3.

## 6.3 Analysis of misclassified samples in the PLS-DA model

In Paper III, 26 misclassified samples were identified in the validation set but as discussed in section 6.2, there may as well be samples among the model samples from non-compliant subjects. To include those, the model samples were used as test set in the final PLS-DA model and the misclassified samples among the model samples were combined with the misclassified samples from the validation set. Even though the misclassified model samples represent an underestimation because the model is already based on these data, it is better to include at least the most extreme samples among the model samples than not to include any information on these samples.

The new selection of misclassified samples in the model is highlighted in Figure 6. In total, 30 samples were misclassified, 25 NND samples and five ADD samples. In order to investigate how misclassified samples deviate from correctly classified samples in the model, a comparable subset of correctly classified samples was randomly selected from all correctly classified samples so that the numbers of samples from the model and the validation set as well as the two dietary groups were the same in the correctly classified subset as for the misclassified samples. Only subjects, for whom all samples were correctly classified, were included in the correctly classified subsets.

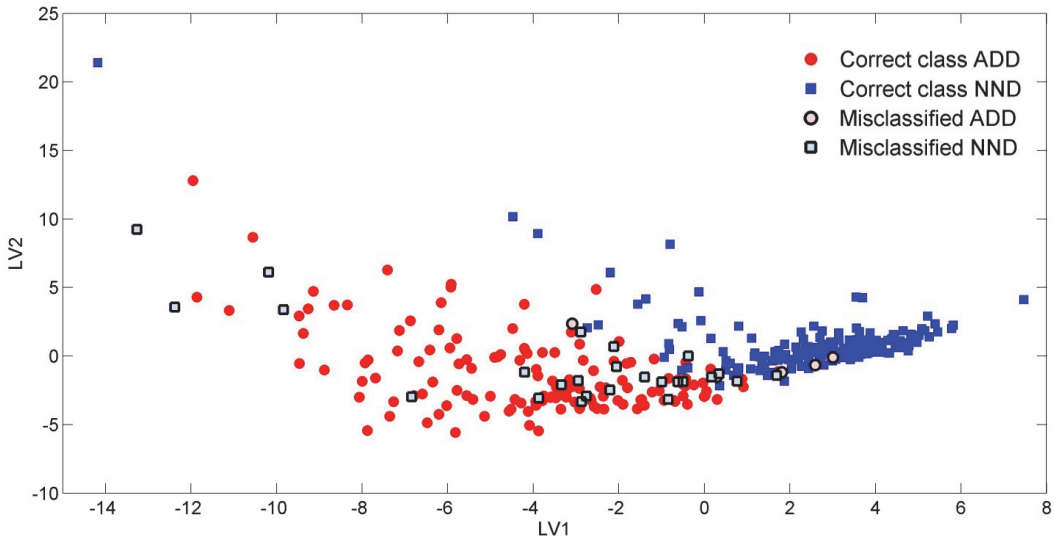



**Figure 6** Score plot of the first two latent variables in the PLS-DA model for compliance. Misclassified samples, when the model was applied on the full dataset, are highlighted.

The selected correctly classified groups of samples from NND and ADD have been compared to the respective misclassified groups for a range of parameters related to: a) Distribution and completeness of urine samples (section 6.3.1), b) Person characteristics and registered foods collected from the shop (section 6.3.2), c) Biological measures (section 6.3.3) and d) Exposure markers of individual foods (section 6.3.4). For categorical variables, a chi square test was used to test, if the misclassified and correctly classified samples were from the same distribution. For continuous variables, an unpaired t-test was used. A list of detailed information on the individual correctly classified and misclassified samples is provided in Appendix A.

### 6.3.1 Distribution and completeness of urine samples

The PLS-DA model is based on 24 h urine samples and the nature of these samples can potentially affect the likelihood of a sample being misclassified. The effect of sample completeness, as estimated by PABA recovery, was compared between correctly classified and misclassified samples. Since NND is a seasonal diet, samples collected during some seasons may be easier to classify in the model than samples from other seasons and this was also tested. Finally, it was investigated, if it was important for the ability of the model to classify a sample correctly, if the sample was from a subject from whom samples had been included to develop the model. An overview of the tests performed is given in Table 3.

**Table 3** Parameters related to the distribution and completeness of urine samples

| | **NND** | | | **ADD** | | |
|---|---|---|---|---|---|---|
| | Misclassified | Correctly classified | P-value | Misclassified | Correctly classified | P-value |
| n | 25 | 25 | | 5 | 5 | |
| Season (Winter/Spring/Summer/Autumn) | 15/7/1/2 | 14/7/1/3 | 0.97 | 4/1/0/0 | 5/0/0/0 | 0.29 |
| **No of incomplete samples (PABA recovery <85 %)** | 14 | 7 | **0.045** | 1 | 2 | 0.49 |
| No of samples for which samples from the subject was part of the model samples | 8 | 12 | 0.25 | 4 | 4 | 1.00 |

No differences were found for ADD in the distribution and completeness of urine samples, while the number of incomplete samples was different between the NND groups. A cut-off of 85 % for PABA was used for completeness as recommended in (Bingham & Cummings, 1983). This result can be interpreted in two ways. Either the model is dependent on complete samples or the misclassified subjects are less careful with their urine collections.

### 6.3.2 Person characteristics and registered foods collected from the shop

The number of subjects in the misclassified and correctly classified NND groups, respectively, was 18 and 21because the samples within each of the groups were in some cases from the same subject (see Appendix A). In the ADD group, all misclassified and correctly classified samples were from different subjects. Baseline characteristics of the subjects are given in Table 4.

**Table 4** Person baseline characteristics

| | **NND** | | | **ADD** | | |
|---|---|---|---|---|---|---|
| | Misclassi-fied | Correctly classified | P-value | Misclassified | Correctly classified | P-value |
| n | 18 | 21 | | 5 | 5 | |
| Age | 43.9 (13.5) | 44.9 (13.6) | 0.84 | 41.9 (15.1) | 44.2 (14.5) | 0.81 |
| No of women | 11 | 15 | 0.50 | 4 | 4 | 1.00 |
| Family | 5 | 9 | 0.33 | 2 | 1 | 0.49 |
| **Waist circumference [cm]** | 106.4 (11.7) | 98.9 (13.0) | 0.069 | 110.1 (21.0) | 104.1 (9.8) | 0.57 |
| **Body weight (week 0) [kg]** | 102.6 (14.5) | 87.7 (14.9) | **0.0034** | 101.5 (25.4) | 90.8 (8.6) | 0.40 |
| **BMI (week 0) [kg/m$^2$]** | 33.7 (4.7) | 29.7 (4.9) | **0.013** | 33.2 (9.5) | 30.5 (3.08) | 0.56 |

Family: Number of subjects from households of two study participants

No differences were found for the two ADD groups. NND subjects with one or more misclassified samples in the PLS-DA model weigh more and have higher BMIs at baseline compared to NND subjects with correctly classified samples. The body weight at the end of the intervention was also significantly higher for this group (p=0.0004, data not shown), while there was no significant difference in the change in body weight from the beginning until the end of the intervention (data not shown).

The macronutrient distribution and intakes of food groups that define NND and ADD are given in Table 5, based on all registered foods in the supermarket at the department. For subjects living in households of two study participants, EI registered in the supermarket was divided between the subjects based on their relative estimated EE per day during the intervention.

**Table 5** Food intake and macronutrient distribution of foods collected from the shop

| | NND | | | ADD | | |
|---|---|---|---|---|---|---|
| | Misclassified | Correctly classified | P-value | Misclassified | Correctly classified | P-value |
| n | 18 | 21 | | 5 | 5 | |
| **Main food groups [g/10 MJ]** | | | | | | |
| Nordic foods | 80.0 (5.4) | 80.5 (4.0) | 0.59 | 30.0 (3.8) | 30.9 (6.7) | 0.80 |
| Organic foods | 53.9 (4.9) | 54.2 (3.4) | 0.82 | 1.83 (1.0) | 1.4 (1.3) | 0.56 |
| Fruit | 418.5 (68.8) | 396.7 (67.9) | 0.33 | 196.4 (25.9) | 215.6 (27.1) | 0.29 |
| Berries | 87.8 (21.3) | 94.1 (27.7) | 0.43 | 4.9 (2.6) | 7.6 (2.6) | 0.13 |
| Vegetables | 465.2 (89.9) | 417.9 (66.9) | 0.068 | 220.0 (17.1) | 224.9 (40.6) | 0.81 |
| Cabbages | 65.4 (21.8) | 65.5 (17.2) | 0.99 | 5.3 (1.8) | 9.0 (8.6) | 0.37 |
| **Root vegetables** | 223.1 (57.9) | 183.5 (22.6) | **0.006** | 14.1 (8.5) | 26.3 (4.7) | **0.02** |
| Legumes | 48.5 (12.2) | 46.5 (18.5) | 0.69 | 4.3 (2.8) | 4.9 (2.0) | 0.72 |
| Potatoes | 133.4 (29.5) | 128.1 (42.0) | 0.65 | 70.7 (26) | 68.9 (33.0) | 0.93 |
| Fresh herbs | 8.6 (3.7) | 7.6 (3.2) | 0.35 | 1.4 (1.0) | 3.0 (3.3) | 0.32 |
| Wild plants and mushrooms | 4.6 (2.3) | 5.7 (0.2) | 0.16 | 0.13 (0.3) | 0.3 (0.4) | 0.47 |
| Nuts | 35.0 (5.9) | 34.0 (4.6) | 0.55 | 7.9 (2.3) | 10.1 (4.9) | 0.39 |
| Wholegrain | 151.8 (23.0) | 165.3 (32.5) | 0.79 | 42.1 (11.7) | 50.3 (23.6) | 0.51 |
| **Meat** | 109.2 (19.3) | 94.2 (17.4) | **0.015** | 158.6 (18.6) | 160.0 (11.7) | 0.89 |
| **Game** | 31.7 (9.6) | 23.3 (6.7) | **0.0027** | - | - | - |
| Fish and seafood | 71.5 (20.7) | 73.2 (24.7) | 0.81 | 17.0 (3.3) | 24.4 (9.4) | 0.14 |
| Seaweed | 0.87 (1.0) | 0.69 (0.8) | 0.88 | - | - | - |
| Milk and milk products | 373.1 (125.2) | 360.4 (124.1) | 0.75 | 344.0 (12.5) | 382.7 (109.4) | 0.45 |
| | | | | | | |
| **Other food groups [g/10 MJ]** | | | | | | |
| Chocolate and sweets | 5.3 (6.2) | 5.6 (7.7) | 0.59 | 36.8 (9.7) | 40.6 (14.1) | 0.64 |
| | | | | | | |
| **Macronutrients [E%]** | | | | | | |
| **Protein** | 18.4 (0.9) | 17.5 (0.8) | **0.0011** | 16.7 (0.9) | 16.6 (0.6) | 0.94 |
| Fat | 31.0 (1.6) | 30.7 (2.5) | 0.66 | 33.6 (2.6) | 34.8 (0.9) | 0.38 |
| SFA | 8.2 (1.1) | 8.2 (1.1) | 0.94 | 12.7 (1.1) | 13.6 (0.7) | 0.15 |
| CHO | 54.1 (2.5) | 54.4 (2.3) | 0.62 | 50.2 (2.4) | 51.0 (0.7) | 0.53 |
| Added sugar | 5.2 (1.5) | 5.9 (1.8) | 0.21 | 11.1 (2.8) | 11.9 (2.2) | 0.62 |
| | | | | | | |
| Energy based calculations | | | | | | |
| **Energy density [kJ/100 g]** | 454.0 (40.7) | 481.0 (32.8) | **0.027** | 559.9 (65.7) | 535.3 (77.4) | 0.60 |
| **EI(shop)/EE** | 0.73 (0.13) | 0.88 (0.16) | **0.0027** | 0.73 (0.06) | 0.78 (0.11) | 0.39 |
| **EI(WDR)/EE** | 0.60 (0.14) | 0.80 (0.20) | **0.0020** | 0.78 (0.20) | 0.90 (0.094) | 0.28 |

E%: Energy percent, SFA: Satuated fatty acids, CHO: Carbohydrates, EI/EE: Ratio of EI to EE using EI data from the supermarket or from WDR. EI (shop) has been calculated based on the energy content of foods collected from the supermarket and the number of intervention days foods were collected for. EI (WDR) has been calculated as the average EI across all days of WDR from week 12 and 26. EE has been calculated by using the online Body Weight Simulator (Hall et al., 2011) including the following parameters: Sex, age, height, weight at the beginning and end of the intervention (week 0 and 26), number of intervention days, %body fat at week 0, and self-reported physical activity based on questionnaire data from week 0.

Assuming that the foods registered in the shop was also consumed by the participants, the subjects in the misclassified NND group, have a significantly higher intake of meat, game and root vegetables, compared to the correctly classified NND group. The higher consumption of meat and game is also reflected in the energy contribution from protein, which is higher for the misclassified NND subjects. Despite the relatively higher meat consumption, there is still a large difference in meat consumption between misclassified NND subjects and subjects in the ADD groups. It is therefore not likely that the misclassification of NND samples is caused by markers of animal protein in urine, even though these are known to be part of the PLS-DA model, unless the subjects have consumed more meat without registering it. Within the category of root vegetables, there was no clear trend that the difference between the NND groups was caused by intake of any individual food (data not shown). In addition to the differences for food groups and macronutrients found for the misclassified NND subjects, these subjects have collected foods from the supermarket with a significantly lower energy density on average (Table 5). The ratio EI/EE is also significantly lower for the misclassified group regardless of EI being estimated from the shop data or from WDR. This indicates that the misclassified NND subjects have consumed more foods than what they have collected and reported in WDR. The calculation of the ratio EI/EE is uncertain. The participants were allowed to take up to twenty-one free days from the intervention diet and in the calculation of EI/EE, it is either assumed that the data from the supermarket, or from WDR, represents the average EI over the whole intervention, which is probably an underestimation. This may explain why EI/EE is lower than one in all groups. For EE, the parameters to estimate physical activity strongly influence the result and because the questions on physical activity in the questionnaire did not correspond completely to the questions used in the Body Weight Simulator (Hall et al., 2011), some error may have been introduced. However, there was still a significant difference between the NND groups when the default score for physical activity of 1.5 was used for all subjects, regardless of EI being based on supermarket data or WDR ($p=0.0034$, data not shown). It is well-known from other studies that subjects with high BMI tend to underreport EI more than normal weight subjects (Weber et al., 2001; Livingstone & Black, 2003). The finding that the misclassified NND subjects have higher BMI, collect less energy dense foods and to a higher degree than correctly classified subjects do not collect and report enough foods to cover their estimated EE therefore resemble the findings in other studies based on self-reporting and indicates non-compliance to the dietary intervention. Other studies have also demonstrated that under-reporters of EI generally report higher intakes of healthy foods, among those meat and vegetables, and lower intakes of unhealthy foods (Livingstone & Black, 2003). Even though such a trend is only statistically significant for meat and root vegetables, other food categories point in the same direction in Table 5. Misclassified NND subjects have collected more fruits and vegetables from the supermarket and the percent energy from added sugar in the foods collected is lower for this group.

The only food group that differed statistically between the ADD groups was root vegetables, which was lower in the misclassified group. In general, there is a tendency in Table 5 that the subjects in the misclassified ADD group have collected less food in the shop from the food groups that defines the dietary patterns, except potatoes. This should in theory make them even more characteristic ADD subjects and does not explain why they are classified as NND. As there are no statistically

significant trends in the collected foods and the ratio EI/EE for the misclassified ADD subjects compared to the correctly classified ADD subjects, it is not possible to explain from Table 5, why some samples from these subjects are misclassified.

### 6.3.3 Biological measures

A range of biological measures were taken in the INTER study (Poulsen et al., 2014). In Table 6, the measures in urine and whole blood related to dietary compliance are listed. The only difference found was in the ADD group, where the increase in the percentage of monounsaturated fatty acids (MUFA) in whole blood over the intervention was higher for the misclassified subjects. The fatty acid composition of whole blood can be used as a measure of the composition of dietary fat (Baylin et al. 2005). However, the difference in MUFA for the ADD groups was not reflected in the dietary intake of MUFA calculated from the registered food intake in the shop (data not shown). Whole blood MUFA and satuated fatty acids (SFA) did not correlate well with dietary intakes in a study based on FFQ data (Baylin et al., 2005). It is known that other factors, such as exercise, total fat intake and the metabolism of individual fatty acids affect the fatty acid composition in blood (Hodson et al., 2008). The finding in the present study may therefore also reflect other dietary differences or other subject characteristics than simply a change in dietary MUFA.

**Table 6** Biological compliance measures

| | **NND** | | | **ADD** | | |
|---|---|---|---|---|---|---|
| | Misclassified | Correctly classified | P-value | Misclassified | Correctly classified | P-value |
| n | 16 | 20 | | 5 | 5 | |
| **Urine measures** | | | | | | |
| Nitrogen [g/day] | 15.4 (2.7) | 14.7 (3.7) | 0.55 | 15.2 (2.8) | 14.6 (3.6) | 0.78 |
| **Whole blood measures** | | | | | | |
| ΔSFA (%) | 0.17 (2.62) | -1.46 (3.86) | 0.16 | -0.24 (3.66) | 1.81 (3.04) | 0.36 |
| **ΔMUFA (%)** | -0.44 (2.10) | 0.071 (2.11) | 0.47 | 2.60 (0.94) | 0.14 (1.54) | **0.016** |
| ΔPUFA (%) | 1.34 (3.78) | 2.44 (5.20) | 0.49 | -1.52 (3.76) | -0.33 (4.52) | 0.66 |
| Δn-6 | 0.82 (2.91) | 1.18 (3.93) | 0.77 | -0.60 (3.16) | 0.40 (3.21) | 0.63 |
| Δn-3 | 0.52 (1.27) | 1.26 (1.77) | 0.17 | -0.92 (0.83) | -0.72 (1.46) | 0.80 |
| Δn-6/n-3 | -0.21 (0.84) | -0.65 (1.01) | 0.18 | 0.85 (0.78) | 0.56 (0.62) | 0.54 |
| ΔDHA+EPA (%) | 0.43 (1.07) | 1.10 (1.46) | 0.14 | -0.73 (0.71) | -0.65 (1.25) | 0.91 |
| ΔWB-HUFA (%) | 0.07 (3.36) | 1.44 (4.14) | 0.29 | -1.85 (1.97) | -0.10 (3.09) | 0.32 |

Nitrogen: average nitrogen excretion for complete samples (PABA>85 %) in week 4, 12, 20 and 26 were used for each individual. Δ: Change from week 0 to week 26, SFA: Saturated fatty acids, MUFA: Monunsaturated fatty acids, PUFA: Polyunsaturated fatty acids, EPA: Eicosapentaenoic acids, DHA: Docosahexaenoic acids, HUFA: Highly unsaturated fatty acids ( ≥20 carbons and ≥3 double bonds)

Urinary nitrogen can be used to estimate protein intake in the diet (Bingham, 2003). Due to the higher percentage of energy from protein in the misclassified NND group (Table 5), a higher nitrogen excretion would be expected for this group compared to the correctly classified NND subjects. However, such as difference was not found. One explanation for this apparent discrepancy between the dietary registrations in the shop and the urinary nitrogen excretion is that the urinary nitrogen excretion was not estimated well enough for the subjects. Only 2.6 complete urine samples out of four possible were available for the misclassified NND subjects on average, while three samples were available for the correctly classified subjects. As previously mentioned, eight complete urine collections are required for an individual to obtain a reasonable estimate of protein intake (Bingham, 2003). Even though group level comparisons are made in Table 6, the variation due to few available urine samples may be too high compared to the actual difference in protein intake. Another possible explanation would be if the protein source affects nitrogen excretion. However, that was not the case in a study by Mikkelsen et al. (2000), for which no difference in 24 h urinary nitrogen excretion was found between two diets based on soy and pork protein, respectively, but with similar macronutrient distributions. Finally, a prerequisite for using nitrogen excretion is that the subjects are in nitrogen balance, which may not be the case in a study where subjects are losing weight. When estimating the change in fat free mass as determined by a DEXA scanning for the misclassified and correctly classified NND and ADD groups, no significant differences were found (data not shown) and differences in fat free mass therefore probably do not explain the results on nitrogen excretion. Reported percent energy from protein has been validated previously by 24 h urinary nitrogen excretion in another study applying the department supermarket to monitor dietary intakes (Skov et al., 1997). It is therefore expected that the difference in energy from protein found for misclassified NND subjects can be trusted, even though no significant differences in nitrogen excretions between the NND groups was found.

### 6.3.4 Exposure markers of individual foods

Use of known exposure markers to individual foods is another approach to investigate compliance in the INTER study. The dietary registrations from the supermarket as well as WDR, can be used to estimate exposure to certain foods and compared to the levels of known exposure markers in urine. If a set of exposure markers that cover relevant foods for the dietary intervention is available, it is possible to investigate if the reported diet of the subjects is supported by presence of these urinary markers. There are too few exposure markers available at present, to reasonably cover NND and ADD. However, the exposure markers of citrus and chocolate in Table 2 are supported by findings in other studies and since these two foods are not part of NND, it is interesting to apply the exposure markers for those as compliance markers in INTER. Due to the semi-quantitative analysis, possible errors during data preprocessing as well as inter-individual variation, it is expected that a combined measure including several markers for chocolate and citrus in urine will perform better than the available measures of each individual marker. For this reason, a combined PEM score was made for citrus and chocolate markers, respectively. The PEMs used to estimate citrus and chocolate exposure were all features of identified or tentatively identified PEMs found in Paper II of citrus and

chocolate. These included, two features from proline betaine ([M+H$^+$ and 2M+H$^+$]) and hesperetin-glucuronide ([M-H$^-$]) for citrus and theobromine ([M+H]$^+$), 7-methyluric acid ([2M-H]$^-$) and 6-AMMU ([M+H]$^+$) for chocolate. To obtain a combined marker level for the PEMs of citrus and chocolate, the percentage of the PEM peak area in a sample out of the sum of PEM peak areas for all samples was first calculated for each PEM. Then, the average percentage of the three citrus PEMs and the three chocolate PEMs in a sample was calculated for each sample and this number was applied to represent the marker level in urine.
Intake levels for citrus were calculated as the sum of reported orange juice, orange, mandarin, lemon, lemon juice, lime and grapefruit in WDR on the same day as the urine collection. Chocolate intake was estimated based on cocoa content, which was calculated for all chocolate containing products (13 in total).

WDR data only cover the day of urine collection and, as mentioned in section 5.1.5, chocolate makers are probably not excreted within 24 h. To investigate if the subjects had reported or collected citrus or chocolate close to the date of urine collection, supermarket data was used. For each misclassified and correctly classified NND and ADD sample, the food consumption registered in the supermarket for the visit leading up to the date of the urine collection was examined. Subjects were asked to report foods that had not been consumed as well as foods bought outside the supermarket between each visit and therefore the list of foods from the supermarket should represent all foods consumed by a household between each collection of foods from the shop.

As chocolate and citrus were an important part of ADD and were to be limited as much as possible in the NND, the reported intakes from the ADD diet are expected to be more correct than for the NND diet. The cut-off values for PEM levels to estimate intake of chocolate and citrus was therefore calculated based on samples from the ADD diet for which a WDR was available. The percentage of PEM levels was plotted as a function of reported citrus and cocoa intake for the ADD samples (plots to the left in Figure 7) and a cut-off for intake of 0.5 % was established to estimate chocolate and citrus intake based on these plots. Assuming that the reported intakes in ADD are correct, the probability of incorrectly classifying a subject as a chocolate and citrus consumer on the day of sampling using this cut-off is 15 % and 14 %, respectively, which is reasonably low.

Compliance for misclassified and correctly classified NND and ADD samples was evaluated according to the following criterion: If the PEM level in the sample for citrus or chocolate was above 0.5 and the subject had not reported the food in WDR or collected the food from the supermarket during the days leading up to the urine sampling, the subject was considered non-compliant. The distribution of chocolate and citrus markers in all NND samples and ADD samples are given in Figure 7 (plots in the middle and to the right). Based on these plots and the reported chocolate and citrus intakes, twelve and eight samples in the misclassified NND group were classified as being from non-compliant subjects to chocolate and citrus, respectively (Appendix A). This was significantly more samples than the corresponding number in the correctly classified NND group ($p=7.1E-5$ for chocolate and $p=0.034$ for citrus).

The probabilities of in-correctly classifying a subject as a non-consumer for citrus and chocolate when applying a cut-off of 0.5 were high (47 % and 61 %, respectively). Most ADD subjects in the correctly classified and misclassified groups had either collected chocolate and citrus from the supermarket or reported intake of chocolate and citrus and compliance therefore could not be evaluated for these subjects (Appendix A).

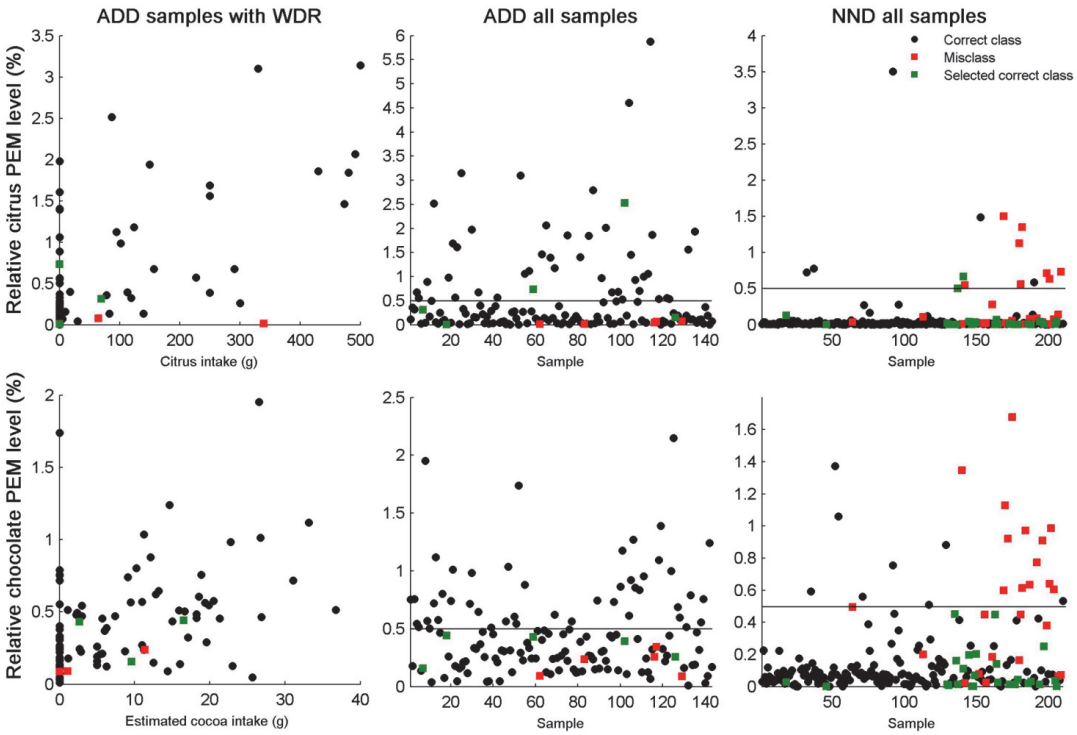



**Figure 7** Relative PEM levels in 24 h urine samples for citrus (top) and cocoa (bottom). Cut-off marker levels for intake are estimated from ADD data (left) and highlighted on the figures in the middle for ADD and to the right for NND. Subsets of misclassified and selected correctly classified samples are highlighted in red and green squares, respectively.

It is clear from the plots of dose-response for ADD samples in Figure 7 that there is a large inter-individual variation in the PEM levels and no clear dose-response relationship for citrus and, in particular, for chocolate. The more pronounced differences between consumers and non-consumers observed for citrus in Figure 7, compared to chocolate is probably because proline betaine and hesperetin-glucuronide are excreted fast (within 24 h), consumed in higher amounts and are com-

pounds that are not present in many other sources than citrus fruits. In addition, the citrus dietary sources in the study were mainly orange juice and orange which contain the highest levels of the two compounds compared to other citrus fruits (de Zwart et al., 2003; Aturki et al., 2004). As opposed to the citrus PEMs, the chocolate PEMs are expected to be relatively slowly excreted, and the calculation of intake of cocoa is subject to large errors due to the broad range of chocolate containing foods. Also, the chocolate PEMs are metabolites of caffeine as well and there is a large inter-individual variation in the excretion of these markers. The metabolism of caffeine and theobromine is largely dependent on the activity of the enzyme CYP1A2, which varies highly both due to genetic polymorphism and several environmental factors (Ghotbi et al., 2007). When comparing WDR for the NND subjects in the misclassified and correctly classified NND groups for which WDR were available, no significant difference was found in the number of subjects consuming caffeinated drinks. Even though it cannot be excluded that some of the suspected non-compliant subjects for chocolate can be explained by caffeine exposure from other dietary sources than chocolate, this is not expected to be the case.

Figure 7 is a good illustration of the challenge in applying exposure markers for evaluation of compliance. In the present case, additional variation is introduced because the measurements of PEMs are semi-quantitative. Proper examination of dose-response by controlled intakes of food and quantitative analysis would without doubt be better to estimate an appropriate cut-off value for compliance. However, even with the data at hand, there is a clear trend in the marker levels for the misclassified NND group compared to the correctly classified NND group (Figure 7) that support the finding that the misclassified NND subjects are likely non-compliant. The fact that non-compliance for citrus and chocolate consumption is more prevalent in the misclassified NND group compared to the correctly classified NND group is expected, since most of the citrus and chocolate exposure markers used to calculate the PEM levels in Figure 7, were also part of the multivariate compliance model. Furthermore, the result that more misclassified NND subjects seem to have consumed chocolate than citrus without reporting it is in accordance with the findings that the majority of misclassified NND samples in Paper III primarily differed from correctly classified NND samples in the level of chocolate related markers.

### 6.4 Validation of the compliance model by evaluation of misclassified samples

The results from the analysis of misclassified samples in the compliance model performed in section 6.3 are summed up in Table 7. Together, the results provide some additional information on the validity of the PLS-DA model. Most differences were found for the NND group, probably because this group of subjects is larger. Only major differences will be statistically detectable with a sample size of five in each group as was the case for the ADD groups. None of the statistical results were adjusted for multiple comparisons and it is therefore likely that some findings are chance findings. This is not very important, however, since the statistical tests in this case are used as a screening tool to characterize the misclassified samples rather than to draw firm conclusions from.

**Table 7** Differences within a dietary pattern between correctly classified and misclassified subjects

| Misclassified NND subjects | Misclassified ADD subjects |
|---|---|
| - Weigh more and have higher BMI<br>- Do not register enough food at the intervention supermarket and in WDR to cover their estimated EE<br>- Have less complete urine samples<br>- Collect more meat and root vegetables from the shop<br>- More energy from protein in foods collected from the shop<br>- Have high levels of chocolate and citrus exposure markers without reporting it | - Report less root vegetables<br>- Have a higher increase in full blood MUFA % over the intervention |

Within the misclassified NND subjects, a subgroup of seven subjects had more than one sample misclassified in the model and another subgroup of five subjects had only one out of three or four samples misclassified (Appendix A). If the compliance model should work properly, it is expected that the first mentioned group would be less compliant than the last mentioned group. However, no clear trends were found for the significant parameters in Table 7, when comparing means and standard deviations of subjects from the two subgroups with the overall means and standard deviations for all the misclassified NND subjects. This may be caused by low sample sizes or some of the parameters being chance findings. A larger dataset with more frequent sampling would be necessary to better understand how many samples from a subject should be misclassified in order to classify a subject as non-compliant. Incomplete sampling may also explain why some NND samples are misclassified. Since most exposure markers are excreted within a relatively narrow time frame missing urine collections, especially after intake of meals, may be important for the classification ability of the model. However, this is not possible to evaluate with the current data.

For misclassified ADD subjects, the significant differences found are not easy to interpret. None of the misclassified ADD subjects had more than one sample misclassified and more than one sample was available for the majority of subjects. Due to few misclassified ADD samples and no clear indication of non-compliance in the present dataset, a higher number of samples would probably be required to understand why some samples are misclassified.

Overall, the findings in the PLS-DA model to estimate compliance to NND are supported by the analyses of the misclassified subjects in comparison to correctly classified subjects. There is a clear indication that subjects with misclassified NND samples consume ADD foods regularly, underreport EI and over-report intake of healthy foods.

# 7. Conclusions

## 7.1 Untargeted metabolomics applied for discovery of PEMs

Untargeted metabolomics can contribute to discovery of new food exposure markers as well as consolidate findings from previous targeted studies on food exposure. The method provides a fast way to elucidate several important biological validation criteria such as sensitivity and specificity and time-course of excretion. With PEMs found in the INTER and MEAL studies as examples, it has been demonstrated that findings by metabolomic fingerprinting are often consistent at least for the strongest markers. The main limitation of untargeted metabolomics is the large number of unknown compounds that are often not reported and also not applied in other studies.

## 7.2 Untargeted metabolomics applied for estimating compliance

Untargeted metabolomics is a promising tool to develop multivariate compliance measures. Analysis of subjects with misclassified samples as compared to subjects with correctly classified samples in the compliance model for ADD and NND, support the validity of such a model to identify non-compliant NND subjects. There were too few subjects with misclassified ADD samples to draw any conclusions on compliance for this group. More studies are needed to determine how true non-compliant subjects can best be identified among subjects with misclassified samples in the model. The importance of complete urine samples for the classification ability of the model should also be investigated further.

# 8. Perspectives

## 8.1 Unresolved questions

The potential of untargeted nutrimetabolomics has been suggested to be almost without limits. This is nicely exemplified by the following quote from a recent review of the field by Llorach et al. (2012):

"Focusing on the nutrition sciences, metabolomics is an interesting tool for assessing the nutritional status of an individual, the food consumption, the biological consequences of following a nutritional intervention, or the study of metabolic mechanisms in response to a diet depending on a particular phenotype."

It is not surprising that metabolomics is forecasted to bring new insights in nutrition. The exploratory work with hundreds of metabolites invites for innumerable ideas, the logic being that so much data must contain an almost unlimited amount of biological information. The work in this thesis concerning exposure and compliance markers in urine witnesses that it is indeed possible to make new discoveries from urine metabolomics fingerprints.

At the same time, however, it is clear that the application of metabolomics for exposure and compliance markers is still far from fully explored. For some identified or tentatively identified PEMs, there are few or no studies to compare the results to. For the unidentified PEMs there is almost no way to validate if they are artifacts or good candidates as future exposure markers. For the compliance model, this study is the first of its kind and therefore serves merely as an example of another promising application of nutrimetabolomics than as a consolidated tool to distinguish non-compliant from compliant subjects.

In other words, while the exploratory work goes on, the findings are rarely followed up on and fully applied and we still know very little about the actual value of markers discovered by metabolomics. There are plenty of unresolved questions. In relation to the topics in this thesis, I think the most important ones to focus on in future studies are the following:

- How can we best identify the most promising dietary exposure markers across different metabolomics studies?
- How strong are the markers found by untargeted metabolomics in a targeted approach?
- Can identification of phenotypes help to discover more and better exposure markers?
- How robust are markers of a dietary pattern or habit and how well do they cover the diet?
- How many individual foods will we be able to find strong exposure markers for?

I will elaborate on how these questions could be addressed in the future in the next section.

## 8.2 The next step

In my opinion, one of the main keys to identify the most promising exposure markers by untargeted metabolomics is to systematize and utilize the knowledge from existing studies better. While many databases exist for known metabolites, it is not possible to search for masses and fragments of unknown metabolites reported in previous studies (if they have been reported at all). It would be time-saving to be able to separate unsupported findings from identified and unidentified markers that are supported by evidence from other studies. Identification is tedious, especially for urinary exposure markers, as these are often not present in existing databases. If a database of unknown markers was available, it would help to discover the strongest marker candidates, also among unknown markers. At present, a lot of useful knowledge is most likely lost because of difficulties with identifications instead of focusing on making information of the large group of unknown markers available for others in a useful form.

The findings across metabolomics studies are not always consistent and it is often not possible to separate if the discovery of a marker is caused by the methodological approach or true biological differences, especially if the marker is not identified. Many metabolomics studies do not apply additional validation steps to take advantage of the study design as has been done in the INTER and MEAL studies. Even though the validation criteria might have been too strict in the INTER and MEAL studies, they were very efficient to select the most promising PEMs. If other previously reported markers for the food or diet of interest could be used to better understand and possibly optimize the data analysis in a new study as well as to validate the previous markers in another study setting, it would be an advantage. For example, some markers from MEAL were present in INTER but did not meet the validation criteria of a sensitivity and specificity above 70 %, which could indicate that the validation criteria in INTER were too strict. Of course, such a procedure would be difficult to establish as it would require a lot of work to correctly match, especially unknown, markers in a new analytical system. However, it would be a way to generate more knowledge on previously reported markers faster.

While untargeted metabolomics is a great tool to screen for possible exposure markers and identify new marker candidates, it can never stand alone. As soon as there is enough evidence in favor of a new exposure marker, it should be investigated in a targeted approach. It is a pity that an excellent exposure marker candidate like proline betaine still has not been explored in a targeted study. There seems to be a large gap between the explorative metabolomics studies conducted and the application of the markers to strengthen future nutrition studies. Targeted studies are needed to provide the final validation as soon as a new marker is supported by enough evidence from existing metabolomics studies.

Other interesting perspectives are to conduct more studies that take into account different phenotypes. In relation to this, it could be interesting to find out if some inter-individual variation could be adjusted for by having a sort of standard diet with known food doses from each subject to compare with the levels in other urine samples. This data was available in the INTER study but it was not used for this purpose. To illustrate how correction for inter-individual differences in INTER

can, in some cases, strengthen the conclusions, when investigating compliance, an example for citrus is given in Figure 8.

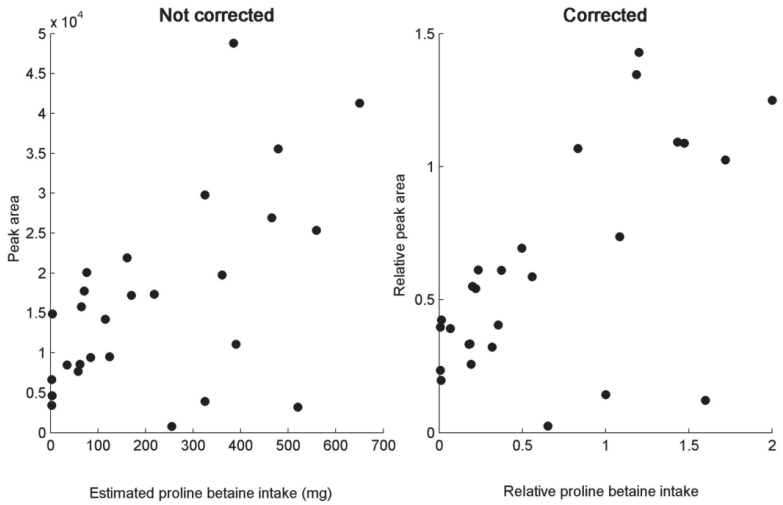



**Figure 8** Evaluation of compliance without (left) and with (right) correction for inter-individual differences in proline betaine excretion. Proline betaine intake has been estimated from reported intakes of citrus fruits in WDR and the proline betaine contents of foods reported in Heinzmann et al. (2010). Only subjects who reported citrus intake in week 12 and/or 26, and for whom a week 3 sample (where the diet was standardized) was available, are included. For proline betaine excretion, peak areas of the ion $[M+H]^+$ were used. Relative proline betaine intake and peak areas have been calculated by dividing peak areas and intakes of proline betaine from week 12 or 26 with week 3.

In Figure 8, there is a stronger trend for a dose-response relationship between the estimated proline betaine intake and the excretion of proline betaine, after correcting for inter-individual differences in proline betaine excretion. Three subjects can be identified as likely non-compliant in the plot with corrected data to the right. These subjects excrete less than 20 % of the amount of proline betaine measured on the standardized diet despite reporting more than 50 % of the reported intake for the standardized diet. Even though the same subjects seem to be non-compliant in the left plot, it is not as clear because they diverge less from the other subjects. It should be mentioned that no improvement in dose-response relationships were found, when similar calculations were done for the three chocolate PEMs used previously as compliance measures in Figure 7. It is therefore not straightforward to correct for inter-individual differences in all cases.

Recurrent urine sampling may be another way to obtain more information on the diet of an individual which can strengthen the evaluation of compliance and exposure. Multiple urine sampling on

the same subject to evaluate compliance has to my knowledge not been investigated yet. Due to the complex interplay of multiple dietary exposures, phenotypic characteristics, timing of urine sampling and intake, it is probably limited how much information on the habitual diet a single urine sample can reveal. Also, it is impossible to know how robust markers of habitual diets or dietary patterns are. To use the example dealt with in this thesis, if a person could be classified to the NND diet after consuming a lot of fruits, vegetables and fish on one day, such a measure would not really be relevant to apply as a compliance measure in a study that runs over six months. Close monitoring of the diet and continuous urine sampling over some time, can help to elucidate how dynamic the urinary metabolome is and if it is possible to separate habitual markers from markers originating from a single food load. Finally, single food studies can probably provide a fast way to screen for exposure markers, to find and compare different food sources of a marker and to understand why some food groups give rise to clear markers in cohort studies, while others do not.

# References


Ajala, O., English, P. & Pinkney, J. (2013): Systematic review and meta-analysis of different dietary approaches to the management of type 2 diabetes. *The American Journal of Clinical Nutrition*. 97, 505-516.

Altmaier, E., Kastenmüller, G., Römisch-Margl, W., Thorand, B., Weinberger, K.M., Illig, T. *et al* (2011): Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics. *European Journal of Epidemiology*. 26, 145-156.

Atkinson, W., Downer, P., Lever, M., Chambers, S.T., George, T. & George, P.M. (2007): Effects of orange juice and proline betaine on glycine betaine and homocysteine in healthy male subjects. *European Journal of Nutrition.* 46, 446-452.

Aturki, Z., Brandi, V. & Sinibaldi, M. (2004): Separation of flavanone- 7- O- glycoside diastereomers and analysis in citrus juices by multidimensional liquid chromatography coupled with mass spectrometry. *Journal of Agricultural and Food Chemistry.* 52, 5303-5308.

Barker, M. & Rayens, W. (2003): Partial least squares for discrimination. *Journal of Chemometrics.* 17, 166-173.

Baylin, A., Kim, M.K., Donovan-Palmer, A., Siles, X., Dougherty, L., Tocco, P. *et al* (2005): Fasting Whole Blood as a Biomarker of Essential Fatty Acid Intake in Epidemiologic Studies: Comparison with Adipose Tissue and Plasma. *American Journal of Epidemiology.* 162, 373-381.

Bere, E. & Brug, J. (2009): Towards health-promoting and environmentally friendly regional diets–a Nordic example. *Public Health Nutrition.* 12, 91-96.

Bertram, H.C., Hoppe, C., Petersen, B.O., Duus, J.Ø, Mølgaard, C. & Michaelsen, K.F. (2007): An NMR-based metabonomic investigation on effects of milk and meat protein diets given to 8-year-old boys. *British Journal of Nutrition.* 97, 758-763.

Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A. *et al* (2006): Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anaytical Chemistry.* 78, 567-574.

Bingham, S. & Cummings, J.H. (1983): The use of 4-aminobenzoic acid as a marker to validate the completeness of 24 h urine collections in man. *Clinical Science.* 64, 629-635.

Bingham, S.A. (2002): Biomarkers in nutritional epidemiology. *Public Health Nutrition.* 5, 821-827.

Bingham, S.A. (2003): Urine nitrogen as a biomarker for the validation of dietary protein intake. *The Journal of Nutrition.* 133, 921S-924S.

Boccard, J., Veuthey, J. & Rudaz, S. (2010): Knowledge discovery in metabolomics: An overview of MS data handling. *Journal of Separation Science.* 33, 290-304.

Bondia-pons, I., Barri, T., Hanhineva, K., Juntunen, K., Dragsted, L.O., Mykkänen, H. *et al* (2013): UPLC-QTOF/ MS metabolic profiling unveils urinary changes in humans after a whole grain rye versus refined wheat bread intervention. *Molecular Nutrition & Food Research*. 57, 412-422.

Brantsæter, A.L., Haugen, M., Rasmussen, S.E., Alexander, J., Samuelsen, S.O. & Meltzer, H.M. (2007): Urine flavonoids and plasma carotenoids in the validation of fruit, vegetable and tea intake during pregnancy in the Norwegian Mother and Child Cohort Study (MoBa). *Public Health Nutrition.* 10, 838-847.

