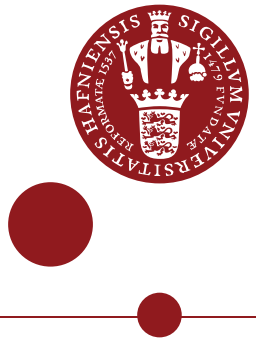
UNIVERSITY OF COPENHAGEN



Discovering Biomarkers 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

ECTS Credits:
30

Study Program:
MSc in Food Science and Technology

Academic Advisors:
Lars Ove Dragsted
Professor
Department of Nutrition, Exercise and Sports (NEXS)
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

List of Tables

Abstract

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

Aims: The first aim is to systematically review the biomarkers of whole grain barley and wheat intake. The second aim is to develop novel bioinformatics tools to streamline the metabolomics data analysis workflow and to validate the toolset by implementing it to identify and validate biomarkers for whole grain barley and wheat intake.

Methods: The systematic review of whole grain barley and wheat followed the Guidelines for Biomarker of Food Intake Reviews (BFIRev). A new data processing workflow was developed by incorporating freely available tools and self-developed functions in the statistical language R. The toolset was validated by implementation and analysis of two intervention studies. The barley bread intervention compared whole grain barley and wheat bread, while the protein source study compared barleyotto with other foods. Potential biomarkers were selected from the barley bread intervention study and further validated in the protein source study. Tandem mass spectrometry was used to elucidate the structures of candidate biomarkers.

Results: PLS-DA modeling and t-test did not reveal any discriminating metabolites from fasting plasma samples from the barley bread intervention, whereas whole grain barley and wheat intake were distinguished from 24-h pooled urine samples. The systematic literature review showed that, no known biomarker could indicate whole grain barley intake. 4-hydroxybenzoic acid-4-sulfate was putatively identified as a potential whole grain barley intake biomarker from the intervention. Both the literature review and the metabolomics analysis showed alkylresorcinol (AR) metabolites may indicate whole grain wheat intake. 2-hydroxy-1,4-benzoxazin-3-one (HBOA) glucuronide was identified as a potential whole grain wheat intake biomarker. A metabolomics data analysis workflow was developed by incorporating freely available tools and self-developed R functions. *m2r* can convert Matlab DataSet to R dataframe. *annotate_kudb* can annotate roughly 100-300 metabolites. *Libra* can calibrate batch effects without changing coefficient of variance of pooled samples.

Conclusions: 4-hydroxybenzoic acid-4-sulfate is proposed as a potential biomarker of whole grain barley intake. AR metabolites and HBOA glucuronide may indicate whole grain wheat intake. Furthermore, the metabolomics data processing workflow has been improved and validated by the barley intervention study.

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

1. Preface and Introduction

In the disaster movie, *2012*, the world comes to an end due to natural disasters. 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 low temperatures. How do the Tibetan people survive in such an extreme environment? What do they eat? Barley.

Barley can grow in the harshest and the most marginal areas from sub-Saharan to the Himalayas. Barley has been cultivated in Tibet for at least 3500 years. 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 its possibility to be used as a food source to cope with food security (*1*). However, we know little about the health beneficial and nutritional effects of barley.

Whole grain has raised the public interest for its health benefits. Dietary guidelines suggest increasing whole grain intake (*2*), because the evidence generally shows that whole grain intake is associated with better health (*3*). However, the health benefits of different types of whole grain have not yet been established. Objective dietary measurement tools are needed to reveal the health benefits of different types of whole grain.

Metabolomics is a powerful tool to measure hundreds, or even thousands of metabolites from human bio-specimens. Some metabolites could be characteristic to a food or a food group and serve as biomarkers of food intake. Biomarkers of food intake could reflect food exposures objectively in populations and resolve some nutritional research questions.

This thesis is a continuation 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 and barley can be discriminated from the urine metabolome. Therefore, the biomarkers of whole grain wheat intake can be identified. In the previous project, I was frustrated by metabolomics data processing. I believe that a more unified and streamlined data processing workflow can facilitate the metabolomics researches. Based on these notions, the research aims of this thesis were formulated.

Aims

The first aim was to systematically review previously reported biomarkers of whole grain barley and wheat intake.

The second aim was to develop novel bioinformatics tools to streamline the metabolomics data processing workflow, and to validate the toolset by applying it to investigate a plasma and urine metabolic profile in order to discover and validate biomarkers of whole grain barley and wheat intake.

2. Background

2.1 Barley

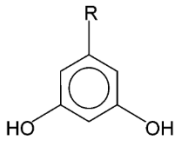
Barley is one of the most produced grains. In 2016, 143 million tons of barley were produced, which ranked it as the fourth-most produced grain after maize, wheat, and rice. EU countries produced 63% of the world's barley. Denmark produced 3.9 million tons, ranking it as the 11th most-produced country of barley (*4*). Barley may be the most widely adapted cereal grain species with good drought, cold, and salt tolerance (*1*). It is cultivated both in highly productive agricultural systems and also in marginal and subsistence environments (*5*). Barley's ability to adapt to multiple biotic and abiotic stresses make it a potential food source to cope with food security (*5*).

Roughly 5% of total barley production is directly consumed by humans (*6*). The rest of the barley production is mainly used for animal feed and brewing (*1*, *5*). 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 (*1*). The quality standard of barley for food use has not been completely established, making it difficult for food manufacturers to select raw materials suitable for use in specific food products (*1*).

2.2 Wheat and Alkylresorcinols

Wheat is the second-most produced grain and widely consumed by humans. Wheat is a major ingredient in a broad range of food, including bread, porridges, pizzas, noodles, and pasta.

Alkylresorcinols (ARs), also known as resorcinolic lipids, are phenolic lipids composed of long aliphatic chains and resorcinol-type phenolic rings (*7*). ARs are relatively rare in nature and are present in high concentration exclusively in the bran part of wheat and rye (0.1-0.3% of dry wheat) (*8*). ARs have different types of homologues. The most common ARs in wheat have odd-number alkyl chain, as shown in **Figure 1** (*9*).



