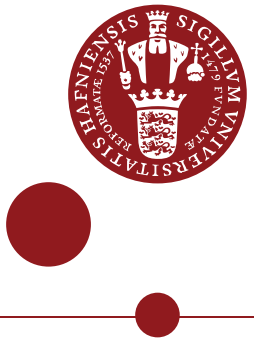
UNIVERSITY OF COPENHAGEN



# Discovering Biomarker of Whole Grain Barley and Wheat Intake by LC-MS Based Untargeted Metabolomics

## Master Thesis

Tu Hu

Supervisors: Lars Ove Dragsted & Gözde Gürdeniz

Submitted on: 2nd Sep 2019

**Author:**
Tu Hu

**Date of Submission:**
2nd, Sep, 2019

**ECPT Credits:**
30

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

Gözde Gürdeniz
Senior Researcher
Copenhagen Prospective Studies on Asthma in Childhood (COPSAC)
Copenhagen University Hospital, Denmark

# Table of contents

# List of Figures and Tables

# Abstract

**Background:** The intake of different types of whole grain might benefit health differently. The discovery of biomarkers for whole grain barley and wheat intake could provide objective tools to measure their exposures and hence reveal the health benefits.

**Aims:** The primary aim is to discover biomarkers to estimate whole grain barley and wheat intake. The secondary aim is to streamline the metabolomics data analysis workflow.

**Methods:** The biomarkers of whole grain barley and wheat intake are discovered by the systematic literature review and LC-MS based metabolomics analysis of two intervention studies. Barley bread intervention uses whole grain barley and wheat bread. Protein source study uses barlyotto and other food. Potential biomarkers are selected from barley bread intervention study, and further validated in protein sources study. Tandem mass spectrometry was used to elucidate the structure of candidate biomarkers. A new data processing workflow is implemented by incorporating free-available tools and self-developed functions in R statistical language.

**Results:** Systematic literature review shows that no biomarker can indicate whole grain barley intake. Alkylresorcinol (AR) and AR metabolites can indicate whole grain wheat intake. PLS-DA modelling and t-test cannot find discriminating metabolites from fasting plasma samples. Whole grain barley and wheat intake can be distinguished from 24-h pooled urine. 4-hydroxybenzoic acid-4-sulphate has been putatively identified as whole grian barley intake biomarker. AR metabolites and HBOA glucuronate can indicate whole grain wheat intake. A metabolomics data analysis workflow is implemented by incorporating free-available tools and self-developed functions. *m2r* can convert Matlab DataSet to R dataframe. *annotate_kudb* can annotate roughly 100-300 metabolites. *Libra* can calibre batch effects without changing coefficient of variance of pooled samples.

**Conclusions:** 4-hydroxybenzoic acid-4-sulphate can indicate whole grain barley intake. AR metabolites and HBOA glucuronate can indicate whole grain wheat intake. A streamlined data processing workflow is implemented.

**Keywords:** metabolomics; biomarkers of food intake; whole grain; barley; wheat; bioinformatics

# Preface and Introduction

In the disaster movie, *2012*, the world comes to an end due to the natural disaster eruptions. A secret ark plan is built in Tibet to ensure the survival of human beings. Tibet Plateau is the highest region on Earth, with an average elevation of 5000 m. In *2012*, the high elevation makes Tibet the last safe place for humans to survive from the disasters. However, the average elevation of 5000 m also brings a low temperature. Large areas of Tibet Plateau have an average temperature below 0 ºC around the year. How do the Tibetan survive in such an extreme environment? What do they eat?

Barley, this magic cereal could be one of the reasons. Barley has been cultivated in Tibet for at least 3500 years. Barley can grow in the harshest and the most marginal areas from sub-Saharan to the Himalayas. In the viewpoint of some anthropologists, barley played a vital role in the transition of human society from a hunting-and-gathering to an agrarian lifestyle. The ability of barley to adapt to the harsh and marginal environment grants it possibility to be used as a food source to cope with food security. However, we know little about the health beneficial and nutritional effects of barley.

Whole grain has raised a lot of public interest for its health benefits. A lot of dietary guidelines have suggested increasing whole grain intake because the scientific evidences generally show that whole grain intake is associated with a better health. But, the association of whole grain intake with individual disease is not clearly clarified. Meanwhile, the health benefits of different types of whole grain are not clear. Both research questions need tools to objectively measure the whole grain intake.

Metabolomics is a powerful tool to measure nearly hundred to thousand metabolites from human biospeciesman. Some metabolites could be characteric to a food or a food group to serve as biomarkers of food intake. Biomarkers of food intake could reflect food exposures objectively in populations and resolve some nutrition research questions.

This thesis is a conitinuation of my previous 15 ECTS research project (Sep – Nov 2018, supervised by Gözde Gürdeniz). In the previous project, I extracted and analyzed the food samples of whole grain barley and wheat, processed the metabolomics data of urine, and putatively identified the discriminating metabolites of whole grain barley intake. The previous project showed that the intake of whole grain wheat can be discriminated from whole grain barley intake. Therefore, the biomarkers of whole grain wheat intake can be identified. In the previous project, I was frustrated by the metabolomics data processing. I believe that a more unified and streamlined

data processing workflow can facilitate the metabolomics research. Therefore, the research aims of this thesis are formulated.

### Aims

The primary aim is to discover biomarkers of whole grain barley and wheat intake through the systematic literature review and metabolomics studies.

The secondary aims include developing new bioinformatics tools to facilitate metabolomics data processing and streamlining metabolomics data analysis workflow.

# Background

## Barley

Barley is one of the most produced grains. In 2016, 143 million tons of barley were produced, which equals to be ranked as the fourth most-produced grain behind maize, rice, and wheat. EU countries produced 63% of the world's barley. Denmark produced 3.9 million tons, which was ranked the 11th most-produced country of barley (*1*). Barley might be the most widely adapted cereal grain species with good drought, cold, and salt tolerance (*2*). It is cultivated both in highly productive agricultural systems and also in marginal and subsistence environments (*3*). Barley's ability to adapt to multiple biotic and abiotic stresses could make it a food source to cope with food security (*3*).

Barley is majorly used for animal feed and brewing (*2*, *3*). Roughly 5% of total production is directly consumed by human (*2*). Barley for direct food use only remains important in a few areas, such as Asia and northern Africa. The reason for barley being rarely used for direct consumption include relatively few efforts devoted to systematically improve its palatable properties as well as improving food processing techniques and product development (*2*). The quality standard of barley for food use has not been well established, making it difficult for food manufacturers to select raw materials suitable for use in specific food products (*2*).

## Whole Grain

Whole grain contains the starchy endosperm, germ, and bran, in contrast to refined grain, which retains only the endosperm. When whole grain exists in the processed forms such as grounded, cracked, or flaked, the three fractions (starchy endosperm, germ, and bran) should be present the same proportions as they exist in native grains (*4*, *5*).

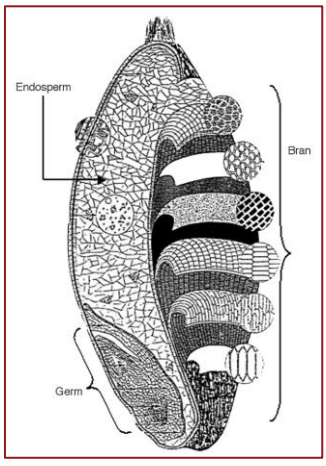



**Figure 1 The Diagram of Grain**

The bran and germ fractions of the grain consist of a wide range of compounds with known health beneficial effects, including dietary fibers, lignans, tocotrienols, phenolic compounds, etc (*5*). The grain-refining removes all or part of the bran. As a result, the grain loses the health-beneficial compounds.

Many observational studies reported that whole grain intake could improve health (*4*, *5*). However, this relies on accurate measurement of whole grain intake. The accurate measurement of dietary intake is a challenging task in general (*6*). With regard to the whole grain, the intake measurement might be particularly prone to errors because there is a considerable variation in whole grain content in different products and consumers may have difficulty in recognising whole grain products among other products (*7*). The intake of different types of whole grain might benefit health differently (*8*). Current dietary measurement tools categorise different types of whole grain as one group. Therefore they cannot distinguish different types of whole grain.

## Biomarkers of Food Intake

Biomarkers of food intake can estimate recent or average intake of a food or a food group by measuring characteristic metabolites from the biological specimen, such as urine, plasma, or tissue (*9*, *10*). The biomarker could be a single metabolite or combination of several metabolites. Biomarkers of food intake can be used as a complement with self-report based instruments to alleviate measurement error and uncertainties of food intake measurement in observational studies (*6*). Biomarkers of food intake can also monitor compliances in intervention studies (*10*).

### Guidelines for Biomarker of Food Intake Reviews

Guidelines for Biomarker of Food Intake Reviews (BFIRev) proposed a systematic approach to conduct an extensive literature search and assess the qualities of food intake biomarkers. The review results can provide the basis to validate the biomarkers and prioritise identifications of novel biomarkers (*11*).

The review process consists of two parts. The first part shares similar frameworks with other systematic reviews, including searching, screening, and selecting articles. The second part differs from other reviews. Reviewers assess the qualities of reported biomarkers for the risk of bias or confounding factors. In the end, reviewers should report the results systematically and summarise the current status of biomarkers.

## Metabolomics

Metabolomics is the complete study of all metabolites, i.e. metabolome, which are small molecules, intermediates or end-products of chemical reactions that continuously go on in the human body. Because blood, urine, and tissues are packed with these compounds, it should be possible to detect and analyse them (*12*, *13*). Though metabolomics is still emerging (*14*), it has shown the potential in the food, plant, environment, nutrition, and health research (*15*, *16*).

Metabolomics can be categorized as targeted or untargeted approaches. Targeted metabolomics identifies and quantifies a limited number of known metabolites (*17*). Untargeted metabolomics analyses all detectable metabolites, including chemical unknowns. This goal is realised by well-conceived sample preparations, analytical methods, and data processing to cover as many metabolites as possible (*12, 18*). Untargeted metabolomics analysis generates large amounts of data, which characterises this approach as data-driven. New hypotheses could be generated by untargeted metabolomics research.

The main analytical techniques of metabolomics are Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS) coupled to a liquid or gas chromatography, i.e., LC-MS and GC-MS (*15*). NMR provides high reproducibility but less sensitivity and limited discovery power. MS is highly sensitive and has strong discovery ability but poor reproducibility (*15*). Because these two techniques offer different strengths, the combination of these two techniques becomes an emerging research direction (*19*, *20*).

### Challenges

***Compound Identifications.*** The major bottleneck of LC-MS based untargeted metabolomics is metabolite identification (*21*). An unknown metabolite can be identified by matching the compound with the database (either experiment or in-silicon) for retention time, mass to charge ratio (m/z), and $MS^2$ spectra (*20*). However, the uncertainties of measurement, the availabilities of reference compound, etc. hurdles the structure elucidation of a lot of metabolites of interest.

***Bioinformatics tools.*** Another challenge of metabolomics study is a lack of complete bioinformatics tools to analyse metabolomics data. Though a lot of open-source software is emerging. Currently, many existing open-source software still lack the ability to capture the complete analysis workflow and tune every parameters (*22*). The harmonisation of different bioinformatics is a challenge task. First, bioinformatics tools might be developed in different programming languages. Second, different open-source software might have different formats for data storage.

# Materials and Methods

## Systematic Review of Biomarkers of Food Intake

This systematic review follows the BFIRev guidelines (*11*). The literature search was performed in 3 databases (PubMed, Web of Science and Scopus). Keywords used for searing barley intake biomarkers in human are: (barley) AND (biomarker* OR marker* OR metabolite* OR biokinetics OR biotransformation OR pharmacokinetics) AND (intake OR meal OR diet OR ingestion OR consumption OR eating OR food) AND (human* OR men OR women OR patient* OR volunteer* OR participant*) AND (trail* or experiment OR study) AND (urine OR plasma OR blood OR serum OR excretion OR hair OR toenail OR faeces OR faecal water). The first element was changed to wheat for the wheat intake biomarker search.

Due to the limited amount of search results for barley, the scope was expanded to include animal studies. The keywords (animal* OR goat OR sheep OR cow OR mice OR mouse* OR animal model* OR dog*) were used to replace the previous 'human*' entry. Besides, 'feed' was added to 'food' entry.

Other databases including HMDB(*23*), FoodDB(*24*), PhenolExplorer(*25*), and Dictionary of Food Compounds(*26*) were also used to search compounds that are present exclusively in WG barley or wheat.

In order to verify the uniqueness of the compounds identified for each food, the same keywords combinations were used but with the compound name instead of 'wheat' and 'barley'.

## Intervention Studies

This study investigated metabolomics data from two intervention studies. **Figure 2** shows the outlines of these two studies.

Barley bread intervention study (*27*) was conducted by Veronica Maria Popovici in 2015. This study is a randomised crossover intervention study, including 14 healthy volunteers. During the intervention period, volunteers ate two bread rolls (227 grams) of whole grain barley or wheat on each day. In each visit, volunteers donated their 24-h pooled urine and fasting plasma.

Protein sources study (*28*) is a crossover intervention study conducted by Maj-Britt Schmidt Andersen. Volunteers received meals including barleyotto, soup or pies with different protein sources. Volunteers donated their urine samples at different time intervals before and after intake of the test meals, covering a total of 24 h.

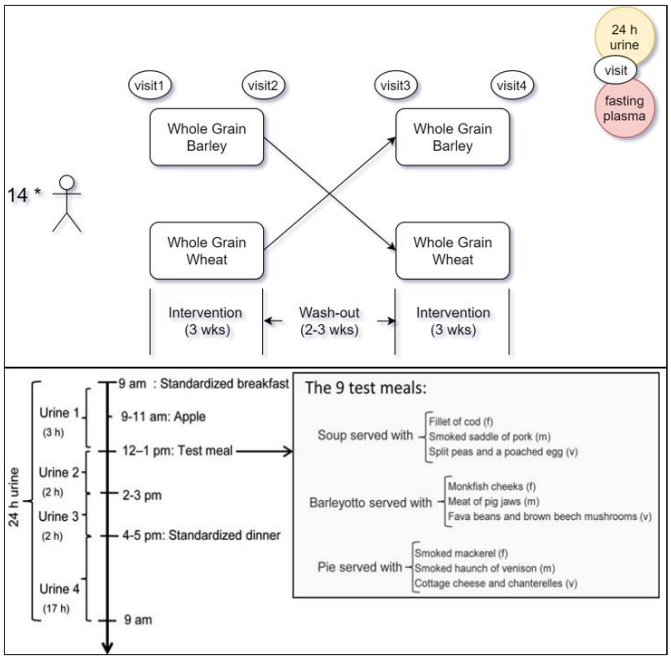



**Figure 2 Outline of study designs for barley bread study (top) and protein source study (bottom)**

## Metabolomics Analysis

### Sample Processing and Data Acquisition

Plasma protein precipitation, urine centrifugation, and dilution were performed as before (*28*, *29*). Metabolomics data were acquired by a UPLC-MS system consisting of reversed-phase C18 liquid chromatography, electrospray ionization and QTOF mass analyzer according to the previous method (*30*)

### Data Processing Workflow

Metabolomics data of barley bread intervention study (*27*) was processed through converting format by DataBridge (Waters Corporation), pre-processing by MZmine (*31*), and multivariable data analysis by PLS toolbox (*32*). Discriminating metabolites are selected to distinguish whole grain barley and wheat intake. Details are described in the **Appendix.**

### Streamlined Data Processing Workflow

All processes are performed in R statistic language 3.5.3 (*33*) if not specifically stated.