Brevik, A., Rasmussen, S.E., Drevon, C.A. & Andersen, L.F. (2004): Urinary excretion of flavonoids reflects even small changes in the dietary intake of fruits and vegetables. *Cancer Epidemiology, Biomarkers & Prevention* 13, 843-849.

Broadhurst, D.I. & Kell, D.B. (2006): Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics.* 2, 171-196.

Chung, S.W.C. & Chan, B.T.P. (2009): Trimethylamine oxide, dimethylamine, trimethylamine and formaldehyde levels in main traded fish species in Hong Kong. *Food Additives and Contaminants: Part B.* 2, 44-51.

Clarke, D.B. (2010): Glucosinolates, structures and analysis in food. *Analytical Methods.* 2, 310-325.

Damsgaard, C.T., Dalskov, S.M., Petersen, R.A., Sørensen, L.B., Mølgaard, C., Biltoft-Jensen, A. *et al* (2012): Design of the OPUS School Meal Study: A randomised controlled trial assessing the impact of serving school meals based on the New Nordic Diet. *Scandinavian Journal of Public Health*. 40, 693-703.

de Zwart, F.J., Slow, S., Payne, R.J., Lever, M., George, P.M., Gerrard, J.A. *et al* (2003): Glycine betaine and glycine betaine analogues in common foods. *Food Chemistry.* 83, 197-204.

Dettmer, K., Aronov, P.A. & Hammock, B.D. (2007): Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews.* 26, 51-78.

Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R. & Griffin, J.L. (2011): Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews.* 40, 387-426.

Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R.*, et al* (2013): Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 9, S44-S66.

Edmands, W.M.B., Beckonert, O.P., Stella, C., Campbell, A., Lake, B.G., Lindon, J.C. *et al* (2011): Identification of human urinary biomarkers of cruciferous vegetable consumption by metabonomic profiling. *Journal of Proteome Research.* 10, 4513-4521.

Erlund, I., Meririnne, E., Alfthan, G. & Aro, A. (2001): Plasma kinetics and urinary excretion of the flavanones naringenin and hesperetin in humans after ingestion of orange juice and grapefruit juice. *The Journal of Nutrition.* 131, 235-241.

Fairweather-Tait, S. (2003): Human nutrition and food research: opportunities and challenges in the post-genomic era. *Philosophical Transactions B.* 358, 1709-1727.

Favé, G., Beckmann, M., Lloyd, A.J., Zhou, S., Harold, G., Lin, W. *et al* (2011): Development and validation of a standardized protocol to monitor human dietary exposure by metabolite fingerprinting of urine samples. *Metabolomics.* 7, 469-484.

Feldman, J.M. & Lee, E.M. (1985): Serotonin content of foods: effect on urinary excretion of 5-hydroxyindoleacetic acid. *The American Journal of Clinical Nutrition.* 42, 639-643.

Floegel, A., von Ruesten, A., Drogan, D., Schulze, M.B., Prehn, C., Adamski, J. *et al* (2013): Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam. *European Journal of Clinical Nutrition.* 67, 1100-1108.

Franke, A.A., Cooney, R.V., Henning, S.M. & Custer, L.J. (2005): Bioavailability and antioxidant effects of orange juice components in humans. *Journal of Agricultural and Food Chemistry.* 53, 5170-5178.

Ghotbi, R., Christensen, M., Roh, H., Ingelman-Sundberg, M., Aklillu, E. & Bertilsson, L. (2007): Comparisons of CYP1A2 genetic polymorphisms, enzyme activity and the genotype-phenotype relationship in Swedes and Koreans. *European Journal of Clinical Pharmacology.* 63, 537-546.

Gibney, M.J., Walsh, M., Brennan, L., Roche, H.M., German, B. & van Ommen, B. (2005): Metabolomics in human nutrition: opportunities and challenges. *The American Journal of Clinical Nutrition.* 82, 497-503.

Gika, H.G., Macpherson, E., Theodoridis, G.A. & Wilson, I.D. (2008a): Evaluation of the repeatability of ultra-performance liquid chromatography-TOF-MS for global metabolic profiling of human urine samples. *Journal of Chromatography B.* 871, 299-305.

Gika, H.G., Theodoridis, G.A. & Wilson, I.D. (2008b): Liquid chromatography and ultra- performance liquid chromatography-mass spectrometry fingerprinting of human urine - Sample stability under different handling and storage conditions for metabonomics studies. *Journal of Chromatography A.* 1189, 314-322.

Goodacre, R., Broadhurst, D., Smilde, A.K., Kristal, B.S., Baker, J.D., Beger, R. *et al* (2007): Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics.* 3, 231-241.

Gürdeniz, G., Kristensen, M., Skov, T. & 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.

Guy, P.A., Tavazzi, I., Bruce, S.J., Ramadan, Z. & Kochhar, S. (2008): Global metabolic profiling analysis on human urine by UPLC-TOFMS: Issues and method validation in nutritional metabolomics. *Journal of Chromatography B.* 871, 253-260.

Hall, K.D., Sacks, G., Chandramohan, D., Chow, C.C., Wang, Y.C., Gortmaker, S.L. *et al* (2011): Quantification of the effect of energy imbalance on bodyweight. *The Lancet.* 378, 826-837.

Harnly, J.M., Doherty, R.F., Beecher, G.R., Holden, J.M., Haytowitz, D.B., Bhagwat, S. *et al* (2006): Flavonoid content of U.S. fruits, vegetables, and nuts. *Journal of Agricultural and Food Chemistry.* 54, 9966-9977.

Heinzmann, S.S., Brown, I.J., Chan, Q., Bictash, M., Dumas, M., Kochhar, S. *et al* (2010): Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. *American Journal of Clinical Nutrition.* 92, 436-443.

Heinzmann, S.S., Merrifield, C.A., Rezzi, S., Kochhar, S., Lindon, J.C., Holmes, E. *et al* (2011): Stability and robustness of human metabolic phenotypes in response to sequential food challenges. *Journal of Proteome Research.* 11, 643-655.

Helander, A., Wikstrom, T., Löwenmo, C., Jacobsson, G. & Beck, O. (1992): Urinary excretion of 5- hydroxyindole-3- acetic acid and 5-hydroxytryptophol after oral loading with serotonin. *Life Sciences.* 50, 1207-1213.

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

Hollands, W., Brett, G.M., Radreau, P., Saha, S., Teucher, B., Bennett, R.N. *et al* (2008): Processing blackcurrants dramatically reduces the content and does not enhance the urinary yield of anthocyanins in human subjects. *Food Chemistry.* 108, 869-878.

Ito, H., Gonthier, M., Manach, C., Morand, C., Mennen, L., Rémésy, C. *et al* (2005): Polyphenol levels in human urine after intake of six different polyphenol-rich beverages. *The British Journal of Nutrition.* 94, 500-509.

Janobi, A.A.A., Mithen, R.F., Gasper, A.V., Shaw, P.N., Middleton, R.J., Ortori, C.A. *et al* (2006): Quantitative measurement of sulforaphane, iberin and their mercapturic acid pathway metabolites in human plasma and urine using liquid chromatography–tandem electrospray ionisation mass spectrometry. *Journal of Chromatography B.* 844, 223-234.

Jenab, M., Slimani, N., Bictash, M., Ferrari, P. & Bingham, S.A. (2009): Biomarkers in nutritional epidemiology: applications, needs and new horizons. *Human Genetics.* 125, 507-525.

Jiao, D., Ho, C.T., Foiles, P. & Chung, F.L. (1994): Identification and quantification of the N-acetylcysteine conjugate of allyl isothiocyanate in human urine after ingestion of mustard. *Cancer Epidemiology, Biomarkers & Prevention.* 3, 487-492.

Katajamaa, M. & Orešič, M. (2007): Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A.* 1158, 318-328.

Key, T.J., Allen, N.E., Spencer, E.A. & Travis, R.C. (2002): The effect of diet on risk of cancer. *The Lancet.* 360, 861-868.

Kjeldahl, K. & Bro, R. (2010): Some common misunderstandings in chemometrics. *Journal of Chemometrics.* 24, 558-564.

Krogholm, K.S., Bredsdorff, L., Knuthsen, P., Haraldsdóttir, J. & Rasmussen, S.E. (2010): Relative bioavailability of the flavonoids quercetin, hesperetin and naringenin given simultaneously through diet. *European Journal of Clinical Nutrition.* 64, 432–435.

Kuhnle, G.G.C. (2012): Nutritional biomarkers for objective dietary assessment. *Journal of the Science of Food and Agriculture.* 92, 1145-1149.

Legido-Quigley, C., Stella, C., Perez-Jimenez, F., Lopez-Miranda, J., Ordovas, J., Powell, J. *et al* (2010): Liquid chromatography-mass spectrometry methods for urinary biomarker detection in metabonomic studies with application to nutritional studies. *Biomedical Chromatography.* 24, 737-743.

Livingstone, M.B.E. & Black, A.E. (2003): Markers of the validity of reported energy intake. *The Journal of Nutrition.* 133, 895S-920S.

Llorach, R., Urpi-Sarda, M., Jauregui, O., Monagas, M. & Andres-Lacueva, C. (2009): An LC-MS- based metabolomics approach for exploring urinary metabolome modifications after cocoa consumption. *Journal of Proteome Research.* 8, 5060-5068.

Llorach, R., Garrido, I., Monagas, M., Urpi-Sarda, M., Tulipani, S., Bartolome, B. *et al* (2010): Metabolomics study of human urinary metabolome modifications after intake of almond (Prunus dulcis (Mill.) D.A. Webb) skin polyphenols. *Journal of Proteome Research.* 9, 5859-5867.

Llorach, R., Garcia-Aloy, M., Tulipani, S., Vazquez-Fresno, R. & Andres-Lacueva, C. (2012): Nutrimetabolomic strategies to develop new biomarkers of intake and health effects. *Journal of Agricultural and Food Chemistry.* 60, 8797-8808.

Llorach, R., Urpi-Sarda, M., Tulipani, S., Garcia-Aloy, M., Monagas, M. & Andres-Lacueva, C. (2013): Metabolomic fingerprint in patients at high risk of cardiovascular disease by cocoa intervention. *Molecular Nutrition & Food Research.* 57, 962-973.

Lloyd, A.J., Beckmann, M., Favé, G., Mathers, J.C. & Draper, J. (2011a): Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. *British Journal of Nutrition.* 106, 812-824.

Lloyd, A.J., Favé, G., Beckmann, M., Lin, W., Tailliart, K., Xie, L. *et al* (2011b): Use of mass spectrometry fingerprinting to identify urinary metabolites after consumption of specific foods. *American Journal of Clinical Nutrition.* 94, 981-991.

Lloyd, A.J., Beckmann, M., Haldar, S., Seal, C., Brandt, K. & Draper, J. (2013): Data-driven strategy for the discovery of potential urinary biomarkers of habitual dietary exposure. *The American Journal of Clinical Nutrition.* 97, 377-389.

Maher, A.D., Zirah, S.F.M., Holmes, E. & Nicholson, J.K. (2007): Experimental and analytical variation in human urine in 1 H NMR spectroscopy-based metabolic phenotyping studies. *Analytical Chemistry.* 79, 5204-5211.

Manach, C., Morand, C., Gil-Izquierdo, A., Bouteloup-Demange, C. & Rémésy, C. (2003): Bioavailability in humans of the flavanones hesperidin and narirutin after the ingestion of two doses of orange juice. *European Journal of Clinical Nutrition.* 57, 235-242.

Manach, C., Hubert, J., Llorach, R. & Scalbert, A. (2009): The complex links between dietary phytochemicals and human health deciphered by metabolomics. *Molecular Nutrition & Food Research.* 53, 1303-1315.

Mann, J. (2002): Diet and risk of coronary heart disease and type 2 diabetes. *The Lancet.* 360, 783-789.

Martin, F.J., Rezzi, S., Pere-Trepat, E., Kamlage, B., Collino, S., Leibold, E.*, et al* (2009): Metabolic Effects of Dark Chocolate Consumption on Energy, Gut Microbiota, and Stress-Related Metabolism in Free-Living Subjects. *Journal of Proteome Research.* Vol. 8:12, pp. 5568-5579.

Martin, F.J., Montoliu, I., Nagy, K., Moco, S., Collino, S., Guy, P. *et al* (2012): Specific dietary preferences are linked to differing gut microbial metabolic activity in response to dark chocolate intake. *Journal of Proteome Research.* 11, 6252-6263.

May, D.H., Navarro, S.L., Ruczinski, I., Hogan, J., Ogata, Y., Schwarz, Y. *et al* (2013): Metabolomic profiling of urine: response to a randomised, controlled feeding study of select fruits and vegetables, and application to an observational study. *The British Journal of Nutrition.* Published online 9 May. DOI:10.1017/S000711451300127X.

Menni, C., Zhai, G., MacGregor, A., Prehn, C., Römisch-Margl, W., Suhre, K. *et al* (2013): Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. *Metabolomics.* 9, 506-514.

Mennicke, W.H., Görler, K., Krumbiegel, G., Lorenz, D. & Rittmann, N. (1988): Studies on the metabolism and excretion of benzyl isothiocyanate in man. *Xenobiotica.* 18, 441-447.

Mente, A., de Koning, L., Shannon, H.S. & Anand, S.S. (2009): A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. *Archives of Internal Medicine.* 169, 659-669.

Mikkelsen, P.B., Toubro, S. & Astrup, A. (2000): Effect of fat-reduced diets on 24-h energy expenditure: comparisons between animal protein, vegetable protein, and carbohydrate. *The American Journal of Clinical Nutrition.* 72, 1135-1141.

Mithril, C., Dragsted, L.O., Meyer, C., Blauert, E., Holt, M.K. & Astrup, A. (2012): Guidelines for the New Nordic Diet. *Public Health Nutrition.* 15, 1941-1947.

Mithril, C., Dragsted, L.O., Meyer, C., Tetens, I., Biltoft-Jensen, A. & Astrup, A. (2013): Dietary composition and nutrient content of the New Nordic Diet. *Public Health Nutrition.* 16, 777-785.

Neveu, V., Perez-Jiménez, J., Vos, F., Crespy, V., du Chaffaut, L., Mennen, L. *et al* (2010): Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database.* bap024, DOI:10.1093/database/bap024.

Nordic Council of Ministers (2004): Nordic Nutrition Recommendations NNR 2004, Integrating Nutrition and Physical Activity. 4th ed. Copenhagen: Nordic Council of Ministers

O'Sullivan, A., Gibney, M.J. & Brennan, L. (2011): Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. *American Journal of Clinical Nutrition.* 93, 314-321.

Penn, L., Boeing, H., Boushey, C.J., Dragsted, L.O., Kaput, J., Scalbert, A. *et al* (2010): Assessment of dietary intake: NuGO symposium report. *Genes & Nutrition.* 5, 205-213.

Peré-Trepat, E., Ross, A.B., Martin, F., Rezzi, S., Kochhar, S., Hasselbalch, A.L. *et al* (2010): Chemometric strategies to assess metabonomic imprinting of food habits in epidemiological studies. *Chemometrics and Intelligent Laboratory Systems.* 104, 95-100.

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

Potischman, N. & Freudenheim, J.L. (2003): Biomarkers of nutritional exposure and nutritional status: an overview. *The Journal of Nutrition.* 133, 873S-874S.

Poulsen, S.K., Due, A., Jordy, A.B., Kiens, B., Stark, K.D., Stender, S. *et al* (2014): Health effect of the New Nordic Diet in adults with increased waist circumference: a 6-mo randomized controlled trial. *American Journal of Clinical Nutrition*. 99, 35-45

Ptolemy, A.S., Tzioumis, E., Thomke, A., Rifai, S. & Kellogg, M. (2010): Quantification of theobromine and caffeine in saliva, plasma and urine via liquid chromatography-tandem mass spectrometry: a single analytical protocol applicable to cocoa intervention studies. *Journal of Chromatography B.* 878, 409-416.

Pujos-Guillot, E., Hubert, J., Martin, J., Lyan, B., Quintana, M., Claude, S. *et al* (2013): Mass spectrometry-based metabolomics for the discovery of biomarkers of fruit and vegetable intake: citrus fruit as a case study. *Journal of Proteome Research.* 12, 1645-1659.

Rajalahti, T., Arneberg, R., Kroksveen, A.C., Berle, M., Myhr, K. & Kvalheim, O.M. (2009): Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical Chemistry.* 81, 2581-2590.

Rasmussen, L.G., Savorani, F., Larsen, T.M., Dragsted, L.O., Astrup, A. & Engelsen, S.B. (2010): Standardization of factors that influence human urine metabolomics. *Metabolomics.* 7, 71-83.

Rasmussen, L.G., Winning, H., Savorani, F., Ritz, C., Engelsen, S.B., Astrup, A. *et al* (2012a): Assessment of dietary exposure related to dietary GI and fibre intake in a nutritional metabolomic study of human urine. *Genes & Nutrition.* 7, 281-293.

Rasmussen, L.G., Winning, H., Savorani, F., Toft, H., Larsen, T.M., Dragsted, L.O. *et al* (2012b): Assessment of the effect of high or low protein diet on the human urine metabolome as measured by NMR. *Nutrients.* 4, 112-131.

Rechner, A.R., Spencer, J.P.E., Kuhnle, G., Hahn, U. & Rice-Evans, C.A. *et al* (2001): Novel biomarkers of the metabolism of caffeic acid derivatives in vivo. *Free Radical Biology and Medicine.* 30, 1213-1222.

Rechner, A.R., Kuhnle, G., Bremner, P., Hubbard, G.P., Moore, K.P. & Rice-Evans, C. (2002a): The metabolic fate of dietary polyphenols in humans. *Free Radical Biology & Medicine.* 33, 220-235.

Rechner, A.R., Kuhnle, G., Hu, H., Roedig-Penman, A., van den Braak, M.H., Moore, K.P. *et al* (2002b): The metabolism of dietary polyphenols and the relevance to circulating levels of conjugated metabolites. *Free Radical Research.* 36, 1229-1241.

Rehman, H.U. (1999): Fish odour syndrome. *Postgraduate Medical Journal.* 75, 451-452.

Rezzi, S., Ramadan, Z., Fay, L.B. & Kochhar, S. (2007): Nutritional metabonomics: applications and perspectives. *Journal of Proteome Research.* 6, 513-525.

Rodopoulos, N., Höjvall, L. & Norman, A. (1996): Elimination of theobromine metabolites in healthy adults. *Scandinavian Journal of Clinical and Laboratory Investigation.* 56, 373-383.

Rodopoulos, N. & Norman, A. (1996): Assessment of dimethylxanthine formation from caffeine in healthy adults: comparison between plasma and saliva concentrations and urinary excretion of metabolites. *Scandinavian Journal of Clinical and Laboratory Investigation.* 56, 259-268.

Ross, A.B., Pere-Trepat, E., Montoliu, I., Martin, F.J., Collino, S., Moco, S. *et al* (2013): A whole-grain-rich diet reduces urinary excretion of markers of protein catabolism and gut microbiota metabolism in healthy men after one week. *The Journal of Nutrition.* 143, 766-773.

Rouzaud, G., Young, S.A. & Duncan, A.J. (2004): Hydrolysis of glucosinolates to isothiocyanates after ingestion of raw or microwaved cabbage by human volunteers. *Cancer Epidemiology, Biomarkers & Prevention.* 13, 125-131.

Ryan, D., Robards, K., Prenzler, P.D. & Kendall, M. (2011): Recent and potential developments in the analysis of urine: a review. *Analytica Chimica Acta.* 684, 17-29.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B.S., van Ommen, B. *et al* (2009): Mass- spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics.* 5, 435-458.

Seline, K. & Johein, H. (2007): The determination of L-carnitine in several food samples. *Food Chemistry.* 105, 793-804.

Shively, C.A. & Tarka, S.M., Jr (1984): Methylxanthine composition and consumption patterns of cocoa and chocolate products. *Progress in Clinical and Biological Research.* 158, 149-178.

Skov, A.R., Toubro, S., Raben, A. & Astrup, A. (1997): A method to achieve control of dietary macronutrient composition in ad libitum diets consumed by free-living subjects. *European Journal of Clinical Nutrition.* 51, 667-672.

Slupsky, C.M., Rankin, K.N., Wagner, J., Fu, H., Chang, D., Weljie, A.M., *et al* (2007): Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Analytical Chemistry.* 79, 6995-7004.

Solanky, K.S., Bailey, N.J., Beckwith-Hall, B.M., Bingham, S., Davis, A., Holmes, E. *et al* (2005): Biofluid $^{1}$H NMR-based metabonomic techniques in nutrition research — metabolic effects of dietary isoflavones in humans. *The Journal of Nutritional Biochemistry.* 16, 236-244.

Spencer, J.P.E., Abd El Mohsen, Manal M, Minihane, A. & Mathers, J.C. (2008): Biomarkers of the intake of dietary polyphenols: strengths, limitations and application in nutrition research. *British Journal of Nutrition.* 99, 12-22.

Stella, C., Beckwith-hall, B., Cloarec, O., Holmes, E., Lindon, J.C., Powell, J. *et al* (2006): Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research.* 5, 2780-2788.

Storey, J.D. (2002): A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 64, 479-498.

Sullivan, L.M. & D'Agostino, R.B. (1992): Robustness of the t test applied to data distorted from normality by floor effects. *Journal of Dental Research.* 71, 1938-1943.

Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C. *et al* (2007): Proposed minimum reporting standards for chemical analysis. *Metabolomics.* 3, 211-221.

Svensson, B., Åkesson, B., Nilsson, A. & Paulsson, K. (1994): Urinary excretion of methylamines in men with varying intake of fish from the baltic sea. *Journal of Toxicology and Environmental Health.* 41, 411-420.

Theodoridis, G., Gika, H.G. & Wilson, I.D. (2008): LC-MS- based methodology for global metabolite profiling in metabonomics/metabolomics. *Trends in Analytical Chemistry.* 27, 251-260.

Tulipani, S., Llorach, R., Jáuregui, O., López-Uriarte, P., Garcia-Aloy, M., Bullo, M. *et al* (2011): Metabolomics unveils urinary changes in subjects with metabolic syndrome following 12-week nut consumption. *Journal of Proteome Research.* 10, 5047-5058.

van den Berg, R.A.,Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. & van der Werf, M.J (2011): Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*. 7, 142

van Dorsten, F.A., Daykin, C.A., Mulder, T.P.J. & van Duynhoven, J.P.M. (2006): Metabonomics approach to determine metabolic differences between green tea and black tea consumption. *Journal of Agricultural and Food Chemistry.* 54, 6929-6938.

van Dorsten, F.A., Grün, C.H., van Velzen, E.J.J., Jacobs, D.M., Draijer, R. & van Duynhoven, J.P.M. (2010): The metabolic fate of red wine and grape juice polyphenols in humans assessed by metabolomics. *Molecular Nutrition & Food Research.* 54, 897-908.

van Velzen, E.J.J., Westerhuis, J.A., van Duynhoven, J.P.M., van Dorsten, F.A., Hoefsloot, H.C.J., Jacobs, D.M. *et al* (2008): Multilevel data analysis of a crossover designed human nutritional intervention study. *Journal of Proteome Research.* 7, 4483-4491.

van Velzen, E.J.J., Westerhuis, J.A., van Duynhoven, J.P.M., van Dorsten, F.A., Grün, C.H., Jacobs, D.M. *et al* (2009): Phenotyping tea consumers by nutrikinetic analysis of polyphenolic end-metabolites. *Journal of Proteome Research.* 8, 3317-3330.

Vermeulen, M., van den Berg, R., Freidig, A.P., van Bladeren, P.J. & Vaes, W.H.J. (2006): Association between consumption of cruciferous vegetables and condiments and excretion in urine of isothiocyanate mercapturic acids. *Journal of Agricultural and Food Chemistry.* 54, 5350-5358.

Vitolins, M.Z., Rand, C.S., Rapp, S.R., Ribisl, P.M. & Sevick, M.A. (2000): Measuring adherence to behavioral and medical interventions. *Controlled Clinical Trials.* 21, 188S-194S.

Walsh, M.C., Brennan, L., Pujos-Guillot, E., Sébédio, J., Scalbert, A., Fagan, A. *et al* (2007): Influence of acute phytochemical intake on human urinary metabolomic profiles. *American Journal of Clinical Nutrition.* 86, 1687-1693.

Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., DuGar, B. *et al* (2011): Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature.* 472, 57-63.

Wang, Y., Tang, H., Nicholson, J.K., Hylands, P.J., Sampson, J. & Holmes, E. (2005): A metabonomic strategy for the detection of the metabolic effects of chamomile (Matricaria recutita L.) ingestion. *Journal of Agricultural and Food Chemistry.* 53, 191-196.

Warrack, B.M., Hnatyshyn, S., Ott, K., Reily, M.D., Sanders, M., Zhang, H. *et al* (2009): Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B.* 877, 547-552.

Weber, J.L., Reid, P.M., Greaves, K.A., DeLany, J.P., Stanford, V.A., Going, S.B. *et al* (2001): Validity of self-reported energy intake in lean and obese young women, using two nutrient databases, compared with total energy expenditure assessed by doubly labeled water. *European Journal of Clinical Nutrition.* 55, 940-950.

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J. *et al* (2008): Assessment of PLSDA cross validation. *Metabolomics.* 4, 81-89.

Wishart, D.S. (2008): Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology.* 19, 482-493.

Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B. *et al* (2009): HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research.* 37, D603-D610.

Wold, S., Sjöström, M. & Eriksson, L. (2001): PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems.* 58, 109-130.

Xu, J., Yang, S., Cai, S., Dong, J., Li, X. & Chen, Z. (2010): Identification of biochemical changes in lactovegetarian urine using $^{1}$H NMR spectroscopy and pattern recognition. *Analytical and Bioanalytical Chemistry.* 396, 1451-1463.

Zeisel, S.H., Mar, M., Howe, J.C. & Holden, J.M. (2003): Concentrations of choline-containing compounds and betaine in common foods. *The Journal of Nutrition.* 133, 1302-1307.

Zhang, A.Q., Mitchell, S.C. & Smith, R.L. (1999): Dietary precursors of trimethylamine in man: a pilot study. *Food and Chemical Toxicology.* 37, 515-520.

Zhang, X., Yap, Y., Wei, D., Chen, G. & Chen, F. (2008): Novel omics technologies in nutrition research. *Biotechnology Advances.* 26, 169-176.

# Appendices

## Appendix A: Information on individual samples in the misclassified and selected correctly classified NND and ADD groups

**Misclassified NND samples**

| ID | Week | Sea-son | PABA Reco-vered | In model | Total no | % Miss-classified | Chocolate WDR | Chocolate supermar-ket | Chocolate PEM level (%) | Chocolate compliance | Citrus WDR | Citrus supermar-ket | Citrus PEM level (%) | Citrus com-pliance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 006-2 | 20 | W | Y | N | 2 | 100 | - | N | 0.022 | Y | - | N | 0.55 | N |
| 006-2 | 26 | W | N | N | 2 | 100 | N | N | 0.074 | Y | N | N | 0.73 | N |
| 081-2 | 12 | W | Y | N | 2 | 100 | N | N | 0.64 | N | N | N | 0.084 | Y |
| 081-2 | 26 | Sp | Y | N | 2 | 100 | N | N | 0.60 | N | N | N | 0.078 | Y |
| 114-1 | 12 | W | N | N | 2 | 100 | N | N | 0.61 | N | N | N | 1.35 | N |
| 114-1 | 26 | Sp | Y | N | 2 | 100 | N | N | 0.38 | Y | N | N | 0.71 | N |
| 114-2 | 12 | W | N | N | 2 | 100 | N | N | 0.97 | N | N | N | 0.012 | Y |
| 114-2 | 26 | Sp | N | N | 2 | 100 | N | N | 0.64 | N | N | N | 0.63 | N |
| 062-1 | 4 | A | N | N | 3 | 67 | - | N | 0.19 | Y | - | N | 0.28 | Y |
| 062-1 | 26 | Sp | N | N | 3 | 67 | N | N | 0.06 | Y | N | N | 0.14 | Y |
| 095-2 | 4 | W | Y | Y | 3 | 67 | - | N | 1.35 | N | N | N | 0.0098 | Y |
| 095-2 | 26 | Sp | Y | Y | 3 | 67 | N | N | 0.49 | Y | N | N | 0.034 | Y |
| 172-1 | 4 | W | N | Y | 3 | 67 | - | N | 1.13 | N | N | N | 0.024 | Y |
| 172-1 | 12 | W | Y | Y | 3 | 67 | N | N | 0.20 | Y | N | N | 0.11 | Y |
| 033-2 | 26 | W | N | N | 1 | 100 | N | N | 0.91 | N | N | N | 0.039 | Y |
| 071-1 | 12 | W | N | N | 1 | 100 | N | N | 1.68 | N | N | Y | 0.053 | Y |
| 080-2 | 12 | W | N | N | 1 | 100 | N | N | 0.16 | Y | N | Y | 1.13 | N |
| 090-2 | 26 | Sp | N | N | 1 | 100 | N | N | 0.98 | N | N | N | 0.011 | Y |
| 175-2 | 12 | W | Y | N | 1 | 100 | N | N | 0.45 | Y | N | Y | 0.56 | N |
| 057-1 | 12 | W | N | N | 2 | 50 | Y | Y | 0.92 | Y | N | N | 0.012 | Y |
| 016-1 | 20 | W | N | Y | 3 | 33 | - | N | 0.026 | Y | - | N | 0.024 | Y |
| 036-2 | 20 | Sp | Y | Y | 3 | 33 | - | N | 0.45 | Y | - | N | 0.017 | Y |
| 187-1 | 26 | S | N | N | 3 | 33 | N | N | 0.77 | N | N | N | 0.090 | Y |
| 007-2 | 20 | W | Y | Y | 4 | 25 | - | N | 0.60 | N | - | N | 1.499 | N |
| 012-2 | 4 | A | Y | Y | 4 | 25 | - | N | 0.079 | Y | - | N | 0.025 | Y |

In model: Sample included as a model sample in the PLS-DA model. Total no: Total number of samples from the subject (including model and validation samples). % Misclassified: The number of misclassified samples from the subject out of the total number of available samples from the subject. Chocolate/Citrus WDR: Chocolate/Citrus containing product reported in WDR. Chocolate/Citrus supermarket: Chocolate/Citrus containing products collected from the supermarket close to the time of sampling. Chocolate/Citrus PEM level (%): Level of chocolate/citrus markers (see section 6.3.4). Y/N: Yes/No, Sp: Spring, A: Autumn, S: Summer, W: Winter.

## Appendix A

**Selected correctly classified NND samples**

| ID | Week | Sea-son | PABA Recovered | In model | Total no | Chocolate WDR | Chocolate supermarket | Chocolate PEM level (%) | Chocolate compliance | Citrus WDR | Citrus Super-market | Citrus PEM level (%) | Citrus com-pliance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 005-2 | 20 | Sp | N | Y | 3 | - | N | 0.45 | Y | - | N | 0.0061 | Y |
| 010-2 | 4 | A | Y | N | 3 | - | N | 0.45 | Y | - | N | 0.0042 | Y |
| 018-2 | 26 | Sp | Y | Y | 3 | N | N | 0.030 | Y | N | N | 0.13 | Y |
| 019-2 | 20 | W | Y | Y | 3 | - | N | 0.014 | Y | - | N | 0.50 | N |
| 020-1 | 4 | W | N | Y | 4 | - | N | 0.11 | Y | - | N | 0.67 | N |
| 020-1 | 20 | W | N | Y | 4 | - | N | 0.20 | Y | - | N | 0.041 | Y |
| 020-2 | 12 | W | N | N | 3 | N | N | 0.042 | Y | N | N | 0.0038 | Y |
| 039-1 | 4 | A | Y | N | 3 | - | N | 0.0081 | Y | - | N | 0.0039 | Y |
| 039-1 | 20 | W | Y | N | 3 | - | N | 0.16 | Y | - | N | 0.0064 | Y |
| 039-2 | 26 | Sp | N | N | 2 | N | N | 0.038 | Y | N | N | 0.010 | Y |
| 056-2 | 20 | W | Y | Y | 3 | N | Y | 0.20 | Y | N | N | 0.027 | Y |
| 064-2 | 20 | W | N | Y | 3 | N | N | 0.014 | Y | N | N | 0.026 | Y |
| 064-2 | 26 | Sp | Y | Y | 3 | N | N | 0.0012 | Y | N | N | 0.0015 | Y |
| 064-4 | 20 | W | Y | Y | 3 | - | N | 0.0008 | Y | - | N | 0.0057 | Y |
| 065-2 | 20 | Sp | Y | Y | 3 | - | Y | 0.068 | Y | - | N | 0.035 | Y |
| 086-2 | 4 | A | Y | Y | 3 | - | N | 0.024 | Y | - | N | 0.0082 | Y |
| 130-1 | 26 | Sp | Y | N | 1 | N | N | 0.0005 | Y | N | N | 0.040 | Y |
| 154-1 | 4 | W | Y | N | 2 | - | N | 0.0001 | Y | - | N | 0.016 | Y |
| 159-2 | 12 | W | Y | N | 1 | N | N | 0.012 | Y | N | N | 0.0037 | Y |
| 167-2 | 12 | W | Y | N | 1 | N | N | 0.032 | Y | N | N | 0.0015 | Y |
| 178-2 | 12 | W | Y | N | 3 | N | N | 0.014 | Y | N | N | 0.014 | Y |
| 178-2 | 26 | S | Y | N | 3 | N | N | 0.013 | Y | N | N | 0.041 | Y |
| 180-2 | 26 | Sp | Y | N | 2 | N | N | 0.249 | Y | N | N | 0.0093 | Y |
| 183-1 | 4 | W | Y | Y | 3 | N | N | 0.14 | Y | N | N | 0.066 | Y |
| 185-1 | 12 | W | N | N | 1 | N | N | 0.019 | Y | N | N | 0.0094 | Y |

In model: Sample included as a model sample in the PLS-DA model. Total no: Total number of samples from the subject (including model and validation samples). Chocolate/Citrus WDR: Chocolate/Citrus containing product reported in WDR. Chocolate/Citrus supermarket: Chocolate/Citrus containing products collected from the supermarket close to the time of sampling. Chocolate/Citrus PEM level (%): Level of chocolate/citrus markers (see section 6.3.4). Y/N: Yes/No, Sp: Spring, A: Autumn, S: Summer, W: Winter.

## Appendix A

**Misclassified ADD samples**

| ID | Week | Sea-son | PABA Reco-vered | In model | Total no | % Miss-classified | Choco-late WDR | Choco-late super-market | Choco-late PEM level (%) | Choco-late compli-ance | Citrus WDR | Citrus Super-market | Citrus PEM level (%) | Citrus com-pliance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 066-1 | 20 | W | Y | Y | 4 | 25 | - | Y | 0.26 | - | - | Y | 0.041 | - |
| 121-2 | 12 | W | Y | N | 1 | 100 | N | Y | 0.086 | - | Y | Y | 0.077 | - |
| 125-2 | 26 | Sp | Y | Y | 2 | 50 | Y | Y | 0.091 | - | N | N | 0.013 | Y |
| 145-2 | 4 | W | Y | Y | 3 | 33 | - | Y | 0.34 | - | - | Y | 0.068 | - |
| 158-2 | 12 | W | N | Y | 3 | 33 | Y | Y | 0.24 | - | Y | Y | 0.013 | - |

In model: Sample included as a model sample in the PLS-DA model. Total no: Total number of samples from the subject (including model and validation samples). % Misclassified: The number of misclassified samples from the subject out of the total number of available samples from the subject. Chocolate/Citrus WDR: Chocolate/Citrus containing product reported in WDR. Chocolate/Citrus supermarket: Chocolate/Citrus containing products collected from the supermarket close to the time of sampling. Chocolate/Citrus PEM level (%): Level of chocolate/citrus markers (see section 6.3.4). Y/N: Yes/No, Sp: Spring, A: Autumn, S: Summer, W: Winter.

**Selected correctly classified ADD samples**

| ID | Week | Sea-son | PABA Reco-vered | In model | Total no | Chocolate WDR | Chocolate supermar-ket | Chocolate PEM level (%) | Chocolate compliance | Citrus WDR | Citrus Super-market | Citrus PEM level (%) | Citrus com-pliance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 028-2 | 12 | W | N | Y | 4 | Y | Y | 0.16 | - | Y | Y | 0.31 | - |
| 040-2 | 20 | W | Y | N | 2 | - | Y | 0.26 | - | - | Y | 0.15 | - |
| 043-1 | 26 | W | N | Y | 4 | Y | Y | 0.44 | - | N | Y | 0.0093 | - |
| 104-2 | 4 | W | Y | Y | 3 | - | Y | 0.39 | - | Y | Y | 2.53 | Y |
| 120-2 | 12 | W | Y | Y | 2 | N | Y | 0.43 | - | N | Y | 0.73 | Y |

In model: Sample included as a model sample in the PLS-DA model. Total no: Total number of samples from the subject (including model and validation samples). Chocolate/Citrus WDR: Chocolate/Citrus containing product reported in WDR. Chocolate/Citrus supermarket: Chocolate/Citrus containing products collected from the supermarket close to the time of sampling. Chocolate/Citrus PEM level (%): Level of chocolate/citrus markers (see section 6.3.4). Y/N: Yes/No, Sp: Spring, A: Autumn, S: Summer, W: Winter.

# List of papers

Paper I:

**Andersen, M.S.**, Reinbach, H.C., Rinnan, Å., Barri, T., Mithril, C. and Dragsted, L.O. (2013): Discovery of exposure markers in urine for *Brassica*-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics. *Metabolomics* 9, 984-997

Paper II:

**Andersen, M.S.**, Kristensen, M., Manach, C., Pujos-Guillot, E., Poulsen, S.K., Larsen, T.M., Astrup, A., Dragsted, L.O. (2014): Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics, *Analytical and Bioanalytical Chemistry*, DOI 10.1007/s00216-013-7498-5 [Epub ahead of print]

Paper III:

**Andersen, M.S.**, Rinnan, Å., Manach, C., Poulsen, S.K., Pujos-Guillot, E., Larsen, T.M., Astrup, A., Dragsted, L. Untargeted metabolomics as a screening tool for estimating compliance to a dietary pattern. Revised and resubmitted for Journal of Proteome Research

# Author contributions to the papers

## Paper I:

HCR, CM and LOD designed the meal study and MSA and LOD designed the single food studies. The meal study was conducted by HCR and CM and the single food studies were conducted by MSA. Urine samples were analyzed by TB. MSA and ÅR carried out the data analysis and MSA and LOD were the main responsible for marker identification and interpretation of the results. MSA drafted the manuscript which was circulated to all co-authors for comments and approval before submission.

## Paper II:

SKP, TML, AA and LOD designed the study and SKP and TML conducted the study. Urine samples were prepared by MSA and analyzed by MK. Data analysis was conducted by MSA and MSA, CM, EP and LOD were involved in marker identification. The results were interpreted by MSA, CM and LOD. MSA drafted the manuscript which was circulated to all co-authors for comments and approval before submission.

## Paper III:

SKP, TML, AA and LOD designed the study and SKP and TML conducted the study. Urine samples were prepared by MSA. Data analysis was conducted by MSA and ÅR. Marker identification was carried out by MSA, CM, EP and LOD. The responsible for the interpretation of results were MSA, ÅR, CM and LOD. MSA drafted the manuscript which was circulated to all co-authors for comments and approval before submission.

# Paper I

Discovery of exposure markers in urine for *Brassica*-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics

Andersen, M.S., Reinbach, H.C., Rinnan, Å., Barri, T., Mithril, C. and Dragsted, L.O.

Metabolomics (2013); 9, 984-997

Metabolomics (2013) 9:984–997
DOI 10.1007/s11306-013-0522-0

ORIGINAL ARTICLE

# Discovery of exposure markers in urine for *Brassica*-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics

**Maj-Britt Schmidt Andersen · Helene Christine Reinbach · Åsmund Rinnan · Thaer Barri · Charlotte Mithril · Lars Ove Dragsted**

Received: 18 September 2012 / Accepted: 15 March 2013 / Published online: 27 March 2013


**Abstract** An untargeted metabolomics approach has been applied to discover and identify exposure markers in urine for nine Nordic meals. A cross-over meal study was carried out in 17 subjects. The meals included a Pie, a Soup and a Barleyotto (pearl barley based risotto), each prepared with three protein sources; meat, fish or vegetarian. Urine samples were collected in different time intervals before and after intake of the test meals, covering a total of 24 h. The samples were analyzed by UPLC-qTOF-MS. Discriminating features for meals and protein sources were selected by use of double cross-validated partial least squares discriminant analysis and two additional validation steps: (1) time-course of excretion and (2) analysis of sensitivity and specificity. In addition, eight meal studies with single foods were carried out to investigate the food sources of the markers. In total 31 potential exposure markers (PEMs) of foods were found for the meals and protein sources. Fifteen of the 31 PEMs were also found in studies with single foods. Ten PEMs were identified or putatively annotated. Among the PEMs were a range of conjugated isothiocyanates from the *Brassica oleracea* species. Trimethylamine *N*-oxide was found as a fish marker. Additional unknown PEMs were found for chicory salad, parsley and fava beans, while other PEMs were dependent on the meal matrix rather than individual foods. The study demonstrates that it is possible to find PEMs in 24 h urine samples even when foods are given as part of a complex meal.

**Keywords** Food exposure markers · Meal study · Metabolomics · UPLC-QTOF-MS · Urine · Multivariate analysis

## 1 Introduction

Measurement of dietary exposure has been claimed to be one of the most difficult and challenging tasks in nutrition (Bingham 2002; Favé et al. 2009). Systematic and random errors in dietary reporting can significantly influence the outcome of a study and potentially mask or modify associations between diet and disease risk (Bingham 2002). Use of dietary exposure markers measured in biofluids has been demonstrated to correlate well with weighed food records and may therefore serve as an objective measurement to estimate food and nutrient intake (Bingham 2002). However, there are only a limited number of validated dietary markers currently available and the development of tools for assessing dietary intake is lacking behind (Penn et al. 2010). Thus, there is a need to discover more exposure biomarkers in order to better explore the associations between dietary exposure, health and disease risk (Favé et al. 2009; Jenab et al. 2009).

One way to search for new exposure biomarkers is an untargeted metabolomics approach, where a global fingerprint of metabolites in biofluids obtained from nutrition studies is used to discover which metabolites represent specific foods or dietary patterns (Walsh et al. 2007;

**Electronic supplementary material** The online version of this article (doi:10.1007/s11306-013-0522-0) contains supplementary material, which is available to authorized users.

M.-B. S. Andersen (✉) · T. Barri · C. Mithril · L. O. Dragsted
Department of Nutrition, Exercise and Sports,
Faculty of Science, University of Copenhagen,
Rolighedsvej 30, 1958 Frederiksberg C, Denmark
e-mail: mbsa@life.ku.dk

H. C. Reinbach · Å. Rinnan
Department of Food Science, Faculty of Science,
University of Copenhagen, Frederiksberg C, Denmark

Llorach et al. 2010; Lloyd et al. 2011; O'Sullivan et al. 2011). Metabolomics enables simultaneous detection of a high number of low molecular weight metabolites in a single sample. As the dietary metabolic pathways are not always known, metabolomics can contribute to elucidating how dietary components are metabolized. Furthermore, metabolites unique to specific foods or food groups can be explored in order to find the most likely biomarker candidates (Koulman and Volmer 2008; Penn et al. 2010).

Application of metabolomics for biomarker discovery is still a relatively new field and many different metabolomics methods and strategies have been explored (Lodge 2010). The metabolic profile depends on a vast number of factors: timing of food intake and sampling, food preparation and origin, food matrix, and intra- and inter-individual differences to mention a few (Favé et al. 2011; Puiggròs et al. 2011). A standardized protocol has been suggested for fingerprinting of urinary metabolites for specific foods (Lloyd et al. 2011). However, the outcome of a metabolomics study will reflect the study design used. Therefore, the potential exposure markers (PEMs) found in one study can supplement or further consolidate findings from other metabolomics studies with different study designs and sampling procedures.

The current study is part of the large 5-year multidisciplinary Danish research project, OPUS, which aims to develop a New Nordic Diet (NND) that is healthy, sustainable and tasty (Mithril et al. 2012a, b). NND is defined by a number of food groups and amounts (g/10 MJ) (Mithril et al. 2012b). It differs from an average Danish diet (ADD) by the level of intake of several foods like: root vegetables, cabbages, fish, nuts and edible plants from the wild landscapes growing in the North. One of the objectives in OPUS is to investigate if it is possible to discover a range of exposure markers by metabolomics that can be used to discriminate between subjects eating NND and ADD. Such exposure markers would be useful for future studies investigating the health potential of NND.