R	Names	[CAS No.]
$C_{15}H_{31}$	**5-n-pentadecylresorcinol**, 5-pentadecyl-1,3-benzendiol, cardol (trivial) (C15:0)	[3158-56-3]
$C_{17}H_{35}$	**5-n-heptadecylresorcinol** 5-heptadecyl-1,3-benzendiol (C17:0)	[41442-57-3]
$C_{19}H_{39}$	**5-n-nonadecylresorcinol** 5-nonadecyl-1,3-benzendiol (C19:0)	[35176-46-6]
$C_{21}H_{43}$	**5-n-heneicosylresorcinol** 5-heneicosyl-1,3-benzendiol (C21:0)	[70110-69-7]
$C_{23}H_{47}$	**5-n-tricosylresorcinol** 5-tricosyl-1,3-benzendiol (C23:0)	[70110-60-0]
$C_{25}H_{51}$	**5-n-pentacosylresorcinol** 5-pentacosyl-1,3-benzendiol (C25:0)	[70110-61-1]

Figure 1 Structure, Names, and CAS No. of Common Alkylresorcinol in wheat. Adapted from J. Agric. Food Chem. 2003, 51.

2.3 Whole Grain

Whole grain contains the starchy endosperm, the germ, and the bran (**Figure 2**), whereas the refined grain retains only the endosperm. The processed form of whole grain, such as grounded, cracked, or flaked, the three fractions (the starchy endosperm, the germ, and the bran) should be present in the same proportions as they exist in native grains (*10*, *11*).

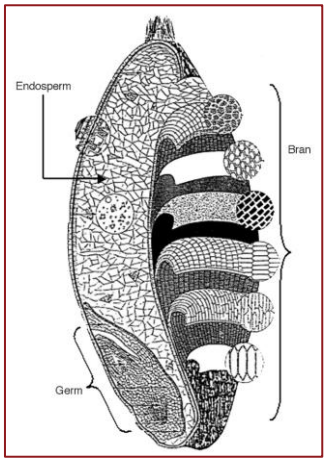


Figure 2 The Diagram of Whole Grain

The bran and germ fractions of the grain consist of a wide range of compounds with known health beneficial effects, for example, dietary fibers, lignans, tocotrienols, and phenolic compounds (*11*). The grain-refining process removes all or part of the bran resulting in the loss of health-beneficial compounds of the grain.

Many observational studies have reported that whole grain intake could improve health (*10, 11*), but such studies rely on self-report based dietary measurement tools. The accurate measurement of dietary intake is a challenging task in general (*12*). With regards to 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 recognizing whole grain products and contents (*13*). The intake of different types of whole grain may benefit health differently (*14*). Current dietary measurement tools categorize different types of whole grain as one group, and therefore they cannot distinguish the consumption of different whole grain types.

2.4 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. Since blood, urine, and tissues are packed with these compounds, it should be possible to detect and measure them (*15, 16*). Though metabolomics is still emerging (*17*), it has shown the potential in food, plant, environment, nutrition, and health research (*18, 19*).

Metabolomics can be categorized as targeted or untargeted approaches. Targeted metabolomics identifies and quantifies a limited number of known metabolites (*20*). Untargeted metabolomics analyses all detectable metabolites, including unknown chemical compounds and correlate metabolites with a certain physiological state (*15, 21, 22*). Untargeted metabolomics consists of well-conceived sample preparations, analytical methods, and data processing to cover as many metabolites as possible (*15, 21*).

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 (*18*). NMR provides high reproducibility but less sensitivity and limited discovery power. MS is highly sensitive and has strong discovery ability but poor reproducibility (*18*). Because these two techniques offer different strengths, the combination of these two techniques becomes an emerging research direction (*23, 24*).

2.4.1 Challenges

Compound Identifications. The major bottleneck of LC-MS based untargeted metabolomics is metabolite identification (*25*). 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 (*24*). However, the uncertainties of measurement and the availabilities of reference compound may hurdle the structure elucidation of the metabolites of interest.

Bioinformatics tools. Another challenge of metabolomics is a lack of complete bioinformatics pipeline to analyze metabolomics data. Though a lot of open-source software is emerging. Currently, many existing open-source software still cannot handle the complete analysis workflow and tune every parameter (*26*). The harmonization of different bioinformatics is a challenging task. Because different bioinformatics tools might be developed in different programming languages. Besides, different open-source software might have different formats to store data.

2.5 Biomarker

2.5.1 Biomarker of Food Intake

Biomarkers of food intake 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 (*27, 28*). The biomarker can be a single metabolite or a combination of several metabolites. Biomarkers of food intake can be used as a complement to self-report based instruments to alleviate measurement error and uncertainties of food intake measurement in observational studies (*12*). Biomarkers of food intake can also be used to monitor compliances in intervention studies (*28*).

2.5.2 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 prioritize the identifications of novel biomarkers (*29*).

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 summaries the current status for applicability of biomarkers.

2.5.3 Other Biomarkers in Dietary and Health Research

Biomarker in dietary and health research can be categorized into three main areas, as proposed by Dragsted (*28*) and Gao (*27*).

Exposure (intake) biomarker reflects the level of extrinsic dietary variables that humans are exposed to. *Biomarker of food intake* belongs to this category. Within this category, *food compound intake biomarker* reflects the exposure to a specific food compound, such as potassium, selenium, and phenol. *Dietary pattern biomarker* consists of a set of *biomarkers of food intake* and *food compound intake biomarkers*, which reflects the average diet of an individual (*27*).

Effect biomarker reflects the functional response of the human body to an exposure. *Susceptibility biomarker* reflects the genetic or acquired host factors and intrinsic factors, influencing the response of an individual to the dietary exposure (*27*).

3. Materials and Methods

3.1 Systematic Review of Biomarkers of Whole Grain Barley and Wheat Intake

This systematic review follows the BFIRev guidelines (*29*). The literature search was performed in 3 databases (PubMed, Web of Science, and Scopus). Keywords used for searching 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 (trial* or experiment OR study) AND (urine OR plasma OR blood OR serum OR excretion OR hair OR toenail OR faeces OR fecal water). For the wheat intake biomarker search, 'barley' was changed to wheat.