***Data Conversion.*** Raw data was converted from *.RAW* to *.mlXML* by ProteoWizard 3.0 MSConvert (*34*). Scan event was set to *1* to remove lock-mass trace.

***Data Pre-processing.*** Converted data was pre-processed by XCMS 3.4.4 (*35–37*) including following steps: read raw file, peak picking, grouping, adjusting retention time, and peak filling. Detailed parameters are shown in the **Appendix**.

***Calibrate Intra- and Inter- Batch Effects (Libra).*** A new algorithm, *Libra*, is developed to correct intra- and inter- batch effects. *Libra* is available on GitLab (https://gitlab.com/nexs-metabolomics/libra).

*Libra* calibrates intra- and inter- batch effects by linear regression models. Within one batch, for each feature, intensity $(I_1, I_2, I_3, \dots, I_n)$ is fit to injection sequence $(1, 2, 3, \dots, n)$ by least-squares linear regression: $I = f_1(x) = \beta x + \alpha$. *Libra* calibrates the intensities if *p*-value is lower than the defined threshold, *0.1*, by the following steps (**Figure 3** shows the outline):

First, *kick-out the elephants*. A few samples, *elephants*, drive the linear model due to their significantly higher or lower intensities. The *elephants* which have top *15% (elephant fraction)* absolute residuals $(|\varepsilon_i|)$are replaced by their estimated values $(\hat{y}_i)$.

Second, *balance the single libra and pick up the sands*. After replacing *elephants* with the estimated values, a new linear model is fit: $I = f_2(x) = \beta x + \alpha$. Then, the intensity of each sample, $I_k$, is calibrated to $f_2\left(\frac{n}{2}\right) + \varepsilon_k$. The linear calibration might introduce negative values, *sands*, for low-intensity samples. These negative values are replaced by their native values.

Third, *equilibrate multiple Libras*. The centers of all *libras* are averaged to a new center. Different batches are scaled to this new center.

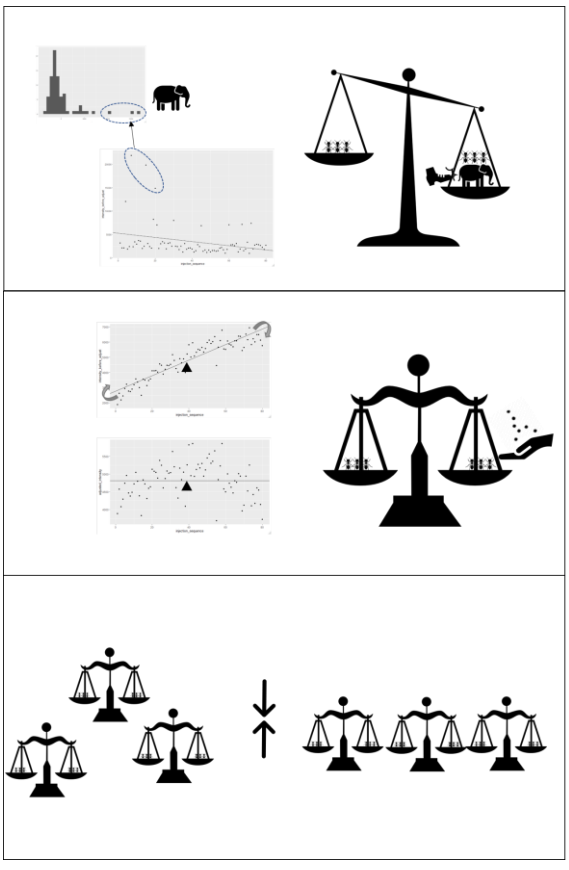


**Figure 3 Diagram of Libra Calibration**

The coefficient of variance (CV%) of pooled samples are calculated to estimate the calibration effects.

***Annotation.*** Metabolites are annotated by CAMERA (1.38.1) and *annotate_kudb*. CAMERA is an R package available on Bioconductor (*38*). It is used to detect pseudo compound groups (PCgroups), isotopes, adducts, multiple charged ions, and cluster ions. Detailed parameters and settings are shown in the appendix. *Annotate_kudb* is a self-developed R function (source code is shown in the Appendix). *Annotate_kudb* matches RT and m/z of features with an in-house database library (*KUDB*). The window for matching was 0.015 (m/z) and 9 s (RT). Sources of annotate_kudb include experiment data and predicted data by PredRet (*39*).

KUDB stores data as *tidy* format (*40*) in a *.csv* file to facilitate computing. One entry stores one compound, including the information of retention time, identity (InChi, InChi key, etc.), and neutral monoisotopic mass, analytical method, and source. The common adducts (**Table 1**) were *expended* before matching.

**Table 1 Common adducts included in annotate_kudb**

| Positive | Negative |
|---|---|
| [M+H]+ | [M-H]- |
| [M+Na]+ | [M+Na-2H]- |
| [M+K]+ | [M+K-2H]- |
| [M-$H_2O$+H]+ | [M+Cl]- |
| [M+FA+Na]+ | [M+FA-H]- |
| | [M+HCOONa-H]- |

FA: formic acid

***Deisotope.*** Isotopes and multiple charged ions are removed based on CAMERA annotation.

***Statistics.*** The t-test is used for univariate data analysis. Benjamini-Yekutieli false discovery rate control method is used to calibrate *p*-values by R stats package. Multivariate data analysis is performed by mixOmics (*41*) including Principle Component Analysis (PCA), Partial Least Squares-Discriminant Analysis (PLS-DA) and sparse Partial Least Squares-Discriminant Analysis (sPLS-DA). The performance of PLS-DA and sPLS-DA is optimized by the 5-fold cross-validation repeated 10 times.

### MATLAB DataSet conversion

Metabolomics data of protein source study (*28*) is processed and stored in a Matlab DataSet object (*42*), which is the standard format for PLS Toolbox to store multivariable data. A self-developed R function *m2r* (source code is shown in the appendix) converts data from Matlab to R.

### Compound Identification

Unknown compounds are identified by comparing retention time (RT), m/z, and $MS^2$ spectra with reference compounds, in-house and public libraries. Level 1 – 4 indicates the level of identification (*43*).

***$MS^2$ experiment.*** Target ions are selected and fragmented by collision energy 14, 28, and 42 eV to acquire $MS^2$ spectra.

***Sitostanol reference compound.*** Sitostanol reference compound is analyzed. The experiment procedures are shown in the **Appendix**.

***Derivatization.*** Sulfates and glucuronates of alkylresorcinol metabolites were synthesized according to an internal SOP (Standard Operation Procedures) by enzymatic reactions[1].

***β-glucuronidase experiment.*** Urine samples were incubated with *β*-glucuronidase for 1.5 h. A positive control was included. Details of the experiment are described in the **Appendix**.

### Other Software

Partition coefficient (logP and ClogP) was predicted by ChemDraw Professional 16.0 (PerkinElmer Information, Inc.). CFM-ID 3.0 (*44*) was used to predict $MS^2$ spectra. BioTransformer (*45*) was used to predict metabolism pathways. Sirius was used to study $MS^2$ spectra.

[1] This was done by Gözde Gürdeniz

# Results

## Systematic Review of Barley Intake Biomarker and Wheat Intake Biomarkers

### Whole Grain Barley

The literature search retrieved 129 records after removing duplicates. **Figure 4** shows the searching process. No biomarker of barley intake has been reported from either human or animal studies.

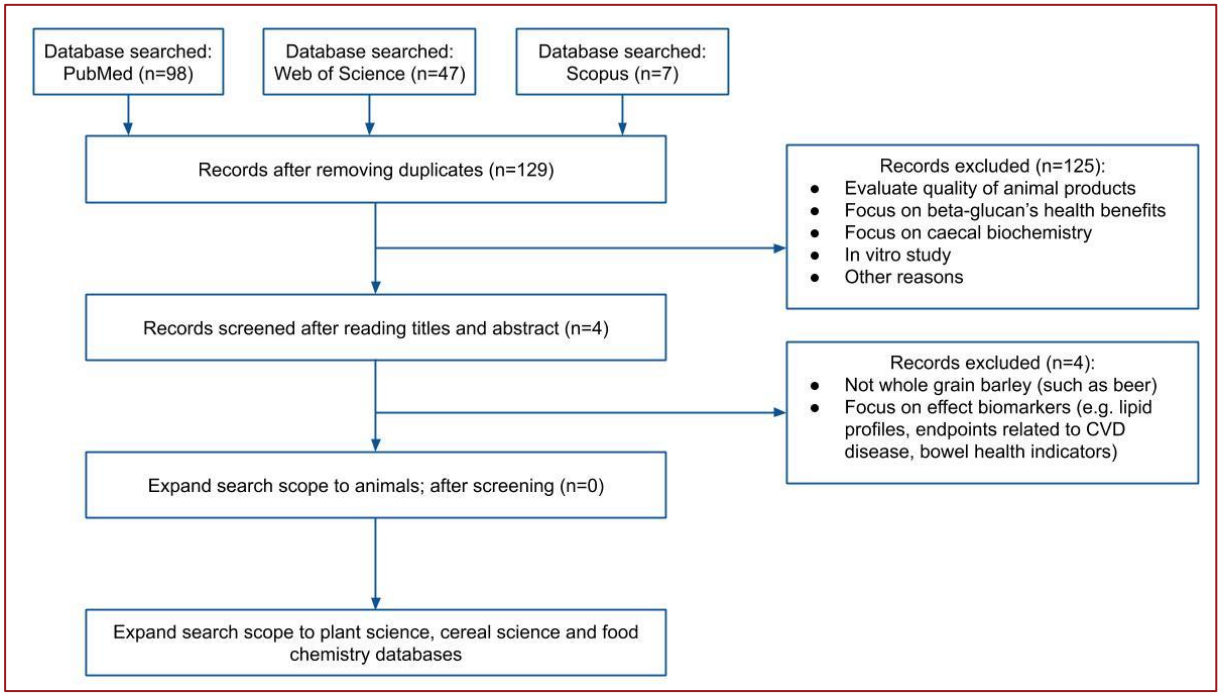



**Figure 4 Diagram of Literature Searching and Screening for Articles of WG Barley Intake Biomarkers**

The term *biomarkers* mentioned in the retrieved papers mostly refer to *effect biomarkers* of barley intake as defined by Dragsted (*10*) and Gao (*9*), such as bowel health indicators (*46*), postprandial glucose and insulin response (*47*), lipid profiles, endpoints related to cardiovascular disease (CVD) risk (*48*), etc. For animal studies, the term, *biomarkers*, mostly refer to the growth of animals or quality indicators of animal-sourced products (*49, 50*), which could also be regarded as *effect biomarkers* in animals. However, for all the studies, the intervention with barley lacked objective biomarkers to monitor the compliance.

Search results in food chemistry, cereal science, and plant science databases show that some compounds are present exclusively in WG barley. These compounds should be investigated as potential biomarkers of barley intake.

**Table 2 Potential Biomarkers for WG Barley Intake**

| No | Candidate biomarker | Formula | Chemical group | Presence in Food | Reference |
|---|---|---|---|---|---|
| 1 | Hordenine | $C_{10}H_{15}NO$ | alkaloid | germinating barley, beer and other plants | (*30*) |
| 2 | Hordatine A | $C_{28}H_{38}N_8O_5$ | alkaloid | only reported in barley | FoodDB (002330) |
| 3 | Hordatine B | $C_{29}H_{40}N_8O_5$ | alkaloid | only reported in barley | FoodDB (002328) |
| 4 | Distichonic acid A | $C_{10}H_{18}N_2O_8$ | gamma amino acids and derivatives | only reported in barley | FoodDB (18164) |
| 5 | Distichonic acid B | $C_{10}H_{18}N_2O_8$ | gamma amino acids and derivatives | only reported in barley | FoodDB (018165) |
| 6 | 14,16-Nona cosanedione | $C_{29}H_{56}O_2$ | ketone | only reported in barley | FoodDB (013891) |
| 7 | N-Norgramine | $C_{10}H_{12}N_2$ | indole | only reported in barley | FoodDB (017815) |

### Whole Grain Wheat

The literature search retrieved 312 references after removing duplicates. Some articles were found from the references of searched results. **Figure 5** shows the searching process. The results (**Table 3**) includ four intervention studies and three observational studies.

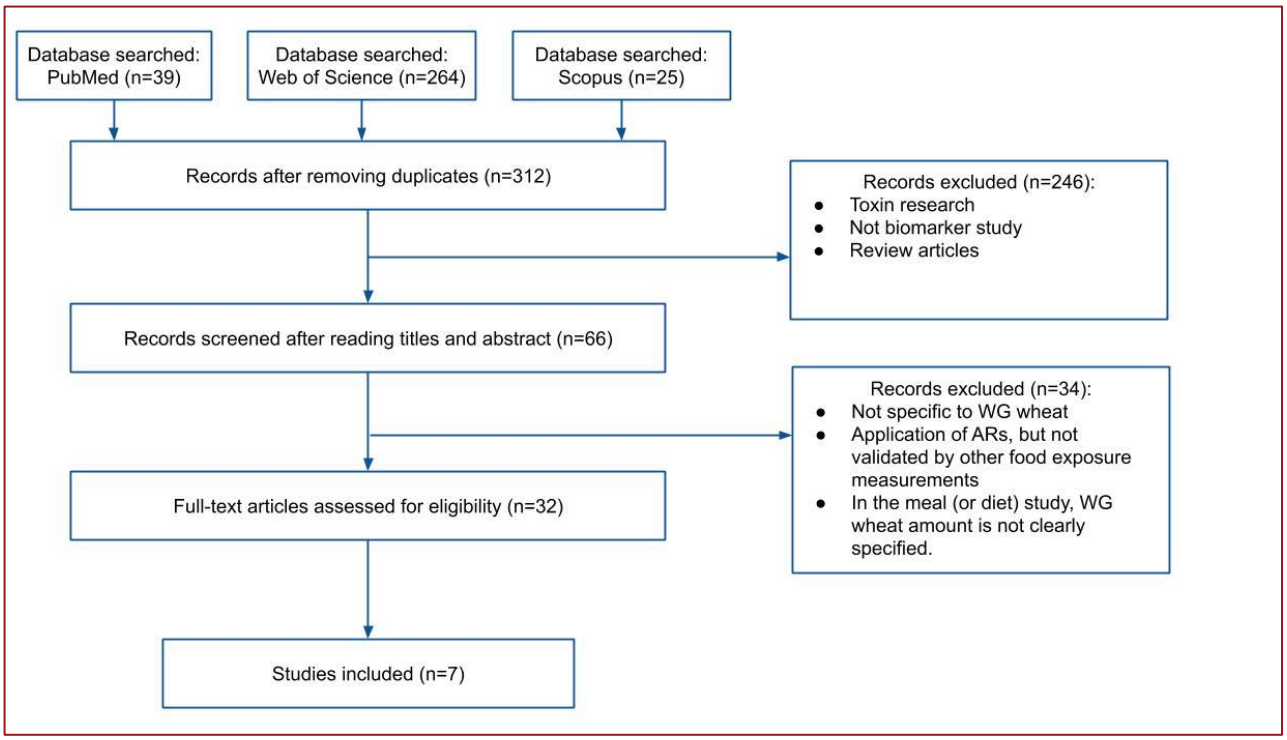



**Figure 5 Diagram of Literature Searching and Screening for Articles of WG Wheat Intake Biomarkers**

**Table 3 Biomarkers for WG Wheat Intake**

| Food items | No. subjects | Study design | Sample type | Analytical method | Candidate biomarker(s) | Identifier | Reference |
|---|---|---|---|---|---|---|---|
| WGs | 2845 | Observational | Fasting and non-fasting plasma | GC-MS | AR-homologue Ratio of C17:0/C21:0 | HMDB0038530<br>HMDB0031035 | (*51*) |
| WG wheat<br>WG rye | 73 | Observational | Fasting plasma | GC-MS | Total ARs (C17:0,C19:0, C21:0,C23:0,C25:0)<br>AR-homologue Ratio of C17:0/C21:0 | HMDB0038530<br>HMDB0030956<br>HMDB0031035<br>HMDB0038524<br>HMDB0038485 | (*52*) |
| WG wheat<br>WG rye | 39 | Intervention (crossover) | Blood | GC-MS | AR-homologue Ratio of C17:0/C21:0 | HMDB0038530<br>HMDB0031035 | (*53*) |
| WG wheat<br>WG rye | 15 | Intervention (cross-over) | Plasma and serum enterolactone | GC-MS | AR-homologue Ratio of C17:0/C21:0 | HMDB0038530<br>HMDB0031035 | (*54*) |
| WG wheat<br>3 or 6 servings | 19 | Intervention (cross-over, dose-response) | Fasting plasma | GC-MS | AR C19:0, C21:0, and C23:0 | HMDB0030956<br>HMDB0031035<br>HMDB0038524 | (*55*) |
| | | | Urine | HPLC-ECD | 3,5-DHBA<br>3,5-DHPPA | HMDB0013677<br>HMDB0125533 | |
| WGs | 40 | Observational | Spot urine | GC-MS | 3,5-DHCA<br>3,5-DHBA glycine<br>3,5-DHPPTA | HMDB0032131<br>HMDB0126654<br>HMDB0125533 | (*56*) |
| WGs | 104 | Observational | Spot urine | GC-MS | 3,5-DHBA<br>3,5-DHPPA | HMDB0013677<br>HMDB0125533 | (*57*) |

***Total Alkylresorcinols and the AR-homologue Ratio C17:0/C21:0***

Combining total Alkylresorcinols (ARs) and AR-homologue ratio C17:0/C21:0 in plasma can potentially be used as biomarkers to indicate WG wheat intake.