In this paper, we report on the findings from a controlled cross-over meal study with nine *Brassica*-containing NND meals (Reinbach et al., unpublished observation) designed to investigate whether it is possible to identify exposure markers for specific NND meals and foods in 24 h urine samples by a metabolomics approach.

## 2 Materials and methods

### 2.1 Study designs

#### *2.1.1 Meal study*

A complete cross-over meal study with a total of nine test meals—three meals each prepared with three different protein sources—was performed, as outlined in Fig. 1. The study has been approved by the Regional Ethics Committee (H-1-2011-016) and the Danish Data Protection Agency (2007-54-0269).

Subjects were recruited from the local area around Frederiksberg campus at University of Copenhagen by poster and website announcements. They were assessed for suitability with the following criteria; healthy normal-weight men and women, between 20 and 50 years of age with no food allergies.

Four males and thirteen females completed the study. They had an average age of 27.8 ± 6.4 and body mass index (BMI) of 22.4 ± 2.1 kg/m$^2$.

On each study day, the subjects had a standardized breakfast at 9 am, an apple between 9 and 11 am and a test meal between 12 and 1 pm. Four hours after the test meal had been served, an ad libitum standardized dinner was given between 4 and 5 pm. There were no dietary restrictions or standardizations before 9 am and after 5 pm on each study day.

Urine was collected on each test day in four time spans covering 24 h in total: (1) from 9 am (excluding first void) until intake of the test meal. (2) From intake of test meal until 2 h after intake of test meal. (3) From 2 h after intake of test meal until dinner. (4) From dinner until 9 am the next morning (including first void). The urine samples collected were analyzed separately and also as a pooled 24 h urine sample. The pooled sample was made by mixing urines 1–4 after removing 1.5 ml from each.

The study was carried out on three consecutive days in three consecutive weeks. Each of the three meals (Soup, Barleyotto and Pie) was served once per week in a randomized order and the participants were randomized to one of three protein types of the meal served on each study day (meat, fish or vegetarian).

#### *2.1.2 Meals*

The standardized breakfast consisted of yoghurt with muesli, a carrot bun, cheese, raspberry jam and a drink of free choice that had to be the same on each test day (tee, coffee, water, juice or milk). The ad libitum standardized dinner consisted of 700 g pizza with ham and cheese corresponding to 7,000 kJ, which was served for each participant. The participants were requested to drink 0.5 l bottled water before and after the test meal. In addition they were asked to consume 2.5 dl water with the test meal and the dinner.

The test meals were a Soup, a Barleyotto and a Pie. The Soup was based on chicken stock with kale, parsley roots and carrots. The Pie was a wholemeal flour base filled with broccoli, pointed cabbage and chicory salad and the Barleyotto consisted of boiled pearl barley with cream cheese

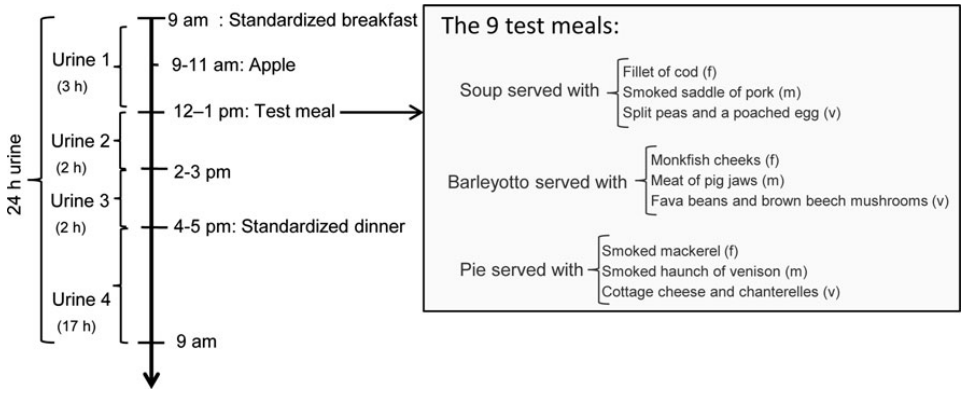



**Fig. 1** Outline of study design and test meals. A schedule of meals and urine sampling for each study day is given on the timeline to the *left*. The *right part* of the figure shows the nine test meals. A Soup, Barleyotto and a Pie meal were each prepared with three different protein sources, one with fish (f), one with meat (m) and one vegetarian (v)

and boullion. The Pie was served with chicory salad while the Soup and the Barleyotto meals were served with a salad of white cabbage, Brussels sprouts and apple. All meals were served with bread. The meat, fish and vegetarian protein sources used in the meals are given in Fig. 1. Details on amounts and ingredients and preparation methods of all meals are given in the supplementary material (Tables S1 and S2).

### *2.1.3 Single-food studies*

To confirm the dietary origin of the PEMs found in the meal study, a range of small single food studies (SFS) were performed with 3–4 subjects in each. In these studies, the subjects were not allowed to eat or drink anything between 9 am and 3 pm, except from water and a test meal, which was served at 12–1 pm. The test meal was served ad libitum and consisted of a single food prepared in accordance with how it was served in the meal study. Urine was collected in two time intervals. From 11 am until intake of the test meal and from intake of the test meal until 3 pm. An outline of the study design for SFS is given in the supplemental material Fig. S1.

The subjects participating in SFS were not the same as the subjects in the meal study. The only inclusion criterion for SFS was willingness to eat the served foods. In total four males and seven females participated in SFS. The subjects had an average age of 32.4 ± 9.4 and BMI of 21.9 ± 1.3 kg/m$^2$ and participated in one to six SFS each.

Foods tested in SFS were: Raw white cabbage (ingredient in the Soup, and the Barleyotto meals) raw Brussels sprouts (ingredient in the Soup, and the Barleyotto meals), boiled carrots (ingredient in the Soup), boiled parsley roots (ingredient in the Soup), boiled kale (ingredient in the Soup), raw chicory salad (ingredient in the Pie meal), brown beech mushrooms fried in butter (ingredient in the vegetarian Barleyotto meal), boiled fresh and dried fava beans (ingredient in the vegetarian Barleyotto meal).

## 2.2 Sample preparation

### *2.2.1 Urine samples*

During sampling, urine was stored below 5 °C in cooler bags and after completion, sample aliquots were kept at −80 °C until the time of analysis. The samples were analyzed both separately (urine 1–4) and combined (pooled samples representing 24 h) as shown in Fig. 1. After centrifugation (4 °C, 4 min, 2,700 RCF), 150 µl of each urine sample was added to a well in 96-well collection plates (Waters, Milford, MA) together with 150 µl of solvent (aqueous 5 % 30:70 (v/v) acetonitrile (ACN):methanol (MeOH), Optima grade LC–MS, Fisher Scientific, US) containing an internal standard mixture similar to that described previously (Barri et al. 2012). For quality control purposes, an external standards mixture and within-plate urine pools were used. Samples from the same person were randomized within one plate to minimize intra-individual variation due to plate differences.

### *2.2.2 Food extracts*

To aid the metabolite identification, two food extracts, one with ethanol and one with water, were prepared for chicory salad, fresh fava beans, parsley root, carrots, white cabbage and Brussels sprouts. Approximately 1 g of prepared food from SFS was crushed in a mortar and mixed with 3 ml water or ethanol for a few minutes. The food samples were

centrifuged and supernatants were removed, filtered (0.2 μm filter, SMI Lab-Hut, UK) and evaporated. The dried samples were redissolved in 200 μl water or ethanol prior to analysis.

#### 2.2.3 *Glucuronidation of standards*

For glucuronidation, the method described by Z. Liu was applied (1984). In brief, diosmetin was added to an aqueous mixture of $MgCl_2$, $KP_i$ buffered liver extract (pH 7.4) and uridinediphosphoglucuronic acid (UDPGA). The mixture was incubated for 1 h at 37 °C, after which methanol was added to terminate the enzymatic reaction. The supernatant was withdrawn, evaporated to dryness and redissolved in aqueous 5 % 70:30 (v/v) MeOH:ACN. The liver extract used was prepared according to the optimized conditions in Nelson et al. (2001).

### 2.3 UPLC-qTOF-MS analysis

Urine samples and food extracts were analyzed by an ultra-performance liquid chromatography quadruple time-of-flight mass spectrometer (UPLC-qTOF-MS), Waters Corporation, Manchester, UK, as described previously (Barri et al. 2012). The UPLC was equipped with an HSS T3 $C_{18}$ column and pre-column (Waters, Milford, MA). The mobile phase was 0.1 % formic acid in milli-Q® water, (Billerica, MA, US) (A) and 0.1 % formic acid (Sigma-Aldrich, Germany) in 30:70 (v/v) MeOH:ACN (B). A mobile phase concentration gradient from (A) to (B) was applied in a total run time of 7 min together with a flow gradient. Positive and negative acquisition modes of electrospray ionization were implemented at probe voltage of 3.2 and 2.8 kV, respectively. The selected *m/z* range was from 50 to 1,000 Da. On-line accurate mass calibration was performed by infusing leucine encephalin as lock-mass solution.

### 2.4 Identification

Identification was mainly done by use of MS/MS data, searches in the Human Metabolome Database (www.hmdb.ca) and by matching of *m/z* and retention times to results obtained from SFS and food extracts. In addition, literature searches were carried out to find information on known metabolites and characteristic compounds present in the investigated single foods. MS/MS fragmentation of markers was performed in product ion scan mode by ramping collision energy from 10 to 40 eV and using cone voltage of 25 V.

Levels of identification are reported in accordance with (Sumner et al. 2007). Compounds identified by matching masses and retention times with authentic standards both alone and spiked into a sample are reported as level I. Compounds only identified by MS/MS spectra and matched to spectra from databases were annotated as Level II. Compounds identified by spectral similarities to a similar compound class and knowledge from previous literature are reported as level III. Unknown compounds are reported as level IV.

The following standards were run for level one identifications: Diosmetin and trimethylamine *N*-oxide dehydrate (Sigma-Aldrich, Germany), D, L-Sulforaphane *N*-Acetyl-L-cysteine and *N*-Acetyl-S-(*N*-benzylthiocarbamoyl)-L-cysteine (AH diagnostics, Denmark). Diosmetin was glucuronated prior to analysis.

### 2.5 Data processing and pre-treatment

All raw spectra were converted to netCDF files using DataBridge (Waters) and imported into MZmine2 (Pluskal et al. 2010) for data processing. Data from the meal study and from the SFS were processed separately in MZmine. The parameters used were optimized for positive and negative mode individually, using a subset of data and the external standards. Applied parameters and batch steps are given in the supplementary material, Table S3.

After processing, the retention times (RT), *m/z* values and peak areas from the generated peak lists in positive and negative mode were imported into MATLAB® version 7.12.0.635, R2011a (Mathworks Inc., Sherborn, MA, US). The term 'feature' will be used to designate a peak from the MZmine peak list throughout the rest of this paper.

Before the statistical analysis, data from the meal study was pre-treated in order to correct for unsystematic and systematic variation caused by: (1) Presence of noisy features. (2) Changes in overall spectra signal intensities across sample runs. (3) Differences in urine concentrations. (4) Individual differences. Pre-treatment of the data was done in four steps as described below (paragraph 2.5.1–2.5.4). The processed data from SFS was not pre-treated. This data was used only to match *m/z* and retention times to PEMs found in the meal study in order to investigate the food source(s) of the markers.

#### 2.5.1 *80 % rule*

First, the 80 % rule (Bijlsma et al. 2006) was applied. The rule was slightly modified since it is not possible to use zero as the threshold when using processing software with a gap filling algorithm. Instead, an iteration procedure with varying thresholds was used to find the optimal threshold. Data was divided into 36 groups, where each group consisted of samples from one urine sampling point (urine 1–4) and one of the nine meals. A threshold was applied and if at least 80 % of the measurements for a feature was

not above the threshold in at least one group, the feature was removed.

#### 2.5.2 Normalization

After applying the 80 % rule, the reduced matrix was normalized to a total area of 1,000 for each sample across the remaining features. This was done in order to correct for differences in urine concentration and ionization levels during sample runs.

#### 2.5.3 Person correction

Individual differences in metabolite levels were corrected by a feature-wise normalization across individuals. For each individual, the measurement of a feature in a given sample was divided by the sum of all sample measurements for the same feature and person:

$$C_{k,i,j} = \frac{X_{k,i,j}}{\sum X_{k,i}}$$

where $x_{k,i,j}$ are the measurements of feature $i$ on person $k$ for the $j$'th meal- and timepoint. $C_{k,i,j}$ is the person corrected value.

#### 2.5.4 Average Euclidian distance

A new matrix was calculated from the normalized and person corrected matrix by taking the mean of all samples from each of the nine meals. This was done for each feature, resulting in a [meal type × number of features] matrix. The Euclidian distance between all pairwise meals was calculated and the average distance between two meals was used to evaluate the optimum threshold for the 80 % rule.

For both positive and negative mode data, all four steps (paragraph 2.5.1–2.5.4) were iterated with thresholds from 0 to 5 and increases of 0.1 per iteration. The normalized and person corrected matrix with the highest average Euclidian distance was used for the statistical analysis.

### 2.6 Statistical analysis

The statistical analysis was in two parts. First, a multivariate partial least squares discriminant analysis (PLS-DA) was performed on 24 h samples from the pre-treated data matrices in negative and positive mode. Next, validity of the discriminating features in PLS-DA as PEMs was investigated by a validation procedure in two steps. The complete statistical analysis was carried out in MATLAB®. For the PLS-DA analyses, the PLS-toolbox (version 6.5.1, Eigenvector Research, Inc., MA, US) for MATLAB® was used.

#### 2.6.1 Multivariate analysis by PLS-DA

Two separate PLS-DA analyses were made. In one, the three meals, Pie, Soup and Barleyotto were used as class vectors and in the other, the three protein sources fish (f), meat (m) and vegetarian (v) were used as class vectors (Fig. 1). All data was autoscaled before applying PLS-DA.

For both class vectors, one optimized PLS-DA model was developed and validated as described in points A–D:

A. Initially, seventeen PLS-DA models were made in which two subjects were left out in each model, thus leaving out each subject from a total of two models. Each of the seventeen models were reduced according to variable importance in projection (VIP) scores (Wold et al. 1993; Chong and Jun 2005) as follows: The PLS-DA model was cross-validated by leaving out data from one person at a time. Features with a VIP score below one for all classes in the model were removed and a new model was developed with the remaining features. This procedure was iterated until all included features in the model had VIP scores above one for at least one class. The number of latent variables used in each iterated model was automatically selected as the lowest number for which the mean of the cross validated classification errors was below 0.02 or, alternatively, the number with the lowest cross validated classification error mean.
B. Features found from the models in A were ranked according to how many of the 17 models included a specific feature.
C. Seventeen new PLS-DA models including all subjects were then made based on the features selected in the initial models. For the first model, features present in at least one of the initial models after feature selection were included. The model was cross-validated with the same segmentation as the one used for the initial models. Then, a second model was build including features present in at least two of the initial models. This procedure was repeated until 17 new models had been made, including features present in one to 17 of the initial models.
D. The model in C with the lowest mean cross-validated classification error was selected as the optimized model and the included features in that model were used for the second part of the statistical analysis.

#### 2.6.2 Validation

For validation, raw data from the selected features in D was used. First, the sensitivity and specificity (percentage of correct classifications for each group) were investigated by looking at the distribution of raw data from 24 h. Only

features for which a specificity and sensitivity of at least 80 % could be obtained for a given threshold, when comparing measurements of the feature for one meal group, protein group, or individual meal, to the rest of the meals, were included for the second part of validation. In the second validation part, raw data for each feature from all urine sampling points were plotted to investigate if the feature had a meaningful time-course. Only features fulfilling both parts of the data validation were considered as PEMs.

## 3 Results and discussion

Overall, 144 complete 24 h urine pools were collected. Processing of the samples in MZmine gave 5,129 and 4,684 features in positive and negative mode, respectively. After pre-treatment of the data, 3,955 and 2,978 features remained for the statistical analysis. An optimal threshold of 1.7 was found for the iteration procedure.

### 3.1 PLS-DA analysis

The results from the PLS-DA models are reported in Table 1. For both positive and negative mode, where meal was used as class vector, the initial PLS-DA models for meals were very robust, while the initial PLS-DA models for protein sources had high classification errors both for cross-validation and for the test sets (14.2–39.1 %, Table 1). The higher classification errors for the protein PLS-DA models can be explained by the protein sources being less comparable than the meals. Whereas each meal class had most ingredients in common and only deviated in the protein source (meat, fish or vegetarian), the grouping of the three fish, meat, and vegetarian meals as class vector is a grouping where each class does not have any ingredients in common. The higher classification errors for protein sources even after variable selection demonstrate that very few features distinguish the individual protein sources well and it can be expected that a high proportion of the features left in these models are not PEMs.

Use of class vectors based on protein sources and meals has the advantage of including the full dataset, thereby increasing the number of observations within each class. However, combining data known to be from different meals into one class will also add unwanted variation to the model. Which approach is the best depends on how much variation is caused by within class differences due to different meal compositions compared to the total variation within a meal or protein class. In order to investigate the within class variation we have performed a principal component analysis (PCA) and a PLS-DA including only data from the three protein sources of one meal at a time. These models were not good (data not shown), suggesting that the main variation in the dataset from the meals is from the meals rather than the protein sources. For that reason, the models including the full dataset were chosen for the data analysis. If an individual meal gives rise to strong PEMs, these will most likely still be included after variable selection in the optimized models, since such markers will classify one third of the samples within a class.

When comparing the optimized PLS-DA models for meals and protein, the protein models are characterized by both a higher number of latent variables and a higher number of features (Table 1). However, the cross-validated classification errors for all the models are acceptable (<4.5 %) and for that reason, all features selected in the PLS-DA models were used for the validation step.

### 3.2 Validation

An example of feature validation for four characteristic PEMs is given in Fig. 2a–d. Figure 2a is a PEM for the Barleyotto and the Soup meals, Fig. 2b is a PEM for fish meals, Fig. 2c is a PEM for Pie, and Fig. 2d is a PEM for the Soup meal. All PEMs illustrated in Fig. 2 show a clear

**Table 1** Results from the PLS-DA models

| PLS-DA models | | Positive mode | | Negative mode | |
|---|---|---|---|---|---|
| | | Meals | Protein | Meals | Protein |
| Initial | CV class error | 0.007 ± 0.005 | 0.142 ± 0.038 | 0.059 ± 0.066 | 0.286 ± 0.0515 |
| | TS class error | 0.008 ± 0.014 | 0.194 ± 0.056 | 0.057 ± 0.103 | 0.391 ± 0.077 |
| | Number of LVs | 2.29 ± 0.47 | 5.47 ± 2.15 | 4.76 ± 2.05 | 6.70 ± 2.44 |
| | Number of features | 108 ± 5 | 32 ± 36 | 79 ± 18 | 333 ± 97 |
| Optimised | CV class error | 0.007 | 0.040 | 0.045 | 0.045 |
| | Number of LVs | 2 | 5 | 3 | 7 |
| | Number of features | 148 | 180 | 156 | 370 |

For the initial models, the mean and the standard deviation of the results from all 17 models are reported

*CV class error*, cross validated classification error; *TS class error*, test set classification error; *LVs*, latent variables

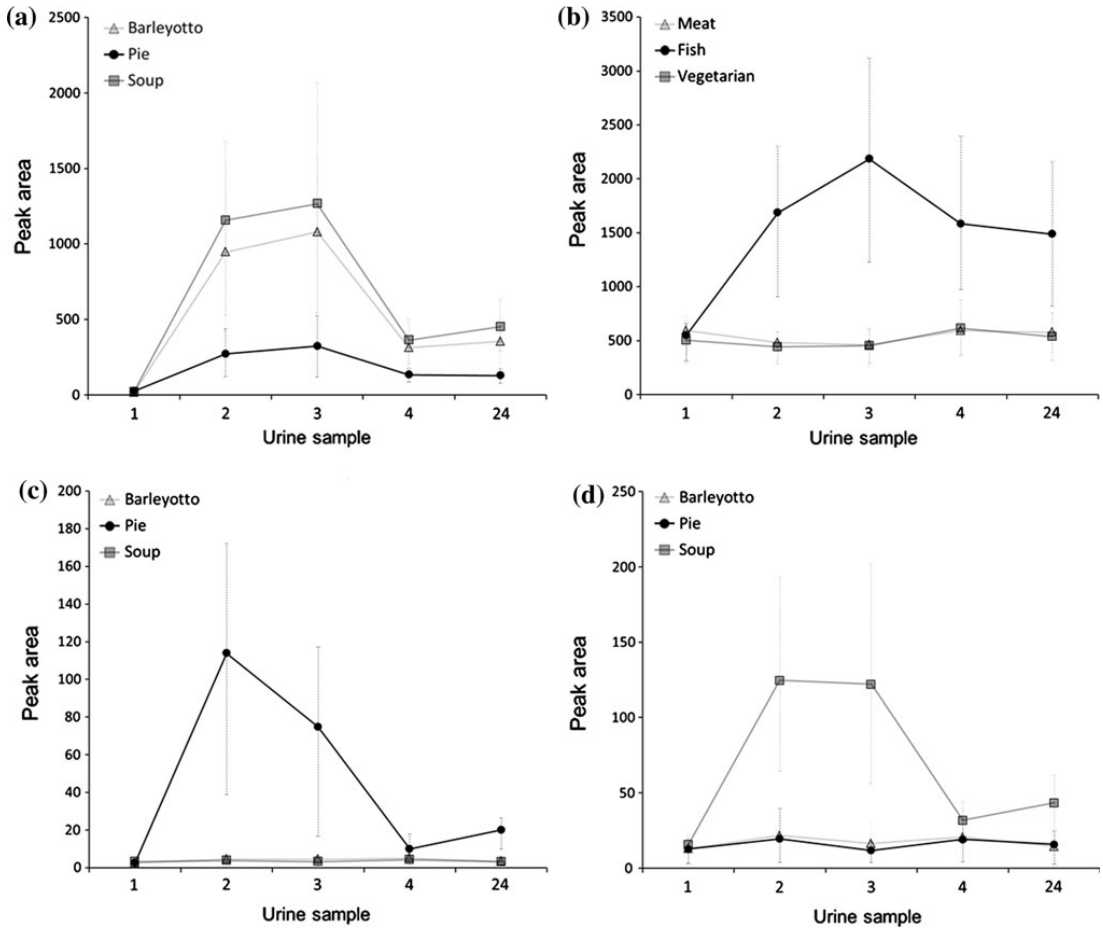



**Fig. 2** Data examples for four characteristic PEMs. Average peak area of each feature for each urine sampling point (1, 2, 3, 4 and 24) is shown for each meal (Barleyotto, Pie and Soup) or protein (Meat, Fish and Vegetarian) type. *Error bars* in the figure are 20 and 80 th percentiles. **a** PEM for Barleyotto and Soup, [$m/z = 327.052$, RT = 2.76], **b** PEM for fish, [$m/z = 76.076$, RT = 0.49], **c** PEM for Pie, [$m/z = 398.129$, RT = 2.30], **d** PEM for Soup [$m/z = 301.073$, RT = 3.70]

time-sequence for excretion and for the 24 h samples, the 20 th percentile of the meal(s) the PEM is a marker for, is higher than the 80 th percentile of the other meals. In general, highest variation is found for urine samples 2 and 3 probably reflecting individual differences in excretion rates and in excretion half-lives.

In total, 45 and 12 PEMs were found in positive and negative mode, respectively. These PEMs corresponded to 30 unique PEMs, as many were fragments or adducts of the same metabolite. The distribution across meals for the PEMs found is given in Table 2. Most PEMs were Barleyotto and Soup markers following the same pattern as in Fig. 2a. In addition, PEMs for fish intake regardless of meal type, PEMs for the Soup meal, and PEMs for the Barleyotto vegetarian and fish meals were found. Overall,

**Table 2** Meal sources of PEMs found in the study

| Meal(s) | Total PEMs | Unique PEMs |
|---|---|---|
| Barleyotto and soup | 22 | 6 |
| Pie | 15 | 7 |
| Soup | 9 | 8 |
| Barleyotto vegetarian meal | 8 | 7 |
| Meals with fish as protein source | 2 | 1 |
| Barleyotto fish meal | 1 | 1 |

The meals in the left column are meals containing higher levels of the PEMs compared to the rest of the meals. Total PEMs is total number of PEMs found for each meal including fragments and adducts. The number of unique PEMs in column three is the corresponding number of metabolites, excluding fragments and adducts

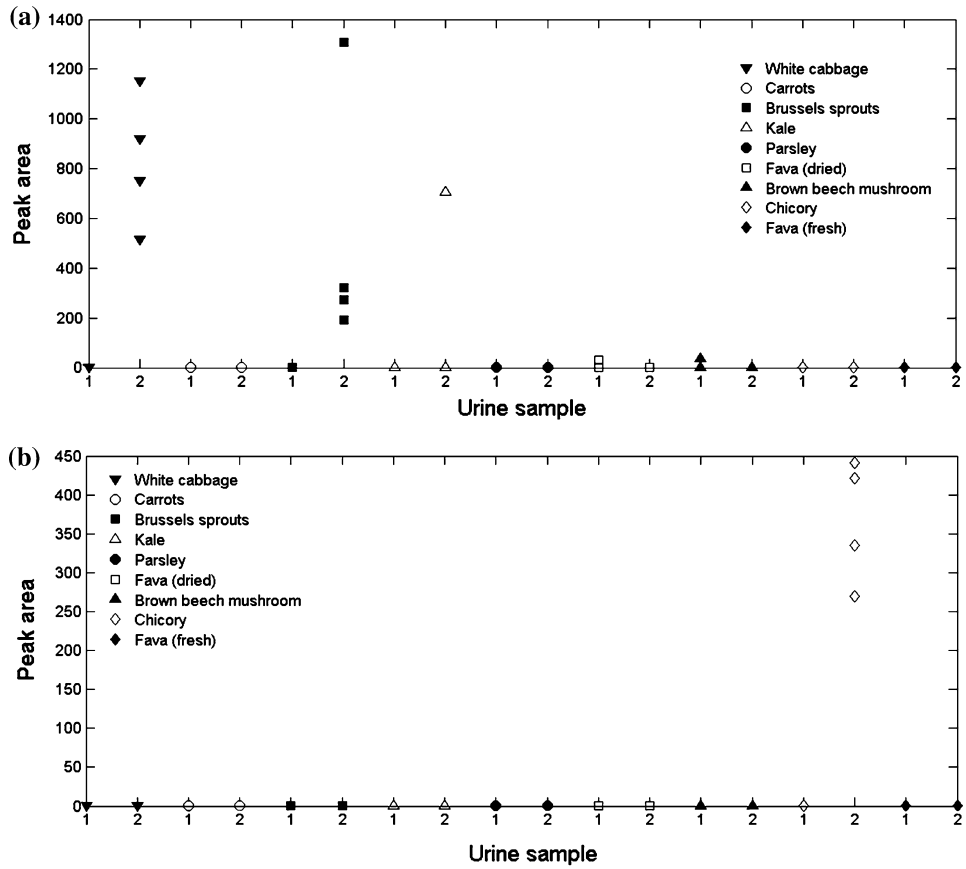



**Fig. 3** Peak area of two PEMs, for each subject before (urine sample 1) and after (urine sample 2) intake of different single foods. **a** PEM found after intake of white cabbage and Brussels sprouts [$m/z = 327.052$, RT = 2.76], **b** PEM found after intake of chicory salad [$m/z = 398.129$, RT = 2.30]

very few markers characterize protein sources compared to meals, which is in accordance with the PLS-DA model on protein classes being inferior to the PLS-DA model on meal classes for predictive strength.

### 3.3 Identification

Fifteen of the unique PEMs in Table 2 were also found in SFS. Eight of the PEMs were found in SFS with both Brussels sprouts and white cabbage while five PEMs were found after intake of parsley root. Intake of chicory salad and fresh fava beans gave rise to one unique PEM each.

In Fig. 3a and b, examples of data from two PEMs found in SFS are given, one for Brussels sprouts and white cabbage and one for chicory salad. The two PEMs are the same as depicted in Fig. 2a and c.

Even though other subjects were used in SFS compared to the meal study and not the same subjects were used in all SFS, there was a clear trend for the PEMs found in SFS. It cannot be excluded that some PEMs found in SFS are present in other food sources from the meals as well, since only a limited number of ingredients were tested in SFS. Furthermore, the short duration of urine collection in SFS compared to the meal study may affect the result even though peak time of excretion for almost all PEM was within 4 h after intake of the meals (Table 3). The only protein source for which a high number of markers were found in the study was the Barleyotto vegetarian meal. Only one out of seven markers of the Barleyotto vegetarian meal was confirmed in SFS, even though all the foods present in the vegetarian version of the meal were investigated (brown beech mushroom and fava beans). This suggests that the rest of the PEMs found are matrix dependent either in their formation or in their time-course of excretion and it confirms the importance of testing the presence of a marker also in individual foods.

**Table 3** List of PEMs in the meal study that have been identified on level I-III and/or found in SFS

| Meal(s) | Found in SFS | *m/z* | RT (min) | Adduct/fragment/parent ion | Identification | Level | Peak time | Present in other meals |
|---|---|---|---|---|---|---|---|---|
| Barleyotto and Soup | White cabbage Brussels sprouts | 333.041 | 3.67 | $[M + Na]^+$ | *N*-acetyl-S-(*N*-3-methylthiopropyl)cysteine | III | 3 | Yes |
| | White cabbage | 285.037 | 3.52 | $[M + Na]^+$ | *N*-acetyl-S-(*N*-allylthiocarbamoyl)cysteine (AITC-NAC) | III | 3 | Yes |
| | Brussels sprouts | 263.054 | 3.54 | $[M + H]^+$ | | | | |
| | | 187.057 | 3.51 | $[M + H\text{-}75.997]^+$ | | | | |
| | | 164.039 | 3.55 | $[M + H\text{-}99.015]^+$ | | | | |
| | | 146.029 | 3.51 | $[M + H\text{-}117.025]^+$ | | | | |
| | | 134.010 | 3.56 | $[M + H\text{-}129.044]^+$ | | | | |
| | | 122.028 | 3.55 | $[M + H\text{-}141.026]^+$ | | | | |
| | | 162.022 | 3.52 | $[M - H\text{-}99.015]^-$ | | | | |
| | | 132.002 | 3.52 | $[M - H\text{-}129.036]^-$ | | | | |
| | | 84.044 | 3.50 | $[M - H\text{-}176.993]^-$ | | | | |
| | White cabbage | 349.035 | 2.76 | $[M + Na]^+$ | Iberin *N*-acetyl-cysteine (IB-NAC) | III | 2–3 | Yes |
| | Brussels sprouts | 327.052 | 2.76 | $[M + H]^+$ | | | | |
| | | 263.053 | 2.77 | $[M + H\text{-}63.998]^+$ | | | | |
| | | 198.009 | 2.74 | $[M + H\text{-}129.043]^+$ | | | | |
| | | 164.024 | 2.76 | $[M + H\text{-}163.028]^+$ | | | | |
| | | 134.012 | 2.73 | $[M + H\text{-}193.040]^+$ | | | | |
| | | 325.045 | 2.81 | $[M - H]^-$ | | | | |
| | | 162.024 | 2.73 | $[M - H\text{-}163.021]^-$ | | | | |
| | White cabbage Brussels sprouts | 382.062 | 1.88 | $[M + H]^+$ | *N*-acetyl-cysteine conjugate | III | 2 | Yes |
| | – | 187.056 | 1.01 | $[M + CH_2O_2]^+$ | 4-iminopentylisothiocyanate | III | 3 | Yes |
| Pie | White cabbage | 347.056 | 3.81 | $[M + Na]^+$ | Erucin *N*-acetyl-cysteine (ERN-NAC) | III | 3 | Yes |
| | Brussels sprouts | 325.074 | 3.80 | $[M + H]^+$ | | | | |
| | | 196.026 | 3.78 | $[M + H\text{-}129.047]^+$ | | | | |
| | White cabbage | 335.050 | 3.76 | $[M + Na]^+$ | *N*-Acetyl-(*N*'-benzylthiocarbamoyl)-cysteine | I | 3 | Yes |
| | Brussels sprouts | 313.066 | 3.74 | $[M + H]^+$ | | | | |
| | | 184.023 | 3.76 | $[M + H\text{-}129.043]^+$ | | | | |
| Pie | White cabbage | 363.051 | 3.16 | $[M + Na]^+$ | Sulforaphane *N*-acetyl-cysteine (SFN-NAC) | I | 3 | Yes |
| | Brussels sprouts | 341.067 | 3.17 | $[M + H]^+$ | | | | |
| | | 212.026 | 3.19 | $[M + H\text{-}129.041]^+$ | | | | |
| | | 178.044 | 3.25 | $[M + H\text{-}163.022]^+$ | | | | |
| | | 114.041 | 3.21 | $[M + H\text{-}227.026]^+$ | | | | |

**Table 3** continued

| Meal(s) | Found in SFS | *m/z* | RT (min) | Adduct/fragment/parent ion | Identification | Level | Peak time | Present in other meals |
|---|---|---|---|---|---|---|---|---|
| | White cabbage Brussels sprouts | 299.058 | 1.81 | $[M + H]^+$ | Sulforaphane *N*-cysteine (SFN-Cys) | III | 2–3 | Yes |
| | Chicory salad | 398.129 | 2.30 | Pos mode | Unknown. Present in aqueous food extract. | IV | 2 | No |
| Soup | Parsley root | 285.047 | 3.85 | Neg mode | Unknown | IV | 2 | No |
| | Parsley root | 259.027 | 3.80 | Neg mode | Unknown | IV | 2 | No |
| | Parsley root | 289.040 | 3.76 | Neg mode | Unknown | IV | 2–3 | No |
| | Parsley root | 475.097 | 3.68 | $[M - H]^-$ | Unknown glucuronide | IV | 2 | No |
| | | 301.073 | 3.70 | $[M + H\text{-}176.0321]^+$ | | | | |
| | Parsley root | 301.037 | 3.64 | Neg mode | Unknown | IV | 2 | No |
| Barleyotto vegetarian meal | Fresh fava beans | 327.116 | 0.69 | Pos mode | Unknown | IV | 3 | No |
| Meals with fish as protein source | NI | 151.145 | 0.49 | $[2M + H]^+$ | Trimethylamine *N*-oxide (TMAO) | I | 3 | No |
| | | 76.076 | 0.49 | $[M + H]^+$ | | | | |

In the meal column, the meals with the highest levels of the feature are listed. Peak time is the urine sample point where the highest mean level of the feature was detected. Foods in the 'Found in SFS' column are foods in SFS for which the feature was present in all subjects after intake. The level of identification is reported in the 'Level' column

*NI* Not investigated

In Table 3, all PEMs that were identified at level I-III and/or found in SFS are listed. The rest of the PEMs that were neither identified nor found in SFS are provided as supplementary material (Table S4). As we have not been able to find a food source of the markers of the PEMs in the supplemental material and we did not find any database matches, they may originate from the meal matrix or other single foods. Identification of these markers would therefore require further experiments and more sensitive equipment.

Four PEMs present in SFS after intake of white cabbage and Brussels sprouts were markers of the Soup and Barleyotto meals (Table 3). This finding is not surprising since these foods were served in the salad accompanying both these meals. However, Brussels sprouts and white cabbage also gave rise to PEMs characterizing the Pie meal even though none of these ingredients were present in that meal. Brussels sprouts and white cabbage are botanically related as they both belong to the plant species, *Brassica oleracea* (Clarke 2010). Kale (ingredient in the Soup), broccoli and pointed cabbage (ingredients in the Pie) belong to this species as well, which might explain why slightly higher levels of these PEMs are found after consuming kale containing Soup compared to Barleyotto in Fig. 2a. Also, this explains why a time-sequence is seen even for the Pie meal in the same figure.

All the PEMs found in white cabbage and Brussels sprouts were identified at level I or III as acetyl-cysteine derivatives of isothiocyanates from *Brassica o.* The assignment of this group of compounds as acetyl-cysteine derivatives was made based on shared fragments or mass losses with the level I identified markers *N*-Acetyl-(*N*'-benzylthiocarbamoyl)-cysteine and Sulforaphane *N*-acetyl-cysteine. Compounds were annotated as acetyl-cysteine derivatives if either the fragment with $m/z = 164.038$, which corresponds to acetyl-cysteine or losses of 163.030 or 129.043, which corresponds to acetyl-cysteine loss with and without sulfur, respectively, were present in the raw LC–MS spectra and/or in MS/MS data. The isothiocyanate moiety was annotated as well, if any of the fragments found corresponded to known isothiocyanates from the *Brassica o.* varieties used in the study.

Unfortunately, it was not possible to identify any of the five PEMs for parsley root found. The feature with $[M - H]^- = 475.097$ is probably a glucuronide of a diosmetin isomer. We have synthesized a mixture of diosmetin (luteolin 4′-methyl ether) glucuronides and obtained the same *m/z* but none of the retention times for the glucuronidated products matched the retention time for the PEM. However, luteolin could also be methylated in another position, which may be the correct marker.

One PEM from Soup and Barleyotto were identified at level III as 4-iminopentylisothiocyanate. This metabolite has not been described in previous literature. In MS/MS, two fragments of this PEM were obtained, 141.049 and 114.040, and annotated as the parent ion ($[M + H]^+$) and loss of hydrogen cyanide ($[M - CNH]^+$), respectively.

Only the PEM from chicory salad was also present in a food extract but despite this, it was not possible to identify the compound. As many food components undergo chemical modifications such as methylation, glucuronidation and sulfation, we also searched the food extract chromatograms for masses corresponding to loss of these chemical groups where it was relevant.

The PEM from the fish was identified as TMAO, while the PEM from fresh fava beans could not be identified. Interestingly, the fava bean marker was not detected after intake of dried fava beans.

### 3.4 Study design

The number of discriminating features found in the meal study is relatively low compared to what was reported in another meal metabolomics study aiming to find new food exposure markers (Lloyd et al. 2011). There are several explanations for this difference. Some of the differences could be caused by the different metabolomics techniques and instruments used for sample analysis. In addition, the study design and data analysis applied in the present study were very conservative, since very strict criteria have been applied in order to define a feature as a PEM. For example, 24 h samples were applied for feature selection in this study, whereas 3 h samples were used in the other meal study. As excretion of the majority of compounds reached maximum within 3–4 h after a meal, large contrasts can be expected when 3 h samples are used. The meals in the present study were also more complex, as they contained more ingredients. In the other study, only one ingredient was replaced at a time in the meal (Lloyd et al. 2011) and the analysis was done in a longitudinal way and not across meals for the same time point. The advantage of increasing the complexity of the meals in the present study and using 24 h samples is that there is a higher chance that the PEMs found are valid exposure markers also in studies with free-living subjects. Using 24 h urine may aid in the detection of both acute and habitual markers since they cover exposure during a longer time-span which is independent of the time since the last meal. There is also a high chance of catching late eluting markers in 24 h samples. For example, some of the markers found in the present study reached a maximum of excretion later than 4 h after the meal (Table S4). Another reason to investigate 24 h urine samples is that such samples are already commonly collected in nutrition studies.

An important limitation of the study is the limited number of subjects. It is likely that not all important phenotypes are represented in such a small study size. In

addition, the subjects were all healthy, normal weight and within a relatively narrow age span. Despite the high sensitivity and specificity obtained for the PEMs, they may therefore not be valid for a larger study population. Another factor contributing to this is the standardization of two meals during 24 h which will reduce variation caused by exposure to other foods than the test meals and increase the possibility of finding meal markers. The gender distribution in the study was skewed with a majority of women. We did not consider a potential effect of the hormonal cycle for the women in the study and we also did not test if there are markers that are only found in women, since the purpose of the study was to discover exposure markers that would potentially be applicable for the general population. The cut-offs for sensitivity and specificity applied in the study ensures that none of the PEMs found are valid for women only.

From the present meal study, it is not possible to evaluate if there is a dose–response relationship between the amount eaten and the amount present in 24 h urine, which is an important criteria for evaluating food exposure markers (Spencer et al. 2008). In the SFS, the levels of excretion often did not exceed the maximum levels found in the meal study even though a much higher dose of the single foods were consumed in these studies. However, this may be explained by the SFS samples representing a 3 h postprandial excretion which is not directly comparable to the sampling points in the meal study.

### 3.5 Knowledge on the PEMs found

For foods belonging to the *Brassica o.*, we have identified or putatively annotated nine PEMs. Most of these PEMs are *N*-acetyl- or cysteine conjugates of isothiocyanates and are endproducts of the mercapturic acid pathway (Janobi et al. 2006). Isothiocyanates are formed from glucosinolates by an enzymatic hydrolysis with myrosinase which takes place upon cell damage, such as chewing or in the intestinal tract by microbial reactions (Zhang 2004). Glucosinolates are almost unique to plants from the Brassicales order of which most of the common food crops are found within the family of Brassicaceae (Clarke 2010).

Allyl isothiocyanate which is excreted as *N*-acetyl-S-(*N*-allylthiocarbamoyl)cysteine (AITC-NAC), is the major isothiocyanate formed from most food crops in the Brassicaceae family, except broccoli, and has also been found in high quantities in human urine after consumption of different *Brassica o.* cultivars (Hwang and Jeffery 2003; Rouzaud et al. 2004; Vermeulen et al. 2006). We confirm here that AITC-NAC is excreted in high quantities after consumption of the Soup and Barleyotto where the intake of Brassicaceae other than broccoli was highest. The major isothiocyanates formed after intake of broccoli are sulforaphane, iberin and erucin (Vermeulen et al. 2006). In accordance with this, sulforaphane and erucin conjugates were found as PEMs for the Pie meal, where broccoli was the main ingredient within the *Brassica o.* species. However, an iberin-derived conjugate was also a PEM for Soup and Barleyotto, probably due to the fact that Brussels sprouts are also rich in iberin (Agudo et al. 2004). In addition, the Pie was baked, whereas the Brussels sprouts and white cabbage were served raw. It has previously been demonstrated that the bioavailability of isothiocyanates is lower, and the peak in excretion later, for heat treated products compared to raw products (Vermeulen et al. 2006).

3-methylthiopropyl is common in glucosinolates from cabbage (Clarke 2010) which explains why we find a 3-methylthiopropyl conjugate as a PEM for Soup and Barleyotto. Benzyl conjugates have also been reported as being present in cabbage (Clarke 2010). We observed that benzyl conjugates were highest after the Pie meals suggesting that pointed cabbage, even after heat treatment, contains higher levels of benzyl isothiocyanate producing glucosinolates compared to white cabbage. The last two PEMs found, an *N*-acetyl-cysteine derivative and 4-iminopentylisothiocyanate, have to our knowledge not been reported in the literature before. Even though 4-iminopentylisothiocyanate was not found in the processed data in SFS, the compound was present in very small quantities when inspecting the raw data rather than the processed data.

As all the PEM isothiocyanate conjugates found in our study were present in all meals, it is not likely that it is possible to distinguish between intake of individual members of the *Brassica o.* cultivars from measuring the isothiocyanate conjugate composition in urine. Even though the excretion of isothiocyanate conjugates has been demonstrated to reflect the glucosinulate content of the cruciferous vegetable eaten, a large variation in bioavailability of isothiocyanates has also been shown depending on preparation method and degree of chewing (Vermeulen et al. 2006).

Trimethylamine *N*-oxide (TMAO), a marker for fish intake in this study, is present in fish and can also be produced in humans by the intestinal microflora from carnitine, a dietary precursor found in meat (Zhang et al. 1999; Xu et al. 2010). Several other studies have found TMAO as a marker in urine for fish consumption (Svensson et al. 1994; Zhang et al. 1999; Lloyd et al. 2011). In addition, TMAO has been found as a marker of meat containing diets compared to vegetarian diets (Xu et al. 2010; Stella et al. 2006) and as a marker which differ in level across populations (Holmes et al. 2008; Zuppi et al. 1998). Increases in TMAO excretion has also been demonstrated following an intervention with soy products in

one study (Solanky et al. 2005). It is likely that the findings in the population studies are mainly due to differences in fish intake even though soy intake is also high in Japan (Nagata et al. 2002) and may contribute to the high level found in a Japanese population (Holmes et al., 2008). The contrasts in TMAO found for diets high in meat in comparison to vegetarian diets are probably due to meat consumption. However, in one of the intervention studies, no difference in TMAO was observed between a low meat diet and a vegetarian diet (Stella et al. 2006). This is in accordance with our results, where no clear difference in TMAO was seen between meals prepared with a vegetarian or a meat source (Fig. 2b). Therefore TMAO probably mainly reflects fish intake, even though the contribution to TMAO levels from soy and from very high meat intakes needs to be investigated further. It has been demonstrated in a previous study that the majority of TMAO is excreted within 24 h after intake (Zhang et al. 1999). To our knowledge this is the first study to report a time-sequence for excretion (Fig. 2b).