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(*30*), FoodDB(*31*), PhenolExplorer(*32*), and Dictionary of Food Compounds(*33*) 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'.

3.2 Intervention Studies

This thesis investigated metabolomics data from two intervention studies. **Figure 3** presents the outlines of these two studies.

The barley bread intervention (*34*) was conducted by at NEXS by Veronica Maria Popovici in 2015 as part of her Master's thesis work. This study was a randomized crossover intervention study, including 14 healthy volunteers. During the intervention period, volunteers ate two bread rolls of whole grain barley or wheat on each day. The bread rolls were 227 g, providing 11 g dietary fibers for barley and 13 g dietary fibers for wheat. In each visit, volunteers donated their 24-h pooled urine and fasting plasma.

The protein sources study (*35*) was a crossover intervention study conducted by Maj-Britt Schmidt Andersen as part of her PhD's thesis work. Healthy volunteers were recruited and 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.

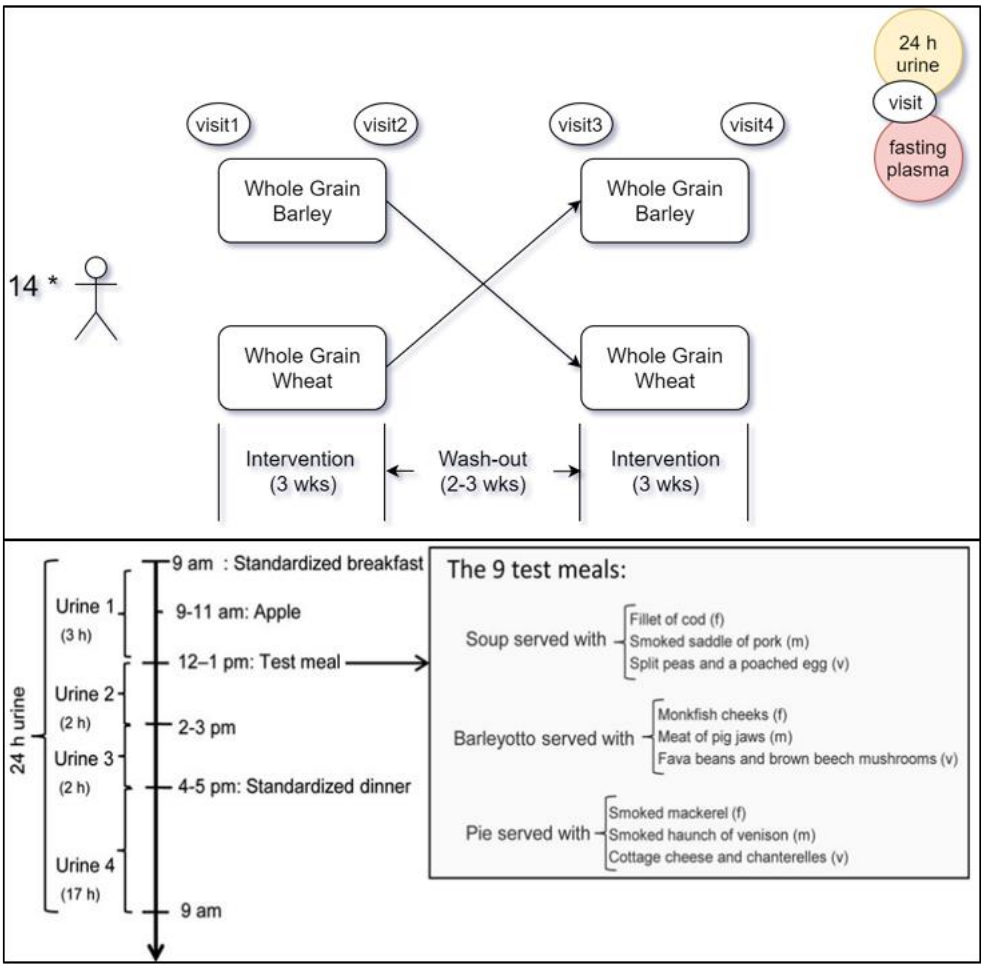


Figure 3 Outline of study designs for the barley bread study (top) and the protein source study (bottom)

3.3 Metabolomics Analysis

3.3.1 Sample Processing and Data Acquisition

Plasma protein precipitation, urine centrifugation, and dilution were performed as described previously (*35*, *36*). 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 (*37*).

3.3.2 Data Processing Workflow

Metabolomics data of barley bread intervention study (*34*) was processed through converting format by DataBridge (Waters Corporation), pre-processing by MZmine (*38*), and multivariate data analysis by PLS toolbox. Discriminating metabolites were selected to distinguish whole grain barley and wheat intake. Details are described in the **Appendix-A**.

3.3.3 Streamlined Data Processing Workflow

All processes were performed in statistic language R 3.5.3 (*39*) if not specifically stated.

Data Conversion. ProteoWizard 3.0 MSConvert (*40*) converted raw data from *.RAW* to *.mlXML*. Scan event was set to *1* to remove lock-mass trace.

Data Pre-processing. Converted data was pre-processed by XCMS 3.4.4 (*41–43*) including following steps: read raw file, peak picking, grouping, adjusting retention time, and peak filling. Detailed parameters are shown in the **Appendix-B**.

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

Libra calibrated intra- and inter- batch effects by linear regression models. Within one batch, for each feature, intensity $(I_1, I_2, I_3, \dots, I_n)$ was 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 the *p*-value is lower than the defined threshold, *0.1*, by the following steps (**Figure 4** shows the outline):

(1) *Kick-out the elephants(outliers)*. 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 ($|\varepsilon_i|$) were replaced by their estimated values $(\hat{y}_i)$.

(2) *Balance the single Libra and pick up the sands*. After replacing *elephants* with the estimated values, a new linear model was fit: $I = f_2(x) = \beta x + \alpha$. Then, the intensity of each sample, I_k, was calibrated to $f_2\left(\frac{n}{2}\right) + \varepsilon_k$. The linear calibration may introduce negative values, *sands*, for low-intensity samples. Their native values replaced the nagative values.