Within commonly consumed plant-based foods, ARs are present in high concentration exclusively in the bran part of wheat and rye. AR-homologue ratio C17:0/C21:0 was first reported by cereal scientists in 2004 to distinguish WG rye and wheat grains (*58*). Rye has the AR-homologue C17:0/C21:0 ratio close to 1.0, while for wheat it is around 0.1, for durum wheat even around 0.01. This marker was therefore proposed by nutritionists to distinguish WG rye and wheat intake. In 2005, Linko (*53*) first investigated this biomarker in human plasma by an intervention study. The results showed that AR-homologue ratio C17:0/C21:0 can distinguish WG wheat and rye diets in healthy postmenopausal women. For a rye-dominated diet, the ratio was 0.84, and for a WG wheat-dominated diet, the ratio was around 0.53. In 2007, Linko-Parvinen validated this marker in healthy adults by an intervention study (*54*). In plasma, the value was 0.1 after WG wheat intake while it was 0.6 after WG rye intake. In erythrocytes, the value was 0.06 and 0.33, respectively, after WG wheat and rye intake. This study also implied that human plasma lipoproteins could transport ARs.

However, the AR-homologue ratio C17:0/C21:0 was unable to differentiate a WG diet from a refined cereal diet but the total plasma AR concentration can distinguish WG and refined diet (*59*).

The EPIC[2] cohort study (*51*) further indicated the usefulness of this marker. This study measured plasma total ARs and the AR-homologue ratio C17:0/C21:0 in subjects from 10 European countries. The result showed that Greek, Italian, Dutch, and UK participants for whom the diet was dominated by wheat, had low C17:0/C21:0 ratios in plasma. In contrast, the Danish, German and Swedish subjects had high C17:0/C21:0 ratios. French and Norwegian subjects had intermediate ratios. The result is in line with descriptive studies on the intake of WG wheat in different countries.

***Alkylresorcinols (C19:0, C21:0, C23:0)***

ARs (C19:0, C21:0, C23:0) constitutes 85% of alkylresorcinols in whole grain wheat (*55*). In a dose-response crossover intervention study, the combination of these three ARs correlates with WG wheat intake. After three and six daily servings of WG wheat, plasma ARs (C19:0, C21:0, C23:0) were ≥ 3.1-fold higher ($p < 0.001$) than run in and washout when adjusted for sex, age, and energy intake (*55*).

***Urinary AR Metabolites***

In a dose-response intervention study, urinary AR metabolites 3,5-dihydroxylbenzoic acid (3,5-DHBA), 3,5-dihydroxyphenoylpropanoic acid (3,5-DHPPA), and the sums significantly increased after whole grain wheat intake ($P < 0.001$). The dose-response is also significant (six vs three servings, $P = 0.004$) (*55*).

In another observational study, 3,5-dihydroxycinamic acid (3,5-DHCA), 3,5-dihydroxyphenylpentanoic acid (3,5-DHPPA), and 3,5-dihydroxybenzamido acetic acid (3,5-DHBA glycine) showed moderate to excellent medium-term (2 wk) reproducibility (intra-class correlation coefficient = 0.35–0.67) (*56*). However, the long-term reproducibility is poor for DHBA, modest for DHPPA (*57*).

***Other Potential Biomarkers***

The search results also include some *food compound intake biomarkers (FCIBs)* (*9*) such as phenolic compounds (*60*), benzoxazinoids (BXOs) (*61*, *62*), phytoestrogen (*63*), phytosterol and lignan (*64*) and *effect biomarkers* such as microbial metabolites (*62*). These compounds are not exclusively present in WG wheat. Therefore, they cannot specifically indicate WG wheat intake. These results have been summarized in the **Appendix**.

One study proposed that a panel of metabolites consisting of 7 AR metabolites, 5 BXO metabolites, and 5 phenolic acid derivatives can objectively assess WG wheat intake (*64*). Because the

[2] European Prospective Investigation into Cancer and Nutrition

concentrations of the precursor vary in different types of whole grain, a combination of their metabolites could potentially indicate intake of different types of whole grain. This proposal needs to be further validated.

## Data Processing

### Pre-processing and Annotation

**Table 4 Data Pre-processing Results**

| Biospecimen | Mode | Detected features | After Deisotope (PCgroups) | Annotated by KuExp | Annotated by PredRet | Annotated PCgroups |
|---|---|---|---|---|---|---|
| Urine | positive | 1508 | 1400 (1056) | 121 | 168 | 196 |
| | negative | 2473 | 2267 (1860) | 143 | 275 | 326 |
| Plasma | positive | 1643 | 1381 (1144) | 78 | 56 | 101 |
| | negative | 1278 | 1088 (861) | 74 | 65 | 99 |

*Annotate_kudb* can roughly suggest identities of 100-300 metabolites for one metabolomics analysis.

### Libra Calibration

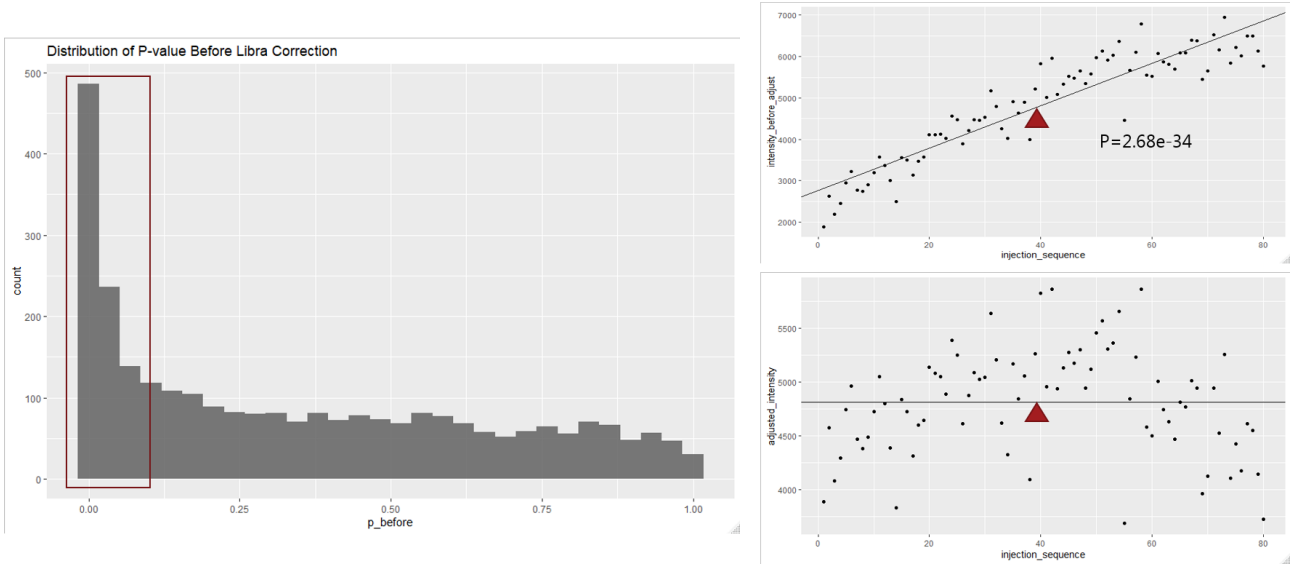



**Figure 6 Distributions of *p*-value before Libra calibration (Left); one example of intensities before-calibration (Top Right) and after-calibration (Bottom Right)**

Before *Libra* calibration, some features are highly correlated with injection sequency. **Figure 6** shows the distributions of p-value and an calibration example. After *Libra* calibration, most significantly correlated features are calibrated.

*Libra* calibration does not change CV% of pooled samples. **Figure 7** shows the distribution of the CV% after and before *Libra* calibration. The median CV% of pooled samples slightly decreases from 35.07 to 34.92 after the calibration.

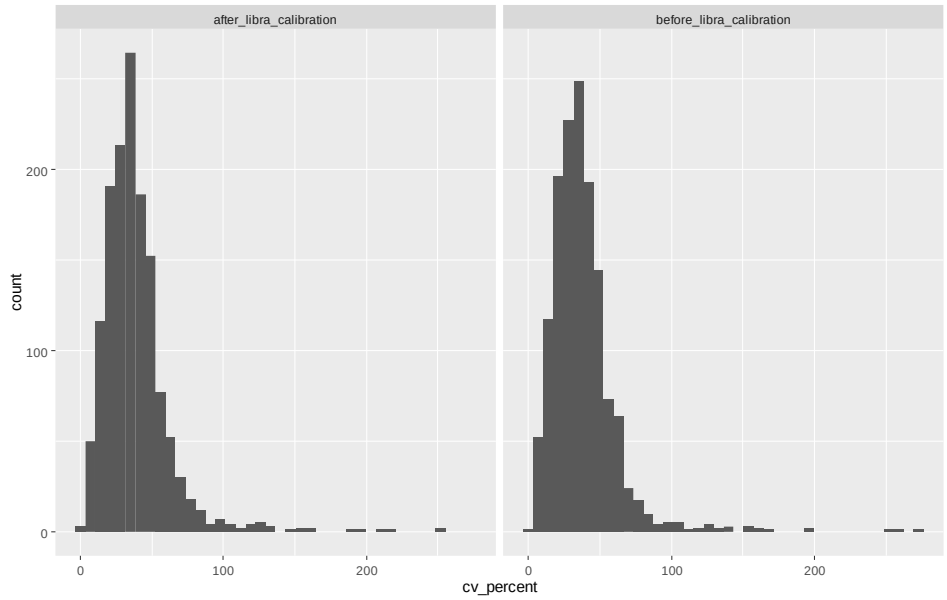



**Figure 7 CV% of pooled sample before (Left) and after (Right) Libra calibration**

### Multivariate Data Analysis

***Plasma.*** PLS-DA modelling cannot distinguish plasma samples after the barley or wheat intervention (**Figure 8**). The t-test result shows that only 18 out of 1088 metabolites are significantly different ($p<0.05$). After multiple comparison adjustment, none of the metabolites changes significantly after whole grain intake. However, PLS-DA modeling can distinguish pooled samples from biological samples. The separation is clearer when the background is calculated (**Figure 9**).


**Figure 8 PLSDA Score Plot of Plasma Samples (PC 1-2)**

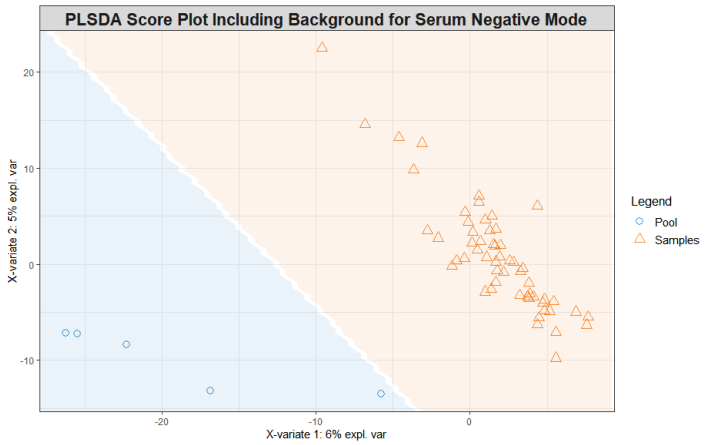



**Figure 9 PLSDA Score Plot of Plasma Samples (Pooled vs Biological samples)**

***Urine.*** The score plot of PLS-DA modelling (**Figure 10**) shows that urine samples can be separated after the barley intake in the negative mode. Meanwhile, pool samples locate tightly in the mid of figure indicating high qualities of data. Compared with plasma samples, more variances of urine samples can be explained from principle components (**Figure 11**).