No exposure markers for parsley roots, fava beans and chicory salad have been published previously. Except from the chicory salad PEM, which was also present in the aqueous extract and the hypothesis on methylated luteolin as one of the parsley root markers, we do not know how the PEMs found for fava bean and the rest of the parsley markers are related to the foods. Further experiments would be needed to investigate this.

## 4 Conclusion

The present study has demonstrated that it is possible to find food exposure markers in 24 h urine samples with a sensitivity and specificity >80 % for a specific meal or protein source with an untargeted semi-quantitative metabolomics approach. About half of the potential exposure markers could also be found in SFS, indicating that some of the PEMs originate from certain foods or food groups. The formation of other PEMs that were not present in SFS is probably dependent on the food matrix. The PEMs found in the current meal study needs to be validated in a larger population in order to investigate dose–response effects as well as specificity and sensitivity in a setting where there is no dietary standardization.

**Acknowledgments** The meal study was conducted as part of the OPUS project and is supported by a Grant from the Nordea Foundation. We would like to thank the study participants and the staff who have been involved in conducting the study, preparing the meals and analyzing the samples: Hanne Lysdal Petersen, Ümmühan Celik, Daniela Rago, Jan Stanstrup, the kitchen staff at the department and Meyers Madhus.

## References

Agudo, A., Bailey, G. S., Bradlow, H. L., et al. (2004). *Cruciferous vegetables, isothiocyanates and indoles*. Lyon: International Agency for Research, on Cancer. IARC Press.

Barri, T., Holmer-Jensen, J., Hermansen, K., & Dragsted, L. O. (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.

Bijlsma, S., Bobeldijk, I., Verheij, E. R., et al. (2006). Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry, 78*, 567–574.

Bingham, S. A. (2002). Biomarkers in nutritional epidemiology. *Public Health Nutrition, 5*, 821–827.

Chong, I., & Jun, C. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems, 78*, 103–112.

Clarke, D. B. (2010). Glucosinolates, structures and analysis in food. *Analytical Methods, 2*, 310–325.

Favé, G., Beckmann, M. E., Draper, J. H., & Mathers, J. C. (2009). Measurement of dietary exposure: A challenging problem which may be overcome thanks to metabolomics? *Genes and Nutrition, 4*, 135–141.

Favé, G., Beckmann, M., Lloyd, A., et al. (2011). Development and validation of a standardized protocol to monitor human dietary exposure by metabolite fingerprinting of urine samples. *Metabolomics, 7*, 469–484.

Holmes, E., Loo, R. L., Stamler, J., et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature, 453*, 396–400.

Hwang, E. S., & Jeffery, E. H. (2003). Evaluation of urinary *N*-acetyl cysteinyl allyl isothiocyanate as a biomarker for intake and bioactivity of Brussels sprouts. *Food and Chemical Toxicology, 41*, 1817–1825.

Janobi, A. A. A., Mithen, R. F., Gasper, A. V., et al. (2006). Quantitative measurement of sulforaphane, iberin and their mercapturic acid pathway metabolites in human plasma and urine using liquid chromatography–tandem electrospray ionisation mass spectrometry. *Journal of Chromatography B, 844*, 223–234.

Jenab, M., Slimani, N., Bictash, M., et al. (2009). Biomarkers in nutritional epidemiology: Applications, needs and new horizons. *Human Genetics, 125*, 507–525.

Koulman, A., & Volmer, D. A. (2008). Perspectives for metabolomics in human nutrition: An overview. *Nutrition Bulletin, 33*, 324–330.

Liu, Z., & Franklin, M. R. (1984). Separation of four glucuronides in a single sample by high-pressure liquid chromatography and its use in the determination of UDP glucuronosyltransferase activity toward four aglycones. *Analytical Biochemistry, 142*, 340–346.

Llorach, R., Garrido, I., Monagas, M., et al. (2010). Metabolomics study of human urinary metabolome modifications after intake of almond (*Prunus dulcis* (Mill.) DA Webb) skin polyphenols. *Journal of Proteome Research, 9*, 5859–5867.

Lloyd, A. J., Fave, G., Beckmann, M., et al. (2011). Use of mass spectrometry fingerprinting to identify urinary metabolites after consumption of specific foods. *American Journal of Clinical Nutrition, 94*, 981–991.

Lodge, J. K. (2010). Symposium 2: Modern approaches to nutritional research challenges: Targeted and non-targeted approaches for metabolite profiling in nutritional research. *Proceedings of the Nutrition Society, 69*, 95–102.

Mithril, C., Dragsted, L. O., Meyer, C., Blauert, E., Holt, M. K., & Astrup, A. (2012a). Guidelines for the new Nordic diet. *Public Health Nutrition, 15*, 1941–1947.

Mithril, C, Dragsted, L. O., Meyer, C., Tetens, I., Jensen, A. B. & Astrup, A. (2012b). Dietary composition and nutrient content of the new Nordic diet, in press.

Nagata, C., Takatsuka, N., & Shimizu, H. (2002). Soy and fish oil intake and mortality in a Japanese community. *American Journal of Epidemiology, 156*, 824–831.

Nelson, A. C., Huang, W., & Moody, D. E. (2001). Variables in human liver microsome preparation: Impact on the kinetics of l-alpha-acetylmethadol (LAAM) n-demethylation and dextromethorphan O-demethylation. *Drug Metabolism and Disposition, 29*, 319–325.

O'Sullivan, A., Gibney, M. J., & Brennan, L. (2011). Dietary intake patterns are reflected in metabolomic profiles: Potential role in dietary assessment studies. *American Journal of Clinical Nutrition, 93*, 314–321.

Penn, L., Boeing, H., Boushey, C., et al. (2010). Assessment of dietary intake: NuGO symposium report. *Genes and Nutrition, 5*, 205–213.

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

Puiggròs, F., Solà, R., Bladé, C., Salvadó, M., & Arola, L. (2011). Nutritional biomarkers and foodomic methodologies for qualitative and quantitative analysis of bioactive ingredients in dietary intervention studies. *Journal of Chromatography A, 1218*, 7399–7414.

Rouzaud, G., Young, S. A., & Duncan, A. J. (2004). Hydrolysis of glucosinolates to isothiocyanates after ingestion of raw or microwaved cabbage by human volunteers. *Cancer Epidemiology, Biomarkers and Prevention, 13*, 125–131.

Solanky, K. S., Bailey, N. J., Beckwith-Hall, B. M., et al. (2005). Biofluid 1H NMR-based metabonomic techniques in nutrition research: Metabolic effects of dietary isoflavones in humans. *Journal of Nutritional Biochemistry, 16*, 236–244.

Spencer, J. P. E., Abd El Mohsen, M. M., Minihane, A., et al. (2008). Biomarkers of the intake of dietary polyphenols: Strengths, limitations and application in nutrition research. *British Journal of Nutrition, 99*(1), 12–22.

Stella, C., Beckwith-hall, B., Cloarec, O., et al. (2006). Susceptibility of human metabolic phenotypes to dietary modulation. *Journal of Proteome Research, 5*, 2780–2788.

Sumner, L., Amberg, A., Barrett, D., Beale, M. H., et al. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics, 3*, 211–221.

Svensson, B. G., Åkesson, B., Nilsson, A., & Paulsson, K. (1994). Urinary excretion of methylamines in men with varying intake of fish from the Baltic sea. *Journal of Toxicology and Environment Health, 41*, 411–420.

Vermeulen, M., Van Den Berg, R., Freidig, A. P., Van Bladeren, P. J., & Vaes, W. H. J. (2006). Association between consumption of cruciferous vegetables and condiments and excretion in urine of isothiocyanate mercapturic acids. *Journal of Agriculture and Food Chemistry, 54*, 5350–5358.

Walsh, M. C., Brennan, L., Pujos-Guillot, E., et al. (2007). Influence of acute phytochemical intake on human urinary metabolomic profiles. *American Journal of Clinical Nutrition, 86*, 1687–1693.

Wold, S., Johansson, E., & Cocchi, M. (1993). PLS—partial least-squares projections to latent structures. In H. Kubinyi (Ed.), *3D QSAR in drug design, theory, methods and applications* (pp. 523–550). Leiden: ESCOM Leiden.

Xu, J., Yang, S., Cai, S., Dong, J., Li, X., & Chen, Z. (2010). Identification of biochemical changes in lactovegetarian urine using $^{1}$H NMR spectroscopy and pattern recognition. *Analytical and Bioanalytical Chemistry, 396*, 1451–1463.

Zhang, Y. (2004). Cancer-preventive isothiocyanates: Measurement of human exposure and mechanism of action. *Mutation Research, 555*, 173–190.

Zhang, A. Q., Mitchell, S. C., & Smith, R. L. (1999). Dietary precursors of trimethylamine in man: A pilot study. *Food and Chemical Toxicology, 37*, 515–520.

Zuppi, C., Messana, I., Forni, F., Ferrari, F. B., Rossi, C. B., & Giardina, B. (1998). Influence of feeding on metabolite excretion evidenced by urine $^{1}$H NMR spectral profiles: A comparison between subjects living in Rome and subjects living at arctic latitudes (Svaldbard). *Clinica Chimica Acta, 278*, 75–79.

# Discovery of exposure markers in urine for Brassica-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics

Maj-Britt Schmidt Andersen[a], Helene Christine Reinbach[b], Åsmund Rinnan[b], Thaer Barri[a], Charlotte Mithril[a], Lars Ove Dragsted[a]

[a]Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen; [c]Department of Food Science, Faculty of Science, University of Copenhagen

Corresponding author:

Maj-Britt Schmidt Andersen
Department of Nutrition, Exercise and Sports,
Faculty of Science, University of Copenhagen
Rolighedsvej 30
DK-1958 Frederiksberg C
Email: mbsa@life.ku.dk
Telephone: +4529176809
Fax: 3533 2483

# Supplemental material

**Table S1** Meal ingredients excluding protein sources

| | Barleyotto | Soup | Pie |
|---|---|---|---|
| Wholemeal bread | 96.7 | 100.7 | 77.7 |
| Rapeseed oil | 5.8 | 2.5 | 8.8 |
| Onion | 15 | 15 | 15 |
| Apple vinegar | 2.5 | 2.5 | 2.5 |
| Apple | 40 | 40 | - |
| Apple juice | 50 | 62.5 | - |
| **White cabbage** | 25 | 25 | - |
| **Brussels sprouts** | 20 | 20 | - |
| Garlic | - | 1.5 | 1 |
| Pearl Barley | 60 | - | - |
| Cream cheese | 12.5 | - | - |
| Bouillon | 150 | - | - |
| **Kale** | - | 125 | - |
| **Carrots** | - | 62.5 | - |
| **Parsley roots** | - | 62.5 | - |
| Chicken stock | - | 125 | - |
| Wheat/rye/oat flour (2:1:1) | - | - | 50 |
| Baking powder | - | - | 1 |
| Yogurt naturel | - | - | 25 |
| Eggs | - | - | 45 |
| Skimmed milk | - | - | 75 |
| Broccoli | - | - | 63 |
| Pointed cabbage | - | - | 38 |
| **Chicory salad** | - | - | 20 |

Ingredients (in grams per single serving) used in the test meals excluding the protein sources. Foods highlighted in bold are food ingredients investigated in single food studies.

**Table S2** Ingredients used as protein sources in the meals

| | | Barleyotto | Soup | Pie |
|---|---|---|---|---|
| **Fish** | Monkfish cheeks | 25 | - | - |
| | Fillet of cod | - | 58 | - |
| | Smoked mackerel | - | - | 26 |
| **Meat** | Pig jaws | 25 | - | - |
| | Smoked saddle of pork | - | 54 | - |
| | Smoked haunch of venison | | | 25 |
| **Vegetarian** | Fava beans | 20 | - | - |
| | Brown beech mushroom | 120 | - | - |
| | Poached egg | - | 75 | - |
| | Split peas | - | 15 | - |
| | Cottage cheese | - | - | 40 |
| | Chanterelles | - | - | 20 |

Ingredients (in grams per serving) used as protein sources for the meat, fish and vegetarian versions of the test meals. Foods highlighted in bold are food ingredients investigated in single food studies.

**Cooking methods for meals:**

Barleyotto

The onion was chopped and sautéed in rapeseed oil before adding pearl barley which was left to sizzle for one minute. Then apple juice was added and one minute after also the bouillon. After simmering for 20 minutes, cream cheese was added.

*Vegetarian version:* Fava beans were added to the pearl barley for the last 6-8 minutes of cooking. The brown beech mushrooms were fried for 2-3 minutes and mixed with the pearl barley before serving.

*Meat version:* Pig jaws were fried in butter before adding apple juice and water. They were then left to simmer for 45 minutes and added to the pearl barley before serving.

*Fish version:* Monkfish cheeks were fried in butter for 2 minutes and added to the pearl barley before serving.

All Barleyotto meals were seasoned with salt and pepper and served with bread and a salad of Brussels sprouts, white cabbage and apple with a mixture of rapeseed oil, apple vinegar, salt and pepper as dressing.

Pie

Flour and baking powder were mixed into a dough with water and yogurt. The dough was rolled, put in a pan and baked for 15 minutes at 160°C. Onion and garlic were chopped and fried in rapeseed oil. Then, broccoli was cut and added to the pan with the onion. After beating eggs and milk, the mixture was poured over the vegetables and the pie was baked for 25 minutes at 150°C. Chicory salad was cut, mixed with the protein source and added on top of the pie before serving.

*Vegetable version*: Cottage cheese was mixed with sliced chanterelles

*Meat version:* Smoked haunch of venison (no preparation)

*Fish version*: Smoked mackerel (no preparation)

All Pie meals were served with bread and seasoned with salt and pepper.

Soup

Onion and garlic were chopped and sautéed in rapeseed oil before adding peeled and chopped carrots and parsley roots. Apple juice and chicken stock were added and the mixture was boiled until the vegetables were tender

*Vegetarian version:* Split peas were boiled until tender. The egg was added to boiling water and poached for 5-8 minutes.

*Meat version*: Smoked haunch of venison was served with chopped parsley.

*Fish version*: Fillet of cod was baked in the oven with rapeseed oil.

All Soup meals were seasoned with salt and pepper and served with bread and a salad of Brussels sprouts, white cabbage and apple with a mixture of rapeseed oil, apple vinegar, salt and pepper as dressing.

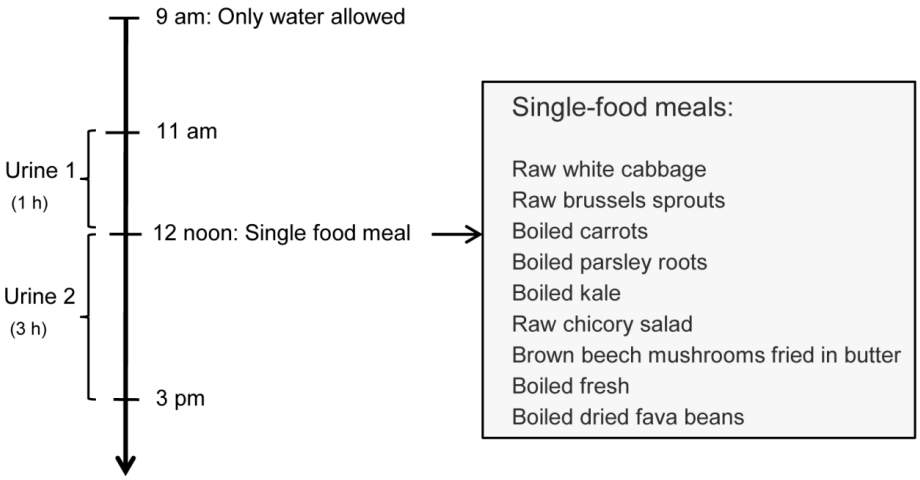



**Fig. S1** Outline of study design for Single-food studies (SFS). A schedule of meals and urine sampling for each study day is given on the timeline to the left. The right part of the figure shows the nine SFS meals.

**Table S3** Parameters and batch steps used for the processing of samples from the meal study and the single-food studies (SFS).

| Batch step | | Parameters |
|---|---|---|
| **Negative mode** | Raw data import | |
| | Chromatogram builder | Noise level: 5.0E (15); Min time span: 0.01 (0.01); Min height: 5.0E1 (4.0E1); m/z tolerance: 0.06 (0.055 or 30 ppm) |
| | Chromatogram deconvolution | Chromatographic threshold: 98% (95%); Search minimum in RT range: 0:01 (0.01); Minimum relative height: 1% (10 %); Minimum absolute height: 5.0E1 (4.0E1); Min ratio of peak/top edge: 1.2 (1.3); (Peak duration range (min): 0.01-0.2) |
| | Isotopic pattern | m/z tolerance: 0.06 (0.06 or 30 ppm); Retention time tolerance: 0:01 (0.01); Monotonic shape; maximum charge: 1 (1) |
| | Join aligner | m/z tolerance: 0.06 (0.06 or 30 ppm); Absolute retention time tolerance: 0:12 (0.15); Weight for both m/z tolerance and retention time tolerance: 10 (10) |
| | Duplicate peak filter | m/z tolerance: 0.06 (0.5 or 600 ppm); RT tolerance: 0:12 (0.15) |
| | Peak finder | Intensity tolerance: 20% (50%); m/z tolerance: 0.06 (0.06 or 30 ppm); Absolute retention time tolerance: 0:12 (0.15) |
| | Export to csv | |
| **Positive mode** | Raw data import | |
| | Chromatogram builder | Noise level: 4.0E1 (15); Min time span: 0.01 (0.01); Min height: 4.0E1 (4.0E1); m/z tolerance: 0.026 (0.055 mz or 30 ppm) |
| | Chromatogram deconvolution | Chromatographic threshold: 98% (97%); Search minimum in RT range: 0:01 (0.01); Minimum relative height: 1% (10%); Minimum absolute height: 4.0E1 (6.0E1); Min ratio of peak/top edge: 1.2 (1.5); (Peak duration range (min): 0.01-0.2) |
| | Isotopic pattern | m/z tolerance: 0.026 (0.06 or 30 ppm); Retention time tolerance: 0:01 (0.01); Monotonic shape; maximum charge: 1 (1) |
| | Join aligner | m/z tolerance: 0.026 (0.06 or 30 ppm); Absolute retention time tolerance: 0:17 (0.15); Weight for both m/z tolerance and retention time tolerance: 10 (10) |
| | Duplicate peak filter | m/z tolerance: 0.026 (0.5 or 600 ppm); RT tolerance: 0:17 (0.15) |
| | Peak finder | Intensity tolerance: 50% (50%); m/z tolerance: 0.026 (0.06 or 30 ppm); Absolute retention time tolerance: 0:17 (0.17) |
| | Export to csv | |

Samples from the meal study were processed in MZmine2.2 and samples from the single-food studies were processed in MZmine 2.7. The numbers used for SFS are given in brackets.

**Table S4** List of unknown PEMs found in the meal study

| Meal(s) | m/z | RT (min) | Adduct/fragment/parent ion | Peak time | Present in other meals |
|---|---|---|---|---|---|
| Barleyotto and Soup | 261.093 | 2.01 | Pos mode | 3 | No |
| Pie | 205.096 | 1.94 | $[M+H]^+$ | 3 | No |
| | 255.048 | 0.93 | Pos mode | 3 | No |
| Soup | 403.229 | 4.05 | $[M+Na]^+$ | 3 | No |
| | 345.188 | 3.83 | $[M+Na]^+$ | 3 | No |
| | 191.143 | 3.58 | Pos mode | 4 | Yes |
| Barleyotto vegetarian meal | 325.167 | 4.14 | Pos mode | 3 | No |
| | 301.157 | 4.20 | Neg mode | | |
| | 287.150 | 4.09 | Neg mode | 3 | No |
| | 311.153 | 4.00 | Pos mode | 3 | No |
| | 297.136 | 3.97 | Pos mode | 3 | Yes |
| | 306.103 | 2.64 | Pos mode | 3-4 | No |
| | 158.088 | 0.77 | Pos mode | 3 | Yes |
| Barleyotto fish meal | 179.011 | 0.49 | Pos mode | 3 | No |

The meals with highest levels of the feature are listed in the first column. Peak time is the urine sampling point, where the highest level of the feature was detected. None of the listed PEMs were confirmed in single food studies.

# Paper II

## Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics

Andersen, M.S., Kristensen, M., Manach, C., Pujos-Guillot, E., Poulsen, S.K., Larsen, T.M., Astrup, A., Dragsted, L.O.

Analytical and Bioanalytical Chemistry (2014); DOI 10.1007/s00216-013-7498-5 (E-pub ahead of print)

Anal Bioanal Chem
DOI 10.1007/s00216-013-7498-5

PAPER IN FOREFRONT

# Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics

**Maj-Britt Schmidt Andersen · Mette Kristensen · Claudine Manach · Estelle Pujos-Guillot · Sanne Kellebjerg Poulsen · Thomas Meinert Larsen · Arne Astrup · Lars Dragsted**

Received: 1 October 2013 / Revised: 4 November 2013 / Accepted: 7 November 2013


**Abstract** While metabolomics is increasingly used to investigate the food metabolome and identify new markers of food exposure, limited attention has been given to the validation of such markers. The main objectives of the present study were to (1) discover potential food exposure markers (PEMs) for a range of plant foods in a study setting with a mixed dietary background and (2) validate PEMs found in a previous meal study. Three-day weighed dietary records and 24-h urine samples were collected three times during a 6-month parallel intervention study from 107 subjects randomized to two distinct dietary patterns. An untargeted UPLC-qTOF-MS metabolomics analysis was performed on the urine samples, and all features detected underwent strict data analyses, including an iterative paired *t* test and sensitivity and specificity analyses for foods. A total of 22 unique PEMs were identified that covered 7 out of 40 investigated food groups (strawberry, cabbages, beetroot, walnut, citrus, green beans and chocolate). The PEMs reflected foods with a distinct composition rather than foods eaten more frequently or in larger amounts. We found that 23 % of the PEMs found in a previous meal study were also valid in the present intervention study. The study demonstrates that it is possible to discover and validate PEMs for several foods and food classes in an intervention study with a mixed dietary background, despite the large variability in such a dataset. Final validation of PEMs for intake of foods should be performed by quantitative analysis.

**Keywords** Untargeted metabolomics · UPLC-MS · Food exposure markers · Urine · Dietary intervention study · Humans

## Introduction

Measurement of dietary exposure is a challenging topic in nutrition research and of crucial importance for the discovery of true associations between dietary intake and effects on health [1]. By far, the most commonly applied tools for estimating dietary exposure are based on self-reporting (food frequency questionnaires, weighed dietary records (WDR), etc.). A major drawback of this methodology, however, is that presence of systematic and random errors can be considerable. Objective dietary exposure markers, measured in biological samples such as urine or plasma, are promising supplements to self-reported food intakes, but very few dietary exposure markers are known and commonly applied in nutrition studies [2,3].

One approach to discover new food exposure markers, which is being increasingly studied, is the possibility of mapping the metabolic fate of foods in plasma and urine. By knowing how foods are metabolized, it may be possible to find unique metabolites for foods or related food groups and use these as food exposure markers [3–5]. In line with this hypothesis, several metabolomics studies have been published demonstrating clear metabolic responses to dietary interventions [6,7]. The idea of developing new dietary exposure markers by untargeted metabolomics is therefore promising even though there is still a long way from discovery of a potential food exposure marker (PEM) to its application in a dietary study.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00216-013-7498-5) contains supplementary material, which is available to authorized users.

M.-B. S. Andersen (✉) · M. Kristensen · S. K. Poulsen · T. M. Larsen · A. Astrup · L. Dragsted
Department of Nutrition Exercise and Sports, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark
e-mail: mbsa@life.ku.dk

C. Manach · E. Pujos-Guillot
INRA, UMR1019, Human Nutrition Unit, University of Auvergne, 63000 Clermont-Ferrand-Theix, France

Published online: 04 January 2014

An important step that needs further investigation in metabolomics studies of different foods and diets is the validation of discriminating metabolites found in a study [4,8]. It is likely that the most discriminating metabolites for a food or food group depend on several factors such as timing of intake, amount, food matrix and characteristics of the study population [9]. In order to understand the strength of a discriminant metabolite as a food exposure marker, it is therefore crucial to investigate the same marker in different study settings that cover different population groups and food exposures (long-term, short-term, different intake levels, dietary backgrounds, etc.).

In a recent paper, we reported on the findings from a controlled meal study with *Brassica*-containing Nordic meals and foods. We introduced sensitivity and specificity analyses as a means to select the most promising PEMs among the discriminant metabolites in 24-h urine samples [10]. As part of the 5-year multidisciplinary research project, OPUS, a New Nordic Diet (NND) has been developed, which differentiates considerably in food composition and amounts of intake from an Average Danish Diet (ADD) [11]. The NND is particularly rich in cabbages, root vegetables, berries and legumes, since these foods can be grown in Nordic countries, while the ADD is characterized by a lower intake and a different selection of fruit and vegetables, including tropical fruits. The present study is a 6-month parallel dietary intervention study with NND and ADD as the two study arms. The diverse food composition of the diets gives a good opportunity to investigate the metabolic response to individual foods and food groups. In comparison to the previously published meal study, the intervention study is less controlled and it represents a wider selection of foods with varied amounts of intake and different preparation methods. By combining LC-MS data from 24 h urine samples and data from WDR in this study, we have applied a metabolomics strategy to (1) discover PEMs for individual plant foods and plant food groups and (2) validate the PEMs found in the previously published meal study [10]. An analysis of the metabolite patterns associated with the NND and ADD based on untargeted metabolomics will be reported elsewhere.

## Materials and methods

### Study design

A parallel, randomized, controlled 6-month intervention study was performed comparing the dietary patterns, NND and ADD. The dietary patterns were defined by intakes of 15 food groups and by the energy distribution of macronutrients, as reported elsewhere [12]. All foods in the study were collected free of charge and ad libitum by the study participants from a small supermarket that was set up at the University of Copenhagen. For each visit to the shop, the subjects could freely choose their own foods as long as the collected foods were in accordance with the defined dietary pattern. Subjects recruited for the study were adults aged 18–65 years with increased waist circumference (>94 cm for men and >80 cm for women) and preferably one or more additional risk factors of the metabolic syndrome (increased plasma triglyceride level, reduced high-density-lipoprotein cholesterol, hypertension and/or impaired fasting glucose). The study has been approved by the Regional Ethics Committee of Greater Copenhagen and Frederiksberg (H-3-2010-058) and the Danish Data Protection Agency (2007-54-0269).

One hundred and eighty-one subjects (71 % women) were randomized to the two diets. At three sampling points (week 0, 12 and 26), 3-day WDR were collected, and on one of the days with dietary recording, one 24-h urine sample was collected as well. The urine collection was from 8 am, excluding the first void, until 8 am the following morning, including the first void. For the WDR, the participants were requested to report intake of all foods and beverages, except water, from early morning until late evening.

During the first WDR, the diet was standardized and was the same for all 3 days. Foods contained in the standardized diet are listed in Table S1 in the Electronic supplementary material. An outline of the study design is given in the Electronic supplementary material, Fig. S1. Urine was stored in cooler bags during collection and otherwise at −80 °C. For the purpose of the current study, data have only been included from study completers who provided a urine collection (1× 24 h) and a food dietary record (1×24 h) on the same day as the urine collection at all sampling points (week 0, 12 and 26). There were 107 participants (74 % women) who fulfilled these criteria, of which 64 followed the NND and 43 followed the ADD. Baseline characteristics of the study completers are given in Table S2 in the Electronic supplementary material.

### Sample preparation and mass spectrometry analyses

#### *UPLC-qTOF-MS analysis*

Urine samples were thawed on ice and centrifuged at 3,000×*g* for 2 min at 4 °C; 150 μL of the resulting supernatants was immediately distributed randomly into a 96-well auto-injector tray, keeping samples from the same subject together within a plate to minimize intra-individual variation. The urine was diluted 1:1 with aqueous 5 % 30:70 (*v*/*v*) acetonitrile (ACN)–methanol (MeOH) (Optima grade LC–MS, Fisher Scientific, USA), and a 5-μL aliquot of each sample was injected into a UPLC (Waters, Manchester, UK) equipped with an HSS T3 $C_{18}$ column (1.8 μm, 2.1×100 mm) and a HSS $C_{18}$ pre-column (1,8 μm, 2.1×5 mm) (Waters, Milford, MA) held at 50 °C. The column was eluted using a 7.0 min gradient gradually changing mobile phase composition and flow (0.5–1.2 mL/min) as described previously [12]. The mobile

phase was 0.1 % formic acid (A) and 0.1 % formic acid in 70 % ACN and 30 % MeOH (B). All samples were kept at 4 °C during the analysis. The eluate was analysed by a qTOF Premier mass spectrometer (Waters, Manchester, UK) equipped with an electrospray ion source operating in either positive or negative ion mode and with a mass resolution of around 8,000. Centroid data were collected from 50 to 1000 $m/z$ with a scan time of 0.08 s and an interscan delay of 0.02 s. Leucine encephalin (500 ng/mL) was infused intermittently every 10 s and used as lock mass to calibrate mass accuracy. A metabolomics standard sample containing 44 different physiologically relevant compounds [13] and a pooled sample of all urines within a plate were analysed two and five times, respectively, in both ionization modes for each plate and served as quality control of the analytical platform. For identification purposes, MS/MS fragmentation of selected compounds was performed in product ion scan mode with collision energies of 10, 20 and 30 eV.

*LTQ Orbitrap Velos™ MS analysis*

To aid identification, a subset of samples was profiled using an LTQ Orbitrap Velos™ MS (Thermo Scientific) operating in full scan mode and with a mass resolution of 30,000. Chromatographic separation was achieved with a Dionex RSLC Ultimate 3000 liquid chromatography module (Dionex). Samples (10 μL) were injected onto a BEH Shield RP18 column (1.7 μm×2.1 mm×100 mm). Mobile phase A was Milli-Q purified water containing 0.1 % formic acid and mobile phase B was acetonitrile containing 0.1 % formic acid. Mobile phases were pumped through the system at a flow rate of 400 μL/min. A 22-min gradient was used, starting at 0 % B, increasing to 10 % between 2 and 7 min and then to 95 % at 22 min, prior to a 4-min post-run at 100 % A to re-equilibrate the column.

Preprocessing and pretreatment of data

Preprocessing of raw data from UPLC-qTOF-MS was done in MZmine2.7 [14] to obtain a list of peak areas, retention times (RT) and mass to charge ratios ($m/z$) in negative and positive ionization modes. Batch steps and parameters applied for preprocessing can be found in Table S3 in the Electronic supplementary material.

Throughout this paper, the term feature will be used to designate an ion in the peak list, while the term potential exposure marker (PEM) will be used for features remaining after performing the statistical analysis that are associated with specific foods.

All features from positive and negative ionization modes were imported into MATLAB® version 7.12.0.635, R2011a (Mathworks Inc., Sherborn, MA, USA), which was used for the subsequent data pre-treatment and statistical analysis.

First, samples were normalized to the same mean total intensity to correct for batch drift and differences in urine concentration. Next, plate differences were corrected by normalising all samples that had been run on the same plate to the same mean value for each feature. The two normalisation steps were done separately for negative and positive mode data, and afterwards, the two datasets were combined into one data matrix (data matrix A).

Data analysis for PEMs

For each subject, a list of foods eaten on the days where urine had been collected was extracted from WDR. If the same food had been eaten twice on the same day, the amounts were added up. The food list extracted from WDR contained 412 different entries. From these, 40 plant based food groups were selected which had been reported to be ingested by volunteers at least ten times. A food group could consist of a single food prepared in different ways (e.g. beetroot and pickled beetroot), different plant foods from the same family (e.g. cranberries, blueberries and lingonberries from the Ericaceae family), foods that were nutritionally related (e.g. walnut and hazelnut) or single foods (e.g. avocado). For each food group, a list of related foods was also made that were taken into account in the data analysis to increase the contrast between exposed and non-exposed groups. These could be nutritionally or botanically related foods (e.g. celeriac for parsley), or food products containing small amounts of a food in the food group (e.g. muesli for nuts). Detailed information on the food groups and related foods can be found in Table 1.

The data analysis to find PEMs for the selected food groups was in four steps. We took advantage of the samples collected from the same person on days where the person had and had not consumed a food. This created exposure and control samples for each food group, which were explored for PEMs. An overview of the data analysis is given in Fig. 1.

For each food group, a data matrix B was generated from data matrix A. Samples from subjects who had eaten at least one food in the food group were included in data matrix B (food group) together with samples from the same subjects on days where they had not eaten foods from the food group or from the related foods (control samples for food group).

*80 % rule*

The number of features in data matrix B was reduced by applying the 80 % rule [15] to non-normalized data. To find the optimum threshold defining the noise level for peak areas, an iterative procedure was performed with peak area thresholds from one to hundred. The number of remaining features was plotted against the noise level thresholds and the threshold for which the curve reached a flattening point was chosen as optimum.

**Table 1** Food groups analysed

| Food group | Amount (g) | Related foods |
|---|---|---|
| Berries | | |
| Blackcurrant (37) | 44±49 | Blackcurrant jam (15), Blackcurrant squash (1), Gooseberry (12), Redcurrant (5) |
| Grossulariaceae (Blackcurrant (37), Gooseberry (12), Redcurrant (5)) | 90±112 | Blackcurrant jam (15), Blackcurrant squash (1) |
| Blueberry (17) | 32±48 | Cranberry (16), Lingonberry (13) |
| Cranberry (16) | 19±18 | Blueberry (17), Lingonberry (13) |
| Lingonberry (13) | 28±19 | Blueberry (17), Cranberry (16), |
| Ericaceae (Cranberry (16), Blueberry (17), Lingonberry (13)) | 32±37 | – |
| Gooseberry (12) | 190±108 | Blackcurrant (37), Lingonberry (13), Redcurrant (5) |
| Sea buckthorn (16) | 41±49 | – |
| Strawberry (118[a]) | 115±42 | Raspberry jam (9), Strawberry icecream (1), Strawberry jam (109[a]), Strawberry yogurt (5) |
| Fruits | | |
| Apple (Apple (110), Apple juice (69)) | 243±194 | Apple cake (2) |
| Pear (140 [a]) | 121±52 | Yogurt with pear (1) |
| Pome fruit (Apple (110), Apple juice (69), Pear (140[a])) | 206±180 | Apple cake (2), Yogurt with pear (1) |
| Banana (14) | 143±78 | Banana yogurt (1) |
| Rutaceae (Grapefruit (3), Lime (2), Lemon (6), Mandarin (8), Orange (5), Orange juice (112[a])) | 251±94 | Lemon juice (1), Orange jam (1), Yogurt with orange juice (4) |
| Orange (Orange (5), Orange juice (112[a])) | 263±70 | Grapefruit (3), Lemon (6), Lemon juice (1), Lime (2), Mandarin (8), Orange jam (1), Yogurt with orange juice (4) |
| Rhubarb (14) | 72±37 | – |
| Vegetables and herbs | | |
| Asparagus (Green asparagus (12), White asparagus (1)) | 88±43 | – |
| Avocado (13) | 91±68 | – |
| Beetroot (Beetroot (24), Pickled beetroot (20)) | 102±107 | Spinach (16) |
| Spinach (16) | 79±53 | Beetroot (24), Pickled beetroot (20) |
| Pointed cabbage (18) | 103±101 | Broccoli (4), Brussels sprouts (9), Cauliflower (6), Red cabbage (12), Savoy cabbage (5), White cabbage (7) |
| Red cabbage (12) | 69±39 | Broccoli (4), Brussels sprouts (9), Cauliflower (6), Pointed cabbage (18), Savoy cabbage (5), White cabbage (7) |
| *Brassica oleracea* (Broccoli (4), Brussels sprouts (9), Cauliflower (6), Pointed cabbage (18), Red cabbage (12), Savoy cabbage (5), White cabbage (7)) | 107±90 | – |
| Celeriac (Celeriac (23), Celeriac leaves (5)) | 75±45 | Dill (5), Parsley (17), Parsley root (13), Parsnip (8) |
| Chives (10) | 8±3 | – |
| Cucumber (Cucumber (119[a]), Pickled cucumber (15)) | 43±19 | – |
| Carrot (188[a]) | 63±65 | Celeriac (23), Celeriac leaves (5), Dill (5), Fennel (2), Parsley (17), Parsley root (13), Parsnip (8) |
| Parsley (Parsley (17), Parsley root (13)) | 44±52 | Carrot (188), Celeriac (23), Celeriac leaves (5), Dill (5), Fennel (2), Parsnip (8) |
| Apiaceae (Carrot (188[a]), Celeriac (23), Celeriac leaves (5), Dill (5), Fennel (2), Parsley (17), Parsley root (13), Parsnip (8)) | 70±76 | – |
| Green beans (14) | 146±92 | Green peas (18), Split peas (11) |
| Green peas (18) | 101±67 | Green beans (14), Split peas (11) |
| Split peas (11) | 41±21 | Green beans (14), Green peas (18) |
| Fabaceae | | |
| (Green beans (14), Green peas (18), Split peas (11)) | 108±81 | – |

**Table 1** (continued)

| Food group | Amount (g) | Related foods |
|---|---|---|
| Tomato (Canned tomatoes (9), Tomato (33)) | 109±103 | Tomato soup (2), Tomato salad (2), Ketchup (11) |
| Oils | | |
| Olive oil (111[a]) | 18±10 | Olives (2) |
| Rapeseed oil (58) | 12±10 | Cress (8), Radish (8), Swede (4), Turnip (2) |
| Nuts | | |
| Hazelnuts (67) | 29±32 | Almonds (6), Muesli (9), Nut based spreads (6), Peanuts (1), Pine nuts (5), Walnuts (32) |
| Walnuts (32) | 28±19 | Almonds (6), Hazelnuts (67), Muesli (9), Nut based spreads (6), Peanuts (1), Pine nuts (5) |
| Nuts (Almonds (6), Hazelnuts (67), Walnuts (32)) | 32±32 | Muesli (9), Nut based spreads (6), Peanuts (1), Pine nuts (5) |
| Chocolate | | |
| Chocolate (cocoa intake[b]≥5 g) (Chocolate (dark (116[a]), milk (11), filled (29) and with marzipan (5)), Cocoa containing nut spread (6), Chocolate covered marshmallows (2), Chocolate milk (14), Cocoa powder (3)) | 15±7 | Chocolate mousse (2), Chocolate cake (2) Cocoa intake[b]<5 g |

Related foods listed in the right column are foods that were botanically or nutritionally related to the food group or foods that contain low amounts of foods in the food group. The number of times the foods were reported in WDR is given in brackets. Amounts are given as average intake (± the standard deviation)

[a] Foods that were part of the standardized diet

[b] Due to the diverse composition of chocolate containing products, the cocoa content of the products was calculated. Samples where subjects had reported intakes of chocolate corresponding to ≥ 5 g cocoa were included in the food group

*Paired t test*

The purpose of the univariate analysis was to find exposure marker candidates among the features in data matrix B for each food group. A paired *t* test was used to take into account inter-individual variation. Subjects who had reported intake of a food in the food group or the group of related foods at all sampling points were excluded from the analysis due to lack of control

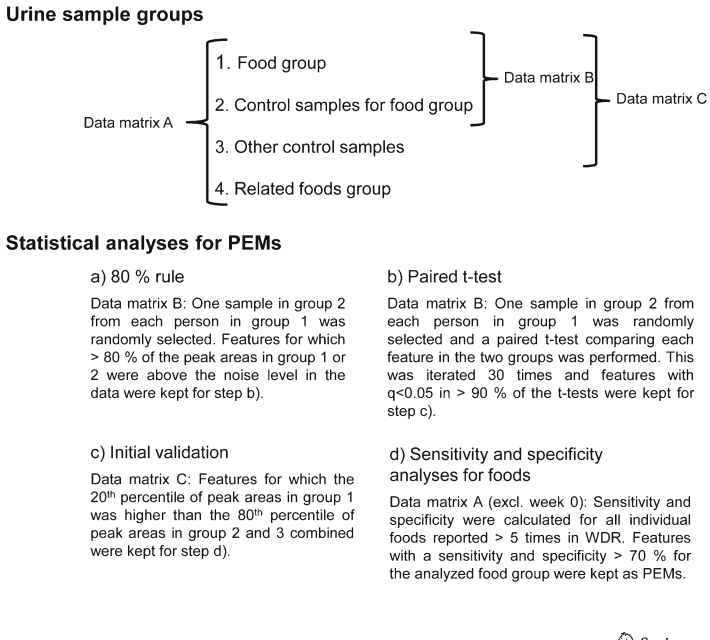



**Fig. 1** Schematic representation of the data analysis for PEMs

samples for these subjects. The paired *t* test was iterated 30 times, and in each test, the control group was changed by randomly selecting a control sample from the one or two available samples from each subject in the food group. To take into account multiple testing, positive-false discovery rates were applied [16].

*Initial validation*

In the initial validation step, the sensitivity and specificity of the features in a larger dataset (data matrix C) were evaluated, as described previously [10].

*Sensitivity and specificity analyses for foods*

The sensitivity and specificity analyses for foods were included to ensure that intake of no other foods correlated to the food group analysed, as this would bias the analysis. Data from week zero were excluded in the sensitivity and specificity analysis because of the dietary standardization at this sampling point. Peak area thresholds to define consumption were determined from the distribution of peak areas for the investigated feature. Peak areas were included in the analysis in an iterative fashion, starting from the peak area of the 5th lowest sample and increasing by two samples (7th lowest sample, 9th lowest sample, etc.) until 102 samples were included. This number corresponds to the maximum number of samples for which it is possible to reach a sensitivity above 70 % for the largest food group size (excluding data from week 0).

For each threshold, the sensitivity and specificity were calculated as follows for each food and food group:

$$\text{I)}\quad \text{Sensitivity} = \frac{N_{\text{food}>Th}}{N_{\text{food}>Th} + N_{\text{food}<Th}}$$

$$\text{II)}\quad \text{Specificity} = \frac{N_{\text{samples}} - N_{\text{other}>Th} - N_{\text{food}>Th} - N_{\text{food}<Th}}{N_{\text{samples}} - N_{\text{food}<Th} - N_{\text{food}>Th}}$$

| | |
|---|---|
| $N_{\text{food>Th}}$ and $N_{\text{food<Th}}$ | Number of samples above and below the threshold, respectively, where the food had been eaten according to the WDR. |
| $N_{\text{other>Th}}$ | Number of samples above the threshold, where the food had not been eaten according to the WDR. |
| $N_{\text{samples}}$ | Total number of samples, excluding week 0 (214). |

Four levels of strengths for a feature as a food exposure marker was defined according to sensitivity and specificity thresholds: sensitivity>80 % and specificity>80 % (grade 1), sensitivity >70 % and specificity >80 % (grade 2), sensitivity >80 % and specificity >70 % (grade 3), sensitivity>70 % and specificity>70 % (grade 4)

If a feature had strength of at least 4, for the food group as a whole or for at least two individual foods in the food group, and for any threshold, it was considered a PEM. A feature was not considered a PEM for a food group if the marker strength was higher for a food outside the food group, unless the food with a higher strength was a related food. The dose–response relationship for the food group was investigated for each PEM by visual inspection of a plot of peak areas as a function of the amounts of food reported in WDR.

Validation of PEMs from a previously published meal study

Several PEMs were identified in a previous crossover-meal study including 17 subjects and nine lunch test meals (a 'barleyotto', a soup and a pie, each prepared in three versions, one with each of the three protein sources: meat, fish and vegetarian). An untargeted metabolomics analysis revealed 30 unique PEMs that differed between the test meals. To identify the food sources of the markers, a range of small meal studies with single foods and three to four volunteers were additionally carried out, including meals with white cabbage, Brussels sprouts, carrots, parsley root, kale, chicory salad, brown beech mushroom and fava beans. Fifteen of the PEMs were found to be present after intake of one or more single foods [10].