(3) *Equilibrate multiple Libras*. The centers of all *Libras* were averaged to a new center. Different batches were scaled to this new center.

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

Annotation. Metabolites were annotated by CAMERA (1.38.1) and *annotate_kudb*. CAMERA is an R package available on Bioconductor (*44*) and was used to detect pseudo compound groups (PCgroups), isotopes, adducts, multiply charged ions, and cluster ions. Detailed parameters and settings are shown in the **Appendix-B**. *Annotate_kudb* is a self-developed R function (source code is shown in the **Appendix-E**). *Annotate_kudb* matched 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 included experiment data and predicted data by PredRet (*45*).

KUDB stored data in the *tidy* format (*46*) in a *.csv* file to facilitate computing. One entry stored one compound, including the information of retention time, identity (including InChi, InChi key, and PubChem Id), and neutral monoisotopic mass, analytical method, and source. The common adducts (**Table 1**) were *expended* before matching.

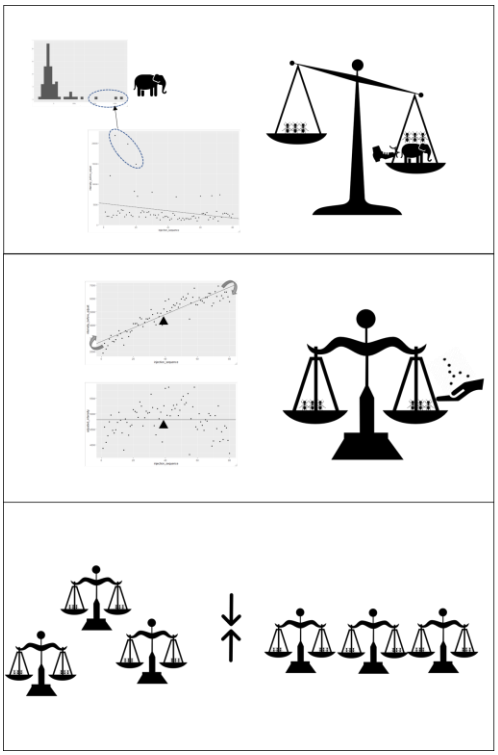


Figure 4 Diagram of Libra Calibration

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, multiply charged ions, and dimers were removed based on CAMERA annotation.

Statistics. The t-test was used for univariate data analysis. Benjamini-Yekutieli false discovery rate control method was used to calibrate *p*-values by R stats package. Multivariate data analysis was performed by mixOmics (*47*) 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 was optimized by the 5-fold cross-validation repeated 10 times.

3.3.4 MATLAB DataSet conversion

The metabolomics data of protein source study (*35*) was processed and stored in a Matlab DataSet object , 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-D**) was used to convert data from Matlab to R.

3.3.5 Compound Identification

Unknown compounds were identified by comparing retention time (RT), m/z, and MS^2 spectra with reference compounds, in-house, and public libraries. Level 1 – 4 indicated the level of identification (*48*).

MS^2 experiment. Target ions were selected and fragmented by collision energy 14, 28, and 42 eV to acquire MS^2 spectra in Vion Q-Tof Mass Spectrometry (Waters).

Sitostanol reference compound. Sitostanol reference compound was analyzed. The experiment procedures are shown in the **Appendix-G**.

Derivatization. Sulfates and glucuronides of AR metabolites were synthesized according to an internal Standard Operation Procedure (SOP) 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-H**.

3.3.6 Other Software

Partition coefficient (logP and ClogP) was predicted by ChemDraw Professional 16.0 (PerkinElmer Information, Inc.). CFM-ID 3.0 (*49*) was used to predict MS^2 spectra. BioTransformer (*50*) was used to predict metabolism pathways. Sirius was used to study MS^2 spectra.

[1] This was done by Gözde Gürdeniz. The results are shown in **Appendix-I**

4. Results

4.1 Systematic Review of Barley Intake Biomarker and Wheat Intake Biomarkers

4.1.1 Whole Grain Barley

The literature search retrieved 129 records after removing duplicates. **Figure 5** shows the searching process. No biomarker of barley intake was reported from either human or animal studies.

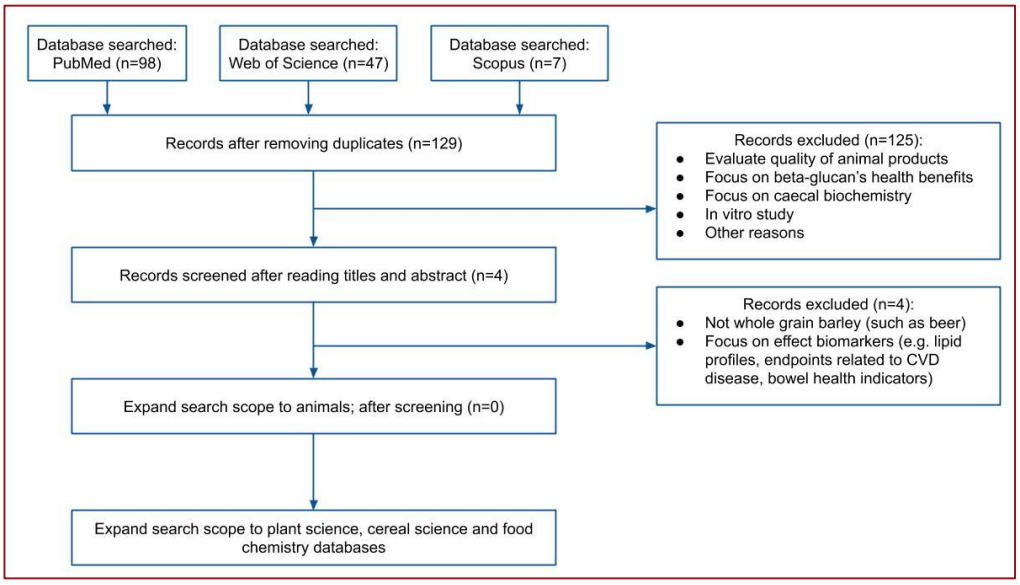


Figure 5 Diagram of Literature Searching and Screening for Articles of WG Barley Intake Biomarkers

The term *biomarkers* mentioned in the retrieved papers mostly refer to *effect biomarkers* of barley intake as defined by Dragsted (*28*) and Gao (*27*), including bowel health indicators (*51*), postprandial glucose and insulin response (*52*), lipid profiles, and endpoints related to cardiovascular disease (CVD) risk (*53*). For animal studies, the term, *biomarkers*, mostly refer to the growth of animals or quality indicators of animal-sourced products (*54*, *55*), 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 showed that some compounds are present exclusively in WG barley (**Table 2**). These compounds should be investigated as potential biomarkers of barley intake.