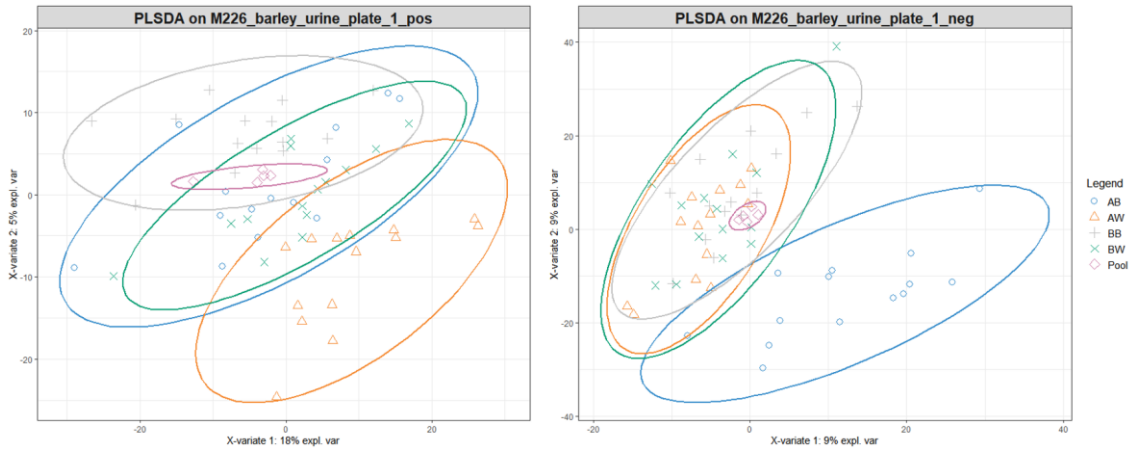



**Figure 10 PLSDA Score Plot of Urine Samples**

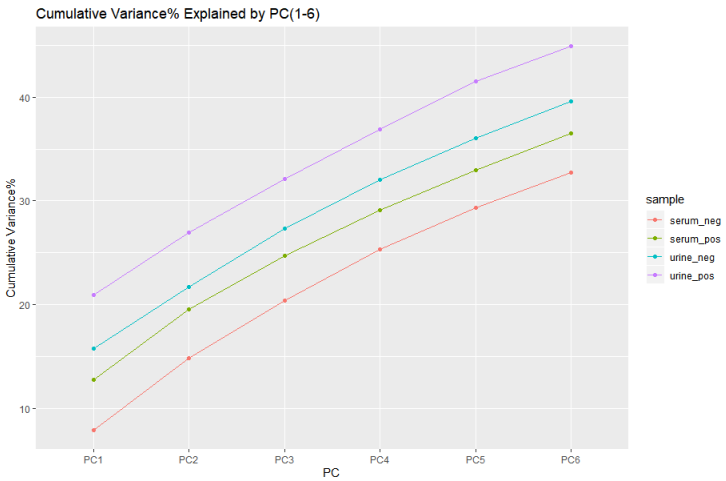



**Figure 11 Cumulative Variance% Explained by PC (1-6)**

## Biomarkers of Whole Grain Barley Intake

### Summary

**Table 5 Biomarkers of Whole Grain Barley Intake**

| No. | m/z | Neutral formula | RT (Quat) | RT (Bi) | MS/MS And Adducts | Annotation | Suggested Compound |
|---|---|---|---|---|---|---|---|
| 1 | 216.99 | $C_7H_6O_6S$ | NA | 2.94 | --- | [M-H]- | 4-hydroxybenzoic acid-4-sulphate[3] |
| 2 | 517.30 | $C_{30}H_{46}O_7$ | 6.48 | 3.99 | 113.023<br>341.267<br>399.272<br>499.289<br>517.303 | [M-glucuronate-H]-<br><br>[M-$H_2O$-H]-<br>[M-H]- | Glucuronate of Unknown I[4] |
| 3 | 341.26 | $C_{20}H_{38}O_4$ | 0.88 | --- | --- | [M-H]- | Unknown I[4] |
| 4 | 501.30 | $C_{30}H_{46}O_6$ | 6.7 | 4.22 | 113.0110<br>157.1222<br>171.1372<br>325.2739<br>383.2629<br>483.2741<br>501.3054 | [M-glucuronate-H]-<br><br>[M-$H_2O$-H]-<br>[M-H]- | Glucuronate of Unknown II[4] |
| 5 | 231.08 | C10H16O6 | NA | 2.10 | --- | [M-H]- | Unknown[4] |
| 6 | 387.16 | $C_{18}H_{28}O_9$ | NA | 3.78 | 387.16<br>775.34 | [M-H]-<br>[2M-H]- | Unknown[4] |
| 7 | 537.33 | Unknown | 6.59 | 3.84 | --- | [M-H]- | Unknown glucuronate[4] |
| 8 | 291.26 | Unknown | 6.71 | 4.21 | --- | [M-H]- | Unknown sulfate[4] |
| * Molecular formula corresponds to the neutral compound.<br>Superscript notations represent level of identifications. | | | | | | | |

### Glucuronates and *β*-glucuronidase Experiments

Ion 501.3054 and 517.3030 could be glucuronates. Both ions showed the loss of 176 in $MS^2$ spectra (**Figure 20** in the **Appendix**), which is characteristic to glucuronate groups (*65*).

*β*-glucuronidase hydrolysed both glucuronate ions. However, the hydrolysates cannot provide more information of the structure. After the treatment, the intensities of these two ions decreased. However, the expected ion 325.27 did not spike. Another expected ion 341.26 spiked on the same day of *β*-glucuronidase treatment. However, after one week storage in -20 ºC freezer, the intensity of this ion 341.26 decreased. Its intensity is not high enough for $MS^2$ experiment.

Ion 537.33 could be a glucuronate as well, as indicated by its decreased intensities after the *β*-glucuronidase treatment.

### Sulfur-Containing Compound

The ion 291.26 possesses a sulfur atom as indicated by its isotopic pattern (**Figure 12**), but another metabolite interferes the structure elucidation. This interfering metabolite elutes close to the biomarker and has a similar m/z: 291.23 (**Figure 13**).

In $MS^2$ experiment, VION has a better resolution of the interfering metabolite. The interfering ion was selected for fragmentation.

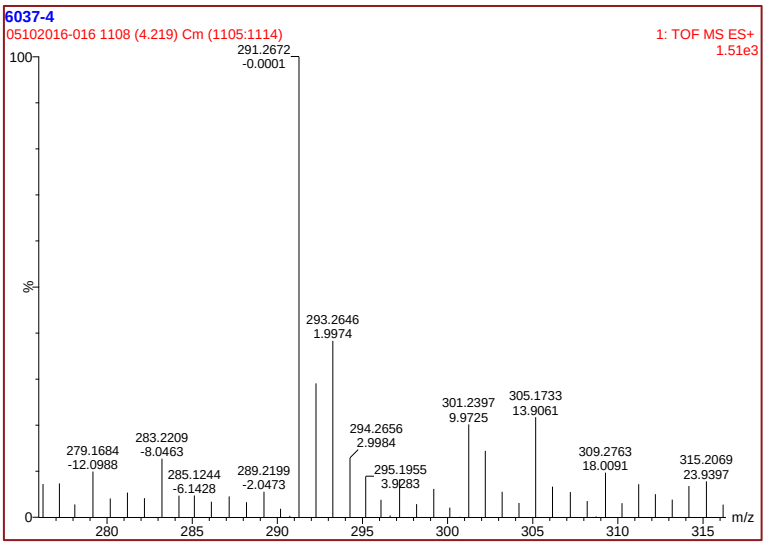



**Figure 12 Isotopic Pattern of Ion 291.26**

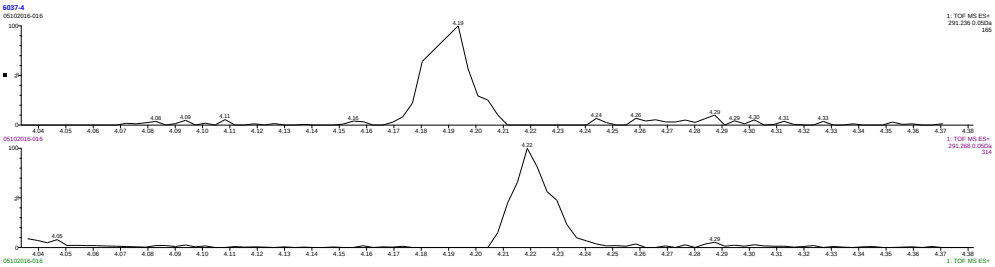


**Figure 13 EIC of Endogenous Metabolite (Top) and Barley Intake Biomarker (Bottom)**

### Comparisons with Protein Source Study and Beer Study

Some discriminating metabolites of whole grain barley intake are also detected in protein source study, six from positive mode, and nine from negative mode. However, none of these features has reasonable excretion profiles for barleyotto intake.

Discriminating metabolites of whole grain barley does not overlap with beer intake biomarkers.

## Biomarkers of Whole Grain Wheat Intake

**Table 6 Biomarkers of Whole Grain Wheat Intake**

| No. | m/z | Molecular Formula* | RT (Quat) | RT (Bi) | MS/MS and Adducts | Annotation | Suggested Compound |
|---|---|---|---|---|---|---|---|
| 1 | 329.05 | $C_{13}H_{14}O_{10}$ | 2.10 | 0.97 | 113.02<br>153.02<br>175.02 | [3,5-dihydroxybenzoic acid - H]$^-$<br>[M-H]$^-$ | 3,5-DHBA glucuronate[1] |
| 2 | 329.05 | $C_{13}H_{14}O_{10}$ | 3.96 | 1.12 | 109.04<br>135.06<br>153.03<br>329.08 | [3,5-dihydroxybenzoic acid - H]$^-$<br><br>[3,5-dihydroxybenzene - H]$^-$<br>[M-H]$^-$ | Glucuronyl dihydroxybenzoate[2] |
| 3 | 232.97 | $C_7H_6O_7S$ | 4.07 | 1.37 | 79.96<br>96.96<br>109.03<br>123.05<br>153.01<br>215.09 | [$SO_3$]$^-$<br><br>[3,5-dihydroxybenzoic acid - H]$^-$<br><br>[M-H-$SO_3$]$^-$<br>[M-$H_2O$-H]$^-$ | 3,5-DHBA sulfate[2] |
| 4 | 210.04 | $C_9H_9NO_5$ | 4.64 | 1.48 | --- | [M-H]$^-$ | 3,5-DHBA glycine[2] |
| 5 | 153.01 | $C_7H_6O_4$ | 1.18 | 1.88 | 109.03<br>153.01 | [3,5-dihydroxybenzoic acid - H]$^-$<br>[M-H]$^-$ | 3,5-DHBA[2] |
| 8 | 357.08 | $C_{15}H_{18}O_{10}$ | 4.08 | 1.91 | 113.03<br>137.07<br>181.07<br>357.08 | [M-glucuronate-H]$^-$<br>[M-H]$^-$ | 3,5-DHPPA Glucuronate[2] |
| 9 | 359.09 | | | | 165.05<br>183.07<br>359.09<br>376.12 | [M+H-glucuronate-$H_2O$]$^+$<br>[M+ glucuronate H-]$^+$<br>[M+H]$^+$<br>[M+$NH_4$]$^+$ | |
| 10 | 261.00 | $C_9H_{10}O_7S$ | 4.32 | 2.10 | 261.00<br>523.02<br>777.14 | [M-H]$^-$<br>[2M-H]$^-$<br>Unknown adduct | 3,5-DHPPA sulfate[2] |
| 11 | 263.02 | $C_9H_{12}O_7S$ | 5.48 | 3.31 | 79.96<br>153.03<br>168.06<br>183.08<br>248.02<br>263.02 | [$SO_3$]-<br>[M-$CH_3O_4S$-H]<br><br>[M-H-$SO_3$]-<br>[M-$CH_3$-H]-<br>[M-H]$^-$ | DHMBA sulfate (dihydroxy-5-methoxybenzoic acid sulfate)[3] |
| 12 | 385.10 | $C_{17}H_{22}O_{10}$ | 5.23 | 3.31 | --- | [M-H]$^-$ | 3,5-DHPPTA glucuronate[3] |
| 13 | 289.03 | $C_{11}H_{14}O_7S$ | 5.18 | 3.37 | --- | [M-H]$^-$ | 3,5-DHPPTA sulfate[3] |
| 14 | 340.06 | $C_{14}H_{15}NO9$ | --- | 2.96 | 326.08<br>340.06 | [M-$CH_2$-H]$^-$<br>[M-H]$^-$ | HBOA glucuronide[3] |
| 15 | 359.12 | --- | --- | 3.1 | --- | [M-H]$^-$ | Unknown[4] |
| 16 | 623.21 | --- | --- | 3.11 | --- | [M-H]$^-$ | Unknown[4] |
| 17 | 548.30 | --- | --- | 3.72 | --- | [M-H]$^-$ | Unknown[4] |
| * Molecular formula corresponds to the neutral compound.<br>Superscript notations represent level of identifications. | | | | | | | |

# Discussions

## The Ambiguity in Using the Term Biomarker

The ambiguity in using the term *biomarker* in the publications is common in the scientific articles retrieved. Most papers mention the term, *biomarker*. However, *biomarker* refers to different concepts in different contexts, e.g. *food intake biomarkers*, *food compound intake biomarkers* and *effect biomarkers*. In the literature search, it has been difficult to quickly extract what was meant by *biomarkers* before reading the full text.

Dragsted (*10*) and Gao (*9*) proposed the new ontology and classification schemes for the use of the term, *biomarker*. The awareness and implementations of the new ontology and classification schema might relief this problem in the future.

## Alkylresorcinol (AR) as Biomarkers for Whole Grain Wheat Intake

### Urinary AR metabolites

Ten of the identified biomarkers are AR metabolites. Barley intervention study identified most of the AR metabolites in the proposed AR metabolism pathway (**Figure 14**). Barley intervention study used 24-h pooled urine samples, which accumulates a broad range of metabolites after the intake.

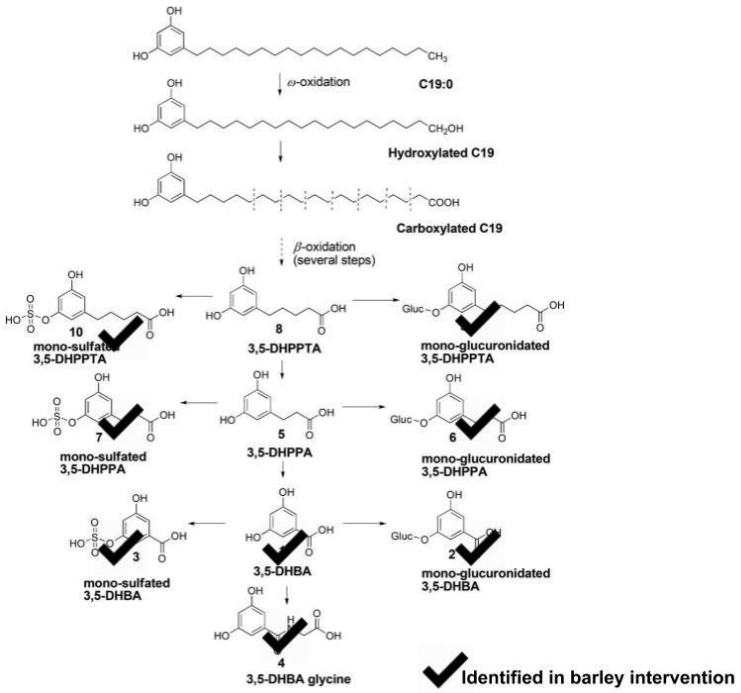



**Figure 14 Detected AR Metabolites. Adapted from *Zhu et al., J Nutr, 2014, 144(2).***

This study also identified two novel AR metabolites. Glucuronyl dihydroxybenzoate is an isomer of 3,5-DHBA glucuronate (**Figure 15**). Glucuronates are produced during phase II metabolism. DHMBA sulfate identified in barley intervention study has a different retention time from Sysdiet study. Therefore, they could be isomers.

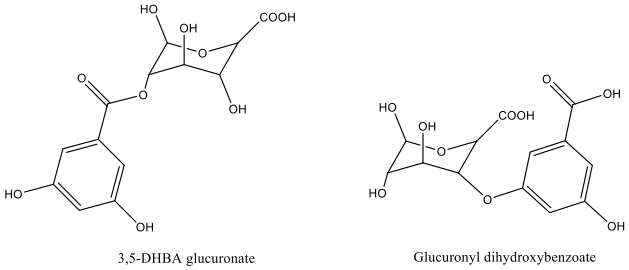



**Figure 15 Structure of 3,5-DHBA Glucuronate and Glucuronyl Dihydroxybenzoate**

Barley intervention is also one of the few studies that detects urinary AR metabolites by LC-MS. Most of the studies used GC-MS to detect urinary AR metabolites.

However, urinary AR metabolites might not be specific to whole grain wheat intake in countries that whole grain rye and wheat are both consumed, which includes Sweden, Denmark, Finland, and Germany. Unlike plasma samples, total AR can indicate the total intake of whole grain rye and wheat. The homologue ratio of AR C17:0/C21:0 can distinguish whole grain rye and wheat. Urinary AR metabolites cannot be traced back to their precursors, because all homologues undergo the same metabolism pathway. In countries that whole grain wheat is the only source of alkylresorcinol (for example, UK and USA), AR metabolites might be good biomarkers to measure whole grain wheat intake. However, this need to be further validated.