To validate the PEMs from the previous cross-over meal study as markers of individual foods, each of the 30 PEMs were searched for in the data from the current intervention study. The analytical conditions were the same in the two studies, and it was therefore possible to simply match the $m/z$ values (absolute tolerance of 0.05) and the retention times (absolute tolerance of 0.2 min). To avoid picking noisy features as matches in the intervention study, processed data were investigated for any match to a feature in the meal study. At least four samples with peaks areas above 15 were required to regard the feature as a correct match. For each matching feature in the intervention study, the sensitivity and specificity analyses for foods were carried out. Foods with marker strengths of 1–4 were used as the basis for performing the initial validation and investigating if there was a dose–response relationship to food intake.

Identification of PEMs

The parent ion was identified by inspecting full-scan raw MS data for dimers, common adducts and fragments. Then, the PEMs were compared to an in-house standards database by matching retention times and $m/z$ ratios. A subset of samples was run on an LTQ Orbitrap Velos™ ('LTQ Orbitrap Velos™ MS analysis') to obtain more accurate $m/z$ of the PEMs, which were used to determine the most probable molecular formula. An MS/MS fragmentation analysis was performed ('UPLC-qTOF-MS analysis') to get more structural information on the markers and identify glucuronide, sulphate, glycine and acetyl-cysteine conjugates. All PEMs were correlated

using Pearson's correlation coefficient and highly correlated PEMs ($r > 0.7$) were grouped as they may belong to the same metabolic pathway.

The parent $m/z$ and $m/z$ of the most intense fragments were searched for in the Human Metabolome Database [17], the METLIN database [18] and in MetFusion [19]. Possible precursor ions of the PEMs were found by searching the CRC Dictionary of Food compounds, Phenol-explorer [20] and KNApSAcK [21] for compounds present in the foods. In addition to the food databases, literature was searched for known metabolites of the foods.

A strawberry extract in water and ethanol was prepared from frozen strawberries according to the procedure described previously [10]. The strawberry extracts, filtered orange juice and a beetroot extract (B-50-WS, Chr. Hansen, Denmark) were run on UPLC-qTOF-MS to investigate if the PEMs or any of their characteristic fragment ions were present in the foods.

The level of identification has been categorized according to Sumner et al. [22]. For level I, $m/z$ pattern and RT have been compared to an authentic standard under identical conditions. For level II, the fragmentation pattern of the PEM corresponds to MS/MS fragments reported in databases or literature. Level III is used for identification of compounds based on similarities to known compound classes and level IV designates unknown compounds and compounds that could not fit into the other categories.

The following standards were analysed: 5-acetylamino-6-amino-3-methyluracil, hydrate (Stratech, UK), 3,4-dihydroxyhydrocinnamic acid (Sigma-Aldrich, Germany), 3-(2,4-dihydroxyphenyl)propionic acid (Sigma-Aldrich, Germany), dopamine hydrochloride (Sigma-Aldrich, Germany), hesperetin (BioNordika, Denmark), 4-hydroxyhippuric acid (Bachem, Germany), 5-hydroxyindole-3-acetic acid (Santa Cruz, USA.), 4-hydroxy-phenyllactic acid (Sigma-Aldrich, Germany), 4-methylpyridine-2-carboxylic acid (Sigma-Aldrich, Germany), 7-methyluric acid (Santa Cruz, USA), stachydrine hydrochloride (Santa Cruz, USA), sulphoraphane *N*-acetyl-cysteine (AHdiagnostics, Denmark), theobromine (Sigma-Aldrich, Germany) and *N*-methyl-cis-4-hydroxy-*L*-proline (Sigma-Aldrich, Germany).

### *Glucuronidation, sulphation and glycination of standards*

Sulphation was performed as described previously [23]. In brief, 90 μL 500 μM TRIS (PH 7.5), 5 μL substrate (beetroot extract, 1,000 ppm dopamine in aqueous 5 % 30:70 (*v*/*v*) ACN–MeOH), 850 μL milli-Q water and 45 μL 2 mM adenosine 3′-phosphate 5′-phosphosulphate lithium salt hydrate was mixed and after 5 min at 37 °C, 10 μL rat liver extract prepared according to the optimized condition described in Nelson et al. [24] was added. The mixture was left for 1 h at 37 °C before adding 1 mL MeOH. Then, the solution was centrifuged and the supernatant evaporated to dryness before redissolving the sample in 200 μL aqueous 5 % 30:70 (*v*/*v*) ACN–MeOH. For glucuronidation of hesperetin, 100 μL 100 mM $MgCl_2$, 260 μL 8 mM uridinediphosphoglucuronic acid, 580 μL milli-Q-water and 5 μL 100 mM hesperetin were mixed before adding 55 μL liver extract. The rest of the procedure was similar to the procedure for sulphation. Glycination of 4-methylpyridine-2-carboxylic acid was performed by adding NaOH (2.1 mg, 0.054 mmol) and acetic anhydride (4.9 mg, 0.048 mmol) to a solution of 4-methylpyridine-2-carboxylic acid (23.95 mg, 0.17 mmol) in water (100 μL). After stirring the mixture at 35 °C for 1 h, 4.6 mg (0.061 mmol) glycine was added and the mixture was stirred at 35 °C for further 18 h. The crude mixture was diluted 500 times before UPLC-MS analysis.

## Results

### Discovery and identification of PEMs for individual foods

After preprocessing, 6,044 and 7,283 features were detected in urine in negative and positive mode, respectively. Details on the number of remaining features after steps a–d of the statistical analysis for all foods and food groups are given in Table 2.

The number of subjects in the food groups in the paired *t* test varied from 10 to 136, and depending on the food group, between 0 and 373 features were significant in at least 90 % of the paired *t* tests. Sixty-two features remained after the initial validation step. However, many of the remaining features were not unique for one food group. The food groups consumed as part of the standardized diet had several features in common, and only after calculating sensitivity and specificity, it was possible distinguish to which food groups these features were related. After removing features that did not have a sensitivity and specificity of at least 70 %, 44 PEMs remained which covered seven different foods. The PEMs had large variation in peak areas for the same reported dose and only a weak tendency for a dose–response relationship was found for some of the PEMs. Four typical examples of dose–response relationships are given in Fig. 2.

The 44 PEMs corresponded to 22 unique metabolites of which 13 were identified at levels I–III. Another four were tentatively identified by their fragmentation patterns, but because no commercial standards were available for those and it was not feasible to synthesize them, they have been marked as level IV identifications. Details on the PEMs are given in Table 3. For all compounds where the identification could not be confirmed with a standard, more detailed information on fragmentation pattern is given in the Electronic supplementary material, Tables S4–S8.

### *PEMs for berries, fruits, nuts and chocolate*

One PEM was found for strawberry for which the parent ion could not be identified. The fragment, $m/z$ 79.957 was found as the most intense, suggesting that the compound is a sulphate metabolite. From the molecular formula and other less intense

**Table 2** Number of features remaining for each food group after steps a–d of the statistical analysis

| Food group | *n* | (a) 80 % rule | (b) Paired *t* test | (c) Initial validation | (d) PEMs |
|---|---|---|---|---|---|
| Berries | | | | | |
| Blackcurrant | 37 | 794 | 114 | 0 | 0 |
| Grossulariaceae | 50 | 868 | 159 | 0 | 0 |
| Blueberry | 17 | 796 | 0 | 0 | 0 |
| Cranberry | 16 | 798 | 0 | 0 | 0 |
| Lingonberry | 13 | 916 | 2 | 0 | 0 |
| Ericeae | 38 | 790 | 53 | 0 | 0 |
| Gooseberry | 12 | 929 | 0 | 0 | 0 |
| Sea buckthorn | 16 | 902 | 0 | 0 | 0 |
| Strawberry[a] | 105 | 816 | 257 | 8 | 1 |
| Fruits | | | | | |
| Apple | 136 | 856 | 373 | 0 | 0 |
| Pear[a] | 119 | 831 | 239 | 0 | 0 |
| Pome fruit[a] | 91 | 806 | 91 | 0 | 0 |
| Banana | 14 | 884 | 0 | 0 | 0 |
| Rutaceae[a] | 121 | 802 | 250 | 6 | 6 |
| Orange[a] | 116 | 854 | 282 | 9 | 7 |
| Rhubarb | 12 | 869 | 0 | 0 | 0 |
| Vegetables and herbs | | | | | |
| Asparagus | 13 | 886 | 0 | 0 | 0 |
| Avocado | 13 | 913 | 0 | 0 | 0 |
| Beetroot | 38 | 865 | 47 | 3 | 3 |
| Spinach | 16 | 832 | 0 | 0 | 0 |
| Pointed cabbage | 18 | 911 | 108 | 2 | 2 |
| Red cabbage | 11 | 989 | 180 | 22 | 19 |
| *Brassica oleracea* | 48 | 843 | 176 | 0 | 0 |
| Celeriac | 27 | 819 | 43 | 0 | 0 |
| Chives | 10 | 784 | 0 | 0 | 0 |
| Cucumber[a] | 119 | 834 | 251 | 1 | 0 |
| Carrot[a] | 103 | 811 | 157 | 0 | 0 |
| Parsley | 11 | 808 | 0 | 0 | 0 |
| Apiaceae[a] | 107 | 778 | 140 | 0 | 0 |
| Green beans | 14 | 838 | 14 | 1 | 1 |
| Green peas | 18 | 846 | 0 | 0 | 0 |
| Split peas | 11 | 906 | 0 | 0 | 0 |
| Fabaceae | 40 | 776 | 40 | 0 | 0 |
| Tomato | 37 | 801 | 41 | 0 | 0 |
| Oils | | | | | |
| Olive oil[a] | 94 | 823 | 217 | 4 | 0 |
| Rapeseed oil | 57 | 823 | 206 | 0 | 0 |
| Nuts | | | | | |
| Hazelnuts | 67 | 914 | 227 | 1 | 0 |
| Walnuts | 32 | 817 | 63 | 2 | 2 |
| Nuts | 107 | 894 | 288 | 0 | 0 |
| Chocolate | | | | | |
| Chocolate[a] | 98 | 846 | 221 | 4 | 3 |

*n* number of samples in the food group after removing samples where the same subject had reported intake of the food or related foods at all sampling points

[a] Foods that were part of the standardized diet

fragments found, we tentatively identified the strawberry PEM as a sulphate ester of 2,5-dimethyl-4-methoxy-3(2H)-furanone (Table S4, Electronic supplementary material), a known aroma compound in strawberry [25]. A small peak with *m/z* 141.056,

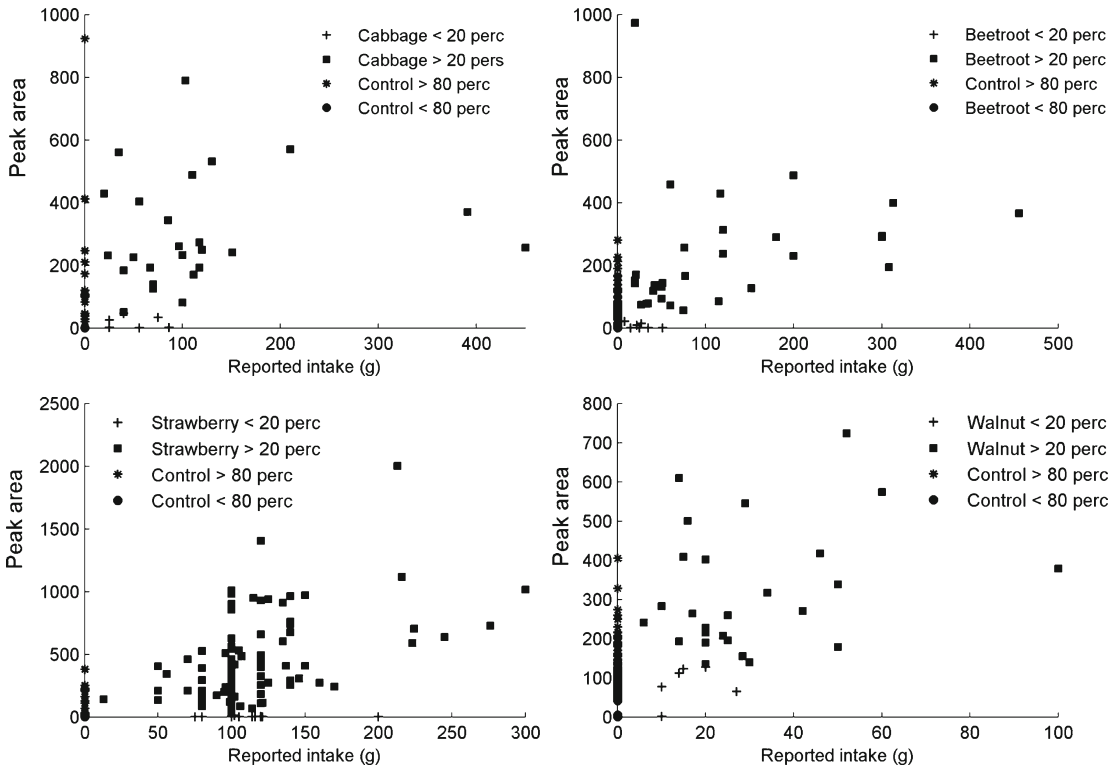



**Fig. 2** Plots of peak areas as a function of reported food intakes for four PEMs. Samples where subjects have eaten related foods have been excluded from the data. Control and food samples below and above the 80th and 20th percentiles, respectively, are highlighted in *black circles* and *squares*, since these percentiles were used as cut-off for the initial marker validation. (**a**) PEM for cabbage (*m*/*z* 327.054, RT 2.72), (**b**) PEM for beetroot (*m*/*z* 246.038, RT 1.11), (**c**) PEM for strawberry (*m*/*z* 221.024, RT 3.43), (**d**) PEM for walnuts (*m*/*z* 146.061, RT 2.80)

which corresponds to loss of sulphate, was detected in the strawberry water extract at RT 1.27. However, it could not be confirmed whether this peak corresponds to 2,5-Dimethyl-4-methoxy-3(2H)-furanone.

Four compounds were PEMs for the orange and citrus food groups. Three of the four PEMs (*m*/*z* 146.083, 144.101 and 142.049) were strongly correlated ($r \approx 0.9$), of which one (*m*/*z* 144.101) was identified as proline betaine. It was not possible to make a tentative identification from MS/MS data of the two other PEMs. For the PEM with *m*/*z* 146.083, *N*-methyl-cis-4-hydroxy-*L*-proline was ruled out by analysing the standard. The last common PEM for citrus and orange was identified as hesperetin glucuronide, while the PEM for orange could not be identified from the fragmentation pattern. Of all citrus and orange markers, only proline betaine was found in orange juice.

Two PEMs, originating from the same compound, was found for walnuts and the parent compound identified at level I as 5-hydroxyindole-3-acetic acid.

Theobromine and 7-methyluric acid from chocolate were identified by standards. The last PEM for chocolate was identified at level II as 6-amino-5-[*N*-methylformylamino]-1-methyluracil (6-AMMU) based on the common fragment *m*/*z* 171.087 reported in Llorach et al. [6]. A standard of the isomer 5-acetylamino-6-amino-3-methyluracil (AAMU) was run to rule out that possibility.

*PEMs for vegetables*

Three unique compounds were found as PEMs for beetroot. Neither of the compounds was present in a beetroot water extract and the molecular formulas did not match any known beetroot compounds or metabolites from the literature. One of the beetroot PEMs was identified at level I as a glycine conjugate of 4-methylpyridine-2-carboxylic acid. The precursor of this compound is most likely betanin, the colouring agent in beetroot, as it contains a tetrahydropyridine moiety. The beetroot PEM with *m*/*z* 234.045 was a sulphate metabolite. All fragments of this compound that did not contain sulphate were also present in the dopamine standard. However, RTs of dopamine and dopamine sulphate were lower than RT of the compound. Betanin is also a likely precursor of dopamine-like compounds due to presence of

**Table 3** Identifications of PEMs for food groups

| Food group | RT qTOF | *m/z* qTOF | *m/z* Orbitrap | Ion | Mass error (ppm) | Molecular formula | *m/z* of main MS/MS fragments | Identification | Strength, dose–response |
|---|---|---|---|---|---|---|---|---|---|
| Berries | | | | | | | | | |
| Strawberry | 3.43 | 221.024 | 221.0121 | Neg mode | 0.4 | C7H10O6S | 79.957 | 2,5-Dimethyl-4-methoxy-3(2H)-furanone sulphate (IV)[c] | 1, y |
| Fruits | | | | | | | | | |
| Orange/citrus | 0.52 | 287.189<br>144.101 | 287.1983<br>144.1024 | $[2M+H]^+$<br>$[M+H]^+$ | 4.2<br>0.2 | C7H13NO2 | 84.080[a] | Proline betaine (I) | 1, y |
| Orange/citrus | 3.57 | 477.108 | 477.1081 | $[M-H]^-$ | 3.3 | C22H22O12 | 301.072[a], 175.023, 113.024[a] | Hesperetin glucuronide (I) | 1, n |
| Orange/citrus | 0.73 | 146.083 | 146.0809 | $[M+H]^+$ | 5.8 | C6H11NO3 | 128.070, 84.080 | Unknown (IV) | 2, y |
| Orange/citrus | 1.08 | 142.049 | 142.0505 | $[M-H]^-$ | 0.7 | C6H9NO3 | None | Unknown (IV) | 2, y |
| Orange | 2.78 | 211.060<br>195.071 | 211.0607<br>195.0646 | $[M-H]^-$<br>$[M-H2O]^+$ | 0.3<br>5.6 | C10H12O5 | Neg: 151.038, 136.016[a]<br>Pos: 153.054, 125.059, 93.033[a] | Unknown (IV) | 1, y |
| Vegetables | | | | | | | | | |
| Beetroot | 0.95 | 234.045 | 234.0428 | $[M+H]^+$ | 3.6 | C8H11NO5S | 217.017, 137.059, 119.049[a] | 4-Ethyl-5-amino-pyrocatechol sulphate (III)[c] | 2, y |
| Beetroot | 1.11 | 246.038 | 246.0444 | $[M-H]^-$ | 3.0 | C9H13NO5S | 166.087, 151.063[a], 79.958 | 4-Ethyl-5-methylamino-pyrocatechol sulphate (IV)[c] | 1, y |
| Beetroot | 3.32 | 195.079 | 195.0757 | $[M+H]^+$ | 3.5 | C9H10N2O3 | 149.071, 120.044, 92.049[a] | 4-Methylpyridine-2-carboxylic acidglycine conjugate (I) | 1, n |
| Green beans | 3.47 | 189.083 | NI | Neg mode | – | C8H14O5 | 129.055, 127.075, 99.081 | Unsaturated aliphatic hydroxy-dicarboxylic acid [c] (IV)[c] | 1, y |
| Red cabbage | 0.53 | 151.009 | 151.0065 | Neg mode | 0.0 | C4H8O4S | 132.996 | 3-Hydroxy-3-(methyl-sulphinyl)propanoic acid (IV)[c] | 1, n |
| Red cabbage | 1.77 | 276.018 | 276.0174 | $[M+H]^+$ | 1.6 | C9H9NO7S | 200.986, 196.060, 121.028 | 3-Hydroxy-hippuric acid sulphate (II) | 1, n |
| Red cabbage | 2.38 | 389.098<br>196.061<br>150.054 | 389.0994<br>196.0600<br>150.0556 | $[2M-H]^-$<br>$[M+H]^+$<br>$[M-HCO2]^-$ | 2.3<br>5.2<br>0.6 | C9H9NO4 | Neg: 150.057, 137.022[a], 93.034<br>Pos: 140.046, 121.027, 93.033 | 3-Hydroxy-hippuric acid (II) | 1-2, n |
| Red cabbage | 2.57 | 363.112 | 363.1075 | $[2M-H]^-$ | 1.3 | C9H10O4 | 121.029, 119.050 | Unknown (IV) | 1, n |
| Red cabbage (brussels sprouts, pointed cabbage) | 2.72 | 327.054<br>164.047 | NI | $[M+H]^+$<br>$[M-C5H8NOS2]^+$ | | C10H19S3O4N2 | 164.020, 134.010, 105.037[a] | Iberin *N*-acetyl-cysteine (IB-NAC) (III) | 1, y |
| Red cabbage (brussels sprouts, pointed cabbage) | 3.64 | 311.055<br>309.076<br>162.022 | NI | $[M+H]^+$<br>$[M-H]^-$<br>$[M-C5H10NS2]^-$ | – | C10H18N2O3S3 | Neg: 162.022, 84.045[a]<br>Pos: 251.130, 164.037, 85.027 | *N*-acetyl-S-(*N*-3-methylthiopropyl)cysteine (III) | 1, y |
| Red cabbage (brussels sprouts, horseradish) | 3.47 | 285.032<br>263.075<br>261.037<br>164.046<br>162.022<br>122.033 | 285.0338<br>263.0519<br>261.0377<br>164.0371<br>162.0222<br>122.0268 | $[M+Na]^+$<br>$[M+H]^+$<br>$[M-H]^-$<br>$[M-C4H4NS]^+$<br>$[M-C4H6NS]^-$<br>$[M-C6H6NOS]^+$ | 2.0<br>2.12<br>3.7<br>6.4<br>1.59<br>6.1 | C9H14N2O3S2 | Neg: 162.021, 84.044<br>Pos: 164.037, 130.050, 122.027 | *N*-acetyl-S-(*N*-allylthiocarbamoyl)cysteine (AITC-NAC) (III) | 1-2, y |
| Red cabbage (Brussels sprouts) | 3.27 | 341.082<br>339.073 | NI | $[M+H]^+$<br>$[M-H]^-$ | – | C11H20N2O4S3 | NI | Sulphoraphane *N*-acetyl-cysteine (SFN-NAC) (I) | 1, n |

**Table 3** (continued)

| Food group | RT qTOF | *m/z* qTOF | *m/z* Orbitrap | Ion | Mass error (ppm) | Molecular formula | *m/z* of main MS/MS fragments | Identification | Strength, dose–response |
|---|---|---|---|---|---|---|---|---|---|
| Nuts | | | | | | | | | |
| Walnut | 2.80 | 146.061<br>146.059 | 146.0611<br>146.0600 | $[M-HCOO]^-$<br>$[M-HCOOH]^+$ | 3.5<br>4.2 | C10H9NO3 | Neg: 146.059, 144.045<br>Pos: 146.060 | 5-Hydroxyindole-3-acetic acid (I) | 2, y |
| Chocolate | | | | | | | | | |
| Chocolate | 0.90 | 199.085 | 199.0822 | $[M+H]^+$ | 4.4 | C7H10N4O3 | 171.087, 156.064[a], 139.041[a] | 6-Amino-5-[N-methylformylamino]-1-methyluracil (6-AMMU) (II) | 1[b], n |
| Chocolate | 2.04 | 181.072 | 181.0711 | $[M+H]^+$ | 7.9 | C7H8N4O2 | 163.061, 138.066, 110.070[a] | Theobromine (I) | 2[b], n |
| Chocolate | 1.01 | 363.080 | 363.0814 | $[2M-H]^-$ | 3.4 | C6H6N4O3 | 166.013, 138.028, 110.035[a] | 7-Methyluric acid (I) | 1[b], y |

Food group: Food group for which the marker was a PEM. A few markers were shared between the citrus and orange group. These have been categorized in citrus if they had high SEN and SPE for other citrus fruits than oranges. Mass error is calculated based on *m/z* from orbitrap data. The main MS/MS fragments are given for the parent ion in the ionization mode where the marker was found and with 10 eV as collision energy. Strength: SEN and SPE level as defined in the sensitivity and specificity analysis for foods. Dose–response: n (no)/y (yes) indicates whether the PEM showed a dose–response relationship to the food group. NI: not investigated

[a] Fragments mainly obtained with 20 or 30 eV

[b] The strength given is for one or more chocolate products in the group

[c] Additional information on the fragmentation pattern to support the tentative identification can be found in the supplemental material

a 3,4-dihydroxyphenylalanine (L-DOPA) structure. From the L-DOPA part of betanin, it was possible to obtain plausible structures for *m/z* 234.045 (4-ethyl-5-amino-pyrocatechol sulphate, Table S5, Electronic Supplementary Material) and *m/z* 246.038 (4-ethyl-5-methylamino-pyrocatechol sulphate, Table S6, Electronic supplementary material) which correspond to the MS/MS fragments and RTs of the compounds.

Intake of red cabbage gave rise to most PEMs. The PEMs corresponded to eight unique compounds of which four were identified as N-acetyl-cysteine conjugates of isothiocyanates. Sulphoraphane *N*-acetyl-cysteine was identified by a standard and the rest were identified by comparing RTs, and characteristic fragments and losses to cabbage markers reported previously [10]. The cabbage marker, 3-hydroxyhippuric acid, was identified by comparing MS/MS data to the standard 4-hydroxyhippuric acid and the fragments *m/z* 150 and 100 reported for 3- and 4-hydroxyhippuric acid, respectively in Gonthier et al. [26]. Sulphated 3-hydroxyhippuric acid was identified by the common fragments to 3-hydroxy-hippuric acid with *m/z* 121.029 and 93.033, the common loss of glycine (*m/z* 75.032) and loss of sulphate (*m/z* 79.958). The PEM with *m/z* 363.112 was correlated to 3-hydroxyhippuric acid ($r \sim 0.8$) and shared two fragments with this marker (*m/z* 121.029, 93.033). The two shared fragments indicate presence of a phenol (*m/z* 93.034) and an additional CO. Three standards were run as likely candidates for this marker, 3-(2,4-dihydroxyphenyl)propionic acid, 3,4-dihydroxyhydrocinnamic acid and 4-hydroxyphenyllactic acid but none of them were correct. The last PEM of red cabbage was tentatively identified as 3-hydroxy-3-(methyl-sulphinyl)-propanoic acid from the fragmentation pattern, Table S7, Electronic supplementary material. This compound has, to our knowledge, not been reported before. When searching the data from our previous single food studies [10], the compound increased following intake of Brussels sprouts and kale. However, no clear trend was found because the baseline levels of the compound in urine were generally high. The compound was not present in a water extract of Brussels sprouts.

The PEM for green beans could not be identified. Based on the fragmentation pattern found with losses of COO, COOH and OH (Table S8, Electronic supplementary material) and the most probable molecular formula, C8H14O5, the compound is an unsaturated aliphatic hydroxyl–dicarboxylic acid. Such a compound would be too polar for the RT found and for this reason, the PEM is probably not the parent ion. One of the carboxylic acid groups may originate from an ester group which has been broken upon fragmentation but further analyses would be needed to identify the exact structure and the parent ion.

## Validation of PEMs from the previous meal study

Seven out of 30 PEMs from the previously published meal study were also markers in the present intervention study (Table 4). Five of the PEMs were markers of various cruciferous vegetables

**Table 4** PEMs found in a previous meal study that were also markers in the intervention study

| PEM | Foods (meal study) | Foods (intervention study) |
|---|---|---|
| *N*-acetyl-S-(*N*-3-methylthiopropyl)cysteine | White cabbage, Brussels sprouts | Red cabbage, Brussels sprouts, pointed cabbage (1), y |
| *N*-acetyl-S-(*N*-allylthiocarbamoyl)cysteine (AITC-NAC) | White cabbage, Brussels sprouts | Red cabbage, Brussels sprouts, horseradish (1–2), y |
| Iberin *N*-acetyl-cysteine (IB-NAC) | White cabbage, Brussels sprouts | Red cabbage, Brussels sprouts, pointed cabbage (1), y |
| Sulphoraphane *N*-acetyl-cysteine (SFN-NAC) | White cabbage, Brussels sprouts | Red cabbage, Brussels sprouts (1), n |
| 4-Iminopentylisothiocyanate | None | Brussels sprouts (2), |
| Trimethylamine *N*-oxide (TMAO) | Fish (cod, smoked mackerel, monkfish) | Cod, Pollock, Halibut (1–2) y[a] |
| Uknown RT 2.01 (2.07[b]), *m/z* 261.093 (261.103[b]), pos mode | None | Brussels sprouts, (2), n |

Foods (meal study): Foods from the meal study for which the PEM was a marker [10]. Foods (intervention study): Foods for which the marker strength of the PEM was 1–4 in the intervention study. The marker strength as defined in the sensitivity and specificity analysis for foods is given in brackets and y/n indicates whether there was a dose–response relationship between amount of intake and marker intensity

[a] Cod roe and smoked cod were excluded in the initial validation analysis as similar foods

[b] Data from intervention study

from the species *Brassica olecearea* in the meal study. Accordingly, a high sensitivity (> 70 %) and specificity (> 80 %) for red cabbage, Brussels sprouts, pointed cabbage and horseradish were found for these markers in the intervention study. Three other markers of *B. olecearea* from the meal study (an unknown N-acetyl-cysteine conjugate, *N*-Acetyl-(*N*′-benzylthiocarbamoyl)-cysteine and sulphoraphane-*N*-cysteine) were found in the intervention study but were not strong enough to be considered markers. Another *B. olecearea* marker from the meal study, Erucin *N*-acetyl-cysteine, did not match any feature in the intervention study. Two markers in the meal study were markers of Brussels sprouts in the intervention study even though they were not found in our previous single food studies with Brussels sprouts [10], Finally, TMAO was a marker of fish intake in the intervention study with marker strengths of 1–2 for cod, pollock and halibut. When all fish and shellfish were included in the analysis for TMAO, the marker strength was 4. The lower marker strength when including all fish is probably mainly reflecting a varied TMAO content for different types of fish and shellfish. When investigating each type of fish separately, high TMAO response was found following intake of cod, pollock and Greenland halibut, and low TMAO response was found following intake of cod roe and tuna, while the remaining types of fish and shellfish were in between. In accordance with this, fish belonging to the *Gadidae* family, i.e. cod and pollock are known to be particularly high in TMAO, while tuna contains low amounts [27].

The rest of the PEMs from the meal study were either not found (10 unknown markers, 2 parsley root markers, 1 chicory salad marker and 1 fresh fava bean marker) in the data from the intervention study or were not strong enough to be markers in the intervention study (2 unknown markers and 3 parsley root markers).

## Discussion

### Discovery and identification of PEMs for foods

The fact that it was only possible to find PEMs for 7 out of 40 food groups is probably mainly because of the strict criteria for sensitivity and specificity applied in the statistical analysis for this study setting. Only the dietary pattern, not the consumption of individual foods, was controlled in the study and the amounts and selections of foods reported in WDR are therefore diverse. Even though WDR perform best in comparison to other dietary assessment methods [28], some recall bias and underreporting are still expected. When considering all the sources of variability in the data together, i.e. errors caused by use of WDR, large variation in the amount of food reported, the low number of consumers for many food groups, inter-individual variability and the semi-quantitative nature of the analytical method, it is not surprising that only very few PEMs were found in the study and that the majority of these PEMs have weak dose–response relationships (Table 3, Fig. 2). If lower thresholds for sensitivity and specificity in the initial validation were applied instead, it may have been possible to find more markers. An analysis of five representative food groups in which sensitivity and specificity thresholds of 60–80 % were applied demonstrated, as expected, that a higher number of features remained for the lower thresholds in the initial validation (Fig. S2, Electronic supplementary material). However, these features did not have the highest sensitivity and specificity for the investigated foods in the last validation step and further studies would therefore be required to investigate the food origin of such features.

Another reason why PEMs are only found for a few foods might be that not all foods give rise to characteristic markers or

that the markers cannot be captured by the analytical method applied. Citrus, chocolate, pear, carrot, cucumber, strawberry and olive oil were all part of the standardized diet and thus consumed by all subjects at one time point or more. This should increase the possibility of finding markers for these food groups but only strawberry, chocolate and, in particular, citrus gave rise to PEMs, suggesting that the composition of the foods is an important determinant for the possibility of finding strong markers. The reason why consumption of citrus resulted in the highest number of PEMs is probably its content of a very diverse group of bioactive compounds, in particular flavonoids, of which several are specific to citrus [29]. Similarly, vegetables belonging to the *Brassica olecearea* species, especially red cabbage, also gave rise to a high number of markers compared to other vegetable species probably due to their content of glucosinolates, which is almost unique to this species [30]. Another factor supporting the notion that the food composition is important for the chance of finding markers is the fact that PEMs were found for some foods consumed in very low amounts. For example, intake of walnuts was reported less frequently and in lower amounts than intake of apples and tomatoes, and despite this, a PEM was found for walnuts and not for the other foods. In a study by Lloyd et al. [31], untargeted metabolomics was also applied to investigate food exposure markers in urine of habitual dietary intake. In accordance with the results from the present study, discriminant features were found for citrus fruit, *Brassica* species and chocolate, and in that study, they also succeeded in finding markers of bananas and tomatoes, for which no PEMs were found in the present study. Unfortunately, only a few of the markers found by Lloyd et al. were reported, among those the markers of tomato, and these were either not specific or related to co-consumed foods which may explain why the same markers are not found in the present study. More studies need to be conducted to understand how many foods give rise to unique markers in urine within habitual consumption levels.

Another factor in the analysis that may influence the marker findings is the selection of related foods. When botanically related foods are excluded from the data, the markers found may characterize a family of plants rather than the individual food which may give a stronger contrast between the food and control group. On the other hand, frequently reported related foods will lower the size of the control group and thereby the power of the statistical analysis. It is not possible in the present study to evaluate how the choice of foods in the food group and the related foods group has influenced marker discovery and more markers may have been found if all analyses had been performed with and without related foods.

Use of food WDR from the same day as the urine collections in the present study limits the exposure markers found to acute markers that are mainly excreted within 24 h and foods giving rise to markers that peak in excretion much later than 24 h, such as urolithins [32], will therefore be missed. The reason for choosing only foods reported in WDR on the day of urine collection was that the day of urine collection varied between the volunteers and was not on the same of the three WDR days in all cases. In addition, the majority of markers in urine are excreted fast and acute markers of frequently consumed foods will also be good candidates as markers of habitual intakes. For example, excretion of proline betaine is known to peak within a few hours after intake and be almost completely excreted within 24 h [33]. Despite this, it has been found as a marker of habitual consumption in several studies [33–35]. Since this study is close to a free-living setting, the markers are most likely valid in observational studies as well, at least as discriminants for recent intake of the foods. Our study therefore provides several PEM candidates for final validation in targeted studies. It should be noted that is not possible to assess if the two overall dietary patterns influence the findings since many foods were mainly consumed within one of the dietary patterns. However, this is not expected since many of the PEMs found have been reported previously in other studies, as evident from the discussion below.

### *PEMs from berries, fruits, nuts and chocolate*

All PEMs of citrus were also reported in a previous untargeted metabolomics study in which citrus markers from a short and medium term intervention study and a cohort study were compared [36]. Interestingly, the four citrus markers found in the present study resembled the previous findings for the cohort study best. This confirms that even though the present study was an intervention study, the exposure markers found are likely valid in a free-living setting. The orange marker with $m/z$ 211.060 has not been reported previously. Proline betaine and hesperetin-glucuronide have been found as acute and habitual citrus markers in more studies [33–35], while the unknown markers with $m/z$ 146.083 and 142.049 have only been reported once before. The fact that they are not present in orange juice and that they correlate very strongly to proline betaine suggests that these markers are metabolites of proline betaine even though this disagrees with the notion that proline betaine is an inert metabolite [33].

For walnut consumption, an increase in 5-hydroxyindole-3-acetic acid has been reported previously [7,37]. 5-Hydroxyindole-3-acetic acid is a metabolite of serotonin and comes out as a PEM of walnuts because of their high serotonin content compared to other foods [37].

The finding of theobromine as a marker of cocoa consumption is in accordance with the fact that cocoa or chocolate containing products are the main dietary sources of theobromine [38]. The other PEMs of chocolate, 6-AMMU and 7-methyluric acid, are endogenous metabolites of theobromine [39]. In an untargeted metabolomics study by Llorach et al. [6], more theobromine metabolites were found as markers of cocoa, and methylxanthines were better discriminants in 24-h

urine samples than methyluric acids. When humans are given theobromine orally, 7-methylxanthine has also been demonstrated to be excreted in higher quantities than 7-methyluric acid [39,40]. It is therefore surprising that 7-methylxanthine was not a marker in the present study. As the metabolic pathway for theobromine is shared with caffeine, intake of coffee and tea may confound the findings for chocolate. This could explain why some theobromine metabolites come out as PEMs and others not, even though the main metabolites of theobromine and caffeine differ [40]. Another possible explanation is differences in excretion patterns. In the study by Rodopoulos et al. [39], it was demonstrated that the excretion of theobromine and its metabolites extended beyond 24 h and was not complete within 48 h. It may therefore be that 7-methyluric acid is a better marker of acute intake. Interestingly 7-methyluric acid was also the only chocolate PEM with some dose–response relationship (Table 3).

The strawberry marker found has not been reported previously and the tentative identification could not be verified.

*PEMs of vegetables*

The PEMs found for red cabbage were mainly acetyl-cysteine conjugates of isothiocyanates and microbial metabolites of polyphenols. Isothiocyanates are formed from glucosinolates upon cell damage. Even though some cabbage varieties are known to be particularly rich in certain glucosinolates, individual glucusinolates are widely distributed in several cabbage varieties and the amount of intake and the preparation method are important for the resulting isothiocyanate composition in urine [30]. An example underlining this is that of sulphoraphane, which is found in high quantities in broccoli but is a PEM of red cabbage and Brussels sprouts in the present study even though these cabbage varieties contain much lower levels of sulphoraphane [41]. In general, only the most frequently consumed cabbages (Table 2) resulted in isothiocynate conjugates with marker strengths of 1–2 (Table 3). All isothiocyanate conjugates found in the intervention study were also found in the previous meal study (Table 4). However, they were not reported in another untargeted metabolomics study comparing diets high and low in cruciferous vegetables, probably because data from NMR was applied in that study [42]. NMR has a lower metabolite coverage than LC-MS due to a lower sensitivity.

The microbial metabolite 3-hydroxy-hippuric acid (sulphated or not) can be formed from various polyphenols [43] and is not specific to red cabbage, even though no other foods in the study had a high marker strength for this metabolite. In comparison to other cabbage varieties, red cabbage contains high levels of polyphenols [44], in particular anthocyanins, which are also metabolized in the colon to phenolic acids [32]. This probably explains why red cabbage gives rise to high levels of microbial products. It is assumed that the PEM with $m/z$ 363.112 may also be of microbial origin due to the strong correlation to 3-hydroxy-hippuric acid. The tentatively identified compound, 3-hydroxy-3-(methyl-sulphinyl)-propanoic acid, has not been reported before, and no likely precursor of such a metabolite was found in cabbage. Even though the marker did not have high marker strength for other cabbage varieties in the intervention study, the fact that it was found also in the previous meal study suggests that the marker is generally cabbage related and not specific to red cabbage.

Betanin is a likely precursor of the PEMs found for beetroot since it is contained in very few vegetables of which only beetroot was consumed as part of the intervention study. Little is known about the metabolism of betanin. It has been demonstrated that less than one percent of ingested betanin is excreted unmetabolized in urine [45] while four betanin metabolites in the same study were not identified. We report here for the first time a glycine conjugate of 4-methylpyridine-2-carboxylic acid as a beetroot PEM and a likely betanin metabolite. Further experiments would be needed to verify the proposed structures of the other two beetroot PEMs. To our knowledge, urinary markers of green beans have not been reported before. No likely precursor of the compound was found in the food databases.

Validation of PEMs from meal study

Only 23 % of the PEMs from the meal study were also found to be markers in the intervention study. This is probably because the meal study was a very controlled study setting in which the diet was standardized, as dietary standardization is known to have a strong impact on the urinary metabolome [46,47]. Unspecific PEMs in particular, would not be valid in studies with a more varied dietary background. Some PEMs in the meal study may also reflect the particular meal matrix rather than intake of individual foods [10]. A few PEMs from the meal study for fresh fava beans and chicory salad could not be validated in the intervention study since almost no subjects reported intake of these foods. Differences in the urine collection times in the previous studies with single foods and the meal study may explain why some PEMs in the meal study were markers of Brussels sprouts in the intervention study even though they were not found in the single foods study with Brussels sprouts.

There is a general agreement that a targeted approach is necessary to validate biomarkers found by untargeted metabolomics studies [4,48], since quantification is important for determining biomarker strength as well as possible effects of other factors such as sex, age and BMI. While the lack of quantitation in the validation procedure of the present study is a limitation, untargeted metabolomics is a fast way to get an indication of the performance of a marker in another study setting. The comparison performed here clearly demonstrates the importance of validating PEMs in different subjects and

study settings. A high sensitivity and specificity are crucial in exposure assessments for a PEM to be useful. The fact that some PEMs were common to both studies suggests that it is possible, at least for some foods or food groups, to find valid and strong urinary exposure markers, while other markers may be valid under certain study conditions only.

## Conclusions

The present study demonstrates that it is possible to find strong acute urinary exposure markers of some individual foods and food groups in an intervention study with a self-selected diet similar to free-living conditions, even though food intake relied on self-reporting and PEM validation was based on very strict criteria. PEMs were found for foods giving rise to characteristic metabolites rather than for frequently consumed foods or foods consumed in large amounts. Validation of markers found in a previous controlled meal study indicated that only a few PEMs were also valid with the present more varied diets. These results highlight the importance of the study setting in the search for urinary food exposure markers and the large variability inherent in data from less controlled study settings analysed with a semi-quantitative method. A targeted analysis is necessary to further validate the PEMs found. The strategy outlined for finding PEMs found in the present study may also be used for optimizing and targeting food exposure marker discovery in observational studies.

**Acknowledgments** The intervention study was conducted as part of the OPUS project. OPUS is an acronym of the Danish title of the project 'Optimal well-being, development and health for Danish children through a healthy New Nordic Diet' and is supported by a grant from the Nordea Foundation, Denmark. The authors would like thank Majbritt Hybholt for providing the food intake data and Daniela Rago, Ümmühan Celik and Bernard Lyan for their contribution to the laboratory work.