Table 2 Unique Compound that are Present in Barley

No	Candidate biomarker	Formula	Chemical group	Presence in Food	Reference
1	Hordenine	$C_{10}H_{15}NO$	alkaloid	germinating barley, beer and other plants	(*37*)
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)

4.1.2 Whole Grain Wheat

The literature search retrieved 312 references after removing duplicates. Some articles were found from the references of searched results. **Figure 6** shows the searching process. The results (**Table 3**) included four intervention studies and three observational studies.

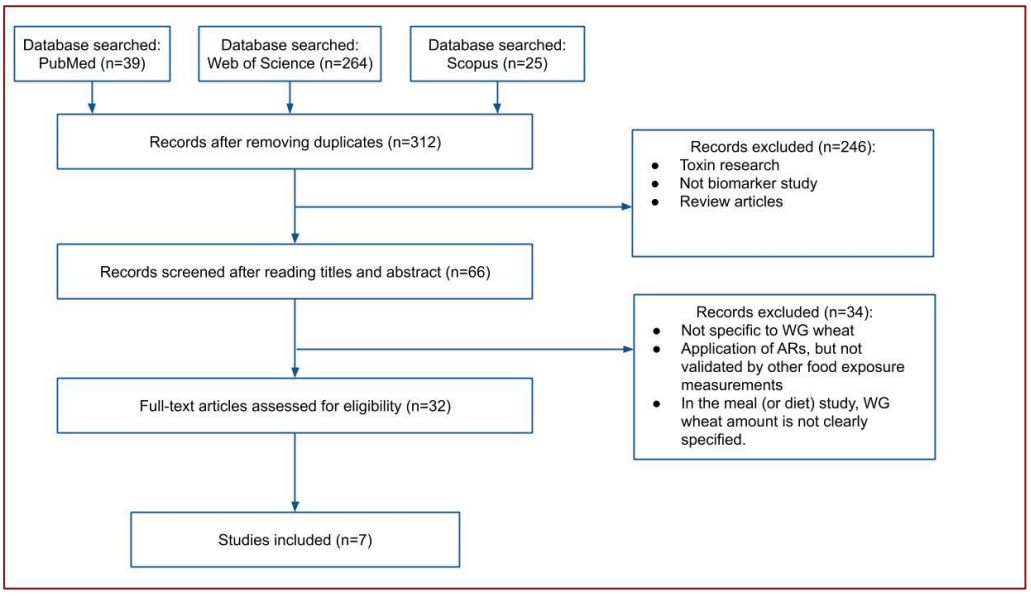


Figure 6 Diagram of Literature Searching and Screening for Articles of WG Wheat Intake Biomarkers

Table 3 Potential Biomarkers for WG Wheat Intake from Systematic Literature Review

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 HMDB0031035	(*56*)
WG wheat WG rye	73	Observational	Fasting plasma	GC-MS	Total ARs (C17:0,C19:0, C21:0,C23:0,C25:0) AR-homologue Ratio of C17:0/C21:0	HMDB0038530 HMDB0030956 HMDB0031035 HMDB0038524 HMDB0038485	(*57*)
WG wheat WG rye	39	Intervention (crossover)	Blood	GC-MS	AR-homologue Ratio of C17:0/C21:0	HMDB0038530 HMDB0031035	(*58*)
WG wheat WG rye	15	Intervention (cross-over)	Plasma and serum enterolactone	GC-MS	AR-homologue Ratio of C17:0/C21:0	HMDB0038530 HMDB0031035	(*59*)
WG wheat 3 or 6 servings	19	Intervention (cross-over, dose-response)	Fasting plasma	GC-MS	AR C19:0, C21:0, and C23:0	HMDB0030956 HMDB0031035 HMDB0038524	(*60*)
			Urine	HPLC-ECD	3,5-DHBA 3,5-DHPPA	HMDB0013677 HMDB0125533	
WGs	40	Observational	Spot urine	GC-MS	3,5-DHCA 3,5-DHBA glycine 3,5-DHPPTA	HMDB0032131 HMDB0126654 HMDB0125533	(*61*)
WGs	104	Observational	Spot urine	GC-MS	3,5-DHBA 3,5-DHPPA	HMDB0013677 HMDB0125533	(*62*)

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

In 2005, Linko (*58*) 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 (*59*). 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 (*63*). The EPIC[2] cohort study (*56*) 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

[2] European Prospective Investigation into Cancer and Nutrition

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)

In a dose-response crossover intervention study, the combination of these ARs (C19:0, C21:0, C23:0) 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 (*60*).

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 was also significant (six vs three servings, $P = 0.004$) (*60*).

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) (*61*). However, the long-term reproducibility was poor for DHBA, and modest for DHPPA (*62*).

4.2 Data Processing

4.2.1 Pre-processing and *Annotate_kudb*

Table 4 Data Pre-processing and Annotate_kudb 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

PCgroups: pseudo compound groups, annotated by CAMERA.

Data pre-processing and *Annotate_kudb* results are shown in the **Table 4**. Taking urine positive mode as an example, 1508 features were detected. After removing isotopes, multiply charged ions, and dimers, 1400 features remained. They were grouped into 1056 PCgroups. Within them, 121 were annotated by experiment database, and 168 were annotated by PredRet database. In sum, 196 grouped features were annotated.

4.2.2 Libra Calibration

Before *Libra* calibration, some features were highly correlated with injection sequence. **Figure 7** showed the distributions of *p*-value and a calibration example. After *Libra* calibration, the most significantly correlated features were calibrated.