#### Plasma AR

Meanwhile, barley intervention study cannot detect intact ARs from fasting plasma. The half-life-time ($t_{1/2}$) for intact ARs are around 5 h (*66*). Therefore, the intact ARs could be fully metabolized and excreted, or partially metabolized to the concentration below the limit of detection.

### Identification of The Interfering Ion: Androsterone

The interfering ion is identified as androsterone with level II confidence. The $MS^2$ spectra (**Figure 22**) show the similarities with sitostanol, the plant-originated steroid alcohol. CFM-ID predicted spectra also show a similar pattern.

Androsterone is an endogenous steroid hormone, neurosteroid, and putative pheromone (*67*). The elevated urinary concentration of androsterone could be associated with several diseases, such as hirsutism (*68*). Androsterone has also been proposed as a potential biomarker to diagnose

schizophrenia together with other endogenous steroids (*69*). In doping test, the ratio of urinary androsterone and etiocholanolone can be used to monitor urine manipulation.

**Figure 16** shows that in barley intervention study, the concentration of androsterone does not seem to vary before and after the intervention, but the concentration seems slightly higher in male than female.

Androsterone is also detectable in Fatmed, Sysdiet, Coffee, and Seaweed study. However, the distribution in different genders is not investigated due to the lack of information.

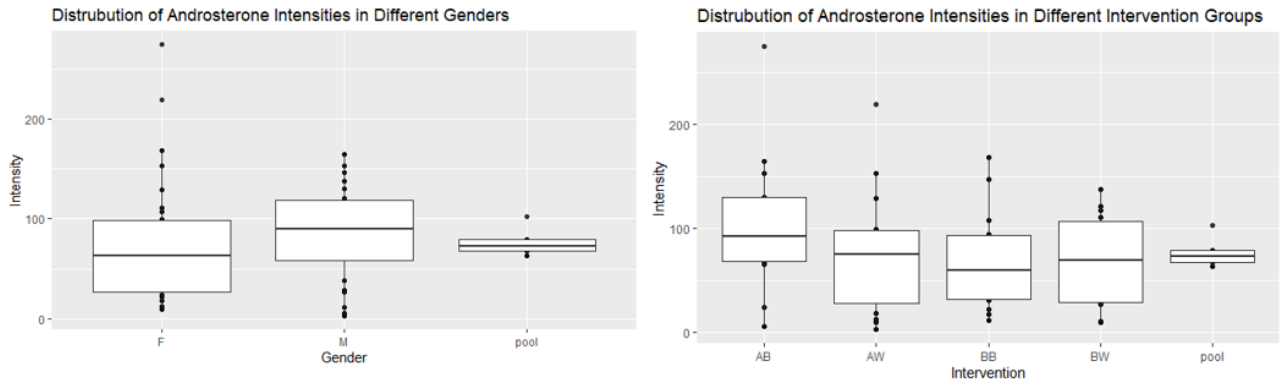



**Figure 16 Distribution of Androsterone Intensities in Different Genders (Left, F=Female, M=Male) and Different Intervention Groups (AB=After Barley, AW=After Wheat, BB= Before Barley, BW= Before Wheat)**

## Data Processing Workflow

This thesis experimentally implemented a data processing workflow to streamline metabolomics data processing.

### *Libra* Calibration

*Libra* calibrates intra- and inter- batch effects by a conservative approach. *Libra* only takes actions to the strong linear tendencies (intensity *vs* injection sequence). The philosophy behind is '*don't be evil'*. When the source of variation is uncertain, it is better not to take actions instead of calibrating the data with the risk of introducing the artifacts.

MS is a less reproducible technique by its nature. The ion trajectories in the flying tube is hard to control and predict. Meanwhile, the instruments are designed according to different principles, which makes it difficult to unify a strategy to calibrate batch effects.

*Libra* only slightly changes CV% of pooled samples, even though strong linear relationships are calibrated. CV% is often used to estimate the data qualities and the performance of signal calibration tools. However, *Libra* does not change CV% much because *Libra* does not take actions to *elephants*. *Elephants* have high leverages for the variance. However, it might not be a good idea

to calibrate the signals of *elephants*. For example, **Figure 17** shows three *elephants* for this metabolite. *Elephants* are *kicked-out* before calibrating the signals for the rest (They are not removed but replaced with the linear estimated values. Therefore, they do not leverage the linear model). When the rest of the data is calibrated. These *elephants* are put back. Therefore, the variance induced by the elephants are still calculated in CV%. *Libra* does not want to take actions to *elephants*, because they might be unique features. In this circumstance, the LOESS model will smooth the intensities of elephants to QC samples. By doing so, the unique features might be lost. Therefore, unlike other calibration tools, which can significantly reduce variances, *Libra* only slightly reduces the variance that are certainly from the batch effects. Instead of targeting at reducing variance, *Libra* respects the naturally occurring variance and only act on variance with known effects.

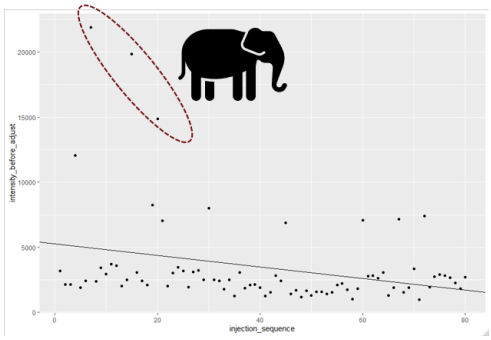



**Figure 17 Elephants in the data**

*Libra* does not use quality control (QC) samples to calibrate the batch effects. Therefore, QC samples can still objectively reflect the data quality. If QC samples are used for both purposes, calibrating the batch effects, and estimating the data quality. In such a circumstance, QC samples lose the objectiveness to estimate the data quality because they play the dual roles both as the *athlete* and *referee*.

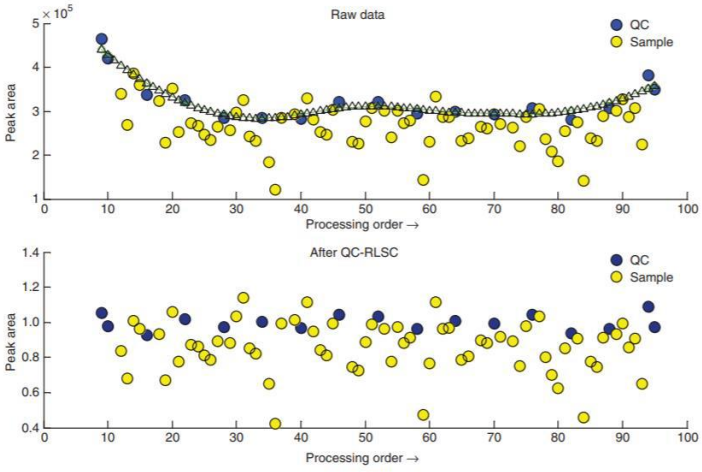



**Figure 18 Intra-batch effects calibrated by LOESS model, adapted from *Dunn et al., Nat Protoc. 2011***

In two scenarios, *Libra* might be the more appropriate calibration tool. First, when the analysis includes too few QC samples, the LOESS might introduce artifacts when smoothing the intensities. The good smooth effect by LOESS can be achieved when 20% of samples are QC samples (*70*) (**Figure 18**). Second, when there are a lot of unique features, for example, the metabolome from plant, food, or environment samples.

#### Annotate_kudb

*Annotate_kudb* can annotate metabolites by matching the m/z and retention time with the database. This annotation method can reduce repeating identifying known metabolites. But, *annotate_kudb* also has its inherent limitations.

First, *annotate_kudb* cannot identify metabolites that do not exist in the database. Though it can increase the coverage by establishing more comprehensive database, *annotate_kudb* can only identify *known unknowns*, but not *unknown unknowns*. Therefore, *annotate_kudb* can only *shoot* the targets in the competitions like a sportsman, but it can never become a hunter to feed himself. Another shortcoming is that, *annotate_kudb* still relies on the expert's manual selections. One reason is, the *expanding algorithm* is not tailored for specific compounds or compound groups. Some adducts or fragments are more abundant due to their chemical properties and the analytical conditions (solvents, ionization, and source parameters etc.). Meanwhile *annotate_kudb* does not consider the physiological feasibility. Some metabolites might not be feasible to be present in the type of human biofluid to be analyzed.

#### Comparisons of New and Old Workflows

**Table 7** compares the new workflow with the old one. The primary advantage of the new data workflow is that it incorporates all data processing steps into the R programming language. This streamlines data processing by minimalizing data format conversion. Besides, the data processing is more reproducible and shareable. Meanwhile, all packages and software that are used in the new workflow are open-sourced and freely available.

However, the new data workflows also have the shortcomings. The new data processing workflow has a steep learning curve because it lacks sufficient visualization functions and requires programming skills. Experienced users might favor the new data workflow when attacking large data analysis tasks. To synergically use both workflows, more bioinformatics tools such as *m2r* can be developed to bridge different workflows.

**Table 7 Comparisons of Two Data Processing Workflows**

| | Old data processing workflow | | | New data processing workflow | | |
|---|---|---|---|---|---|---|
| | Method /tools | Strength | Weakness | Method/tools | Strength | Weakness |
| Data format conversion | DataBridge | --- | - Prone to errors (when each batch consists more than 120 samples, any samples after 120 are dropped without any notifications.)<br>- only be able to only convert waters *.raw* to *.cdf*<br>- Proprietary software<br>- Only available in Windows OS | ProteoWizard MSconvert | - Robust<br>- Able to convert from multiple formats to multiple formats<br>- Open-source and freely-available<br>- Available on all platforms<br>- User friendly interface | --- |
| Data preprocessing | MZmine | - Open-source<br>- Good visualisation functions<br>- Informative manuals and documentations<br>- Easy-to-use with graphic user interface | - High demand of hardware (at least 32 Gb ROM)<br>- Programmed in Java. Difficult to incorporate into the workflows seamlessly | R package XCMS | - Open-source<br>- Easy to incorporate into workflow seamlessly | - Poor visualisation functions<br>- Hard-to-learn |
| | MATLAB | --- | - Proprietary software<br>- Detailed official documents<br>- Rich training resources | | | |
| Annotation and identification | Manual Selection | - The capability of identifying novel compounds | - Time-consuming<br>- Needs a lot of experience and knowledge | R package CAMERA | - Open-source | - Only compatible with XCMS |
| | | | | R function *annotate_kudb* | - Directly suggest level-I identifications | - Unable to identify new compounds<br>- Not delivered as a packaged. Difficult to be re-used by other users.<br>- An extensive database is needed |
| Statistics | PLS Toolbox | - Powerful multivariable data analysis<br>- Good customer support<br>- User friendly interface | - Programmed in proprietary language (MATLAB)<br>- Proprietary software | R package stats<br><br>R package mixOmics | - Open-source<br>- In active development<br>- Contributions from other peer-users | --- |

# Conclusions

Whole grain barley intake can be indicated by 4-hydroxybenzoic acid-4-sulphate and other metabolites. In the areas where whole grain wheat is not the only alkylresorcinol source, the combination of total plasma AR and AR homologue ratio C17:0/C21:0 can indicate whole grain wheat intake. While in the areas where whole grain wheat is the only alkylresorcinol source, plasma alkylresorcinol, AR (C19:0, C21:0, and C23:0), and AR urinary metabolites can indicate whole grain wheat intake.

A streamlined data processing workflow is implemented by incorporating free-available tools and self-developed functions into a unified programming language, R.

# Perspectives

The identification of barley intake biomarkers is hurdled by the limited knowledge of barley phytochemicals. Therefore, the systematic research of barley phytochemicals could facilitate the discovery of novel intake biomarkers.

More efforts should be devoted to the development of metabolomics bioinformatics tools, especially in harmonizing different bioinformatics tools to streamline the data processing workflow.

*Annotate_kudb* can be improved from following angles. First, the *expanding* algorithm can be improved to tailor the adducts or fragments for different groups of compounds. For example, long chain fatty acids are most abundantly present as formic acid adducts. Therefore, formic acid adducts of fatty acids should be the main target to *shoot* when identifying fatty acids. In order to significantly improve the annotation performance, the best staring point might be building a systematic database by using a machine-based algorithm to investigate the spectra. Second, more computational strategies can be integrated to expand the metabolite coverage, such as PredRet. Third, the public database, such as HMDB and Metlin might be included as the source for matching.

Batch effect calibration is a critical step to guarantee data qualities (*71*). However, few bioinformatics tools and research have been devoted to this area. *Libra* could be a potentially useful tool. However, *Libra* should be systematically tested with data and the performance should be compared with tools.

# Acknowledgements

I sincerely acknowledge my dear supervisors Lars Ove Dragsted, and Gözde Gürdeniz for their guidance during my thesis. Lars impresses me as a good teacher. He can always explain the complex concepts in a vivid and understandable manner. And, Gözde offers her babysitting-styled supervision and helps me in every detail. I also acknowledge Susanne Bügel offering me the chance to analyze the data and the examiner Ashfaq Ali for assessing the thesis.

I acknowledge my colleagues from Metabolomics group, Jan, Giorgia, Xiaomin, Muyao, Natalia, Ceyda, Catalina, Sarah, Cecillie, Cristian, and Henrik. All of them offered their help and support. Without them, I cannot have such a happy experience during my thesis. I really appreciate the companion of my friend Kristina. We have sat in the same office and experienced such a wonderful year with metabolomics!

The researchers from the Section of Preventive and Clinical Nutrition are also acknowledged. Steen, and Inge share the knowledge of fiber and cardiovascular disease. I really acknowledge my friend Simon. Besides the help in science, he also inspires me in several aspects of life. And my friend Kåre, thank you. I also benefited from the efficient administrative work by Claude, Geske, Krystyna and Randy.

I also acknowledge my family, home friends and my girlfriend who live 6-7 hours ahead of me but keep encouraging and supporting me.

# Appendix

## Data Processing Parameters

Raw data was first converted to *.cdf* format by DataBridge (Waters). Data was pre-processed by MZmine (2.31) by the following steps: peak detection, deisotoping, alignment, and gap filling. Positive mode and negative mode were pre-processed separately. In the end, the detected features, including mass to charge ratio (m/z), retention time (rt) and intensities were output as *.csv* files.

Data analysis was performed in Matlab R2018a. Data was analyzed by Principle Component Analysis (PCA), and Partial Least Squares Discriminant Analysis (PLS-DA) in PLS Toolbox (v8.6.2, Eigenvector Research Inc). PCA modeling used autoscale and Probabilistic Quotient Normalization (PQN) as pre-processing methods. Data was randomly separated as 80% for training set and 20% for test set. PLS-DA was used to differentiate the whole grain barley and wheat intake. Discriminating metabolites were selected by repeatedly removing variables with low selectivity ratio, and variable importance in projection (VIP) values until no further increase in the cross-validation classification errors can be observed. Final models with selected variables were evaluated using test set misclassification. The variables that were selected in at least 75% of the models were recorded for further investigation. Potential barley intake biomarkers were selected manually from these variables based on the criterials: metabolites should have low intensities in group Before Barley (BB), Before Wheat (BW), After Wheat (AW). And, they should have high intensities in the intervention group After Barley (AB).