## References

1. Favé G, Beckmann ME, Draper JH, Mathers JC (2009) Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics? Genes Nutr 4:135–141
2. Bingham SA (2002) Biomarkers in nutritional epidemiology. Public Health Nutr 5:821–827
3. 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
4. Llorach R, Garcia-Aloy M, Tulipani S, Vazquez-Fresno R, Andres-Lacueva C (2012) Nutrimetabolomic strategies to develop new biomarkers of intake and health effects. J Agric Food Chem 60:8797–8808
5. Penn L, Boeing H, Boushey CJ, Dragsted LO, Kaput J, Scalbert A, Welch A, Mathers J (2010) Assessment of dietary intake: NuGO symposium report. Genes Nutr 5:205–213
6. Llorach R, Urpi-Sarda M, Jáuregui O, Monagas M, Andres-Lacueva C (2009) An LC-MS- based metabolomics approach for exploring urinary metabolome modifications after cocoa consumption. J Proteome Res 8:5060–5068
7. Tulipani S, Llorach R, Jáuregui O, López-Uriarte P, Garcia-Aloy M, Bullo M, Salas-Salvadó J, Andrés-Lacueva C (2011) Metabolomics unveils urinary changes in subjects with metabolic syndrome following 12-week nut consumption. J Proteome Res 10:5047–5058
8. Lodge JK (2010) Symposium 2: Modern approaches to nutritional research challenges: targeted and non-targeted approaches for metabolite profiling in nutritional research. Proc Nutr Soc 69:95–102
9. Spencer JPE, Abd El Mohsen MM, Minihane A, Mathers JC (2008) Biomarkers of the intake of dietary polyphenols: strengths, limitations and application in nutrition research. Br J Nutr 99:12–22
10. Andersen MS, Reinbach HC, Rinnan Å, Barri T, Mithril C, Dragsted LO (2013) Discovery of exposure markers in urine for Brassica-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics. Metabolomics 9:984–997
11. Mithril C, Dragsted LO, Meyer C, Tetens I, Biltoft-Jensen A, Astrup A (2013) Dietary composition and nutrient content of the New Nordic Diet. Public Health Nutr 16:777–785
12. Poulsen SP, Due A, Jordy AB, Stark KD, Stender S, Holst C, Astrup A, Larsen TM (2013) Health effect of the New Nordic Diet in adults with increased waist circumference: a 6-mo randomized controlled trial. Am J Clin Nutr. doi:10.3945/ajcn.113.069393
13. 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
14. Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinforma 11: 395–404
15. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, van Ommen B, Smilde AK (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem 78:567–574
16. Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc B 64:479–498
17. 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 J, 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 Y, 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
18. 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
19. Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. J Mass Spectrom 48:291–298
20. Neveu V, Perez-Jiménez J, Vos F, Crespy V, du Chaffaut L, Mennen L, Knox C, Eisner R, Cruz J, Wishart D, Scalbert A (2010) Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. Database (Oxford). doi:10.1093/database/bap024
21. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K, Kanaya S (2012) KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. Plant Cell Physiol 53:e1–e12
22. 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

23. Rago D, Mette K, Gürdeniz G, Marini F, Poulsen M, Dragsted LO (2013) A LC-MS metabolomics approach to investigate the effect of raw apple intake in the rat plasma metabolome. Metabolomics. doi: 10.1007/s11306-013-0534-9
24. Nelson AC, Huang W, Moody DE (2001) Variables in human liver microsome preparation: impact on the kinetics of l-alpha-acetylmethadol (LAAM) n-demethylation and dextromethorphan O-demethylation. Drug Metab Dispos 29:319–325
25. Wein M, Lavid N, Lunkenbein S, Lewinsohn E, Schwab W, Kaldenhoff R (2002) Isolation, cloning and expression of a multifunctional O-methyltransferase capable of forming 2,5-dimethyl-4-methoxy-3(2H)-furanone, one of the key aroma compounds in strawberry fruits. Plant J 31:755–765
26. Gonthier M, Cheynier V, Donovan JL, Manach C, Morand C, Mila I, Lapierre C, Rémésy C, Scalbert A (2003) Microbial aromatic acid metabolites formed in the gut account for a major fraction of the polyphenols excreted in urine of rats fed red wine polyphenols. J Nutr 133:461–467
27. Bianchi F, Careri M, Musci M, Mangia A (2007) Fish and food safety: determination of formaldehyde in 12 fish species by SPME extraction and GC–MS analysis. Food Chem 100:1049–1053
28. Bingham SA, Cassidy A, Cole TJ, Welch A, Runswick SA, Black AE, Thurnham D, Bates C, Khaw KT, Key TJA (1995) Validation of weighed records and other methods of dietary assessment using the 24 h urine nitrogen technique and other biological markers. Br J Nutr 73:531–550
29. González-Molina E, Domínguez-Perles R, Moreno DA, García-Viguera C (2010) Natural bioactive compounds of Citrus limon for food and health. J Pharm Biomed Anal 51:327–345
30. Vermeulen M, Van Den Berg R, Freidig AP, Van Bladeren PJ, Vaes WHJ (2006) Association between consumption of cruciferous vegetables and condiments and excretion in urine of isothiocyanate mercapturic acids. J Agric Food Chem 54:5350–5358
31. Lloyd AJ, Beckmann M, Haldar S, Seal C, Brandt K, Draper J (2013) Data-driven strategy for the discovery of potential urinary biomarkers of habitual dietary exposure. Am J Clin Nutr 97:377–389
32. González-Barrio R, Edwards CA, Crozier A (2011) Colonic catabolism of ellagitannins, ellagic acid, and raspberry anthocyanins: in vivo and in vitro studies. Drug Metab Dispos 39:1680–1688
33. Heinzmann SS, Brown IJ, Chan Q, Bictash M, Dumas M, Kochhar S, Stamler J, Holmes E, Elliott P, Nicholson JK (2010) Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. Am J Clin Nutr 92:436–443
34. Favé G, Beckmann M, Lloyd A, Zhou S, Harold G, Lin W, Tailliart K, Xie L, Draper J, Mathers J (2011) Development and validation of a standardized protocol to monitor human dietary exposure by metabolite fingerprinting of urine samples. Metabolomics 7:469–484
35. Lloyd AJ, Beckmann M, Favé G, Mathers JC, Draper J (2011) Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. Br J Nutr 106:812–824
36. Pujos-Guillot E, Hubert J, Martin J, Lyan B, Quintana M, Claude S, Chabanas B, Rothwell JA, Bennetau-Pelissero C, Scalbert A, Comte B, Hercberg S, Morand C, Galan P, Manach C (2013) Mass spectrometry-based metabolomics for the discovery of biomarkers of fruit and vegetable intake: citrus fruit as a case study. J Proteome Res 12:1645–1659
37. Feldman JM, Lee EM (1985) Serotonin content of foods: effect on urinary excretion of 5-hydroxyindoleacetic acid. Am J Clin Nutr 42: 639–643
38. Shively CA, Tarka SM Jr (1984) Methylxanthine composition and consumption patterns of cocoa and chocolate products. Prog Clin Biol Res 158:149–178
39. Rodopoulos N, Höjvall L, Norman A (1996) Elimination of theobromine metabolites in healthy adults. Scand J Clin Lab Invest 56: 373–383
40. Cornish HH, Christman AA (1957) A study of the metabolism of theobromine, theophylline, and caffeine in man. J Biol Chem 228: 315–323
41. Farag MA, Motaal AA (2010) Sulforaphane composition, cytotoxic and antioxidant activity of crucifer vegetables. J Adv Res 1:65–70
42. Edmands WMB, Beckonert OP, Stella C, Campbell A, Lake BG, Lindon JC, Holmes E, Gooderham NJ (2011) Identification of human urinary biomarkers of cruciferous vegetable consumption by metabonomic profiling. J Proteome Res 10:4513–4521
43. Rechner AR, Smith MA, Kuhnle G, Gibson GR, Debnam ES, Srai SKS, Moore KP, Rice-Evans CA (2004) Colonic metabolism of dietary polyphenols: influence of structure on microbial fermentation products. Free Radic Biol Med 36:212–225
44. Podsędek A, Sosnowska D, Redzynia M, Anders B (2006) Antioxidant capacity and content of *Brassica oleracea* dietary antioxidants. Int J Food Sci Technol 41:49–58
45. Frank T, Stintzing FC, Carle R, Bitsch I, Quaas D, Stra G, Bitsch R, Netzel M (2005) Urinary pharmacokinetics of betalains following consumption of red beet juice in healthy humans. Pharmacol Res 52: 290–297
46. Walsh MC, Brennan L, Malthouse JPG, Roche HM, Gibney MJ (2006) Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. Am J Clin Nutr 84:531–539
47. Rasmussen LG, Savorani F, Larsen TM, Dragsted LO, Astrup A, Engelsen SB (2011) Standardization of factors that influence human urine metabolomics. Metabolomics 2011(7):71–83
48. Koulman A, Volmer DA (2008) Perspectives for metabolomics in human nutrition: an overview. Nutr Bull 33:324–330


**Maj-Britt Schmidt Andersen** is about to finalize her PhD studies at Department of Nutrition, Exercise and Sports at the University of Copenhagen. Her work is focused on the discovery of new exposure and compliance markers in urine for intake of foods and dietary patterns by application of untargeted LC-MS based metabolomics.

# Electronic supplementary material

Analytical and Bioanalytical Chemistry

**Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics**

Maj-Britt Schmidt Andersen[1], Mette Kristensen[1], Claudine Manach[2], Estelle Pujos-Guillot[2], Sanne Kellebjerg Poulsen[1], Thomas Meinert Larsen[1], Arne Astrup[1], Lars Dragsted[1]

[1]Department of Nutrition Exercise and Sports, Faculty of Science, University of Copenhagen.
[2]INRA, UMR1019, Human Nutrition Unit, University of Auvergne, Clermont-Ferrand-Theix, France

Corresponding author:

Maj-Britt Schmidt Andersen
Department of Nutrition, Exercise and Sports,
Faculty of Science, University of Copenhagen
Rolighedsvej 30
DK-1958 Frederiksberg C
Email: mbsa@life.ku.dk
Telephone: +45353 31086
Fax: 3533 2483

Caption ESM 1: The supplementary material contains an outline of the study design, the baseline characteristics for the included subjects, the parameters applied for preprocessing of data, and further structural information on six tentatively identified metabolites.

**Table S1** List of foods in the standardized diet

The amounts given are guidelines for participants with a daily energy intake of 9-11 MJ.

| **Food** | **Amount** |
|---|---|
| Oatmeal | 60 g |
| Skimmed milk (0.5 % fat) | 160 g |
| Sugar | 16.5 g |
| Orange juice | 250 ml |
| Crispbread | 2 slices |
| Pear | 1 (≈100 g) |
| Rye bread | 2 slices |
| Cold chicken meat | 20 g |
| Tinned tuna | 20 g |
| Mayonnaise | 4 g |
| Cheese spread (25 % fat) | 15 g |
| Cucumber | 45 g |
| Salad leaves | 4 leaves |
| Wholemeal bun | 1 |
| Butter | 30 g |
| Hard cheese (25 % fat) | 1 slice |
| Strawberry jam | 15 g |
| Dark chocolate | 25 g |
| Chicken filet (fried) | 120 g |
| Olive oil | 2 spoons |
| Potatoes | 100 g |
| Courgette | 40 g |
| Onion | 50 g |
| Carrot | 30 g |
| Greek yogurt (10 % fat) or crème-fraîce (9 % fat) | 50 g |
| Wholemeal French loaf | 1.5 pieces |
| Strawberries | 100 g |
| Natural yogurt | 150 ml |

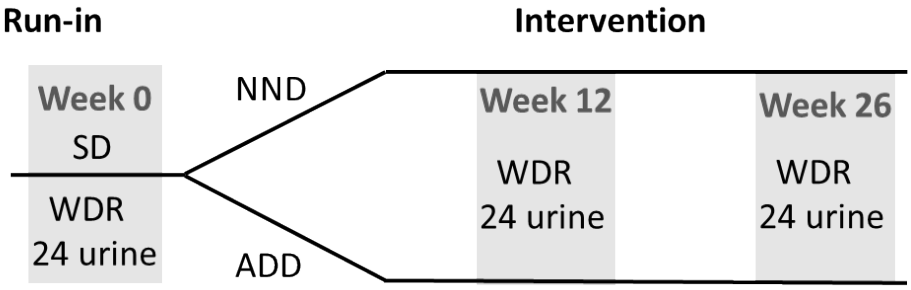



**Fig. S1** Outline of study design. After one week run-in where the subjects had a standardized diet (SD) for 3 days, participants were randomized to follow either a New Nordic Diet (NND) or an Average Danish Diet (ADD) for 26 weeks. At three sampling points during the study, 3-day weighed dietary records (WDR) were made and one 24 h urine sample was collected. The first dietary record was for the standardized diet and the other two were for either NND or ADD.

**Table S2** Baseline characteristics of the 107 included subjects

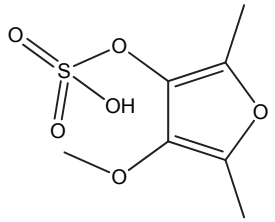
| | **NND (n=64)** | **ADD (n=43)** |
|---|---|---|
| Age | 44.0 (13.3) | 41.2 (13.2) |
| Sex | 16 M, 48 F | 12 M, 31 F |
| BMI (week 0) | 29.9 (4.5) | 29.4 (4.6) |
| Waist circumference (week 0) | 99.2 (11.6) | 99.0 (13.3) |

**Table S3** Batch steps and parameters used for preprocessing of raw data in MZmine2.7

| | Batch step | Parameters |
|---|---|---|
| **Negative mode** | Raw data import | |
| | Mass detection | Noise level: 15 |
| | Chromatogram builder | Min time span (min): 0.01;<br>Min height: 4.0E1; m/z tolerance: 0.055 mz or 30 ppm |
| | Chromatogram deconvolution | Chromatographic threshold: 95%; Search minimum in RT range (min): 0.01; Minimum relative height: 10%; Minimum absolute height: 4.0E1; Min ratio of peak/top edge: 1.3; Peak duration range (min): 0.01-0.2 |
| | Isotopic pattern | m/z tolerance: 0.06 or 30 ppm; Retention time tolerance: 0.01; Monotonic shape; maximum charge: 1 |
| | Join aligner | m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.15; Weight for both m/z tolerance and retention time tolerance: 10 |
| | Duplicate peak filter | m/z tolerance: 0.5 or 600 ppm; RT tolerance: 0.15 |
| | Peak list rows filter | Min peaks in a row: 5<br>Minimum peaks in an isotope pattern: 1; m/z range: 50-1000; RT range: 0-7; peak duration range: 0.01-0.2 |
| | Peak finder | Intensity tolerance: 50%; m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.15 |
| | Export to csv | None required |
| **Positive mode** | Raw data import | |
| | Mass detection | Noise level: 15 |
| | Chromatogram builder | Min time span (min): 0.01;<br>Min height: 4.0E1; m/z tolerance: 0.055 mz or 30 ppm |
| | Chromatogram deconvolution | Chromatographic threshold: 97%; Search minimum in RT range (min): 0.01; Minimum relative height: 10%; Minimum absolute height: 6.0E1; Min ratio of peak/top edge: 1.5; Peak duration range (min): 0.01-0.2 |
| | Isotopic pattern | m/z tolerance: 0.06 or 30 ppm; Retention time tolerance: 0.01; Monotonic shape; maximum charge: 1 |
| | Join aligner | m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.15; Weight for both m/z tolerance and retention time tolerance: 10 |
| | Duplicate peak filter | m/z tolerance: 0.5 or 600 ppm; RT tolerance: 0.15 |
| | Peak list rows filter | Min peaks in a row: 5<br>Minimum peaks in an isotope pattern: 1; m/z range: 50-1000; RT range: 0-7; peak duration range: 0.01-0.2 |
| | Peak finder | Intensity tolerance: 50%; m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.17 |
| | Export to csv | None required |

**Table S4** PEM for strawberry with RT 3.43 and m/z 221.024, neg mode

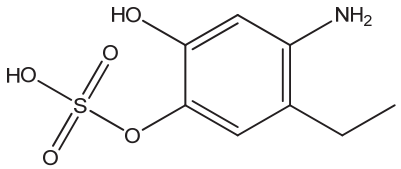
Proposed structure of the molecule (molecular formula C7H10O6S):




| Fragments from MS/MS | Annotation (ESI-) | Loss |
|---|---|---|
| 221.024 | C7H9O6S | |
| 205.989 | C6H6O6S | -CH3 |
| 141.056 | C7H9O3 | -SO3 |
| 126.031 | C6H6O3 | -CH3, -SO3 |
| 79.957 | SO3 | |

**Table S5** PEM for beetroot with RT 0.95 and m/z 234.045, pos mode

Proposed structure of the molecule (molecular formula C8H11NO5S):

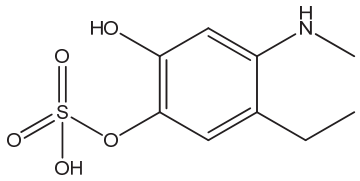



| Fragments from MS/MS | Annotation (ESI+) | Loss |
|---|---|---|
| 234.045 | C8H12NO5S | |
| 217.017 | C8H9O5S | -NH3 |
| 154.087<br>(same m/z as dopamine) | C8H12NO2 | -SO3 |
| *137.059 | C8H9O2 | -NH3, -SO3 |
| *119.049 | C8H7O | -NH3, -SO3, -H2O |
| *91.054 | C7H7 | -NH3, -SO3, -H2O, -CO |

*The m/z was also found as a fragment of dopamine

**Table S6** PEM for beetroot with RT 1.11 and m/z 246.038, neg mode

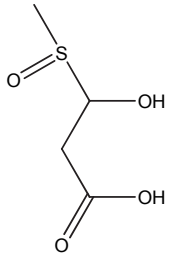
Proposed structure of the molecule (molecular formula C9H13NO5S):




| Fragments from MS/MS | Annotation (ESI-) | Loss |
|---|---|---|
| 246.038 | C9H12NO5S | |
| 203.007 | C7H7O5S | -CH2, -CH3-N (with rearrangements) |
| 166.087 | C9H12NO2 | -SO3 |
| 151.063 | C8H9NO2 | -SO3, -CH3 |
| 121.028 | C7H5O2 | -SO3, -CH3, -CH3-NH |
| 79.958 | SO3 | |

**Table S7** PEM for red cabbage with RT 0.53 and m/z 151.009, neg mode

Proposed structure of the molecule (molecular formula C4H8SO4):

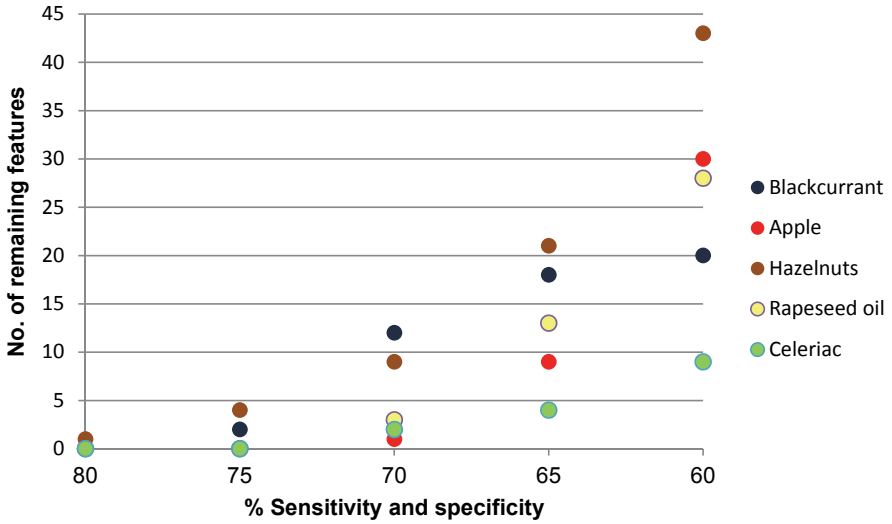



| Fragments from MS/MS | Annotation (ESI-) | Loss |
|---|---|---|
| 151.009 | C4H7O4S | |
| 135.986 | C3H4O4S | -CH3 |
| 132.996 | C4H5O3S | -H2O |
| 111.021 | C2H7O3S | -C2O (with rearrangements) |
| 87.008 | C3H3O3 | -CH3, -SOH |
| 62.989 | CH3SO | - CHOH-CH2-CO2 |

**Table S8** PEM for green beans with RT 3.47 and m/z 189.083, neg mode

| Fragments from MS/MS | Annotation (ESI-) | Loss |
|---|---|---|
| 189.083 | C8H13O5 | |
| 145.085 | C7H13O3 | -CO2 |
| 129.056 | C6H9O3 | -COOH-CH3 |
| 127.075 | C7H11O2 | -CO2-H2O |
| 99.081 | C6H11O | -CO2-HCOOH |

**Figure S2** Number of remaining features after initial validation for different levels of sensitivity and specificity




The number of remaining features after initial validation is given for five representative food groups and five cut-off levels for sensitivity and specificity. The food groups chosen all had a high number of significantly different features in the paired t-test (43-373) of which none remained following initial validation and sensitivity and specificity analysis for foods.

# Paper III

## Untargeted metabolomics as a screening tool for estimating compliance to a dietary pattern

Andersen, M.S., Rinnan, Å., Manach, C., Poulsen, S.K., Pujos-Guillot, E., Larsen, T.M., Astrup, A., Dragsted, L.

Revised and resubmitted for Journal of Proteome Research

# Untargeted metabolomics as a screening tool for estimating compliance to a dietary pattern

*Maj-Britt S. Andersen,* [†] *Åsmund Rinnan,* [‡] *Claudine Manach,* [§] *Sanne K. Poulsen,* [§] *Estelle Pujos-Guillot,* [†] *Thomas M. Larsen,* [†] *Arne Astrup ,* [†] *Lars O. Dragsted*[†]

[†]Department of Nutrition Exercise and Sports, Faculty of Science, University of Copenhagen.
[‡]Department of Food Science, Faculty of Science, University of Copenhagen.
[§]INRA, UMR1019, Human Nutrition Unit, University of Auvergne, Clermont-Ferrand-Theix, France

KEYWORDS: metabolomics, dietary compliance, biomarkers, dietary assessment, dietary patterns, multivariate model, New Nordic Diet, Average Danish Diet

**ABSTRACT**

There is a growing interest in studying the nutritional effects of complex diets. For such studies measurement of dietary compliance is a challenge since the currently available compliance markers only cover limited aspects of a diet. In the present study, an untargeted metabolomics approach was used to develop a compliance measure in urine to distinguish between two dietary patterns. A parallel intervention study was carried out in which 181 participants were randomized to follow either a New Nordic Diet (NND) or an Average Danish Diet (ADD) for six months. Dietary intakes were closely monitored over the whole study period and 24 h urine samples as well as weighed dietary records were collected several times during the study. The urine samples were analysed by UPLC-qTOF-MS and a partial least squares discriminant analysis with feature selection was applied to develop a compliance model based on data from 214 urine samples. The optimised model included fifty-two metabolites and had a misclassification rate of 19 % in a validation set containing 139 samples. The metabolites identified in the model were markers of individual foods such as citrus, cocoa containing products and fish as well as more general dietary traits such as high fruit and vegetable intake or high intake of heat-treated foods. It was easier to classify the ADD diet than the NND diet probably due to seasonal variation in the food composition of NND and indications of lower compliance among the NND subjects. Untargeted metabolomics is a promising approach to develop compliance measures that cover the most important discriminant metabolites of complex diets.

## Introduction

Untargeted metabolomics is a well-established screening tool to explore changes and patterns for a large number of metabolites in biofluids[1]. A common application of untargeted metabolomics within nutrition research is for biomarker discovery of dietary exposure, where the method has proved successful in finding new biomarker candidates for a range of foods and food groups, such as citrus[2], coffee[3], cocoa[4], cabbage[5] and nuts[6]. The advantage of an explorative approach for gaining insight in dietary exposure is that it opens up for studying the food metabolome under various conditions. For example the effect of dietary standardization[7, 8], the timing of sampling[9], and the influence of the food matrix[4] can be studied in a holistic fashion.

In nutrition research, the health effects of complex diets are gaining increasing interest. One reason for this is the fact that the common approach, in which effects of individual nutrients and foods are investigated separately, has limited capability to explain associations between diet and disease[10]. It is likely that different foods in a dietary pattern can exhibit synergistic and antagonistic effects which should be considered to understand which dietary components may be preventive for disease development[11, 12]. In addition, habitual diets are per definition complex and health effects of complex diets are therefore possibly easier to translate into health messages. Performing measurements of dietary intakes and compliance for complex diets is challenging. Few compliance markers of individual foods are available[13] and even the ones that are may not be the best choice for studying whole diets.

Untargeted metabolomics could likely contribute to understanding how humans respond to a complex diet[14]. By performing simultaneous measures of a high number of metabolites in urine it may be possible to develop a compliance measure for different dietary patterns that could be used to identify non-compliant subjects or groups of individuals with certain dietary responses. While some metabolomics studies have tried to identify urinary markers associated with habitual consumption of individual foods[2, 15, 16] or characteristic dietary traits in habitual diets such metabolites associated with diets high or low in meat[17, 18], fruit and vegetables[19], energy percentage from protein[20] or glycemic index[21], few studies have investigated complex dietary patterns. To our knowledge complex dietary patterns based on habitual dietary data have been investigated in five metabolomics studies[11, 22, 23, 24, 25], of which urine samples were only included in one. Even though these studies are diverse in the choice of analytical approach a common characteristic is that habitual dietary patterns seem to be linked to levels of certain metabolites in serum and urine. In the present study,two well-defined complex dietary patterns, a New Nordic Diet (NND) and an Average Danish Diet (ADD)[26], have been compared in a six-month parallel intervention study[27]. Due to a marked difference in food composition between the dietary patterns, it is expected that they will be clearly distinguishable in the urinary metabolome of the subjects. The study therefore provides an excellent opportunity to study the urinary metabolic response to well-defined dietary patterns consumed as habitual diets.

In the present work, we have applied untargeted metabolomics to develop a multivariate model, including the most discriminative urinary metabolites of NND and ADD. This was done in order to

investigate the value of the model as a compliance measure and to identify how the metabolites in such a model are related to the dietary patterns.

## Materials and methods

*Study design*

A six months parallel intervention study was conducted with two dietary patterns, an Average Danish Diet (ADD) and a New Nordic Diet (NND). The diets were defined by macronutrient composition and intake of fifteen food groups. Target ranges and actual intakes for foods and macronutrients in the study are given in Table 1. During the study, the participants selected all their foods free of charge from a small supermarket based at the department of Nutrition, Exercise and Sports, University of Copenhagen. Dietary intakes were monitored continuously throughout the study by registering all foods collected by each household. A household consisted of one subject or a couple who both participated in the study. For each visit to the supermarket, it was ensured that the participants in a household only brought home foods that were in accordance with the target ranges for their dietary pattern. Any food bought in other supermarkets or not consumed by participants also had to be reported in order to achieve a good estimation of dietary intakes during the whole study period. In addition, 3-day weighed dietary records (WDR) were made at week 0, 12 and 26. For the WDR, participants were requested to report all intakes of food and beverages, except water, from early morning until late evening. More details on the study can be found in Poulsen et al. (2013)[27].

**Table 1.** Target ranges and intake data from the supermarket of food groups and macronutrients in the NND and ADD

| | **NND** | | | **ADD** | | |
|---|---|---|---|---|---|---|
| | Target | Model set | Validation set | Target | Model set | Validation set |
| No with household size of one | | 40 | 30 | | 32 | 26 |
| **Macronutrients [Energy %]** | | | | | | |
| Total protein | 15-23 | 18 (16-21) | 18 (17-20) | 10-20 | 17 (15-20) | 17 (15-20) |
| Total carbohydrate (incl. fibre) | 48-56 | 54 (48-58) | 55 (50-58) | 45-55 | 51 (47-55) | 51 (45-55) |
| Added sugar | <10 | 5 (1-10) | 5 (1-9) | ≥12 | 12 (8-17) | 12 (8-17) |
| Total fat | 25-35 | 31 (28-35) | 30 (28-35) | 33-37 | 35 (31-38) | 34 (31-39) |
| SFA | <10 | 8 (6-11) | 8 (6-10) | 10-20 | 14 (11-16) | 14 (11-16) |
| **Food groups [g/10MJ]** | | | | | | |
| Total fruit (incl. berries) | 250-350 | 411 (300-554) | 425 (294-567) | 150-250 | 198 (149-297) | 200 (154-263) |
| Berries | 50-100 | 90 (34-151) | 97 (50-191) | 2-6 | 6 (1-13) | 7 (2-15) |
| Total vegetables (incl. cabbages, root vegetables and legumes) | 350-450 | 438 (269-704) | 463 (330-704) | 150-210 | 201 (125-298) | 210 (147-293) |
| Cabbages | 25-35 | 66 (31-127) | 68 (36-127) | ≤10 | 8 (1-24) | 7 (1-24) |
| Root vegetables | ≥150 | 195 (113-399) | 212 (153-399) | 25-35 | 18 (4-39) | 21 (7-41) |
| Legumes | ≥30 | 48 (21-122) | 48 (27-108) | ≤1 | 5 (0-10) | 6 (2-19) |
| Potatoes | 140-160 | 117 (45-260) | 123 (67-260) | 90-110 | 76 (30-235) | 75 (21-235) |
| Fresh herbs | As much as possible | 11 (3-23) | 9 (3-19) | ≤1 | 2 (0-9) | 2 (0-9) |
| Wild plants and mushrooms | 3-7 | 6 (0-12) | 6 (1-12) | 0 | 0 (0-3) | 0 (0-1) |
| Nuts | ≥30 | 36 (24-46) | 35 (24-43) | ≤1 | 8 (1-18) | 9 (1-18) |
| Wholegrain | ≥75 | 155 (103-220) | 160 (107-248) | 25-45 | 44 (26-92) | 47 (30-92) |
| Total meat (incl. game meat) | 90-110 | 98 (66-179) | 101 (66-137) | 130-150 | 155 (118-207) | 156 (123-247) |
| Game meat | ≥4 | 26 (3-49) | 29 (13-49) | 0 | 0 (0-1) | 0 (0-1) |
| Fish and shellfish | >43 | 75 (40-130) | 75 (42-152) | 15-25 | 20 (12-41) | 21 (12-41) |
| Seaweed | 3-7 | 1 (0-3) | 1 (0-3) | 0 | 0 (0-0) | 0 (0-0) |

Mean intakes for food groups (in grams per 10 megajoule) and macronutrients (in energy %) for subjects with a household size of one, who completed the study and whose samples were used as model and/or validation samples in the partial least squares discriminant analysis (PLS-DA). Intakes are calculated for the whole study period based on supermarket data. Minimum and maximum observations are given in brackets. Target intakes in the study are given in the first columns within NND and ADD.

Twenty-four hour urine samples were collected five times (week 0, 4, 12, 20 and 26). Each urine collection was from 8 am until 8 am, excluding and including the first void, respectively. Urine was stored in cooler bags during collection and an aliquot was transferred to -80°C after delivery to the study unit. An outline of the study is given in Figure 1. The study has been approved by the Regional Ethics Committee of Greater Copenhagen and Frederiksberg (H-3-2010-058) and the Danish Data Protection Agency (2007-54-0269).

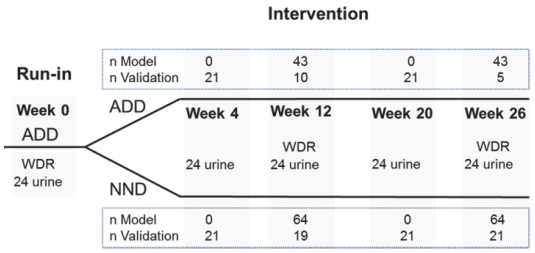



**Figure 1.** Outline of the study design with sampling points for 24 h urine samples (24 urine) and 3-day weighed dietary records (WDR). The numbers of samples used to develop and validate the compliance model (designated Model and Validation) are given for each sampling point. Boxes at the top and bottom are for the Average Danish Diet (ADD) and the New Nordic diet (NND), respectively.

*Subjects*

Subjects recruited were males and females aged 18-65 years with increased waist circumference (>94 cm for men and >80 cm for women). Participants with additional risk factors of the metabolic syndrome were preferred but it was not an inclusion criterion in itself. In total, 181 subjects (71 % women) were randomized in a ratio 3:2 to follow NND or ADD. For the purpose of this study, a subgroup of 107 subjects (74 % women, 60 % NND) was selected, who had provided a urine sample and a WDR on the same day at week 0, 12 and 26. The samples from this subgroup (214 urine samples) were used to develop a compliance model. From the rest of the urine samples, 139 samples (63 % women, 59 % NND) were selected at random to validate the model. Forty-five percent of the validation samples were from the same subjects whose samples had been included to develop the model but from other time points. The rest of the validation samples were from other subjects. A WDR was available for 55 samples in the validation set (Figure 1). Baseline characteristics for the participants are provided in the supporting information Table S1. The distribution of model and validation samples across the sampling points is given in Figure 1.

*UPLC-qTOF analysis*

Urine samples were centrifuged, diluted 1:1 with aqueous 5 % 30:70 (v/v) acetonitrile (ACN):methanol (MeOH) (Optima grade LC–MS, Fisher Scientific, US) and distributed randomly into 96-well auto-injector trays, keeping all samples from the same subject immediately after each other within a plate to minimize intra-individual variation. The samples were analysed using an UPLC-qTOF-MS (Waters, Manchester, UK) fitted with an electrospray ion source and an HSS T3 C18 (Waters, Milford, MA) column and pre-column. Samples were run in positive and negative ionization mode. The run time per sample was 7 minutes and centroid data was collected from 50 to 1000 *m*/*z*. A metabolomics standard solution including 44 different physiologically relevant compounds[28] and a pooled urine sample including all samples on a plate were run four and ten times, respectively, per plate for quality control purposes. For more experimental details see method II in Barri et al. (2012)[28].

*Data preprocessing and pre-treatment*

Raw spectra from the UPLC-qTOF-MS were preprocessed in MZmine2[29] to obtain a list of detected features in urine, characterised by a mass to charge ratio (m/z), a retention time (RT) and an ionisation mode (negative or positive). The batch steps and parameters applied for preprocessing are listed in the supporting information Table S2. After preprocessing, integrated peak areas for the detected features were imported to MATLAB® version 7.12.0.635, R2011a (Mathworks Inc., Sherborn, MA, US) which was used for the following pre-treatment and statistical analysis of the data. The obtained features in positive and negative mode, respectively, were normalized across samples to the same mean sum to correct for urine concentration differences and batch drift. Then, plate correction was performed to remove analytical variation due to plates. All samples run on the same plate were adjusted across each feature to obtain the same mean value on each plate. After normalisation and plate correction, data from negative and positive mode were merged into one data matrix and baseline data (week 0, Figure 1) were excluded. Features with peak areas above a noise level of seven which were present in less than five percent of the samples from the dietary groups NND and ADD were removed as these features are either noise or unrelated to the dietary patterns of interest. Before the statistical analysis, data was visually explored for outliers in PCA..

*Statistical analysis*

*Development of a PLS-DA compliance model for ADD and NDD based on urine samples.*

Partial least squares discriminant analysis (PLS-DA) with feature selection was applied to develop a compliance measure for ADD and NND based on the most discriminative features detected in urine. For the development of the model, the pre-treated data matrix from the 214 model samples were used (Figure 1) and the two dietary patterns were given as class information. Data was autoscaled and feature selection was made based on variable importance in projection (VIP) scores[30, 31] in an iterative procedure as described previously[32]. First, 16 initial PLS-DA models were developed with different subsets of the model samples. For each of the initial models, eight

randomly chosen subjects (four NND and four ADD) were excluded as test set and the rest of the samples were used as training set. Cross-validation was applied to determine the number of latent variables (LVs) in the initial models and the most discriminant features were selected by excluding features with VIP < 0.8 for both dietary classes.To evaluate the performance of the models after feature selection, the test set was used . An optimised final PLS-DA compliance model was developed based on the features selected in all initial models by making 16 new PLS-DA models which included features present in from one to 16 of the initial models, respectively. Among the new models, the one with the lowest CV error was chosen as the final model. The relative importance of the features in the final model was evaluated from the VIP scores and the prediction strength for dietary pattern of the final model was evaluated by applying the validation samples (Figure 1). More detailed information on how the PLS-DA model was developed can be found in the supporting information in the statistical analysis section.

*PCA on urine for individual dietary patterns.*

To explore how misclassified urine samples among the validation samples in PLS-DA differed from correctly classified samples within a dietary pattern, two principal component analyses (PCA) models were made including all NND and ADD urine samples, respectively. The features selected in the final PLS-DA model were used as variables in the PCA model and data was autoscaled prior to analysis.

*PCA on food intake.*

PCA was also applied to investigate the dominant food patterns in NND and ADD. Data from WDR and from the supermarket were used. However, as supermarket data was recorded per household, a subgroup of 72 participants with a household size of one was selected for these data out of the 107 subjects included for development of the PLS-DAcompliance model. From the WDR, only foods reported on the same days as the subjects collected urine were used. This was done to have a food recording consistent with the timing of urine sampling, as the participants had not collected urine on the same day within the three days of WDR in all cases. Foods reported in WDR and the supermarket data were grouped into comparable categories and for each dataset, foods recorded less than five times were excluded. Three PCA models were made: One for WDR, one for the supermarket data (including the 72 participants with a household size of one), and an additional one for WDR (including the same subset of participants as in the supermarket data). For all models, data was first normalised row-wise to unit length to exclude variation caused by differences in amount of food reported between subjects. Then data was autoscaled to give all food categories equal weight in the analysis. The most explanatory foods for each diet were selected in each model. For the PCA based on WDR, 55 WDR from the validation set (Figure 1) were available and these were used to explore if any of the WDR in the validation set had a diverging food pattern from the WDR of the model samples. More details on the PCA of food intake is provided in the supporting information.

*Identification of features in the PLS-DA compliance model*

Identification of the features in urine remaining in the final PLS-DA model is crucial to understand which urinary metabolites are characteristic of the dietary pattern and how they are related to the diets. In order to obtain as much information as possible on the features to generate hypotheses for identification, several steps were performed. The dietary pattern with highest levels of each feature was determined from the distribution of peak areas of the feature across all samples to understand which diet the feature characterised. All features were correlated using Pearson's correlation coefficient (r = 0.6 was used as cut-off) to identify clusters of features that were likely related. For example highly correlated features could be from the same compound (in-source fragments and adducts)or related by other means, such as a shared metabolic pathway. Potential food sources for the features were investigated by calculating the specificity and sensitivity of each feature for all self-reported foods on the day of urine sampling extracted from WDR which were reported more than five times (207 food items). The calculation of sensitivity and specificity is described in the supporting information and has also been applied previously for another study on the same dataset[33]. All food items that reached a sensitivity and specificity above 0.7 for a feature were taken into account as possible dietary origin of the feature but are only reported here if the relation to the food source was supported by identification of one or more features.

A few samples that were representative for the features in the model were analysed in full scan mode on an LTQ Orbitrap Velos$^{TM}$ MS equipped with a BEH Shield RP18 column. The m/z of each feature was searched for in the resulting chromatograms to obtain a better mass accuracy of the features which was used to determine the most probable molecular formulas. The parent ion was identified from in-source fragments and adducts, and additional structural information on the features was obtained by performing MS/MS fragmentation on relevant ions using LC-qTOF-MS in product ion scan mode with collision energies of 10, 20 and 30 eV.

All information on the features was used to search a number of chemical and metabolite databases: The Human Metabolome Database[34], METLIN[35], MetFusion[36]. For features with a high sensitivity and specificity for a food, food databases (The CRC Dictionary of Food compounds[37], Phenol-explorer[38] and KNApSAcK[39]) and literature were searched for possible precursor compounds in the foods.

When standards were available, these were analysed to confirm tentative identifications. Four levels of identification for metabolites are reported[40]. For level I, the identification has been confirmed by an authentic standard, level II identifications are based on a comparison of MS/MS fragmentation patterns to previously published data. Level III is used when it was possible to identify the compound class based on similarities to published fragmentation patterns, while level IV is used for unknown compounds.

*Standards analysed:* 3,7-dimethyluric acid, indole-3-carboxylic acid, theobromine and trimethylamine *N*-oxide, *N*-methyl-cis-4-hydroxy-*L*-proline (purchased from Sigma-Aldrich, Germany), (2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid, 7-methyluric acid, 7-methylxanthine, stachydrine hydrochloride (purchased from Santa cruz, USA), 5-acetylamino-6-amino-3-

methyluracil, hydrate (purchased from Stratech, UK), Hippuric acid (purchased from Buchem BV, the Netherlands), Cyclo(-Pro-Val) (purchased from Bachem, Germany), Pyrraline (purchased from PolyPeptide Group, France).

*Synthesized standards analysed:* Hydroquinone, 3-indoleacetic acid, limonene-1,2-diol, octanoic acid, perillic acid and pyrocatechol, all purchased from Sigma-Aldrich, Germany, underwent enzymatic glucuronidation, as described previously[32], prior to analysis. Pyroglutamyl proline was synthesized by solid-phase peptide synthesis following the procedure described by Stanstrup et al. (2013)[41].

## Results

*Development of a PLS-DA compliance model for ADD and NDD based on urine samples*

After data preprocessing and pre-treatment a total of 4369 features (2104 and 2265 in positive and negative mode, respectively) remained and from these the most discriminant features were selected by PLS-DA analysis. No samples were removed as outliers prior to analysis. The initial 16 PLS-DA models had an average classification error for cross-validation of 0.04, and an average classification error for the test sets of 0.13. Nine LVs and 52 features were included on average in the models (data for the individual models are provided in Table S3). The lowest cross-validated classification error in the models, obtained from the selected features in the initial models, was 0.033 (Figure S1). In this final PLS-DA compliance model, 67 features and four LVs were included (Figure S2). In Figure 2A and B, a score- and a loading plot, respectively, for the first two LVs in the final PLS-DA model are shown. In the score plot (Figure 2A) the two dietary patterns are separated mainly along LV 1. When the predictive performance of the final PLS-DA compliance model was tested with the validation samples, a total of 26 samples corresponding to 19 % were misclassified (23 NND samples and three ADD samples). The proportion of misclassified samples from subjects whose samples had been included to develop the model was 35 %. The misclassified NND samples were from 18 different subjects. Five of the subjects had more than one misclassified sample in the validation set. Four other NND subjects with more than one sample in the validation set only had one sample misclassified each. The misclassified ADD samples were from three different subjects.

*Identification of features in the PLS-DA compliance model*

Six clusters of correlating features were found in WDR of which five had high sensitivities and specificities to certain foods (chocolate, citrus, limonene, various heat treated foods and fish) . The feature clusters are highlighted in the loading plot of Figure 2B. Four of the feature clusters (chocolate, citrus, limonene and the heat treated foods cluster) were related to the ADD diet while two clusters were related to NND (a fish cluster and an NND cluster which did not have high sensitivity and specificity for any particular foods in WDR).

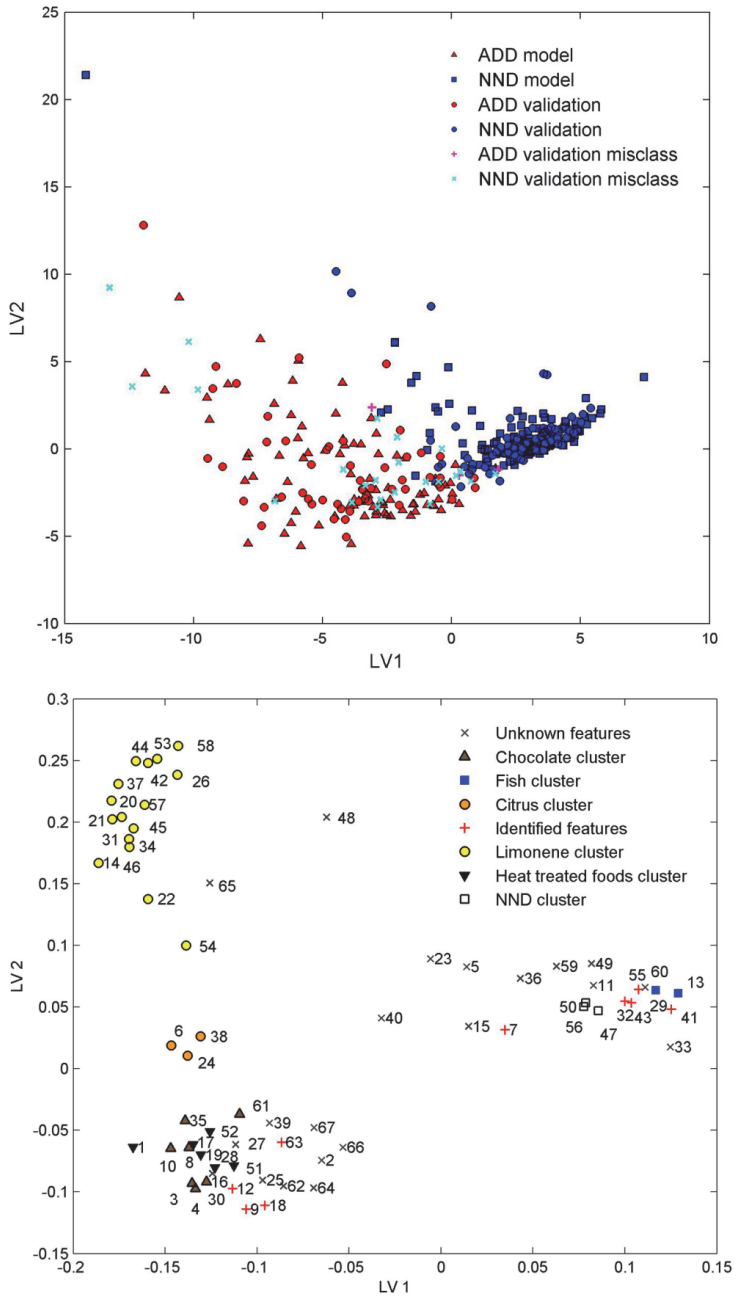



**Figure 2**. (A) Score plot of latent variable (LV) 1 and 2 for the final PLS-DA model. Urine samples used to develop the model ('model'), urine samples used to validate the model ('validation'), and misclassified urine samples in the validation set ('validation misclass') are highlighted for both dietary patterns. (B) Loading plot of LV 1 and 2 for the final PLS-DA model. The features in the loading plot are numbered according their rank in the model based on VIP scores. Groups of correlated feature clusters are highlighted, as well as unknown and identified features outside feature clusters.

The 67 features in the loading plot corresponded to 52 unique metabolites. Fifteen metabolites were identified at level I and six metabolites were tentatively identified at level II-III. Detailed information on the identified and tentatively identified metabolites is given in Table 2, which also includes all unknown features that were part of the correlated clusters in Figure 2B. Supporting information on tentatively identified features is provided in Table S4-S14. A list of all remaining unknown compounds in the PLS-DA is given in Table S15.