Libra calibration did not change CV% of pooled samples. **Figure 8** showed the distribution of the CV% after and before *Libra* calibration. The median CV% of pooled samples slightly decreased from 35.07 to 34.92 after the calibration.

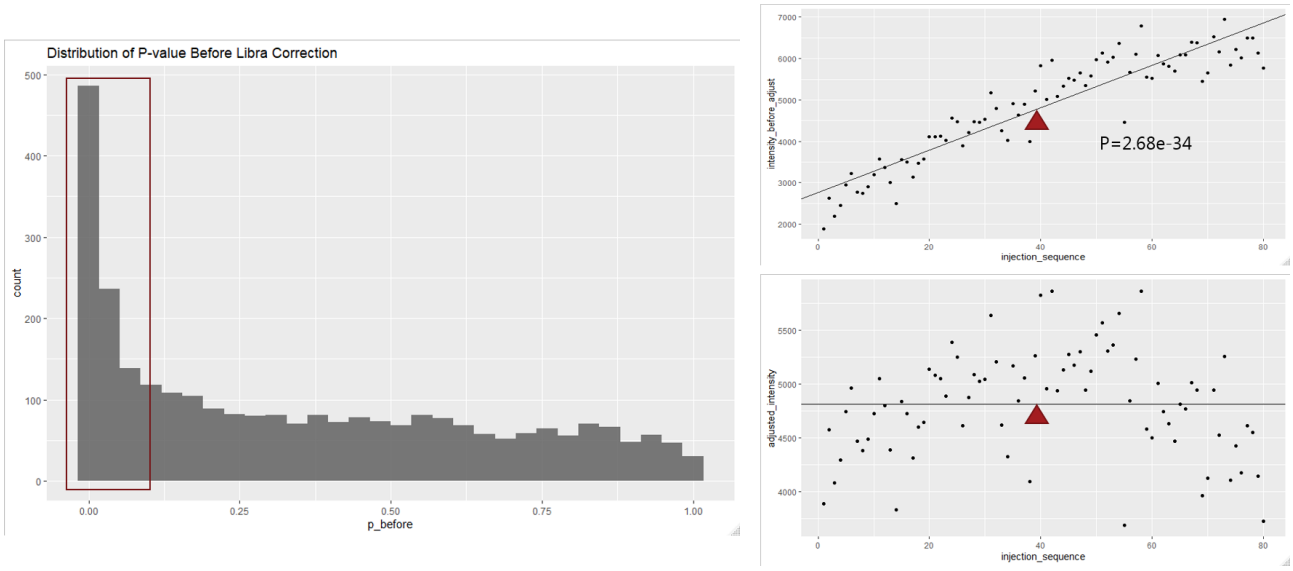


Figure 7 Distributions of p-value before Libra calibration (Left); one example of intensities before-calibration (Top Right) and after-calibration (Bottom Right)

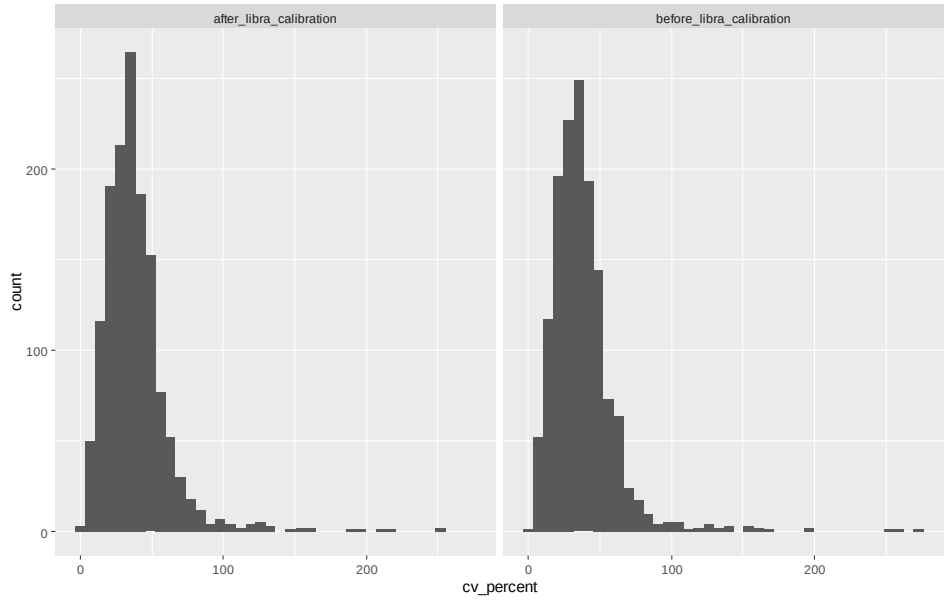


Figure 8 CV% of pooled sample before (Left) and after (Right) Libra calibration

4.2.3 Statistics

Plasma. PLS-DA modelling cannot distinguish plasma samples after the barley or wheat intervention (**Figure 9**). The t-test result showed that only 18 out of 1088 metabolites were significantly different ($p<0.05$). After multiple comparison adjustment, none of the metabolites changed significantly after whole grain intake. PCA score plot (**Figure 10**) showed that pooled samples were separated from each other from the first two components.

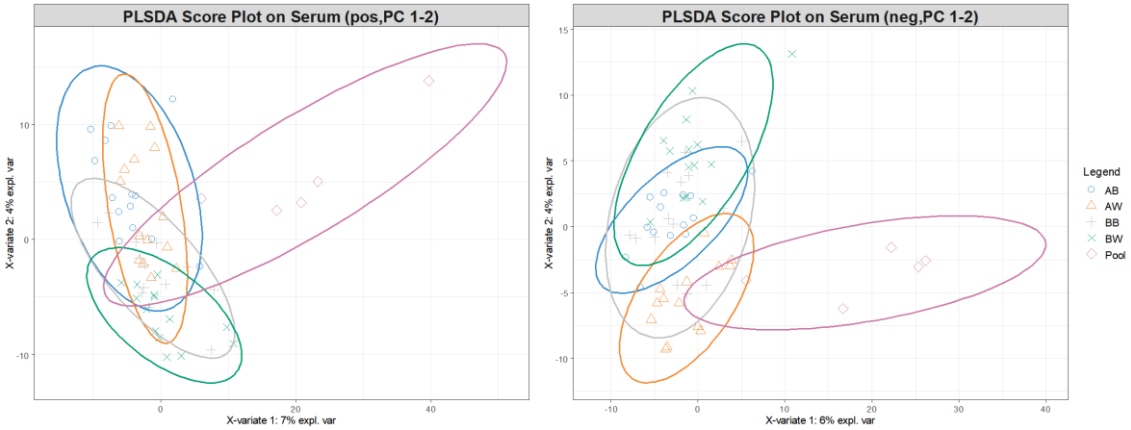


Figure 9 PLSDA Score Plot of Plasma Samples (PC 1-2)