## Streamlined Data Processing Workflow

**Table 8 XCMS parameters**

| Steps | Parameters |
|---|---|
| Read Raw Data | mode = "onDisk", msLevel. = 1 |
| Peak picking | CentWaveParam(ppm = 30, peakwidth =c(0.025*60,0.30*60),snthresh = 10, noise = 0, prefilter = c(3,50), integrate = 2, mzdiff = -0.001, verboseColumns = TRUE, fitgauss = TRUE) |
| Peak grouping-1 | binSize = 0.01, bw = 0.2*60, minSamples=1, minFraction=0.15, MaxFeatures=10 |
| Adjust retention time-1 | smooth = "loess", span = 0.6, minFraction= 0.9, family = "gaussian", extraPeaks = 3 |
| Peak grouping-2 | minSamples=1, minFraction=0.2, MaxFeature=5 |
| Adjust retention time-2 | PeakGroupsParam( smooth = "loess", span = 0.6, minFraction = 0.9, family = "gaussian",extraPeaks = 3) |
| Peak filling | FillChromPeaksParam(expandMz = 0, expandRt = 0, ppm = 30) |

**Table 9 CAMERA parameters**

| Step | Parameters |
|---|---|
| Grouping | groupFWHM(xsa, perfwhm =0.1, intval = "into", sigma = 6) |
| Group correlation | groupCorr(xsaF,<br>calcIso = FALSE,<br>calcCiS = TRUE,<br>calcCaS = TRUE,<br>cor_eic_th=0.7,<br>cor_exp_th=0.7,<br>pval= 0.000001,<br>graphMethod="lpc",<br>intval="into") |
| Find isotopes | findIsotopes(xsaC, ppm = 10, mzabs= 0.01, intval = "into") |
| Find adducts | findAdducts(xsaFI,<br>ppm=10,<br>mzabs=0.01,<br>multiplier=4,<br>polarity=mode,<br>rules=rules) |

## The data structure of kudb

```
> suppressMessages(read_csv("db/db_ku.csv"))
# A tibble: 1,105 x 11
   name          recorded_rt predicted_rt ci_lower ci_upper pubchem inchi                                    inchi_key       exact_mass source method
   <chr>               <dbl>        <dbl>    <dbl>    <dbl>   <dbl> <chr>                                    <chr>                <dbl> <chr>  <chr>
 1 cotinine            0.790           NA       NA       NA      NA InChI=1S/C10H12N2O/c1-12-9(4-5-10(12)13)8-3-2-~ UIKROCXWUNQSPJ~      176. kudb   metho~
 2 2-oxindole-~        3.41            NA       NA       NA      NA InChI=1S/C10H9NO3/c12-9(13)5-7-6-3-1-2-4-8(6)1~ ILGMGHZPXRDCCS~      191. kudb   metho~
 3 hydroxyphen~        2.55            NA       NA       NA      NA InChI=1S/C9H10O4/c10-7-3-1-6(2-4-7)5-8(11)9(12~ JVGVDSSUAVXRDY~      182. kudb   metho~
 4 2-hydroxybu~        1.17            NA       NA       NA      NA InChI=1S/C4H8O3/c1-2-3(5)4(6)7/h3,5H,2H2,1H3,(~ AFENDNXGAFYKQO~      104. kudb   metho~
 5 perillic ac~        4.20            NA       NA       NA      NA InChI=1S/C10H14O2/c1-7(2)8-3-5-9(6-4-8)10(11)1~ CDSMSBUVCWHORP~      166. kudb   metho~
 6 1-methyl hi~        0.45            NA       NA       NA      NA InChI=1S/C7H11N3O2/c1-10-3-5(9-4-10)2-6(8)7(11~ BRMWTNUJHUMWMS~      169. kudb   metho~
 7 1-methylade~        1.62            NA       NA       NA      NA InChI=1S/C11H15N5O4/c1-15-3-14-10-6(9(15)12)13~ GFYLSDSUCHVORB~      281. kudb   metho~
 8 1-methyluri~        1.27            NA       NA       NA      NA InChI=1S/C6H6N4O3/c1-10-4(11)2-3(9-6(10)13)8-5~ QFDRTQONISXGJA~      182. kudb   metho~
 9 1-Methyluri~        1.27            NA       NA       NA      NA InChI=1S/C6H6N4O3/c1-10-4(11)2-3(9-6(10)13)8-5~ QFDRTQONISXGJA~      182. kudb   metho~
10 1,3-dimethy~        1.92            NA       NA       NA      NA InChI=1S/C7H8N4O3/c1-10-4-3(8-6(13)9-4)5(12)11~ OTSBKHHWSQYEHK~      196. kudb   metho~
# ... with 1,095 more rows
```

## Source Code of R Function m2r

## Source Code of R Function annotate_kudb

```
annotate_kudb <- function(data=..., mz_window=..., rt_window=..., polarity=...){
  require(dplyr)
  require(readr)
  require(purrr)
  require(stringr)
  #######load mum db########
  db_ku_mum <- read_csv(file.path("I:","SCIENCE-NEXS-
NyMetabolomics","db","db_ku.csv")) %>% filter(is.na(exact_mass)==FALSE) %>%
mutate(polarity="neutral")

  #######positive mode#####
  db_ku_p_h <- db_ku_mum %>% mutate(name=paste0(name,"+H"),
                                    exact_mass=exact_mass+1.007,
                                    polarity="pos")
  db_ku_p_na <- db_ku_mum %>% mutate(name=paste0(name,"+Na"),
                                     exact_mass=exact_mass+22.99,
                                     polarity="pos")
  db_ku_p_k <- db_ku_mum %>% mutate(name=paste0(name,"+K"),
                                    exact_mass=exact_mass+38.9637)
  db_ku_p_h2ol <- db_ku_mum %>% mutate(name=paste0(name,"-H2O+H"),
                                       exact_mass=exact_mass-18.011+1.007,
                                       polarity="pos")
  db_ku_p_fana <- db_ku_mum %>% mutate(name=paste0(name,"+FA+NA"),
                                       exact_mass=exact_mass+22.99+46.0054,
                                       polarity="pos")

  ######negative mode######
  db_ku_n_h <- db_ku_mum %>% mutate(name=paste0(name,"-H"),
                                    exact_mass=exact_mass-1.007,
                                    polarity="neg")
  db_ku_n_na <- db_ku_mum %>% mutate(name=paste0(name,"+Na-2H"),
                                     exact_mass=exact_mass+22.99-2*1.007,
                                     polarity="neg")
  db_ku_n_k <- db_ku_mum %>% mutate(name=paste0(name,"+K-2H"),
```

```
                              exact_mass=exact_mass+38.9637-2*1.007)
  db_ku_n_h2ol <- db_ku_mum %>% mutate(name=paste0(name,"-H2O-H"),
                                exact_mass=exact_mass-18.011-1.007,
                                polarity="neg")
  db_ku_n_cl <- db_ku_mum %>% mutate(name=paste0(name,"+Cl"),
                              exact_mass=exact_mass+35.45,
                              polarity="neg")
  db_ku_n_fa <- db_ku_mum %>% mutate(name=paste0(name,"+FA-H"),
                              exact_mass=exact_mass+46.005-1.007,
                              polarity="neg")
  db_ku_n_hcoonah <- db_ku_mum %>% mutate(name=paste0(name,"+HCOONa-H"),
                                  exact_mass=exact_mass+66.98,
                                  polarity="neg")

db_ku <- bind_rows(db_ku_p_h,db_ku_p_na,db_ku_p_k,db_ku_p_h2ol,db_ku_p_fana,
db_ku_n_h,db_ku_n_na,db_ku_n_k,db_ku_n_h2ol,db_ku_n_cl,db_ku_n_fa,db_ku_n_hcoonah)

  ############################

  mzrt <- tibble(mz=data %>% pull(mz),
                 rt=data %>% pull(rt))
  n <- 1:nrow(mzrt)

  #Predret annotation
  result <- sapply(n,function(n){
    mz_1 <- mzrt[n,1] %>% as.numeric()
    rt_1 <- mzrt[n,2] %>% as.numeric()
    db_ku %>%
      filter(source=="predret") %>%
      filter(polarity == mode) %>%
      filter(exact_mass>mz_1-mz_window) %>%
      filter(exact_mass<mz_1+mz_window) %>%
      filter(predicted_rt>rt_1-rt_window) %>%
      filter(predicted_rt<rt_1+rt_window)} %>% pull(name))
  result_1 <- data %>% mutate(id_predret=map(result,function(x)paste(x,collapse = ",")) %>%
map_chr(.,1))

  #kudb annotation
  result_2 <- sapply(n,function(n){
    mz_1 <- mzrt[n,1] %>% as.numeric()
    rt_1 <- mzrt[n,2] %>% as.numeric()
    db_ku %>%
      filter(source=="kudb") %>%
      filter(polarity == mode) %>%
      filter(exact_mass>mz_1-mz_window) %>%
      filter(exact_mass<mz_1+mz_window) %>%
      filter(recorded_rt>rt_1-rt_window) %>%
      filter(recorded_rt<rt_1+rt_window)} %>% pull(name))
  result_3 <- result_1 %>% mutate(id_kudb=map(result_2,function(x)paste(x,collapse = ","))
%>% map_chr(.,1))
```

```
  result_4 <- result_3 %>% select(mz,rt,pcgroup,adduct,id_predret:id_kudb,data %>% colnames()
%>% str_match_all("X\\d{1,5}") %>% unlist())
  return(result_4)
}
```

## Biomarkers of Whole Grain Barley and Wheat Intake

**Table 10 Potential Biomarkers for WG Barley Intake**

| No | Candidate biomarker | Formula | Chemical group | Presence in Food | Reference |
|---|---|---|---|---|---|
| 1 | Hordenine | $C_{10}H_{15}NO$ | alkaloid | germinating barley, beer and other plants | (*30*) |
| 4 | Hordatine A | $C_{28}H_{38}N_8O_5$ | alkaloid | only reported in barley | FoodDB (002330) |
| 4 | Hordatine B | $C_{29}H_{40}N_8O_5$ | alkaloid | only reported in barley | FoodDB (002328) |
| 2 | Distichonic acid A | $C_{10}H_{18}N_2O_8$ | gamma amino acids and derivatives | only reported in barley | FoodDB (18164) |
| 3 | Distichonic acid B | $C_{10}H_{18}N_2O_8$ | gamma amino acids and derivatives | only reported in barley | FoodDB (018165) |
| 5 | 14,16-Nona cosanedione | $C_{29}H_{56}O_2$ | ketone | only reported in barley | FoodDB (013891) |
| 6 | N-Norgramine | $C_{10}H_{12}N_2$ | indole | only reported in barley | FoodDB (017815) |

**Table 11 Putative Biomarkers for Whole Grain Wheat Intake**

| Dietary factor | No. subjects | Study design | Sample type | Analytical method | Candidate biomarker(s) | Reference |
|---|---|---|---|---|---|---|
| Wheat bran, Wheat aleurone | 14+13 | Intervention | plasma | LC-MS/MS (Microbiology assay for folate) | Betaine, choline folate, dimethylglycine (DMG) | (*72*) |
| None-bread, White bread, WG bread | 155 | Observational | urine | HPLC-qTOF-MS | Benzoxazinoid-related metabolites(HHPAA, HBOA glycoside) ARs-related metabolites(DHPPA glucuronide, DHPPTA sulphate, microbial-derived metabolites | (*62*) |

## Experiment Procedure: Sitostanol reference compound

### Materials

Sitostanol >99%, 5 mg (CAS Number: 83-45-4, Avanti Polar Lipids INC., USA) was transported and stored in -20 °C; ethanol and methanol (chromatographic grade)

### Procedures

1. Prepare sitostanol stock solution (1 mg/mL) in 100% ethanol
    - Weigh sitostanol 0.6 mg (=0.0006 g) and transfer to 1.5 mL Eppendorf tube
    - Add ethanol (0.6 mL)
    - Vortex mix until completely dissolved
2. Prepare sitostanol working solutions (0.02 mg/mL)
    - Transfer 10 µL to an eppendorf tube
    - Dilute with 490 µL MeOH
    - Label and store in -20°C until use
3. Inject in VION by Quad method
4. Compare $MS^2$ with whole grain barley samples and urine samples

## Experiment Procedure: *β*-glucuronidase Treatment

### Materials and Apparatus:

- Enzyme: *β*-glucuronidase (CAS Number 9001-45-0, E.C. number 3.2.1.31, Sigma-Aldrich, from E. Coil, optimal pH 6-7)
- Chemicals: phosphate buffer (pH=7.2) prepared by disodium phosphate (CAS Number 7558-79-4) and monosodium phosphate (CAS Number 7558-80-7), methanol
- Apparatus: pH meter, water bath, and centrifuge

### Procedures:

1. Prepare phosphate buffer[3] (0.1 M, pH=6.8, 50 mL):
    - Prepare 40 mL of distilled water in a volumetric bottle.
    - Add 0.656 g of monosodium phosphate to the solution.
    - Add 0.352 g of disodium phosphate to the solution.

[3] Calculated by AAT Bioquest

- Adjust pH using HCl or NaOH.
- Add distilled water until volume is 0.05 L.

2. Prepare enzyme solution (5 mg/mL): dissolve 0.0075 g *β*-glucuronidase in 1.5 mL phosphate buffer. Vortex mix for 1 min. Store in -20 ◦C freezer[4]
3. Prepare urine samples: thaw samples in the fridge and centrifuge (3000 rpm, 2 min). Prepare 2 Eppendorf tubes, one labeled as 'blank', one as 'treatment'. Transfer 100 µL to each Eppendorf tube.
4. Enzymatic hydrolysis reaction: add 50 µL phosphate solution to 'blank', 50 µL enzyme solution to 'treatment', incubate in 37 ◦C for one hour.
5. Denature enzymes to terminate the reaction
   - Add 50 µl MeOH to the solution and vortex mix for 1 min
   - Centrifuge at 3000 rpm for 3 min.
   - Transfer supernatant to a vial
6. Dilute with 300 µl Solvent A for further MS2 analysis

## RT of alkylresorcinol metabolites

| Metabolites | Neutral formula | Retention time (min) | m/z* |
|---|---|---|---|
| 3,5 DHPPA glucuronide | $C_{15}H_{18}O_{10}$ | 1.98 | 357.09 |
| 3,5 DHBA glucuronide | $C_{13}H_{14}O_{10}$ | 0.93 | 329.051 |
| 3,5 DHPPA | $C_9H_{10}O_4$ | 2.66 | 181.041 |
| 3,5 DHBA | $C_7H_6O_4$ | 1.94 | 153.018 |
| DHBA sulfate | $C_7H_6O_7S$ | 2.91 | 232.977 |
| DHMBA sulfate (dihydroxy-5-methoxybenzoic acid) | $C_8H_8O_5$ | 0.78 | 262.9874 |
| UI conjugate of DHBA glucuronide | $C_{11}H_{12}O_7$ | 1.19 | 511.0971 |
| 3,5 DHBA glycine | $C_9H_9O_5N$ | 1.63 | 210.041 |

m/z*: monoisotopic mass of $[M-H]^-$

[4] Sigma Product Information: A solution in 75 mM phosphate buffer, pH 6.8, >5 mg/ml may be stored at -20 ◦C for up to 2 months with little or no loss of activity.