**Table 2.** List of identified features at level (I-IV) and features included in the correlated clusters (highlighted in Figure 2B)

| Diet | Feature cluster | Rank PLS-DA | RT qTOF | m/z qTOF | m/z orbitrap | Error (ppm) | Ion | Elemental formula | Main MS/MS fragments | Identification |
|---|---|---|---|---|---|---|---|---|---|---|
| ADD | Heat treatment cluster | 1 | 2.17 | 253.117 | 253.1189 | 4.8 | $[M-H]^-$ | C12H18N2O4 | 223.108; 124.040; 94.032[a] | Pyrraline (I) |
| ADD | Heat treatment cluster | 17 | 2.09 | 148.112 | 148.1118 | 5.8 | Pos mode | C10H14N | 132.081[a] | Unknown |
| ADD | Heat treatment cluster | 51<br>19 | 3.39 | 295.128<br>279.134 | 295.1295<br>279.1351 | 0.4<br>2.1 | $[M-H]^-$<br>$[M-H2O+H]^+$ | C14H20N2O5 | Neg mode: 253.119; 124.039[a]; 94.029[a]<br>Pos mode: 279.134 | [b]Unknown pyrraline derivative (III) |
| ADD | Heat treatment cluster | 28 | 1.03 | 319.066 | 319.0672 | 2.1 | $[M-H]^-$ | C12H16O10 | 175.024; 143.033; 113.024 | [b]Unknown glucuronide (IV) |
| ADD | Heat treatment cluster | 52 | 3.46 | 297.144 | 297.1444 | 2.2 | $[M+H]^+$ | C14H20N2O5 | 239.139; 192.102 | [b]Unknown pyrraline derivative (III) |
| ADD | Chocolate cluster | 3 | 2.04 | 181.0719 | 181.0711 | 7.9 | $[M+H]^+$ | C7H8N4O2 | 163.061, 138.066, 110.070[a] | Theobromine (I) |
| ADD | Chocolate cluster | 4<br>16 | 1.01 | 181.035<br>363.080 | 181.0364<br>363.0814 | 1.5<br>3.4 | $[M-H]^-$<br>$[2M-H]^-$ | C6H6N4O3 | 166.013, 138.028, 110.035[a] | 7-methyluric acid (I) |
| ADD | Chocolate cluster | 8 | 0.90 | 199.085 | 199.0822 | 4.4 | $[M+H]^+$ | C7H10N4O3 | 171.087, 156.064[a], 139.041[a] | 6-amino-5-[N-methylformylamino]-1-methyluracil (II) |
| ADD | Chocolate cluster | 10<br>35 | 1.39 | 197.068<br>195.050 | 197.0666<br>195.0522 | 4.3<br>2.0 | $[M+H]^+$<br>$[M-H]^-$ | C7H8N4O3 | Neg mode: 180.029[a]; 137.023[a]; 124.050[a]<br>Pos mode: 182.041[a]; 139.041[a]; 126.068[a] | 3,7-dimethyluric acid (I) |
| ADD | Chocolate cluster | 61 | 1.13 | 333.104 | 333.1050 | 2.8 | $[2M+H]^+$ | C6H6N4O2 | 150.030; 124.056; 96.055[a] | 7-methylxanthine (I) |
| ADD | Citrus cluster | 6 | 0.52 | 144.101 | 144.1024 | 0.2 | $[M+H]^+$ | C7H13NO2 | 84.080[a] | Proline betaine (I) |
| ADD | Citrus cluster | 24 | 0.73 | 146.083 | 146.0809 | 5.8 | $[M+H]^+$ | C6H11NO3 | 128.070, 84.080 | Unknown (IV) |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ADD | Citrus cluster | 38 | 1.08 | 142.049 | 142.0505 | 0.7 | $[M-H]^-$ | C6H9NO3 | None | Unknown (IV) |
| ADD | - | 9 | 1.90 | 225.087 | 225.0880 | 1.9 | $[M-H]^-$ | C10H14N2O4 | 181.098; 112.039[a]; 82.030 | Pyroglutamyl proline (I) |
| ADD | Limonene cluster | 14 | 2.91 | 361.147 | 361.1490 | 2.4 | $[M-H]^-$ | C16H26O9 | 113.026[a], 85.029[a], 75.009[a] | [b]p-menth-1-ene-6,8,9-triol (IV) |
| ADD | Limonene cluster | 20<br>57 | 2.89 | 375.128<br>394.171 | 375.1289<br>394.1701 | 0.7<br>3.2 | $[M-H]^-$<br>$[M+NH4]^+$ | C16H24O10 | Neg mode: 357.118; 331.140; 113.024<br>Pos mode ($[M+NH4]^+$): 341.121; 183.100; 165.091 | [b]Perillic acid-8,9-diol-glucuronide (IV) |
| ADD | Limonene cluster | 21<br>22<br>31<br>34<br>37<br>44<br>45<br>54<br>58 | 3.62 | 345.154<br>364.197<br>153.122<br>135.114<br>347.179<br>691.318<br>715.318<br>107.083<br>693.332 | 345.1552<br>364.1960<br>153.1269<br>135.1165<br>347.1695<br>691.3134<br>715.3141<br>107.0854<br>693.3319 | 0.7<br>3.0<br>6.9<br>6.8<br>3.1<br>5.9<br>1.7<br>6.1<br>0.3 | $[M-H]^-$<br>$[M+NH4]^+$<br>$[M+H-C6H10O7]^+$<br>$[M+H-C6H12O8]^+$<br>$[M+H]^+$<br>$[2M-H]^-$<br>$[2M+Na]^+$<br>$[M+H-C8H16O8]^+$<br>$[2M+H]^+$ | C16H26O8 | Neg mode: 327.143; 113.024; 75.009<br>Pos mode: 311.148; 153.128; 135.117<br>Table S5 | [b]Limonene-8,9-diol-glucuronide (II) |
| No pattern | Limonene cluster | 26 | 3.78 | 341.124 | 341.1228 | 2.6 | $[M-H]^-$ | C10H14O2 | 175.023; 165.090; 113.023 | Perillic acid glucuronide (I) |
| ADD | Limonene cluster | 42 | 3.83 | 343.135 | 343.1392 | 0.4 | $[M-H]^-$ | C16H24O8 | 175.025; 167.105; 113.023 | [b]Dihydroperillic acid glucuronide (IV) |
| ADD | Limonene cluster | 46 | 3.47 | 345.153 | 345.1542 | 2.0 | $[M-H]^-$ | C16H26O8 | 327.146; 113.024; 75.009 | Limonene-1,2-diol glucuronide (I) |
| ADD | Limonene cluster | 53 | 2.41 | 377.143 | 377.1442 | 1.5 | $[M-H]^-$ | C16H26O10 | 201.112; 175.024; 113.024 | [b]Unknown glucuronide (IV) |
| ADD | - | 18 | 3.88 | 319.1340 | NF | - | $[M-H]^-$ | C14H24O8 | 175.023; 143.107; 113.024 | Octanoyl-glucuronide (I) |
| ADD | - | 63 | 3.43 | 350.087 | 350.0874 | 0.5 | $[M-H]^-$ | C16H17NO8 | 193.033; 175.025; 113.024 | 3-indoleacetic acid glucuronide (I) |
| No pattern | - | 12 | 3.02 | 197.1258 | 197.12798 | 5.2 | $[M+H]^+$ | C10H16N2O2 | 169.135; 124.113; 72.080 | Cyclo(-Pro-Val) (I) |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| NND | - | 7 | 0.85 | 285.064 | 285.0608 | 1.0 | $[M-H]^-$ | C12H13O8 | 175.020; 113.023; 109.029[a] | Hydroquinone glucuronide (I) |
| NND | Fish | 13<br>29 | 0.48 | 76.0751<br>151.143 | BDL<br>151.1434 | <br>8.5 | $[M+H]^+$<br>$[2M+H]^+$ | C3H9NO | - | Trimethylamine *N*-oxide (TMAO) (I) |
| NND | - | 41 | 3.05 | 180.066 | 180.0647 | 7.6 | $[M+H]^+$ | C9H9NO3 | - | Hippuric acid (I) |
| NND | - | 32<br>43 | 3.39 | 146.060<br>192.074 | 146.0597<br>192.0651 | 6.1<br>5.1 | $[M+H-CO2H2]^+$<br>$[M+H]^+$ | C10H9NO3 | 174.055; 146.053; 128.050 | (2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid (I) |
| NND | NND cluster | 47 | 2.65 | 340.103 | 340.1021 | 3.4 | $[M+H]^+$ | C15H17NO8 | 164.069 | [b]Unknown glucuronide (IV) |
| NND | NND cluster | 50 | 2.10 | 242.011 | 242.0120 | 1.5 | Neg mode | C9H8NO5S | 162.055; 134.025; 79.957 | Unknown sulfate (IV) |
| NND | NND cluster | 56 | 2.41 | 336.072 | 336.0726 | 2.0 | $[M-H]^-$ | C15H15NO8 | 175.025; 160.039; 113.024 | [b]Unknown glucuronide (IV)‡ |
| NND | - | 55<br>60 | 4.46 | 109.066<br>184.098 | 109.0648<br>184.0964 | 5.3<br>5.4 | $[M+H-C2H5NO2]^+$<br>$[M+H]^+$ | C9H13NO3 | 109.063; 81.070[a] | [b]3,4,5,6-tetrahydrohippurate (II) |

Features not found (NF). Features below detection limit (BDL). [a]Fragments that were only intense for collision energies >10 eV

‡[b]Additional fragments to support the tentative identifications are given in the supporting information Table S4-S14.

*PCA on urine for individual dietary patterns*

When the NND samples were investigated in a PCA including the features in the PLS-DA model as variables, 12 of the 23 misclassified samples were clearly separated from the majority of the other NND samples along the first principal component (PC) (four samples), and PC 2 (eight samples) in the score plot. Features in the citrus and, in particular, the limonene cluster had highest loadings for PC 1, while the chocolate cluster had highest loadings in PC 2 (Figure 3). The misclassified ADD samples did not diverge from the rest of the ADD samples in PCA (data not shown).

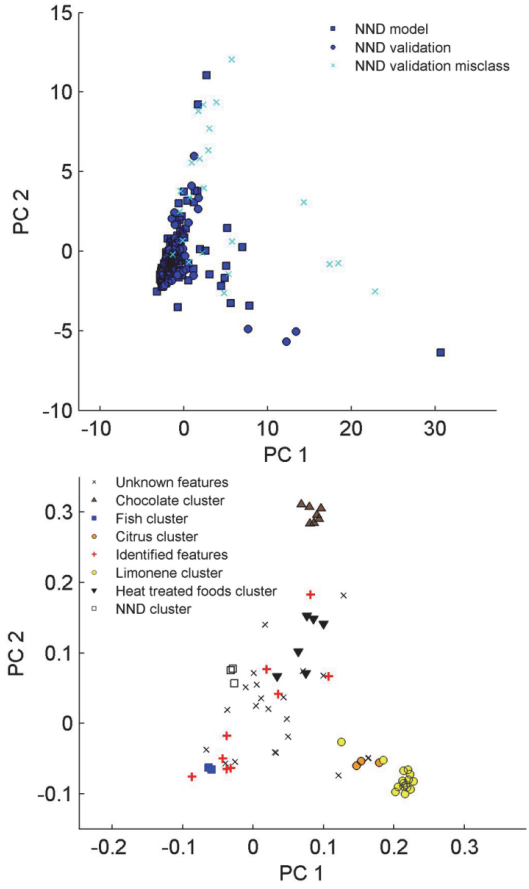



**Figure 3.** (A) Score plot and (B) loading plot for PC 1 and 2 in a PCA including all New Nordic Diet (NND) samples (model and validation) and the selected features from partial least squares discriminant analysis (PLS-DA).

*PCA on food intake*

All dietary registrations including both supermarket data and WDR were grouped into 250 food categories. This number was reduced to 119 for WDR and 237 for the supermarket data after removing food categories registered less than five times in each dataset. PCA models of data from WDR, including the validation set, and for the supermarket data are presented in Figure 4 A and B, respectively. For both PCA models, the two diets are clearly separated along PC 1 but with a much more pronounced separation for the supermarket data. For the 55 samples in the validation set where WDR were available (Figure 1), all WDR were classified to the correct dietary pattern regardless of the samples being misclassified in the PLS-DA model of urine samples or not. The WDR from misclassified samples in PLS-DA are highlighted in Figure 4A. There was no trend that these WDR were generally closer to the ADD pattern than the WDR from other correctly classified validation samples.

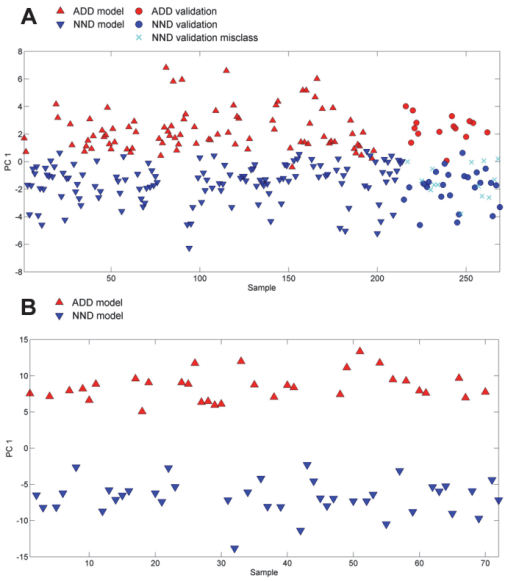



**Figure 4.** (A) Principal component (PC) 1 scores of data from the PCA model of foods reported in WDR. (B) PC 1 scores from the PCA model of reported foods in supermarket data.

The foods in the loading plots that mainly explain the separation between the diets in Figure 4 are listed in Figure 5. All the explanatory foods were consumed in larger amounts and/or by a higher percentage of the subjects in the dietary pattern they characterised. The percentage of consumers

and average amounts reported for the foods are given in the supporting information Table S17-18. When comparing the most explanatory foods between the models based on WDR and supermarket data, eleven foods were in common for the ADD diet while only six foods were in common for NND (Figure 5).

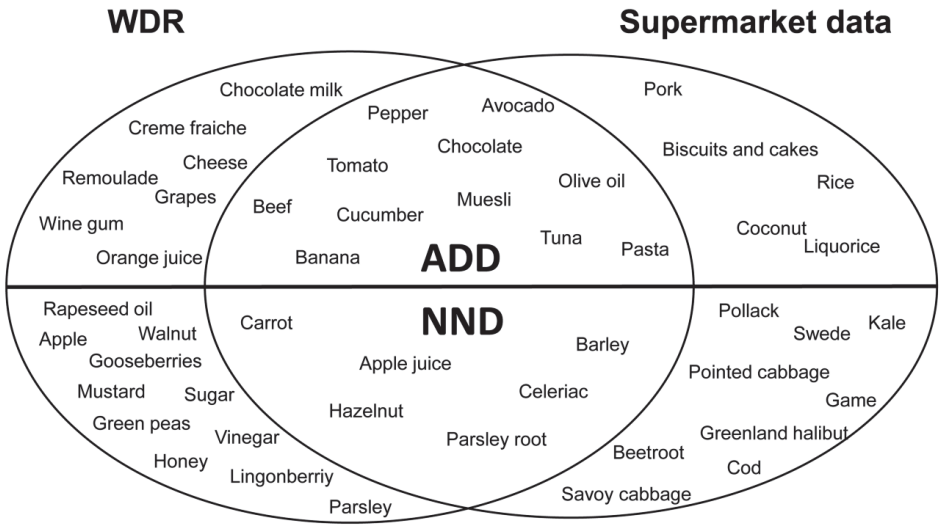



**Figure 5.** Most explanatory foods for the ADD and NND dietary patterns in PCA (Figure 4). Left circle: Data from WDR (the results from the model including all samples and the model including samples from subjects represented in the supermarket data only are merged). Right circle: Data from the supermarket. Foods listed in the overlapping areas of the circles were explanatory foods in both WDR and supermarket data.

## Discussion

*Discrimination of ADD and NND dietary patterns in urine*

The separation between the diets in the final PLS-DA model demonstrates that the two dietary patterns are clearly reflected in the urinary metabolome (Figure 2A). The misclassification rate for the validation samples of 19 % is low, especially when it is considered that it is not possible with the present study design to distinguish between true lack of compliance, which would be a correct misclassification, and limitations of the model to capture individual variations within a dietary pattern. It is an advantage for the evaluation of model performance that a high proportion of the

total number of samples was used as validation samples and that these samples represent new subjects and time points. However, inclusion of repeated measures in the validation set from the same subjects whose samples were used to develop the model has also affected the misclassification rate. The proportion of samples from subjects whose samples had been included as model samples was 10 % lower among the misclassified samples compared to the validation set as a whole, indicating that the model classifies samples from subjects whose samples were included for model development slightly better. The misclassification rate found is therefore probably underestimated compared to what would be expected for an independent study population.In another study on urine fingerprints of habitual diets, the misclassification rate was 32 % for two dietary patterns with contrasting intakes of nine food groups[11]. The higher misclassification rate in that study was probably caused by less well defined diets, use of spot urine samples and lack of variable selection in the PLS-DA model. In a targeted study applying a similar approach to investigate compliance to self-reported analgesic use, the rate of underreporting was estimated to be 15-17 % in a cohort of 496 participants from two Western populations[42]. It is therefore not unrealistic to have a non-compliance rate around 19 % even though compliance to a dietary pattern is of course not necessarily comparable to underreporting of analgesic use.

From the misclassification rates, NND is much more difficult to classify from the urine fingerprints. The fact that the misclassification error for NND is almost eight times higher than for ADD is not surprising. It is clear from the loading plot (Figure 2B) that a higher number of features characterize ADD. Also, the features related to ADD have higher ranks in the model. Since the available WDR from subjects with misclassified NND samples were all within the correct dietary pattern (Figure 4A), the metabolites characterising NND in the PLS-DA model, are not completely representative of the NND related foods in the PCA. However, the metabolites characterising the misclassified NND subjects in the PCA model of the NND urine samples (Figure 3) also suggest that these subjects may not have been fully compliant to NND. Citrus, soft drinks and wine gums, the main food sources related to limonene metabolites, as well as chocolate were not part of NND and should be limited as much as possible in this diet. When inspecting WDR separately, for subjects with misclassified NND samples, only one of the seven NND subjects with high loadings for chocolate related markers in PCA and WDR available had reported intake of chocolate on the day of urine sampling. For the four misclassified NND samples with high loadings for citrus and limonene, none of the three subjects for which WDR were available, reported intake of soft drinks, wine gums or citrus fruits on the day of urine sampling. This observation indicates that the dietary reporting from the misclassified NND subjects is not completely reliable, which would be expected since the NND diet is more demanding to follow than the ADD. However, theobromine and limonene metabolites are not fully excreted within 24 h[43, 44] and the subjects could therefore have consumed large amounts of chocolate or limonene containing products on the day before the urine collection and still have considerable levels of these metabolites in urine the following day. In addition, chocolate metabolites can be produced from caffeine and consumption of other products that do not contain cocoa, such as coffee and tea[45], may therefore also contribute to the high levels of theobromine metabolites found in urine of some subjects. Since the model is based on patterns of urinary metabolites, the misclassified validation samples cannot be explained fully from individual

metabolites. There are also examples of NND subjects reporting soft drink and chocolate consumption without having their samples misclassified in the PLS-DA model. More frequent sampling could be a solution to better distinguish the true non-compliant NND subjects and understand why they are misclassified. It is more likely that the subjects who had all their samples misclassified have been non-compliant than subjects with both correctly classified and misclassified samples. However, they may also have individual metabotypes that were not integrated into the model because these individuals had no samples included in the model samples.

*Metabolites in the PLS-DA model characterising the ADD diet*

The identified features characterising the ADD diet were mainly from four clusters of correlating features (heat treated foods, chocolate, citrus and limonene, Figure 2B). In addition, four metabolites that did not have high sensitivity or specificity for any individual foods were identified: 3-indoleacetyl-glucuronide, octanoyl-glucuronide, pGlu-Pro and cyclo-Pro-Val. The first three of these metabolites were generally found in higher levels for ADD, while cyclo-Pro-Val did not have a clear trend for any diet and therefore probably is part of a multivariate pattern together with other ADD features. The finding of two modified dipeptides as markers of ADD may reflect the protein sources in the diet which were primarily of animal origin (Table 1). The amino acids proline, valine and glutamic acid are all found in high levels in animal protein compared to proteins from fruits and vegetables[46] and a high concentration of pyroglutamyl (bound and unbound) have been found in cheese[47]. Cyclic dipeptides are typically formed during heat treatment and cyclo-Pro-Val have been identified in various food products such as coffee and cacao[48, 49]. Neither cyclo-Pro-Val nor pGlu-Pro has to our knowledge been found as dietary markers in urine before. 3-indoleacetyl-glucuronide is a microbial tryptophan metabolite which has previously been reported to be present in urine of healthy subjects[50]. The tryptophan content is high in animal protein[46] and the finding of 3-indoleacetyl-glucuronide may therefore also reflect differences in the protein sources between the diets. Octanoyl-glucuronide has previously been reported in urine from children on a diet high in medium chain fatty acid triglycerides[51]. Main dietary sources of medium chain fatty acids are dairy products, coconut and palm kernel oil, which is accordant with coconut and cheese being among the common explanatory foods for ADD in Figure 5.

*The heat treated foods cluster.* Five features from four different metabolites, of which one was identified as pyrraline, were moderately correlated (r = 0.6-0.75) and had high sensitivity and specificity for a diverse group of heat treated foods, mainly digestive biscuits, cornflakes and roast beef. Pyrraline belongs to the group of advanced glycation end-products (AGEs) which are formed as end products following heat treatment in a series of reactions between amino acid moieties and reducing sugars, commonly known as Maillard reactions or non-enzymatic browning[52]. It has been demonstrated in two studies that pyrraline excretion in urine decreases when subjects are put on a pyrraline restricted diet[53, 54]. In the same studies, it was concluded that urinary pyrraline is mainly of dietary origin and that the excretion of dietary pyrraline is almost complete. Since the ADD diet contains more heat-treated processed foods, this may explain why pyrraline is a marker of ADD. Two of the other compounds in the cluster are probably isomers of a pyrraline derivative, as they

shared several m/z fragments with pyrraline in both negative and positive ionization modes (Table S9-10). Another compound from the cluster was a glucuronide conjugated product. This compound may be generated from deoxyglucosulose, an intermediate in the Maillard reaction, by dehydration. Three candidate structures formed from 1-deoxyglucosulose have been identified in model systems that would match the molecular formula: 2,4-dihydroxy-2,5-dimethylfuran-3(2H )-one, 4-hydroxy-2-(hydroxymethyl)-5-methylfuran-3(2H )-one and 3,5-dihydroxy-6-methyl-2,3-dihydro-4H-pyran-4-one[55, 56]. Unfortunately, standards were not available to further elucidate if any of these were correct. The last compound in the cluster could not be identified.

*The chocolate cluster.* Chocolate was only allowed in ADD and in line with this, six correlating features (r>0.7) with high sensitivity and specificity for chocolate in WDR were markers of the ADD diet. The metabolites identified in the chocolate cluster: theobromine, 7-methyluric acid, 3,7-dimethyluric acid, 7-methylxanthine and 6-amino-5-[*N*-methylformylamino]-1- methyluracil, have all been reported in a previous metabolomics study on cocoa powder[4].

*The citrus and limonene clusters.* The limonene cluster consisted of seventeen correlated features (r>0.7) from seven unique metabolites. When carrying out the sensitivity and specificity analysis for foods, the cluster was related to intake of orange juice, wine gums and soft drinks in WDR. Accordant with this, limonene is naturally present in citrus oils and is also used widely in the food industry as an additive to various foods such as sweets and beverages[57]. The three features in the citrus cluster only had high sensitivity and specificity for citrus. They were highly correlated (r ~ 0.9) and have all been found as markers of habitual citrus intake previously[2].

Intake of citrus was only allowed in ADD and the consumption of sweets and soft drinks was higher in ADD. It therefore makes sense that both limonene and citrus metabolites characterize the ADD diet. One of the citrus metabolites was identified with a standard as proline betaine. The other two citrus metabolites may be metabolites of proline betaine due to the strong correlation. It has been demonstrated that they, as opposed to proline betaine, are not present in orange juice[33]. The likely identity of the compound with m/z 146.083 as *N*-methyl-cis-4-hydroxy-*L*-proline, was ruled out by analysing the standard.

The metabolism of limonene has been investigated in humans in several studies and the limonene metabolites identified at level I and II in the present study (glucuronides of limonene-1,2-diol, limonene-8,9-diol and perillic acid) are well-known[44, 58, 59]. The rest of the limonene metabolites were tentatively identified as glucuronides of perillic acid-8,9-diol, dihydroperillic acid and p-menth-1-ene-6,8,9-triol which have all been reported previously[44, 58, 59]. In general, the fragmentation in negative mode, where the limonene metabolites were most intense, revealed no characteristic structural features, except from the glucuronide moiety. In positive mode, only perillic acid-8,9-glucuronide and limonene-8,9-diol-glucuronide were intense enough for MS/MS fragmentation and the obtained fragments supported the identification for these compounds. A complete list of MS/MS fragments for the tentatively identified limonene metabolites is given in the Table S4-S8. Limonene-8,9-diol has been found as a marker of citrus consumption in a previous metabolomics study[2]. However, the clustering of the limonene metabolites away from the citrus

metabolites in Figure 2B, suggests that use of limonene metabolites as markers of citrus consumption are biased by other limonene containing foods, possibly citrus-flavoured sweets and soft drinks.

*Metabolites in the PLS-DA model characterising the NND diet*

Markers identified at level I-III of the NND diet were: Trimethylamine *N*-oxide, hippuric acid, hydroquinone-glucuronide, (2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid 3,4,5,6-tetrahydrohippurate (Table S12). Trimethylamine *N*-oxide is present in fish and has been demonstrated to be a urinary marker of fish intake in several studies[32, 60]. In the present study, trimethylamine *N*-oxide reflects the high fish intake in NND compared to ADD (Table S1). Several fish species were among the foods characteristic for NND (Figure 5) and the sensitivity and specificity of this NND marker for fish intake reported in WDR was high. Hippuric acid is a well-known microbial metabolite of various polyphenols in the diet[19, 61]. The finding of hippuric acid as a marker of NND therefore reflects a diet high in plant foods rather than intake of any individual foods. Another microbial metabolite, 3,4,5,6-tetrahydrohippurate, has been found in rats following intraperitoneal injection of shikimate[62]. Shikimate is ubiquitously present in plants[63] and may also be a general marker of fruit and vegetable intake even though it, to our knowledge, has not been reported in human studies before. Arbutin, a hydroquinone glycoside, is present in various foods, particularly wheat and pear[64]. A previous study has demonstrated that urinary excretion of hydroquinone-glucuronide increases following a meal high in arbutin[64]. In the present study, intake of whole wheat was higher in NND. However, since the arbutin content of foods is not well studied, there may be other dietary sources as well, explaining why hydroquinone-glucuronide is a marker of NND. The last identified NND marker, (2-oxo-2,3-dihydro-1H-indol-3-yl)acetic acid was not related to intake of any particular foods and to our knowledge this is the first time the compound is reported as a urinary metabolite. According to the CRC dictionary of food compounds, it has been isolated from redcurrants, sunflower and Brassica spp. which would be in line with the NND. Three other correlated unidentified NND markers (the NND cluster in Figure 2B) were not related to intake of any individual food. Two of the compounds in this cluster may be a sulphate and a glucuronide conjugate of the same compound since the molecular formula was the same for the unconjugated form (m/z 242.011 and m/z 340.103 in Table 2). Indole-3-carboxylic acid was analysed as a possible match for the marker with m/z 336.072 but was not correct.

*The dietary fingerprint in urine*

Based on the identified urinary metabolites, it seems that the ADD diet is better reflected in urine. A higher number of metabolites from ADD are found and the ADD metabolites are both specific to individual foods, that are only allowed in the ADD diet, and representing more general features of the diet such as higher intakes of animal protein and heat-treated foods. For the NND diet, the metabolites found are mainly reflecting a high intake of fish, fruit and vegetables. The reason why fewer markers are found for NND is probably that NND is a seasonal diet[65]. This also explains why much fewer foods are in common for NND in PCA when comparing WDR and supermarket data in

Figure 5. The subjects initiated the study during autumn and winter and the sampling points were mainly placed in winter and spring. Since the sampling points included in the model therefore represent two seasons for the NND, the chance of finding markers of seasonal foods is lower than for foods available all year round. It might have been advantageous to subdivide the NND samples according to season to investigate if markers of more NND foods would then have been found. For example, a separate analysison short-term exposure markers in the same dataset revealed markers of cabbage and beetroot (autumn and winter foods in NND) that are not present in the PLS-DA model[33]. Low meat or vegetarian diets in comparison to diets high in meat have been investigated in several other metabolomics studies. For these studies, some features are commonly found as markers such as hippuric acid, creatine and trimethylamine *N*-oxide[11, 16, 17, 18, 19, 32], while other metabolites seem to be more specific markers for certain diets. This suggests that it is possible to identify urinary metabolites of very general dietary traits as well as metabolites that are specific for more strictly defined dietary patterns like ADD and NND. It remains to be investigated how specific a compliance measure is, like the one developed in the present study. It would be interesting to see how a selection of samples from outside the study population would have been situated in the PLS-DA model and if a removal of samples from suspected non-compliant subjects in the dietary intervention can actually strengthen the study outcome. For the present study, more frequent sampling would have been required to obtain clearer indications of which NND subjects have been non-compliant or have an unusual metabolic phenotype. Even though some of the subjects with misclassified NND samples responded less to the dietary intervention, it could not be justified to remove any subjects based solely on the observations of compliance in PLS-DA.

There is no doubt that untargeted metabolomics can be used to strengthen the available range of compliance measures. At the same time, however, identification remains a major obstacle in untargeted metabolomics[66]. For the present PLS-DA model, only a few features seemed to be chance findings but due to the difficulties in identification, many potentially interesting features remain unknown and the impact of those cannot be assessed. The compliance model is therefore not fully transparent. Despite this, the low misclassification rate of the model and the clear association between the identified metabolites and the dietary patterns demonstrate the high potential in applying multivariate measures to estimate compliance in nutrition studies. The findings need to be followed up with a targeted analysis of the metabolites in the model. The accuracy of the measurements in the untargeted analysis is low and an even stronger compliance measure, possibly including fewer metabolites, could probably be obtained if the metabolites in the model were quantitated and the model further developed. It is also possible to apply a quantitative model in a broader context for validation, whereas this cannot be done easily in untargeted metabolomics due to the dependency on preprocessing and pretreatment in a new set of samples for detection and quantification of features.

*Conclusion*

This study is, to our knowledge, the first to explore the potential of using a metabolomics approach to estimate compliance to a dietary pattern. From 4369 features detected in urine, a PLS-DA model was developed and optimised for which a misclassification rate for two dietary patterns

in a validation set with 139 samples was only 19 % based on 67 selected features in urine. The metabolites in the PLS-DA model cover both general dietary traits such as animal compared to vegetable protein and more specific traits of a dietary pattern such as intake of fish, citrus fruits or cocoa containing products. The study demonstrates that untargeted metabolomics can be used to discover which metabolites are the strongest predictors of compliance to complex diets. Discovery of such metabolites should be followed up by quantitative measurements to further optimise and validate the model. Eventually, development of multivariate compliance measures may lead to a better understanding of the outcomes of dietary intervention studies.

## ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

File name: SupportingInformation

## AUTHOR INFORMATION

Corresponding Author:

Maj-Britt Schmidt Andersen, Department of Nutrition Exercise and Sports, Faculty of Science, University of Copenhagen. E-mail: mbsa@life.ku.dk. Phone: +45 35331086. Fax: +45 3533 2483

## ACKNOWLEDGEMENTS

The study was conducted as part of the OPUS project which is supported by a grant from the Nordea Foundation, Denmark. OPUS is an acronym of the Danish title of the project 'Optimal well-being, development and health for Danish children through a healthy New Nordic Diet'. The authors would like thank Majbritt Hybholt, Mette Kristensen, Daniela Rago, Ümmühan Celik and Bernard Lyan for their contribution to the data collection and laboratory work.

## References


(1) Puiggròs F.; Solà R.; Bladé C.; Salvadó M.; Arola L. Nutritional biomarkers and foodomic methodologies for qualitative and quantitative analysis of bioactive ingredients in dietary intervention studies. *J. Chromatogr. A.* **2011**, 1218 (42), 7399-7414.

(2) Pujos-Guillot, E.; Hubert, J.; Martin, J. F.; Lyan, B.; Quintana, M.; Claude, S.; Chabanas, B.; Rothwell, J. A.; Bennetau-Pelissero, C.; Scalbert, A.; Comte, B.; Hercberg, S.; Morand, C.; Gelan, P.; Manach, C. Mass Spectrometry-based Metabolomics for the Discovery of Biomarkers of Fruit and Vegetable Intake: Citrus Fruit as a Case Study. *J. Proteome Res.* **2013**, 12(4), 1645-1659.

(3) Lang, R.; Wahl, A.; Stark, T.; Hofmann, T. Urinary N-methylpyridinium and trigonelline as candidate dietary biomarkers of coffee consumption. *Mol. Nutr. Food Res*. **2011**, 55(11), 1613-1623.

(4) Llorach, R.; Urpi-Sarda, M.; Jauregui, O.; Monagas, M.; Andres-Lacueva, C. An LC-MS- based metabolomics approach for exploring urinary metabolome modifications after cocoa consumption. *J. Proteome Res.* **2009**, 8(11), 5060-5068.

(5) Edmands, W. M.; Beckonert, O. P.; Stella, C.; Campbell, A.; Lake, B. G.; Lindon J. C.; Holmes, E.; Gooderham, N. J. Identification of human urinary biomarkers of cruciferous vegetable consumption by metabonomic profiling. *J. Proteome Res.* **2011**, 10(10), 4513-4521.

(6) Tulipani, S.; Llorach, R.; Jáuregui, O.; López-Uriarte, P.; Garcia-Aloy, M.; Bullo, M.; Salas-Salvadó, J.; Andrés-Lacueva, C. Metabolomics unveils urinary changes in subjects with metabolic syndrome following 12-week nut consumption. *J. Proteome Res.* **2011**, 10(11), 5047-5058.

(7) Walsh, M. C.; Brennan, L.; Malthouse, J. P.; Roche, H. M.; Gibney, M. J. Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *Am. J. Clin. Nutr.* **2006**, 84(3), 531-539.

(8) Rasmussen, L. G.; Savorani, F.; Larsen, T. M.; Dragsted, L. O.; Astrup, A.; Engelsen, S. B. Standardization of factors that influence human urine metabolomics. *Metabolomics* **2011**, 7(1), 71-83.

(9) Lloyd, A. J.; Beckmann, M.; Favé, G.; Mathers, J. C.; Draper, J. Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. *Br. J. Nutr.* **2011**, 106(6), 812-824.

(10) Slimani, N.; Fahey, M.; Welch, A. A.; Wirfält, E.; Stripp, C.; Bergström, E.; Linseisen, J.; Schulze, M. B.; Bamia, C.; Chloptsios, Y.; Veglia, F.; Panico, S.; Bueno-de-Mesquita, H. B.; Ocké, M. C.; Brustad, M.; Lund, E.; González, C. A.; Barcos, A.; Berglund, G.; Winkvist, A.; Mulligan, A.; Appleby, P.; Overvad, K.; Tjønneland, A.; Clavel-Chapelon, F.; Kesse, E.; Ferrari, P.; Van Staveren, W. A.; Riboli, E. Diversity of dietary patterns observed in the European Prospective Investigation into Cancer and Nutrition (EPIC) project. *Public Health Nutr.* **2002**, 5(6B), 1311-1328.

(11) O'Sullivan, A.; Gibney, M. J.; Brennan, L. Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. *Am. J. Clin. Nutr.*, **2011**, 93(2), 314-321.

(12) Jacques, P. F.; Tucker, K. L. Are dietary patterns useful for understanding the role of diet in chronic disease? *Am. J. Clin. Nutr.* **2001**, 73(1), 1-2.

(13) Bingham, SA. Biomarkers in nutritional epidemiology. *Public health Nutr.* **2002**, 5(6A), 821-827.

(14) Koulman, A.; Volmer, D. A. Perspectives for metabolomics in human nutrition: an overview. *Nutr. Bull*. **2008**, 33(4), 324-330.

(15) Heinzmann, S. S.; Brown, I. J.; Chan, Q; Bictash, M.; Dumas, M.; Kochhar, S.; Stamler, J.; Holmes, E.; Elliott, P.; Nicholson, J. K. Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. Am. J. Clin. Nutr. **2010**, 92, 436-443.

(16) Lloyd, A. J.; Beckmann, M.; Haldar, S.; Seal, C.; Brandt, K.; Draper, J. Data-driven strategy for the discovery of potential urinary biomarkers of habitual dietary exposure. Am. J. Clin. Nutr. **2013**, 97, 377-389.

(17) Xu, J.; Yang, S.; Cai, S.; Dong, J.; Li, X.; Chen, Z. Identification of biochemical changes in lactovegetarian urine using 1 H NMR spectroscopy and pattern recognition. *Anal. Bioanal. Chem.* **2010**, 396(4), 1451-1463.

(18) Stella, C.; Beckwith-hall, B.; Cloarec, O.; Holmes, E.; Lindon, J. C.; Powell, J.; van der Ouderaa, F.; Bingham, S.; Cross, A. J.; Nicholson, J. K. Susceptibility of human metabolic phenotypes to dietary modulation. *J. proteome Res.* **2006**, 5(10), 2780-2788.

(19) May, D. H.; Navarro, S. L.; Ruczinski, I.; Hogan, J.; Ogata, Y.; Schwarz, Y.; Levy, L.; Holzman, T.; McIntosh, M. W.; Lampe, J. W. Metabolomic profiling of urine: response to a randomised, controlled feeding study of select fruits and vegetables, and application to an observational study. *Br. J. Nutr.* **2013**, May 9, 1-11, DOI:10.1017/S000711451300127X

(20) Rasmussen, L. G.; Winning, H.; Savorani, F.; Toft, H.; Larsen, T. M.; Dragsted, L. O.; Astrup, A.; Engelsen, S. B. Assessment of the effect of high or low protein diet on the human urine metabolome as measured by NMR. *Nutrients* **2012**, 4(2), 112-131.

(21) Rasmussen, L. G.; Winning, H.; Savorani, F.; Ritz, C.; Engelsen, S. B.; Astrup, A.; Larsen, T. M.; Dragsted, L. O. Assessment of dietary exposure related to dietary GI and fibre intake in a nutritional metabolomic study of human urine. *Genes Nutr.* **2012**, 7(2), 281-293.

(22) Menni, C.; Zhai, G.; MacGregor, A.; Prehn, C.; Römisch-Margl, W.; Suhre, K.; Adamski, J.; Cassidy, A.; Illig, T.; Spector, T. D.; Valdes, A. M. Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. Metabolomics **2013**, 9, 506-514.

(23) Floegel, A.; von Ruesten, A.; Drogan, D.; Schulze, M. B.; Prehn, C.; Adamski, J.; Pischon, T.; Boeing, H. Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam. Eur. J. Clin. Nutr. **2013**, 67, 1100-1108.

(24) Peré-Trepat, E.; Ross, A. B.; Martin, F.; Rezzi, S.; Kochhar, S.; Hasselbalch, A. L.; Kyvik, K. O.; Sørensen, T. I. A. Chemometric strategies to assess metabonomic imprinting of food habits in epidemiological studies. *Chemometrics Intellig. Lab. Syst.* **2010**, 104(1), 95-100.

(25) Altmaier, E.; Kastenmüller, G.; Römisch-Margl, W.; Thorand, B.; Weinberger, K. M.; Illig, T.; Adamski, J.; Döring, A.; Suhre, K. Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics. *Eur. J. Epidemiol.* **2011**, 26(2), 145-156.

(26) Mithril, C.; Dragsted, L. O.; Meyer C.; Tetens, I.; Biltoft-Jensen, A.; Astrup, A. Dietary composition and nutrient content of the New Nordic Diet. *Public Health Nutr.* **2013**, 16(5), 777-785.

(27) Poulsen, S. K.; Due, A.; Jordy, A. B.; Kiens, B; Stark, K. D ; Stender, S.; Holst, C.; Astrup, A.; Larsen, T. M. Health effect of the New Nordic Diet in adults with increased waist circumference: a 6-mo randomized controlled trial . Am. J. Clin Nutr. **2014**, 99, 35-45

(28) Barri, T.; Holmer-Jensen, J.; Hermansen, K.; Dragsted, L. O. 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* **2012**, 718, 47-57.

(29) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, 11, 395-405.

(30) Chong, I.; Jun, C. Performance of some variable selection methods when multicollinearity is present. *Chemometrics Intellig. Lab. Syst.* **2005**, 78, 103-112.

(31) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intellig. Lab. Syst.* **2001**, 58(2), 109-130.

(32) Andersen, M. S.; Reinbach, H. C.; Rinnan, Å.; Barri, T.; Mithril, C.; Dragsted, L. O. Discovery of exposure markers in urine for Brassica-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics. *Metabolomics* **2013**, 9(5), 984-997.

(33) Andersen, M. S.; Kristensen, M.; Manach, C.; Pujos-Guillot, E.; Poulsen, S. K.; Larsen, T. M.; Astrup, A.; Dragsted, L. O. Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics. *Anal. Bioanal. Chem*. **2014**, January, 1-16, DOI: 10.1007/s00216-013-7498-5.

(34) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychoqios, N., Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; 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, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, J. J.; Forsythe, I. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. **2009**, 37, D603-D610.

(35) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abaquan, R.; Siuzdak, G. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **2005**, 27(6), 747-751.

(36) Gerlich, M.; Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **2013**, 48(3), 291-298.

(37) Dictionary of food compounds with CD-ROM, Second Edition, Yannai, S, Ed. CRC Press, **2012**, Chapter 4, pp. 435

(38) Neveu, V.; Perez-Jiménez, J.; Vos, F.; Crespy, V.; du Chaffaut, L.; Mennen, L.; Knox, C.; Eisner, R.; Cruz, J.; Wishart, D.; Scalbert, A. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database (Oxford)* **2010**, bap024 (1-9).

(39) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K.; Saito, K.; Kanaya, S. KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research. *Plant Cell Physiol.* **2012**, 53(2), e1 (1-12).

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

(41) Stanstrup, J.; Rasmussen, J. E.; Ritz, C.; Holmer-Jensen, J.; Hermansen, K.; Dragsted, L. O. Intakes of whey protein hydrolysate and whole whey proteins are discriminated by LC–MS metabolomics. Metabolomics **2013**. November 17, 1-18, DOI: 10.1007/s11306-013-0607-9

(42) Loo, R. L.; Chan, Q.; Brown, I. J.; Robertson, C. E.; Stamler, J.; Nicholson, J. K.; Holmes, E.; Elliott, P. A comparison of self-reported analgesic use and detection of urinary ibuprofen and acetaminophen metabolites by means of metabonomics: the INTERMAP Study. *Am. J. Epidemiol.* **2012**, 175(4), 348-358.

(43) Rodopoulos, N.; Höjvall, L.; Norman, A. Elimination of theobromine metabolites in healthy adults. *Scand. J. Clin. Lab. Invest.* **1996**, 56(4), 373-383.

(44) Kodama, R.; Yano, T.; Furukawa, K.; Noda, K.; Ide, H. Studies on the Metabolism of d -Limonene (p-mentha-1, 8-diene). IV Isolation and Characterization of New Metabolites and Species Differences in Metabolism. *Xenobiotica* **1976**, 6(6), 377-389.

(45) Heckman, M. A.; Weil, J.; Gonzalez, dM. E. Caffeine (1,3,7-trimethylxanthine) in foods: a comprehensive review on consumption, functionality, safety, and regulatory matters. *J. Food Sci.* **2010**, 75(3), R77-R87.

(46) Sosulski, F. W.; Imafidon, G. I. Amino acid composition and nitrogen-to-protein conversion factors for animal and plant foods. *J. Agric. Food Chem.* **1990**, 38(6), 1351-1356.

(47) Mucchetti, G.; Locci, F.; Gatti, M.; Neviani, E.; Addeo, F.; Dossena, A.; Marchelli, R. Pyroglutamic Acid in Cheese: Presence, Origin, and Correlation with Ripening Time of Grana Padano *Cheese. J. Dairy Sci.* **2000**, 83(4), 659-665.

(48) Stark, T.; Hofmann, T. Structures, Sensory Activity, and Dose/Response Functions of 2,5-Diketopiperazines in Roasted Cocoa Nibs (Theobroma cacao). *J. Agric. Food Chem.* **2005**, 53(18), 7222-7231.