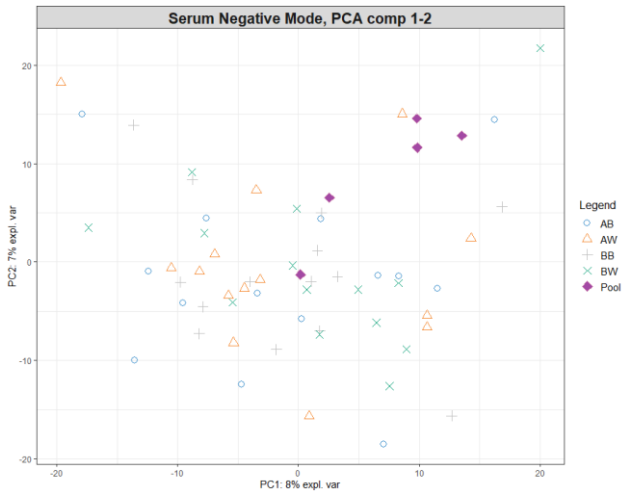


Figure 10 PCA Score Plot of Plasma Samples (PC 1-2)

Urine. The score plot of PLS-DA modelling (**Figure 11**) showed that urine samples were separated after the barley intake in the negative mode. Compared with plasma samples, more variances of urine samples were explained by the first six principal component (**Figure 12**).

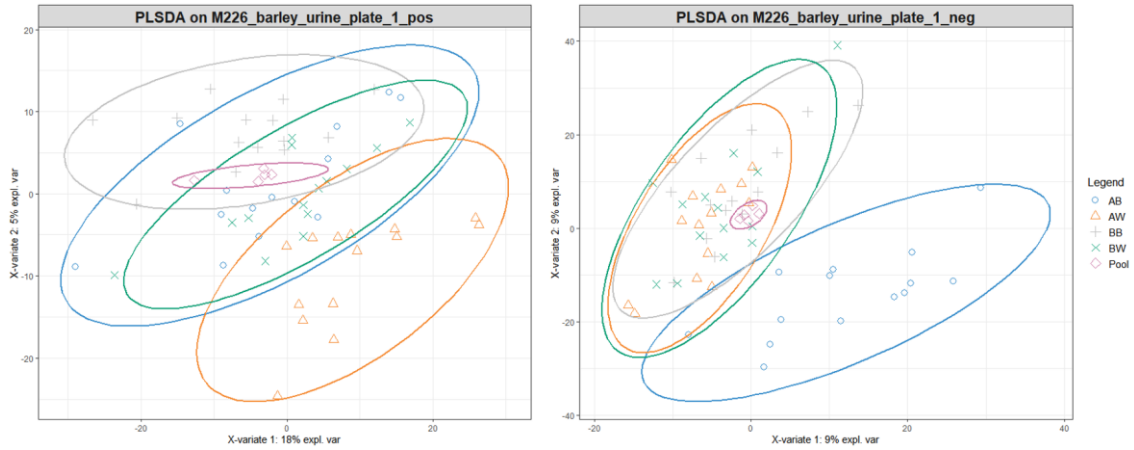


Figure 11 PLSDA Score Plot of Urine Samples

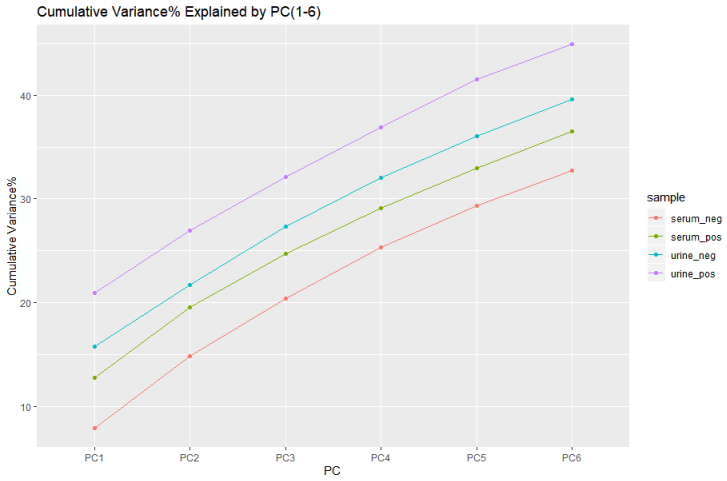


Figure 12 Cumulative Variance% Explained by PC (1-6)

25 metabolites were selected by PLS-DA modelling. Eight of them can indicate whole grain barley intake, and 17 can indicate whole grain wheat intake. The intensities of two potential biomarkers are shown in **Figure 13**. The left is for whole grain barley. The right is for whole grain wheat.