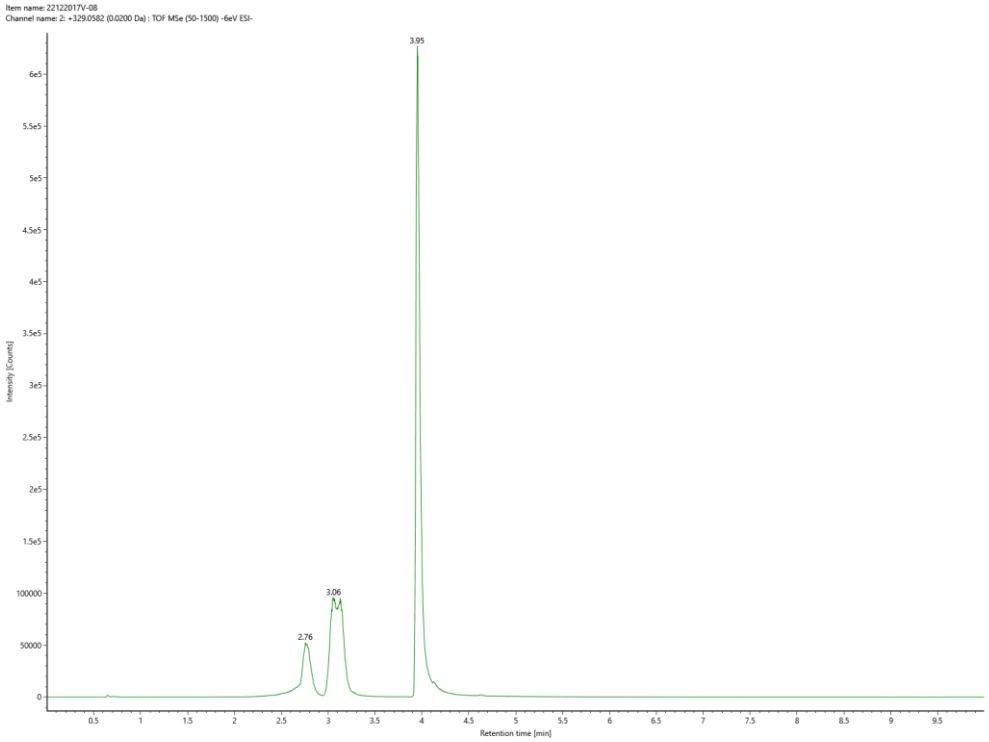
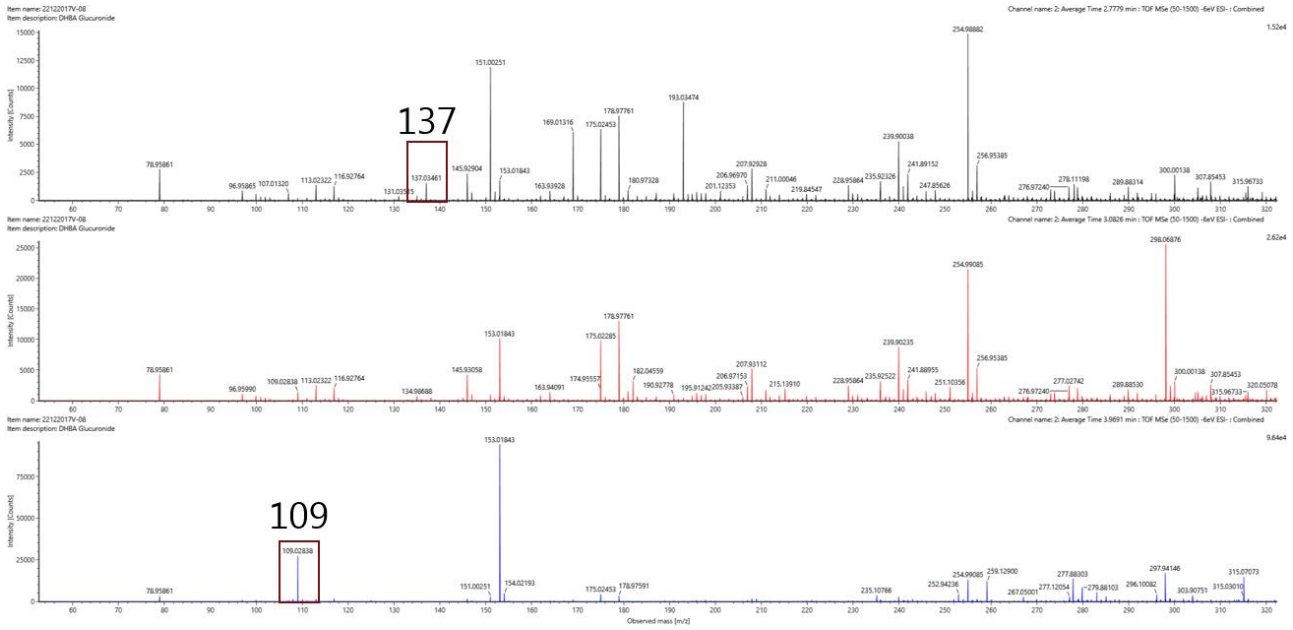