(49) Ginz, M.; Engelhardt, U. H. Identification of Proline-Based Diketopiperazines in Roasted Coffee. *J. Agric. Food Chem.* **2000**, 48(8), 3528-3532.

(50) Sprince, H.; Parker, C.; Dawson, J. T. Jr., Jameson, D.; Dohan, F. C. Paper chromatography of urinary indoles extracted under alkaline and acid conditions. *J. Chromatogr.* **1962**, 8, 457-464.

(51) Kuhara, T.; Matsumoto, I.; Ohno, M.; Ohura, T. Identification and quantification of octanoyl glucuronide in the urine of children who ingested medium-chain triglycerides. *Biol. Mass Spectrom.* **1986**, 13, 595-598.

(52) Henle, T. Protein-bound advanced glycation endproducts (AGEs) as bioactive amino acid derivatives in foods. *Amino Acids* **2005**, 29(4), 313-322.

(53) Foerster, A.; Henle, T. Glycation in food and metabolic transit of dietary AGEs (advanced glycation end-products): studies on the urinary excretion of pyrraline. *Biochem. Soc. Trans.* **2003**, 31(6), 1383-1385.

(54) Förster, A.; Kühne, Y.; Henle, T. O. Studies on Absorption and Elimination of Dietary Maillard Reaction Products. *Ann. N Y Acad. Sci.* **2005**, 1043, 474-481.

(55) Voigt, M.; Glomb, M. A. Reactivity of 1-Deoxy-d-erythro-hexo-2,3-diulose: A Key Intermediate in the Maillard Chemistry of Hexoses. *J. Agric. Food Chem.* **2009**, 57(11), 4765-4770.

(56) Ames, J. M.; Bailey, R. G.; Mann, J. Analysis of furanone, pyranone, and new heterocyclic colored compounds from sugar-glycine model Maillard systems. *J. Agric. Food Chem.* **1999**, 47(2), 438-443.

(57) Sun J. D-Limonene: safety and clinical applications. *Altern. Med. Rev.* **2007**, 12(3), 259-264.

(58) Poon, G. K.; Vigushin, D.; Griggs, L. J.; Rowlands, M. G.; Coombes, R. C.; Jarman, M. Identification and characterization of limonene metabolites in patients with advanced cancer by liquid chromatography/mass spectrometry. *Drug metab. Dispos.* **1996**, 24(5), 565-571.

(59) Vigushin, D. M.; Poon, G. K.; Boddy, A.; English, J.; Halbert, G. W.; Pagonis, C.; Jarman, M.; Coombes, R. C. Phase I and pharmacokinetic study of d-limonene in patients with advanced cancer. *Cancer Chemother. Pharmacol.* **1998**, 42(2), 111-117.

(60) Lloyd, A. J.; Favé, G.; Beckmann, M.; Lin, W.; Tailliart, K.; Xie, L.; Mathers, J. C.; Draper, J. Use of mass spectrometry fingerprinting to identify urinary metabolites after consumption of specific foods. *Am. J. Clin. Nutr.* **2011**, 94(4), 981-991.

(61) van Dorsten, F. A.; Grün, C. H.; van Velzen, E. J.; Jacobs, D. M.; Draijer, R.; van Duynhoven, J. P. The metabolic fate of red wine and grape juice polyphenols in humans assessed by metabolomics. *Mol. Nutr. Food Res.* **2010** 54(7), 897-908.

(62) Brewster, D.; Jones, R. S.; Parke, D. V. The metabolism of shikimate in the rat. *Biochem. J.* **1978**, 170, 257-264.

(63) Bochkov, D. V.; Sysolyatin, S. V.; Kalashnikov, A. I.; Surmacheva, I. A. Shikimic acid: review of its analytical, isolation, and purification techniques from plant and microbial sources. *J. Chem. Biol.* **2012**, 5(1), 5-17.

(64) Deisinger, P. J.; Hill, T. S.; English, J. C. Human exposure to naturally occurring hydroquinone. *J. Toxicol. Environ. Health* **1996** 47(1), 31-46.

(65) Mithril, C.; Dragsted, L. O.; Meyer, C.; Blauert, E.; Holt, M. K.; Astrup, A. Guidelines for the New Nordic Diet. *Public Health Nutr.* **2012**, 15(10), 1941-1947.

(66) Scalbert, A.; Brennan, L.; Fiehn, O.; Hankemeier, T.; Kristal, B. S.; van Ommen, B.; Pujos-Guillot, E.; Verheij, E.; Wishart, D.; Wopereis, S. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics* **2009**, 5(4), 435-458.

# Untargeted metabolomics as a screening tool for estimating compliance to a dietary pattern

## Supporting Information

*Maj-Britt S. Andersen,[†] Åsmund Rinnan,[‡] Claudine Manach,[§] Sanne K. Poulsen,[§] Estelle Pujos-Guillot,[†] Thomas M. Larsen,[†] Arne Astrup ,[†] Lars O. Dragsted[†]*

[†]Department of Nutrition Exercise and Sports, Faculty of Science, University of Copenhagen.
[‡]Department of Food Science, Faculty of Science, University of Copenhagen.
[§]INRA, UMR1019, Human Nutrition Unit, University of Auvergne, Clermont-Ferrand-Theix, France

**Table S1.** Baseline characteristics for subjects

| | NND | | ADD | |
|---|---|---|---|---|
| | **Model** | **Validation** | **Model** | **Validation** |
| No. of subjects | 64 | 55 | 43 | 39 |
| Age | 44.0 (13.3) | 43.7 (12.8) | 41.2 (13.2) | 40.2 (14.1) |
| Sex | 16 M, 48 F | 21 M, 34 F | 12 M, 31 F | 11 M, 29 F |
| BMI [$kg/m^2$] | 29.9 (4.5) | 31.1 (5.2) | 29.4 (4.6) | 31.0 (6.0) |
| Waist circumference [cm] | 99.2 (11.6) | 101.8 (13.1) | 98.9 (13.3) | 101.5 (14.5) |
| No. with household size of one | 40 | 33 (30 completers) | 32 | 27 (26 completers) |

NND: New Nordic Diet; ADD: Average Danish Diet. Model and Validation are groups of subjects whose samples were used for the model and validation set, respectively, in the compliance model.

**Table S2.** Batch steps and parameters used for preprocessing of raw data in MZmine2.7.

| | Batch step | Parameters |
|---|---|---|
| **Negative mode** | Raw data import | |
| | Mass detection | Noise level: 15 |
| | Chromatogram builder | Min time span (min): 0.01;<br>Min height: 4.0E1; m/z tolerance: 0.055 mz or 30 ppm |
| | Chromatogram deconvolution | Chromatographic threshold: 95%; Search minimum in RT range (min): 0.01; Minimum relative height: 10%; Minimum absolute height: 4.0E1; Min ratio of peak/top edge: 1.3; Peak duration range (min): 0.01-0.2 |
| | Isotopic pattern | m/z tolerance: 0.06 or 30 ppm; Retention time tolerance: 0.01; Monotonic shape; maximum charge: 1 |
| | Join aligner | m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.15; Weight for both m/z tolerance and retention time tolerance: 10 |
| | Duplicate peak filter | m/z tolerance: 0.5 or 600 ppm; RT tolerance: 0.15 |
| | Peak list rows filter | Min peaks in a row: 5<br>Minimum peaks in an isotope pattern: 1; m/z range: 50-1000; RT range: 0-7; peak duration range: 0.01-0.2 |
| | Peak finder | Intensity tolerance: 50%; m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.15 |
| | Export to csv | |
| **Positive mode** | Raw data import | |
| | Mass detection | Noise level: 15 |
| | Chromatogram builder | Min time span (min): 0.01;<br>Min height: 4.0E1; m/z tolerance: 0.055 mz or 30 ppm |
| | Chromatogram deconvolution | Chromatographic threshold: 97%; Search minimum in RT range (min): 0.01; Minimum relative height: 10%; Minimum absolute height: 6.0E1; Min ratio of peak/top edge: 1.5; Peak duration range (min): 0.01-0.2 |
| | Isotopic pattern | m/z tolerance: 0.06 or 30 ppm; Retention time tolerance: 0.01; Monotonic shape; maximum charge: 1 |
| | Join aligner | m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.15; Weight for both m/z tolerance and retention time tolerance: 10 |
| | Duplicate peak filter | m/z tolerance: 0.5 or 600 ppm; RT tolerance: 0.15 |
| | Peak list rows filter | Min peaks in a row: 5<br>Minimum peaks in an isotope pattern: 1; m/z range: 50-1000; RT range: 0-7; peak duration range: 0.01-0.2 |
| | Peak finder | Intensity tolerance: 50%; m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.17 |
| | Export to csv | |

**Statistical analyses**

*Development of the PLS-DA model*

Each of the sixteen initial models was made as follows:

1) Samples from eight randomly chosen subjects, four from NND and four from ADD, were excluded as test set. The subjects were chosen in a way where each subject would be removed as test set in at least two of the sixteen models.
2) A model was built and cross-validated by leaving out data from one person at a time. All features in the model with variable importance in projection (VIP) scores[25,26] below 0.8 for both groups were removed.
3) A new model was build and cross-validated based on the remaining features. Again, features with VIP scores below 0.8 were removed and this step was repeated until all remaining features had VIP scores above 0.8.
4) The reduced model was validated by applying the test set of excluded samples from 1).

In order to select the most discriminative features across the sixteen initial PLS-DA models, sixteen new PLS-DA models were made including features that were present in from one to 16 of the initial PLS-DA models, respectively. The new models were cross-validated and the model with the lowest cross-validated error was chosen. From this model, the number of LVs to include in the final PLS-DA model was decided on as a trade-off between low CV error and a low number of LVs.

*PCA on food intake*

The most discriminant foods were selected in PCA as follows for the three datasets (WDR, Supermarket data and WDR with the same subset of samples as used for the supermarket data):

1) 20 % of the samples (or subjects for the supermarket data) from each diet were selected at random and excluded from the dataset.
2) A PCA was performed with the rest of the data, applying cross-validation (nine iterations, from which random subsets of 10 % of the samples were excluded from each).
3) The twenty most extreme loadings in each direction of PC 1 were selected (PC 1 was completely separating the diets in all models)
4) Step 1-3 was repeated five times until all samples (or subjects) had been excluded once.

Foods selected in at least four out of five PCA models for a dataset were considered the strongest discriminants.

**Identification**

The calculation of sensitivity and specificity was done in an iterative fashion. First, the obtained peak areas of a feature for all 214 samples were sorted from the highest to the lowest value. Then the sensitivity and specificity for each food item, as reported in WDR, was calculated (see equation I and II below) applying the peak area from the 5$^{th}$ lowest sample as threshold for positive detection in urine. The threshold for positive detection in urine was then changed to include two more samples (7$^{th}$ lowest peak area) and the sensitivity and specificity for each food item was calculated again. This was continued, including two more urine samples per time, until 157 samples were included, which corresponds to the maximum number of samples for which it is possible to reach a sensitivity of 0.7. All food items that reached a sensitivity and specificity above 0.7 for a feature were taken into account as possible dietary origin of the feature.

I) $\text{Sensitivity} = \frac{n_{\text{Food>Th}}}{n_{\text{Food>Th}} + n_{\text{Food<Th}}}$

II) $\text{Specificity} = \frac{n_{\text{Samples}} - n_{\text{Other>Th}} - n_{\text{Food>Th}} - n_{\text{Food<Th}}}{n_{\text{Samples}} - n_{\text{Food<Th}} - n_{\text{Food>Th}}}$

$n_{Food>Th}$ and $n_{Food<Th}$: Number of samples above and below the threshold, respectively, where the food had been reported on the day of urine sampling according to WDR.
$n_{other>Th}$: Number of samples above the threshold, where the food had not been reported on the day of urine sampling according to WDR.
$n_{samples}$: Total number of samples, excluding week 0 (214).

**Table S3.** Model parameters for the 16 initial PLD-DA models

| Model no. | No. of LVs | CV error | Test set error | Number of features |
|---|---|---|---|---|
| 1 | 9 | 0.06 | 0.13 | 43 |
| 2 | 7 | 0.05 | 0.06 | 52 |
| 3 | 10 | 0.07 | 0.31 | 46 |
| 4 | 10 | 0.05 | 0.19 | 41 |
| 5 | 8 | 0.05 | 0.19 | 41 |
| 6 | 10 | 0.03 | 0.19 | 64 |
| 7 | 10 | 0.03 | 0.31 | 45 |
| 8 | 7 | 0.03 | 0 | 60 |
| 9 | 10 | 0.04 | 0.06 | 49 |
| 10 | 10 | 0.03 | 0.13 | 63 |
| 11 | 10 | 0.04 | 0 | 54 |
| 12 | 9 | 0.05 | 0.06 | 57 |
| 13 | 10 | 0.05 | 0.13 | 63 |
| 14 | 10 | 0.04 | 0 | 43 |
| 15 | 5 | 0.04 | 0.25 | 48 |
| 16 | 7 | 0.05 | 0.13 | 65 |
| Average | 9 | 0.04 | 0.13 | 52 |

LV: Latent variables; CV error: Classification error for cross-validation; Test set error: Classification error for test set.

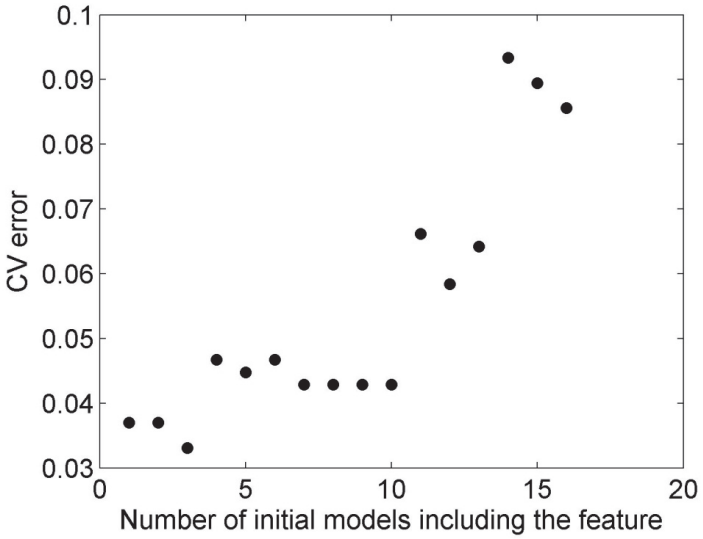



**Figure S1.** Cross-validation classification errors (CV error) for all partial least squares discriminant analysis (PLS-DA) models including features present in 1-16 of the initial models. The PLS-DA model with the lowest value (0.033) was used as the final model to predict New Nordic Diet (NND) and Average Danish Diet (ADD) dietary patterns.

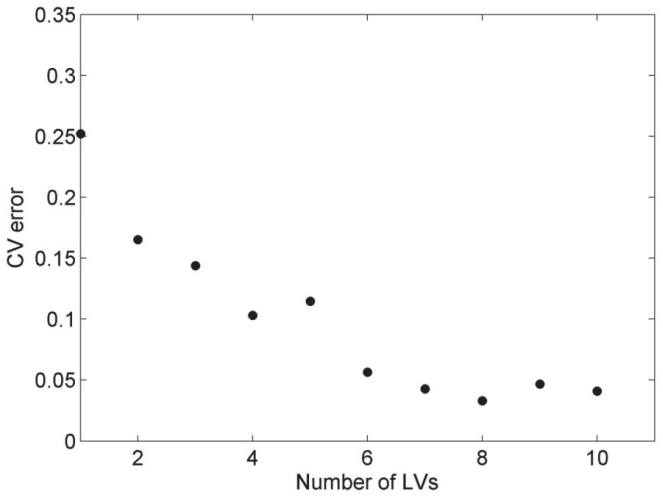
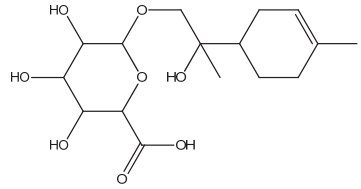



**Figure S2** CV error as a function of the number of LVs for the PLS-DA model with the lowest CV error in Figure S2. Four LVs were chosen for this model.

**Table S4.** Tentative identification: Limonene-8,9-diol-glucuronide (C16H26O8)

RT 3.62, $[M-H]^-$ = 345.154, $[M+H]^+$ =347.179




The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV

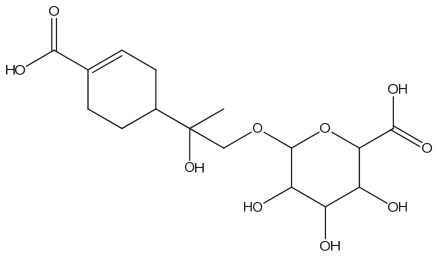
| **Fragment** | **Annotation (ESI+)** | **Loss** |
|---|---|---|
| 347.170** | C16H27O8 | |
| 329.156** | C16H25O7 | -H2O |
| 311.148** | C16H23O6 | -2H2O |
| 293.136** | C16H21O5 | -3H2O |
| 283.157 | C15H23O5 | -2H20, -CO |
| 275.126 | C16H19O4 | -4H2O |
| 265.145 | C15H21O4 | -3H2O, -CO |
| 247.136 | C15H19O3 | -4H2O, -CO |
| 231.139 | C15H19O2 | -4H2O,-CO2 |
| 199.075 | C13H11O2 | |
| 177.040** | C6H9O6 | |
| 171.137 | C10H19O2 | -C6H8O6 |
| 165.129 | C11H17O | |
| 159.029** | C6H7O5 | |
| 153.128** | C10H17O | -C6H8O6, -H20 |
| 141.018 | C6H5O4 | |
| 135.117** | C10H15 | -C6H8O6, -2H20 |
| 113.021 | C5H5O3 | |
| 107.086 | C8H11 | -C6H8O6, -2H20, -C2H4 |
| 95.085 | C7H11 | -C6H8O6, -2H20, -C3H4 |
| 93.069 | C7H9 | -C6H8O6, -2H20, -C3H6 |
| **Fragment** | **Annotation (ESI-)** | **Loss** |
| 345.155 | C16H25O8 | |
| 327.143 | C16H23O7 | -H2O |
| 285.133 | C14H21O6 | -C2H4O2 |
| 269.142 | C14H21O5 | -C2H4O3 |
| 175.024* | C6H7O6 | |
| 169.128 | C10H17O2 | -C6H8O6 |
| 157.012* | C6H5O5 | |
| 129.020* | C5H5O4 | |
| 131.034* | C5H7O4 | |

or 30 eV). Masses marked with (**) corresponds to MSMS fragments reported in Poon et al.[1] for $[M+Na]^+$.
m/z fragments marked with (*) are from the glucuronide moiety


1) Poon, G.K., Vigushin, D., Griggs, L.J., Rowlands, M.G., Coombers, R.C. and Jarman, M., 1996. Identification and characterization of limonene metabolites in patients with advanced cancer by liquid chromatography/mass spectrometry. *Drug metabolism and disposition: the biological fate of chemicals,* 24, 565.

**Table S5.** Tentative identification: Perillic acid-8,9-diol-glucuronide (C16H24O10)

RT 2.89, $[M-H]^-$ = 375.128, $[M+NH4]^+$= 394.171

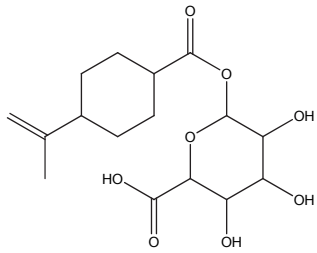



| Fragment | Annotation (ESI+) | Loss |
|---|---|---|
| 394.170 | C16H28O10N | |
| 359.136 | C16H23O9 | -NH3, -H2O |
| 341.121 | C16H21O8 | -NH3, -2H2O |
| 323.113 | C16H19O7 | -NH3, -3H2O |
| 183.100 | C10H15O3 | -NH3,- C6H8O6,-H2O |
| 165.091 | C10H13O2 | -NH3,- C6H8O6,-2H2O |
| 159.029* | C6H7O5 | |
| 141.018* | C6H5O4 | |
| 137.095 | C9H13O | -NH3,- C6H8O6, -2H2O,-CO |
| 113.021* | C5H5O3 | |
| 109.100 | C8H13 | -NH3,- C6H8O6, -2H2O,-2CO |
| 107.049 | C7H7O | -NH3,- C6H8O6, -H2O, -C3H8O2 |
| 85.027* | C4H5O2 | |
| 79.053 | C6H7 | -NH3,- C6H8O6, -H2O, -C3H8O2,-CO |
| **Fragment** | **Annotation (ESI-)** | **Loss** |
| 375.129 | C16H23O10 | |
| 357.118 | C16H21O9 | -H2O |
| 331.140 | C15H23O8 | -CO2 |
| 315.109 | C14H19O8 | -C2H4O2 |
| 199.096 | C10H15O4 | -C6H8O6 |
| 181.089 | C10H13O3 | -C6H8O6, -H2O |
| 175.022* | C6H7O6 | |
| 157.011* | C6H5O5 | |
| 129.018* | C5H5O4 | |
| 125.060 | C7H9O2 | -C6H8O6, -C3H6O2 |
| 113.024* | C5H5O3 | |
| 99.008* | C4H3O3 | |
| 95.015* | C5H3O2 | |
| 87.009* | C3H3O3 | |
| 85.029* | C4H5O2 | |
| 75.008* | C2H3O3 | |
| 73.029* | C3H5O2 | |
| 71.013* | C3H3O2 | |
| 59.015* | C2H3O2 | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S6.** Tentative identification: Dihydroperillic acid-glucuronide (C16H24O8)

RT 3.83, $[M-H]^-$ = 343.135

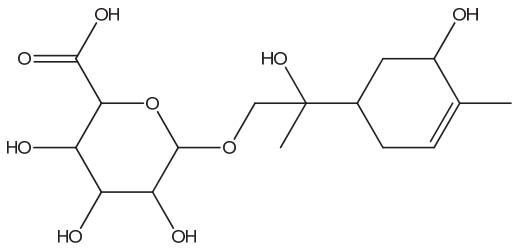



| Fragment | Annotation (ESI-) | Loss |
|---|---|---|
| 343.139 | C16H23O8 | |
| 325.123 | C16H21O7 | -H2O |
| 263.128 | C15H19O4 | -2H20, -CO2 |
| 193.038* | C6H9O7 | |
| 181.121 | C11H17O2 | -H2O,-CO,-2C2H2O2 |
| 175.025* | C6H7O6 | |
| 167.105 | C10H15O2 | -C6H8O6 |
| 157.014* | C6H5O5 | |
| 133.014* | C4H5O5 | |
| 113.023* | C5H5O3 | |
| 103.007* | C3H3O4 | |
| 99.008* | C4H3O3 | |
| 95.013* | C5H3O2 | |
| 89.024* | C3H5O3 | |
| 85.029* | C4H5O2 | |
| 75.009* | C2H3O3 | |
| 72.993* | C3H5O2 | |
| 71.014* | C3H3O2 | |
| 59.014 | C2H3O2 | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S7.** Tentative identification: p-menth-1-ene-6,8,9-triol-glucuronide (C16H26O9)

RT 2.91, $[M-H]^-$ = 361.147




| Fragment | Annotation (ESI-) | Loss |
|---|---|---|
| 361.149 | C16H25O9 | |
| 301.135 | C14H21O7 | -C2H4O2 |
| 113.026* | C5H5O3 | |
| 99.009* | C4H3O3 | |
| 87.009* | C3H3O3 | |
| 85.029* | C4H5O2 | |
| 75.009* | C2H3O3 | |
| 71.014* | C3H3O2 | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S8.** Unknown-glucuronide (C16H26O10)

RT 2.41, $[M-H]^-$ = 377.143

| Fragment | Annotation (ESI-) | Loss |
|---|---|---|
| 377.144 | C16H25O10 | |
| 201.112 | C10H17O4 | -C6H8O6 |
| 193.034* | C6H9O7 | |
| 183.104 | C10H15O3 | -C6H8O6, -H2O |
| 175.024* | C6H7O6 | |
| 113.024* | C5H5O3 | |
| 103.004* | C3H3O4 | |
| 99.008* | C4H3O3 | |
| 95.014* | C5H3O2 | |
| 85.028* | C4H5O2 | |
| 75.010* | C2H3O3 | |
| 73.012* | C3H5O2 | |
| 71.014* | C3H3O2 | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S9.** Unknown pyrraline derivative (C14H20N2O5)

RT 3.39, $[M-H]^-$ = 295.128

| Fragment | Annotation (ESI-) | Loss |
|---|---|---|
| 295.128 | C14H19N2O5 | |
| 253.119* | C12H17N2O4 | -C2H2O |
| 223.109* | C11H15N2O3 | -C2H2O, -CH2O |
| 124.039* | C6H6NO2 | -C2H2O,-C6H11NO2 |
| 94.029* | C5H4NO | -C2H2O,-C6H11NO2,-CH2O |
| **Fragment** | **Annotation (ESI+)** | **Loss** |
| 297.146 | C14H21N2O5 | |
| 279.134 | C14H19N2O4 | -H2O |
| 237.135* | C12H17N2O3 | -H2O, -C2H2O |
| 219.112* | C12H15N2O2 | -2H2O, -C2H2O |
| 175.125* | C11H15N2 | -H2O, -C2H2O, -CO2 |
| 148.111* | C10H14N | -H2O, -C2H2O, -CO2,-CHN |
| 122.057 | C7H8NO | |
| 84.081* | C5H10N | -H2O, -CH2O,-2CO, -C2H2O,-NH3, -C4H2 |
| 82.063* | C5H8N | -H2O, -CH2O,-2CO, -C2H2O,-NH3, -C4H2 |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). Fragments marked with a (*) were also found for the pyrraline standard

**Table S10.** Unknown pyrraline derivative (C14H20N2O5)

RT 3.46, $[M-H]^+$ = 297.145

| Fragment | Annotation (ESI+) | Loss |
|---|---|---|
| 297.144 | C14H21N2O5 | |
| 279.133 | C14H19N2O4 | -H2O |
| 255.139* | C12H19N2O4 | -C2H2O |
| 251.142 | C13H19N2O3 | -H2O, -CO |
| 239.139 | C12H19N2O3 | -CO, -CH2O |
| 209.130 | C11H17N2O2 | -H2O, -CO, -C2H2O |
| 197.126 | C10H17N2O2 | -CO, -CH2O, -C2H2O |
| 192.102 | C11H14NO2 | -H2O, -CO, -C2H2O,-NH3 |
| 146.097* | C10H12N | -2H2O, -2CO, -C2H2O,-NH3 |
| 134.096* | C9H12N | -H2O, -CH2O,-2CO, -C2H2O,-NH3 |
| 84.080* | C5H10N | -H2O, -CH2O,-2CO, -C2H2O,-NH3, -C4H2 |
| **Fragment** | **Annotation (ESI-)** | **Loss** |
| 295.129 | C14H19N2O5 | |
| 253.124* | C12H17N2O4 | -C2H2O |
| 124.039* | C6H6NO2 | -C2H2O,-C6H11NO2 |
| 66.034* | C4H4N | -C2H2O,-C6H11NO2,-2CHO |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). Fragments marked with a (*) were also found for the pyrraline standard

**Table S11.** Unknown glucuronide (C12H16O10)

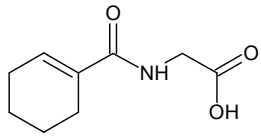
RT 1.03, $[M-H]^-$ = 319.066

| **Fragment** | **Annotation (ESI-)** | **Loss** |
|---|---|---|
| 319.067 | C12H15O10 | |
| 301.055 | C12H13O9 | -H2O |
| 277.061 | C10H13O9 | |
| 193.035* | C6H9O7 | |
| 175.024* | C6H7O6 | |
| 143.033 | C6H7O4 | -C6H8O6 |
| 113.024* | C5H5O3 | |
| 103.001* | C3H3O4 | |
| 101.024 | C4H5O3 | -C6H8O6, -C2H2O |
| 99.009* | C4H3O3 | |
| 95.012* | C5H3O2 | |
| **Fragment** | **Annotation (ESI+)** | **Loss** |
| 321.082 | C12H17O10 | |
| 145.051 | C6H9O4 | -C6H8O6 |
| 132.100 | C6H12O3 | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S12.** Tentative identification: 3,4,5,6-tetrahydrohippurate (C9H14NO3)

RT 4.46, $[M-H]^+$ =184.098




| **Fragment** | **Annotation (ESI+)** | **Loss** |
|---|---|---|
| 184.098 | C9H14NO3 | |
| 109.063* | C7H9O | -glycine (C2H5NO2) |
| 81.070* | C6H9 | -glycine, -CO |
| 79.054* | C6H7 | -glycine, -HCO |
| **Fragment** | **Annotation (ESI-)** | **Loss** |
| 182.082 | C9H12NO3 | |
| 138.089 | C8H12NO | -CO2 |
| 136.077 | C8H10NO | -HCOOH |
| 74.023 | C2H4NO2 | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z with (*) were also reported in Brewster et al.[1]

1) Brewster, D., Jones, R.S. and Parke, D.V., 1977. The metabolism of cyclohexanecarboxylate in the rat. *The Biochemical journal,* 164, 595.

**Table S13.** Unknown glucuronide

RT 2.65, $[M-H]^+$ =340.103

| **Fragment** | **Annotation (ESI+)** | **Loss** |
|---|---|---|
| 340.102 | C15H18NO8 | |
| 229.067 | C10H13O6 | -C5H5NO2 |
| 206.081 | C11H12NO3 | -C4H6O5 |
| 188.072 | C11H10NO2 | - C4H6O5, -H2O |
| 164.070 | C9H10NO2 | -C6H8O6 |
| 146.059 | C9H8NO | -C6H8O6, -H2O |
| 141.018 | C6H5O4* | |
| 122.059 | C7H8NO | -C6H8O6,-C2H2O |
| 113.023 | C5H5O3* | |
| 110.059 | C6H8NO | -C6H8O6,-C3H2O |
| 101.026 | C4H5O3* | |
| 95.008 | C5H3O2 | |
| 85.028 | C4H5O2* | |
| 73.030 | C3H5O2 | |
| **Fragment** | **Annotation (ESI-)** | **Loss** |
| 338.089 | C15H16NO8 | |
| 320.079 | C15H14NO7 | -H2O |
| 175.023 | C6H7O6* | |
| 162.055 | C9H8NO2 | -C6H8O6 |
| 157.015 | C6H5O5* | |
| 129.020 | C5H5O4* | |
| 120.046 | C7H6NO | -C6H8O6,-C2H2O |
| 117.018 | C4H5O4* | |
| 113.024 | C5H5O3* | |
| 103.004 | C3H3O4* | |
| 99.009 | C4H3O3* | |
| 95.014 | C5H3O2* | |
| 89.022 | C3H5O3* | |
| 87.009 | C3H3O3* | |
| 85.029 | C4H5O2* | |
| 75.009 | C2H3O3* | |
| 71.014 | C3H3O2* | |
| 59.015 | C2H3O2* | |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S14.** Unknown glucuronide

RT 2.41, $[M-H]^-$ = 336.072

| **Fragment** | **Annotation (ESI-)** | **Loss** |
|---|---|---|
| 336.073 | C15H14NO8 | |
| 175.025 | C6H7O6* | |
| 160.039 | C9H6NO2 | -C6H8O6 |
| 113.024 | C5H5O3* | |
| 99.008 | C4H3O3* | |
| 95.013 | C5H3O2* | |
| 85.028 | C4H5O2* | |
| **Fragment** | **Annotation (ESI+)** | **Loss** |
| 338.087 | C15H16NO8 | |
| 162.054 | C9H8NO2 | -C6H8O6 |
| 144.045 | C9H6NO | -C6H8O6,-H2O |
| 120.042 | C7H6NO | -C6H8O6,-C2H2O |
| 116.055 | C8H6N | -C6H8O6, -H2O,-CO |

The fragments included in the table are all clear fragments found regardless of the fragmentation energy (10 eV, 20 eV or 30 eV). m/z fragments marked with (*) are from the glucuronide moiety

**Table S15.** Unknown markers in the model

| Diet | Rank | RT qTOF | m/z qTOF | m/z orbitrap | Error (ppm) | Ion | Elemental formula | Main MS/MS fragments |
|---|---|---|---|---|---|---|---|---|
| ADD | 2 | 1.06 | 215.102 | 215.1038 | 3.1 | $[M-H]^-$ | C9H16N2O4 | 173.093; 131.083*; 129.104* |
| ADD | 25 | 1.00 | 506.116 | 506.1152 | - | Neg mode | - | None |
| ADD | 30 | 1.02 | 501.110 | NF | - | Neg mode | - | None |
| ADD | 39 | 1.88 | 123.053 | 123.0562 | - | Neg mode | - | - |
| ADD | 62 | 3.20 | 130.084 | 130.0868 | 0.3 | Neg mode | C6H12NO2 | None |
| ADD | 64 | 1.93 | 279.098 | 279.0988 | 2.3 | $[M-H]^-$ | C13H16N2O5 | 145.061; 127.051*; 109.040* |
| ADD | 67 | 0.85 | 319.067 | 319.0664 | - | $[M-H]^-$ | - | None |
| NND | 5 | 0.52 | 209.056 | 209.0670 | - | Neg mode | - | 129.019; 85.029 |
| NND[1] | 11 | 0.62 | 190.116 | NF | - | Pos mode | - | - |
| NND | 15 | 0.85 | 170.092 | NF | - | Pos mode | - | - |
| NND | 23 | 2.65 | 414.123 | NF | - | Neg mode | - | - |
| NND | 33 | 3.61 | 245.136 | 245.1393 | - | Neg mode | - | 227.129; 209.116* |
| NND | 36 | 2.71 | 306.037[4] | 305.0335 | - | $[M-H]^-$ | C11H13O8S | 163.075*; 123.045*; 101.023* |
| NND[1] | 48 | 3.87 | 113.006 | NF | - | Neg mode | - | - |
| NND[1] | 49 | 3.52 | 328.212 | 328.2116 | 2.6 | Neg mode | C17H29NO5 | - |
| NND | 59 | 4.36 | 267.126 | NF | - | $[M-H]^-$ | - | 223.133; 179.144* |
| No pattern[1] | 27 | 4.10 | 417.213 | NF | - | - | - | - |
| No pattern | 40 | 0.59 | 487.130 | NF | - | $[2M-H]^-$ | [2]C9H12O6N2 | 183.042; 153.030; 140.036 |
| No pattern[3] | 65 | 3.50 | 383.134 | NF | - | Pos mode | - | - |
| No pattern[1] | 66 | 1.01 | 229.154 | NF | - | - | - | - |

NF not found; ADD: Average Danish Diet; NND: New Nordic Diet

[1]Not well resolved peak.

[2]Calculated from $[M-H]^-$ which was visible in orbitrap

[3]Very low peak.

[4]Isotope of 305.0338

**Table S17.** Percentage of consumers and average amount of consumption for the most discriminant foods for ADD (Fig. 4)

| | Supermarket data | | | | Dietary records | | | |
|---|---|---|---|---|---|---|---|---|
| | **No ADD [%]** | **ADD amount [g]** | **No NND [%]** | **NND amount [g]** | **No ADD [%]** | **ADD amount [g]** | **No NND [%]** | **NND amount [g]** |
| **Common ADD** | | | | | | | | |
| Avocado | **91** | **2616** | 23 | 121 | **15** | **91** | 0 | 0 |
| Pepper | **100** | **4072** | 30 | 53 | **20** | **58** | 0 | 0 |
| Chocolate | **100** | **2933** | 85 | 500 | **58** | **43** | 7 | 67 |
| Tomato | **100** | **10964** | 58 | 252 | **52** | **124** | 3 | 34 |
| Olive oil | **97** | **1203** | 10 | 47 | **28** | **14** | 0 | 0 |
| Muesli | **97** | **1990** | 3 | 1562 | **33** | **34** | 4 | 69 |
| Tuna | **100** | **908** | 25 | 102 | **7** | **94** | 0 | 0 |
| Pasta | **100** | **2688** | 25 | 298 | **19** | **125** | 2 | 115 |
| Cucumber | **100** | **2516** | 83 | 571 | **28** | **41** | 6 | 28 |
| Beef | **100** | **9332** | 100 | 1390 | **40** | **139** | 13 | 74 |
| Banana | **97** | **6882** | 23 | 244 | **15** | **146** | 1 | 100 |
| **Dietary records ADD** | | | | | | | | |
| Chocolate milk | 94 | 4529 | 28 | 4212 | **16** | **265** | 0 | 0 |
| Creme fraiche | 94 | 1964 | 90 | 516 | **13** | **50** | 4 | 36 |
| Cheese | 100 | 7053 | 100 | 3698 | **74** | **40** | 48 | 26 |
| Remoulade | 72 | 415 | 25 | 56 | **10** | **16** | 2 | 5 |
| Grapes | 94 | 2535 | 25 | 207 | **10** | **87** | 0 | 0 |
| Wine gum | 94 | 616 | 25 | 185 | **19** | **39** | 0 | 0 |
| Orange juice | 88 | 4552 | 35 | 1599 | **12** | **259** | 1 | 400 |
| **Supermarket data ADD** | | | | | | | | |
| Pork | **100** | **11273** | 100 | 3983 | 51 | 106 | 31 | 123 |
| Biscuits and cakes | **97** | **3735** | 100 | 556 | 22 | 64 | 9 | 68 |
| Rice | **97** | **2444** | 28 | 208 | 13 | 95 | 1 | 120 |
| Coconut (milk and flour) | **88** | **732** | 10 | 106 | - | - | - | - |
| Liquorice | **97** | **1579** | 85 | 410 | 16 | 24 | 4 | 31 |

ADD: Average Danish Diet; NND: New Nordic Diet; No: Percentage of times the food was reported within a dietary group. Amounts for supermarket data are given as the average consumption over 6 months for all consumers within a dietary group who reported the food at least once. Amounts for dietary records are given as the average consumption reported per day for the dietary records where the food was reported within a dietary group. Common ADD foods were discriminant foods in supermarket data and in the dietary records. Dietary records ADD and supermarket data ADD, where only among the most discriminant in the dietary records and the supermarket data, respectively (Figure 5). Numbers in bold are the diet and data source where the foods were found to be discriminant.

**Table S18.** Percentage of consumers and average amount of consumption of the most discriminant foods for NND

| | Supermarket data | | | | Dietary records | | | |
|---|---|---|---|---|---|---|---|---|
| | No ADD [%] | ADD amount [g] | No NND [%] | NND amount [g] | No ADD [%] | ADD amount [g] | No NND [%] | NND amount [g] |
| **Common NND** | | | | | | | | |
| Carrot | 97 | 2348 | **100** | **17376** | 12 | 89 | **60** | **134** |
| Apple juice | 84 | 2953 | **100** | **18502** | 12 | 268 | **46** | **220** |
| Barley | 0 | 0 | **100** | **1303** | 0 | 0 | **13** | **58** |
| Celeriac | 19 | 254 | **100** | **5882** | 1 | 170 | **16** | **68** |
| Hazelnut | 66 | 229 | **100** | **3701** | 1 | 25 | **52** | **28** |
| Parsley root | 3 | 130 | **100** | **1310** | 0 | 0 | **10** | **81** |
| **Dietary records NND** | | | | | | | | |
| Rapeseed oil | 50 | 517 | 100 | 1446 | 1 | 10 | **44** | **12** |
| Apple | 100 | 6695 | 100 | 20680 | 27 | 105 | **67** | **168** |
| Walnut | 56 | 192 | 98 | 1335 | 6 | 17 | **21** | **30** |
| Gooseberries | 3 | 68 | 98 | 1412 | 0 | 0 | **9** | **190** |
| Mustard | 84 | 260 | 95 | 769 | 3 | 8 | **20** | **11** |
| Sugar | 100 | 1373 | 100 | 2022 | 31 | 17 | **45** | **20** |
| Green peas | 63 | 354 | 100 | 3841 | 2 | 116 | **21** | **97** |
| Vinegar | 75 | 775 | 100 | 2201 | 3 | 8 | **35** | **22** |
| Honey | 84 | 155 | 100 | 714 | 7 | 20 | **22** | **13** |
| Lingonberriy | 0 | 0 | 98 | 1049 | 0 | 0 | **10** | **27** |
| Parsley | 88 | 115 | 100 | 488 | 2 | 14 | **12** | **12** |
| **Supermarket data NND** | | | | | | | | |
| Pollack | 31 | 294 | **100** | **2463** | 0 | 0 | 9 | 100 |
| Swede | 3 | 250 | **98** | **2149** | - | - | - | - |
| Kale | 3 | 100 | **98** | **561** | - | - | - | - |
| Pointed cabbage | 31 | 339 | **100** | **2603** | 0 | 0 | 14 | 103 |
| Game | 6 | 113 | **100** | **3837** | - | - | - | - |
| Greenland halibut | 0 | 0 | **98** | **722** | 0 | 0 | 8 | 60 |
| Cod | 19 | 230 | **100** | **1934** | 0 | 0 | 11 | 106 |
| Beetroot | 69 | 638 | **98** | **5216** | 5 | 75 | 27 | 104 |
| Savoy cabbage | 0 | 0 | **100** | **1430** | - | - | - | - |

ADD: Average Danish Diet; NND: New Nordic Diet; No: Percentage of times the food was reported within a dietary group. Amounts for supermarket data are given as the average consumption over 6 months for all consumers within a dietary group who reported the food at least once. Amounts for dietary records are given as the average consumption reported per day for the dietary records where the food was reported within a dietary group. Common NND foods were discriminant foods in supermarket data and in the dietary records. Dietary records NND and supermarket data NND, where only among the most discriminant in the dietary records and the supermarket data, respectively (Figure 5). Numbers in bold are the diet and data source where the foods were found to be discriminant.

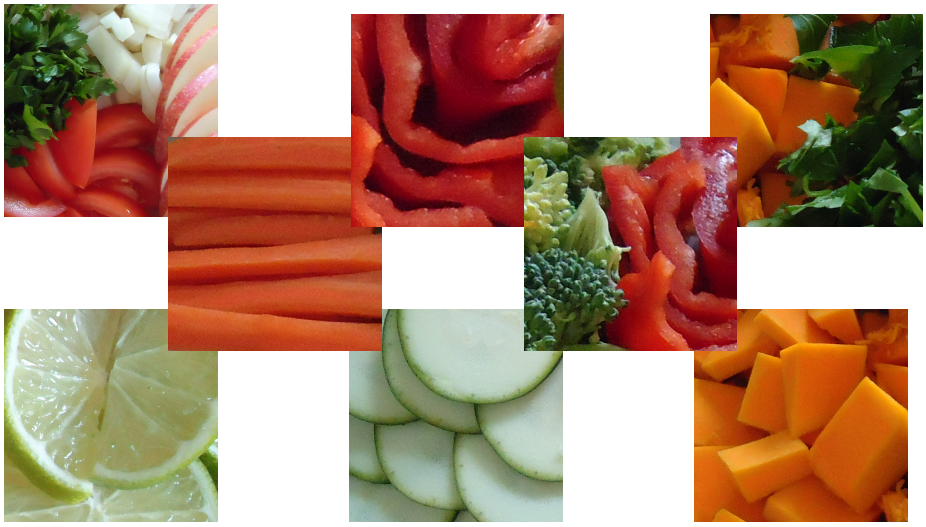
DEPARTMENT OF NUTRITION, EXERCISE AND SPORTS
FACULTY OF SCIENCE · UNIVERSITY OF COPENHAGEN
PHD THESIS 2014 · ISBN 978-87-7611-686-6

MAJ-BRITT SCHMIDT ANDERSEN

# Discovery of food exposure markers in urine and evaluation of dietary compliance by untargeted LC-MS metabolomics

Accurate measurement of dietary intake in nutrition studies is crucial to investigate relationships between diet and health. The common tools used for assessment of dietary exposure in humans rely almost solely on self-reporting, which is associated with a range of random and systematic errors. Biomarkers measured in biological samples, such as urine or plasma, provide a promising supplement to self-reporting, as they are objective measures. However, the few currently available biomarkers cover the diet poorly and more markers, in particular for intake of individual foods, are needed.

In this thesis, untargeted metabolomics, a relatively new method within nutrition research, has been applied to find new potential food exposure markers in urine for intake of a range of foods. In addition, it has been investigated if it is possible to distinguish two dietary patterns, a New Nordic Diet (NND) and an Average Danish Diet (ADD), in urine samples from a controlled intervention study.



SL grafik