**Figure 19 Extracted Ion Chromatogram of Glucuronidation Products of 3,5-DHBA**



## $MS^2$ spectra

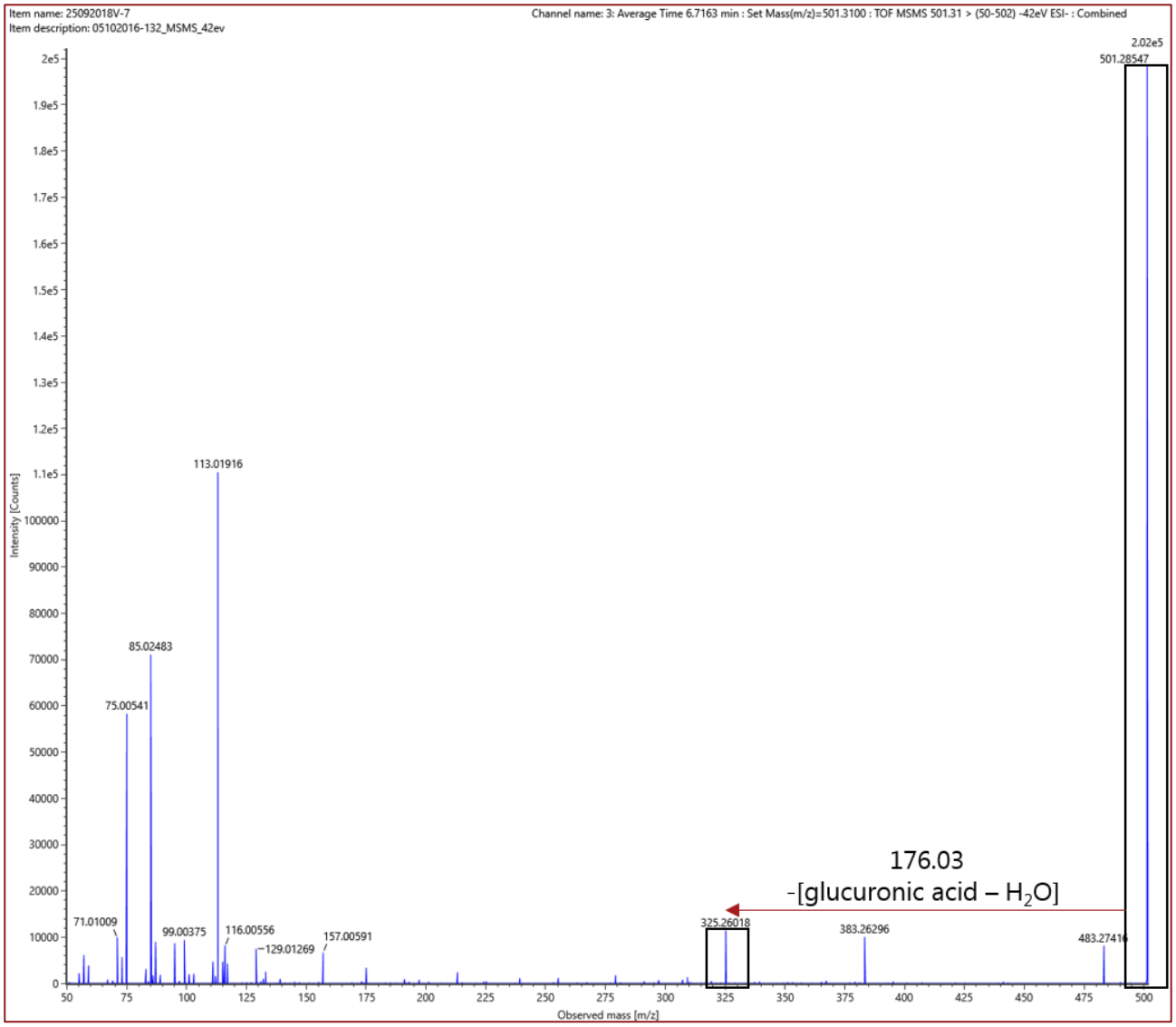



**Figure 20 MS/MS spectra of glucuronate ion 501**

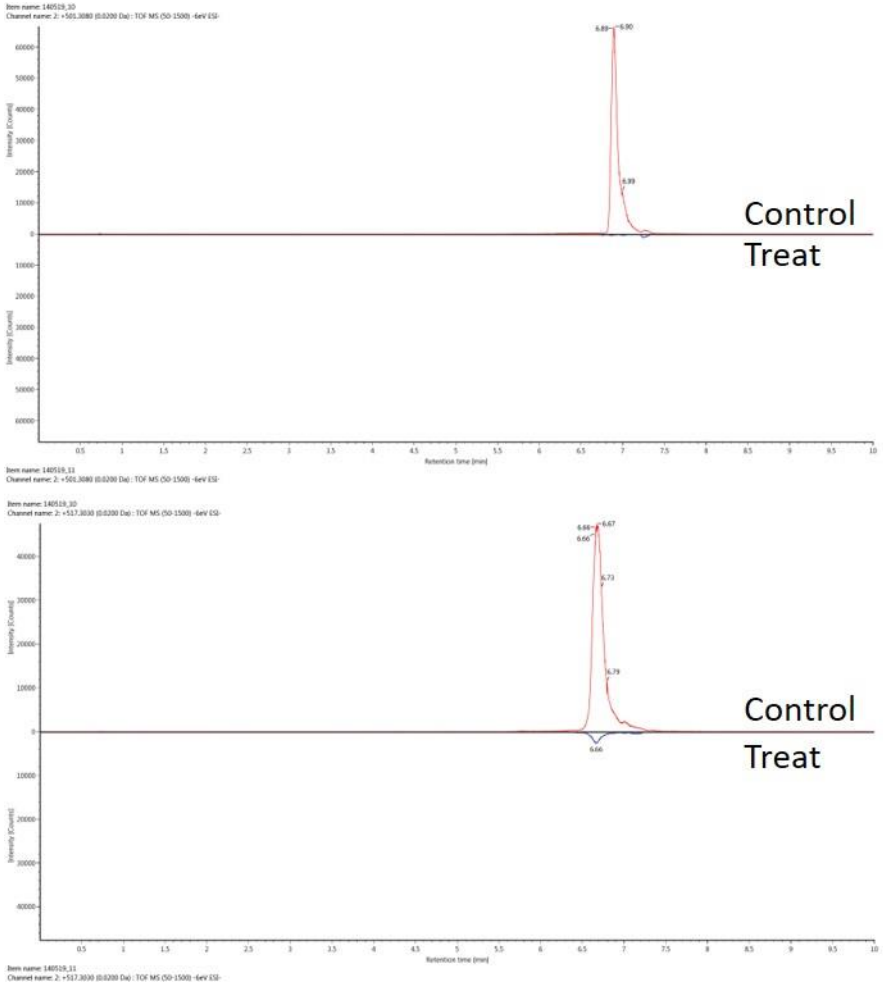



**Figure 21 Extracted Ion Chromatogram of Glucuronate Ions (Top: 501, Bottom: 517)**

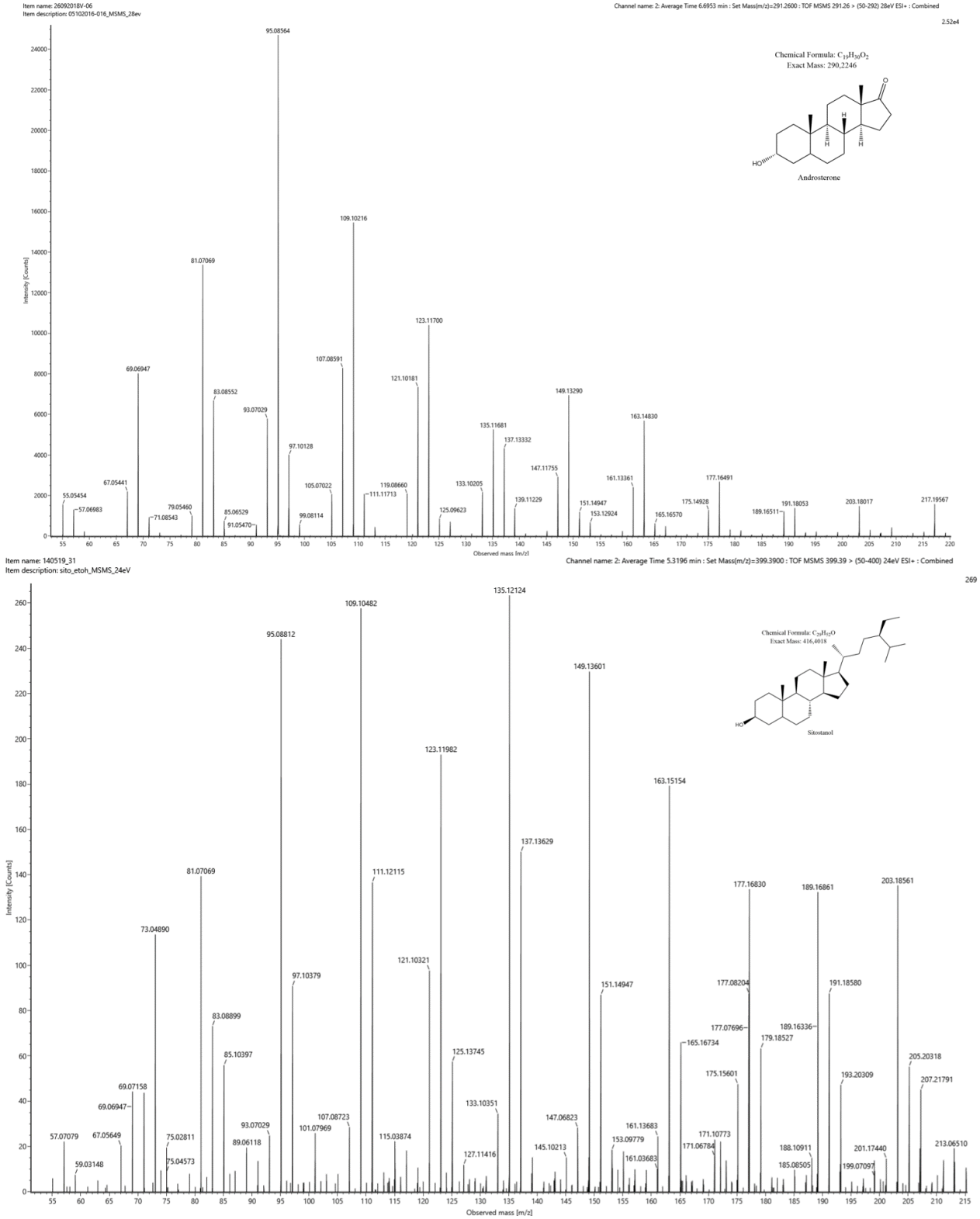



**Figure 22 MS/MS Spectra of Androsterone (Top) and Sitostanol (Bottom)**

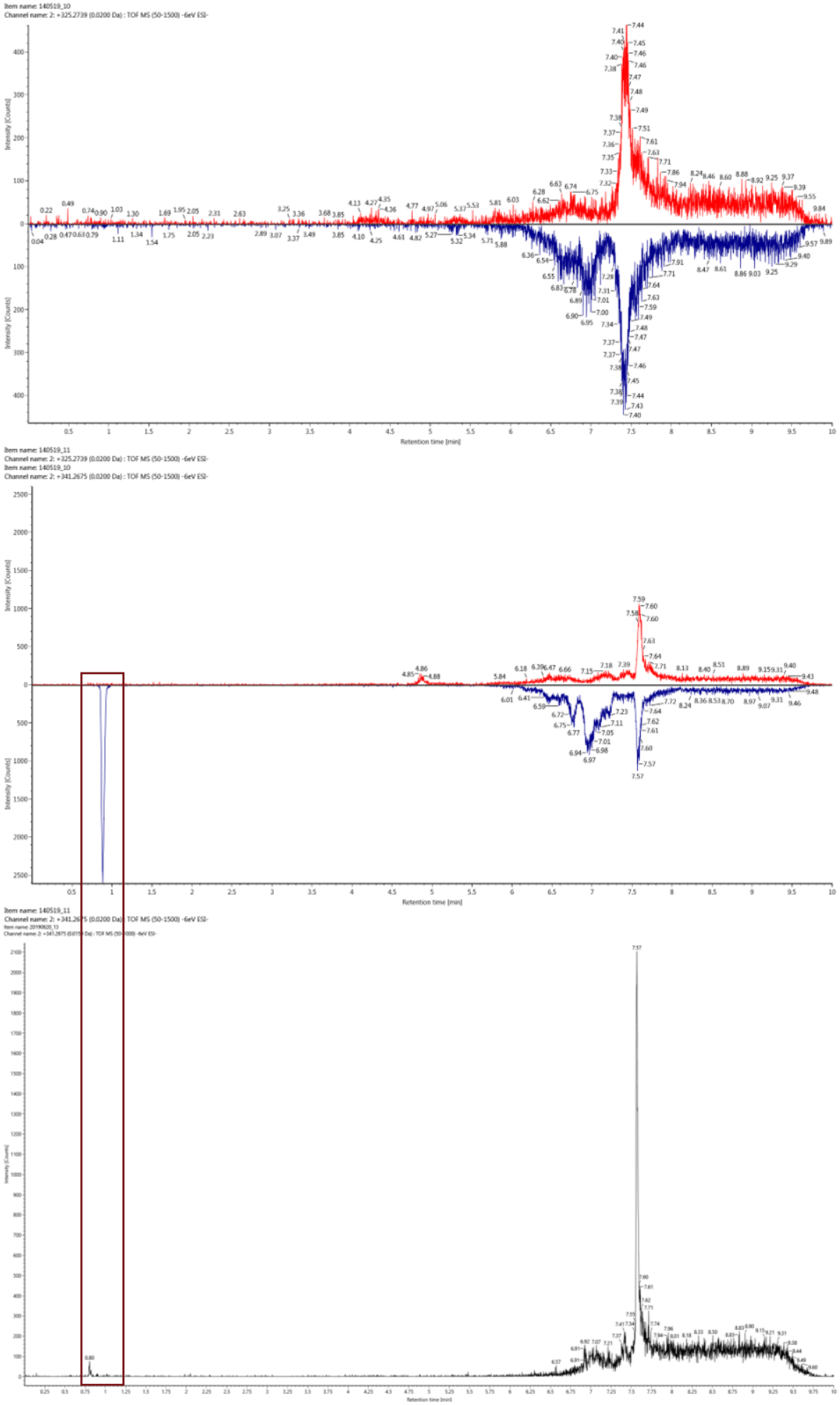



**Figure 23 EIC of expected ions**

# Reference


1. FAO, *FAO Statistics* (2016).
2. S. E. Ullrich, *Barley: Production, improvement, and uses* (John Wiley & Sons, 2010), vol. 12.
3. A. C. Newton *et al.*, Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security. *FOOD Secur.* **3**, 141–178 (2011).
4. J. Slavin, Whole grains and human health. *Nutr. Res. Rev.* **17**, 99–110 (2004).
5. J. Slavin, Why whole grains are protective: biological mechanisms. *Proc. Nutr. Soc.* **62**, 129–134 (2003).
6. L. S. Freedman, A. Schatzkin, D. Midthune, V. Kipnis, Dealing With Dietary Measurement Error in Nutritional Cohort Studies. *JNCI J. Natl. Cancer Inst.* **103**, 1086–1092 (2011).
7. R. M. van Dam, F. B. Hu, Are alkylresorcinols accurate biomarkers for whole grain intake? *Am. J. Clin. Nutr.* **87**, 797–798 (2008).
8. S. S. Jonnalagadda *et al.*, Putting the whole grain puzzle together: health benefits associated with whole grains--summary of American Society for Nutrition 2010 Satellite Symposium. *J. Nutr.* **141**, 1011S–22S (2011).
9. Q. Gao *et al.*, A scheme for a flexible classification of dietary and health biomarkers. *Genes Nutr.* **12**, 34 (2017).
10. L. O. Dragsted *et al.*, Dietary and health biomarkers—time for an update. *Genes Nutr.* **12**, 1–7 (2017).
11. G. Praticò *et al.*, Guidelines for Biomarker of Food Intake Reviews (BFIRev): How to conduct an extensive literature search for biomarker of food intake discovery. *Genes Nutr.* **13** (2018), doi:10.1186/s12263-018-0592-8.
12. O. Fiehn, Metabolomics - The link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
13. Anonymous, Signs of a long life. *Economist*. **387** (2008).
14. C. B. Clish, Metabolomics: an emerging but powerful tool for precision medicine. *Mol. Case Stud.* **1**, a000588 (2015).
15. D. González-Peña, L. Brennan, Recent Advances in the Application of Metabolomics for Nutrition and Health. *Annu. Rev. Food Sci. Technol.* **10**, 479–519 (2019).
16. D. K. Trivedi, K. A. Hollywood, R. Goodacre, Metabolomics for the masses: The future of metabolomics in a personalized world. *New horizons Transl. Med.* **3**, 294–305 (2017).
17. A. C. Schrimpe-Rutledge, S. G. Codreanu, S. D. Sherrod, J. A. McLean, Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **27**, 1897–1905 (2016).
18. T. Cajka, O. Fiehn, Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal. Chem.* **88** (2016), pp. 524–545.
19. J. L. Markley *et al.*, The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **43** (2017), pp. 34–40.
20. K. Bingol *et al.*, Emerging new strategies for successful metabolite identification in metabolomics. *Bioanalysis*. **8**, 557–573 (2016).
21. S. Beisken, M. Eiden, R. M. Salek, Getting the right answers: understanding metabolomics challenges. *Expert Rev. Mol. Diagn.* **15**, 97–109 (2015).

22. R. Spicer, R. M. Salek, P. Moreno, D. Cañueto, C. Steinbeck, Navigating freely-available software tools for metabolomics analysis. *Metabolomics*. **13**, 1–16 (2017).
23. D. S. Wishart *et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
24. FooDB, (available at http://foodb.ca/).
25. A. Medina-Remón *et al.*, Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*. **2013** (2013), doi:10.1093/database/bat070.
26. Dictionary of food compounds with CD-ROM. *Choice Rev. Online* (2013), doi:10.5860/choice.51-1824.
27. V. Popovici, thesis, University of Copenhagen (2016).
28. M.-B. S. Andersen *et al.*, 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 (2013).
29. G. Gürdeniz, M. Kristensen, T. Skov, L. O. Dragsted, The effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats. *Metabolites* (2012), doi:10.3390/metabo2010077.
30. M. G. Jensen, S. Meier, L. Bech, E. Lund, L. O. Dragsted, Detecting Beer Intake by Unique Metabolite Patterns (2016), doi:10.1021/acs.jproteome.6b00635.
31. T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* (2010), doi:10.1186/1471-2105-11-395.
32. Eigenvector Research, PLS_Toolbox, (available at https://eigenvector.com/software/pls-toolbox/).
33. R core team, R: A language and environment for statistical computing (2013).
34. M. C. Chambers *et al.*, A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
35. R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*. **9**, 504 (2008).
36. C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779–787 (2006).
37. H. P. Benton, E. J. Want, T. M. D. Ebbels, Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data. *Bioinformatics*. **26**, 2488–2489 (2010).
38. C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, S. Neumann, CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).
39. J. Stanstrup, S. Neumann, U. Vrhovšek, PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal. Chem.* **87**, 9421–9428 (2015).
40. H. Wickham, Tidy Data. *J. Stat. Software; Vol 1, Issue 10* (2014) (available at https://www.jstatsoft.org/v059/i10).
41. F. Rohart, B. Gautier, A. Singh, K.-A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).
42. Eigenvector Research, DataSet object, (available at https://eigenvector.com/software/dataset-object/).
43. E. L. Schymanski *et al.*, Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
44. Y. Djoumbou-Feunang *et al.*, CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. *Metabolites*. **9** (2019), doi:10.3390/metabo9040072.

45. Y. Djoumbou-Feunang *et al.*, BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminform.* **11**, 2 (2019).
46. A. R. Bird *et al.*, Wholegrain foods made from a novel high-amylose barley variety (Himalaya 292) improve indices of bowel health in human subjects. *Br. J. Nutr.* **99**, 1032–1040 (2008).
47. N. Ames *et al.*, A double-blind randomised controlled trial testing the effect of a barley product containing varying amounts and types of fibre on the postprandial glucose response of healthy volunteers. *Br. J. Nutr.* **113**, 1373–1383 (2015).
48. N. Marungruang, J. Tovar, I. Bjorck, F. F. Hallenius, Improvement in cardiometabolic risk markers following a multifunctional diet is associated with gut microbial taxa in healthy overweight and obese subjects. *Eur. J. Nutr.* **57**, 2927–2936 (2018).
49. S. Abidi, H. Ben Salem, V. Vasta, A. Priolo, Supplementation with barley or spineless cactus (Opuntia ficus indica f. inermis) cladodes on digestion, growth and intramuscular fatty acid composition in sheep and goats receiving oaten hay. *SMALL Rumin. Res.* **87**, 9–16 (2009).
50. A. P. Foster *et al.*, Serum IgE and IgG responses to food antigens in normal and atopic dogs, and dogs with gastrointestinal disease. *Vet. Immunol. Immunopathol.* **92**, 113–124 (2003).
51. C. Kyro *et al.*, Plasma Alkylresorcinols, Biomarkers of Whole-Grain Wheat and Rye Intake, and Incidence of Colorectal Cancer. *JNCI-JOURNAL Natl. CANCER Inst.* **106** (2014), doi:10.1093/jnci/djt352.
52. R. Landberg, P. Aman, G. Hallmans, I. Johansson, Long-term reproducibility of plasma alkylresorcinols as biomarkers of whole-grain wheat and rye intake within Northern Sweden Health and Disease Study Cohort. *Eur. J. Clin. Nutr.* **67**, 259–263 (2013).
53. K. J. Raninen *et al.*, Fiber content of diet affects exhaled breath volatiles in fasting and postprandial state in a pilot crossover study. *Nutr. Res.* **36**, 612–619 (2016).
54. H. Adlercreutz, J. L. Peñalvo, A.-M. Linko-Parvinen, M. J. Tikkanen, R. Landberg, Alkylresorcinols from Whole-Grain Wheat and Rye Are Transported in Human Plasma Lipoproteins. *J. Nutr.* **137**, 1137–1142 (2007).
55. N. M. McKeown *et al.*, Comparison of plasma alkylresorcinols (AR) and urinary AR metabolites as biomarkers of compliance in a short-term, whole-grain intervention study. *Eur. J. Nutr.* **55**, 1235–1244 (2016).
56. R. Landberg *et al.*, New alkylresorcinol metabolites in spot urine as biomarkers of whole grain wheat and rye intake in a Swedish middle-aged population. *Eur. J. Clin. Nutr.* **72**, 1439–1446 (2018).
57. R. Landberg *et al.*, Alkylresorcinol Metabolite Concentrations in Spot Urine Samples Correlated with Whole Grain and Cereal Fiber Intake but Showed Low to Modest Reproducibility over One to Three Years in U.S. Women. *J. Nutr.* **142**, 872–877 (2012).
58. Y. Chen, A. B. Ross, P. Åman, A. Kamal-Eldin, Alkylresorcinols as Markers of Whole Grain Wheat and Rye in Cereal Products. *J. Agric. Food Chem.* **52**, 8242–8246 (2004).
59. R. Landberg, A. Kamal-Eldin, A. Andersson, B. Vessby, P. Aman, Alkylresorcinols as biomarkers of whole-grain wheat and rye intake: plasma concentration and intake estimated from dietary records. *Am. J. Clin. Nutr.* **87**, 832–838 (2008).
60. L. Bresciani *et al.*, Bioavailability and metabolism of phenolic compounds from wholegrain wheat and aleurone-rich wheat bread. *Mol. Nutr. Food Res.* **60**, 2343–2354 (2016).
61. B. M. Jensen *et al.*, Quantitative analysis of absorption, metabolism, and excretion of benzoxazinoids in humans after the consumption of high- and low-benzoxazinoid diets with similar contents of cereal dietary fibres: a crossover study. *Eur. J. Nutr.* **56**, 387–397 (2017).
62. M. Garcia-Aloy *et al.*, Nutrimetabolomics fingerprinting to identify biomarkers of bread exposure in a free-living population from the PREDIMED study cohort. *METABOLOMICS*.

**11**, 155–165 (2015).
63. S. Nybacka, H. B. Forslund, M. Hedelin, Validity of a web-based dietary questionnaire designed especially to measure the intake of phyto-oestrogens. *J. Nutr. Sci.* **5** (2016), doi:10.1017/jns.2016.28.
64. Y. Zhu, P. Wang, W. Sha, S. Sang, Urinary Biomarkers of Whole Grain Wheat Intake Identified by Non-targeted and Targeted Metabolomics Approaches. *Sci. Rep.* **6**, 36278 (2016).
65. K. Levsen *et al.*, Structure elucidation of phase II metabolites by tandem mass spectrometry: an overview. *J. Chromatogr. A.* **1067**, 55–72 (2005).
66. R. Landberg *et al.*, Human Plasma Kinetics and Relative Bioavailability of Alkylresorcinols after Intake of Rye Bran. *J. Nutr.* **136**, 2760–2765 (2006).
67. Wikipedia, Androsterone, (available at https://en.wikipedia.org/wiki/Androsterone).
68. S. Gilad, R. Chayen, K. Tordjman, E. Kisch, N. Stern, Assessment of 5α-reductase activity in hirsute women: comparison of serum androstanediol glucuronide with urinary androsterone and aetiocholanolone excretion. *Clin. Endocrinol. (Oxf).* **40**, 459–464 (1994).
69. M. Bicikova, M. Hill, D. Ripova, P. Mohr, R. Hampl, Determination of steroid metabolome as a possible tool for laboratory diagnosis of schizophrenia. *J. Steroid Biochem. Mol. Biol.* **133**, 77–83 (2013).
70. W. B. Dunn *et al.*, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6**, 1060–1083 (2011).
71. R. Wehrens *et al.*, Improved batch correction in untargeted MS-based metabolomics. *Metabolomics*. **12**, 88 (2016).
72. E. M. Keaveney *et al.*, Postprandial plasma betaine and other methyl donor-related responses after consumption of minimally processed wheat bran or wheat aleurone, or wheat aleurone incorporated into bread. *Br. J. Nutr.* **113**, 445–453 (2015).