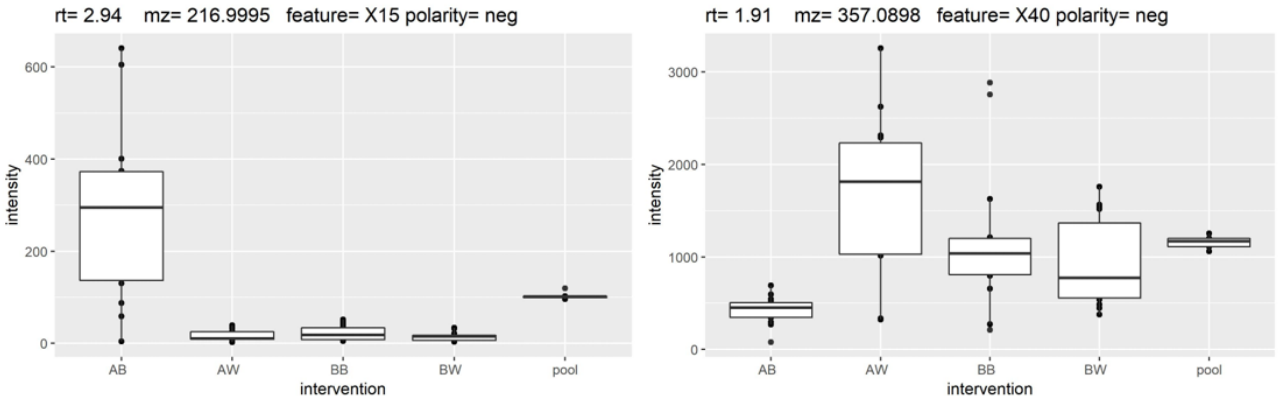


Figure 13 The Intensities of Two Potential Biomarkers for Whole Grain Barley Intake (Left) and Whole Grain Wheat Intake (Right). AB=After Barley, AW=After Wheat, BB=Before Barley, BW=Before Wheat, Pool=pooled samples.

4.3 Biomarkers of Whole Grain Barley Intake

4.3.1 Summary

The identified metabolites for whole grain barley intake is summarized in **Table 5**.

Table 5 Potential Biomarkers of Whole Grain Barley Intake Identified from the Barley Bread Intervention Study

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	113.023 341.267 399.272 499.289 517.303	[M-glucuronate-H]$^-$ [M-H_2O-H]$^-$ [M-H]$^-$	Glucuronide 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 157.1222 171.1372 325.2739 383.2629 483.2741 501.3054	[M-glucuronate-H]$^-$ [M-H_2O-H]$^-$ [M-H]$^-$	Glucuronide 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	387.16 775.34	[M-H]- [2M-H]-	Unknown[4]
7	537.33	Unknown	6.59	3.84	---	[M-H]-	Unknown glucuronide[4]
8	291.26	Unknown	6.71	4.21	---	[M-H]-	Unknown sulfate[4]
* Molecular formula corresponds to the neutral compound. Superscript notations represent level of identifications.							

4.3.2 Glucuronides and *β*-glucuronidase Experiments

Ion 501.3054 and 517.3030 could be glucuronides. Both ions showed the loss of 176 in MS^2 spectra (**Figure 22** in the **Appendix-J**), which is characteristic to glucuronide groups (*64*).

β-glucuronidase hydrolysed both glucuronide 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 and after one week of storage in -20 ºC freezer, the intensity of this ion 341.26 decreased. Its intensity was not high enough for MS^2 experiment.

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

4.3.3 Sulfur-Containing Compound

The ion 291.26 possessed a sulfur atom as indicated by its isotopic pattern (**Figure 14**), but another metabolite interfered the structure elucidation. This interfering metabolite eluted close to the biomarker and had a similar m/z: 291.23 (**Figure 15**).

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

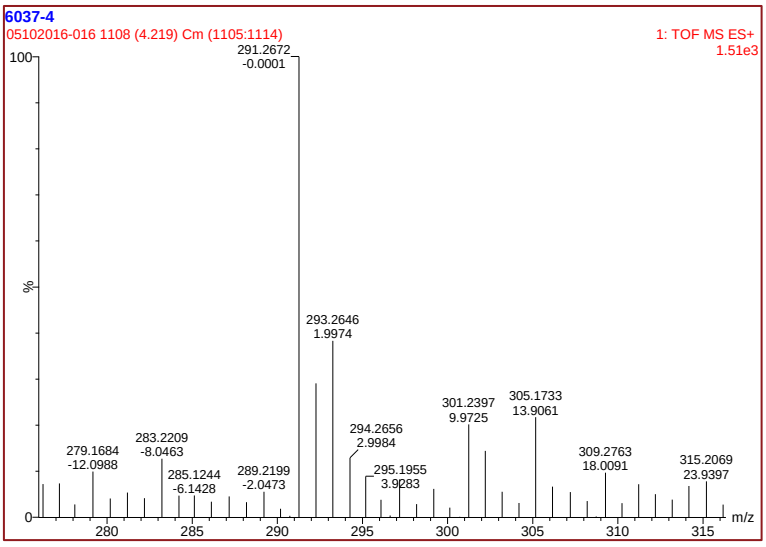


Figure 14 Isotopic Pattern of Ion 291.26

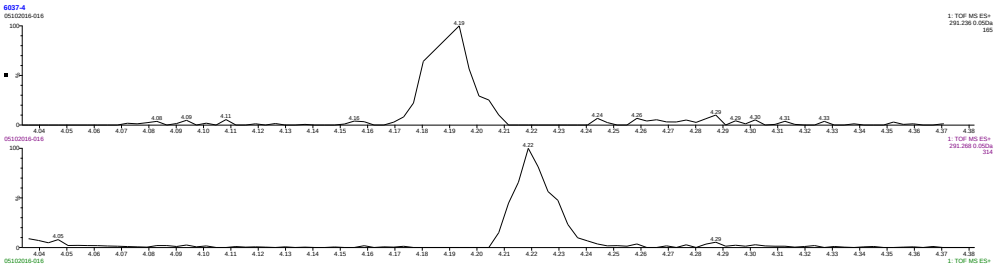


Figure 15 EIC of Endogenous Metabolite (Top) and Barley Intake Biomarker (Bottom)

4.3.4 Comparisons with Protein Source Study and Beer Study

Some discriminating metabolites of whole grain barley intake were also detected in the protein source study: six from positive mode, and nine from negative mode. However, none of these features showed reasonable excretion profiles after barleyotto intake.

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

4.4 Biomarkers of Whole Grain Wheat Intake

14 metabolites were identified as potential biomarkers for whole grain wheat intake from the intervention study. They were summarized in **Table 6**. Within these metabolites, 13 are AR metabolites, and 2-hydroxy-1,4-benzoxazin-3-one glucuronide (HBOA glucuronide) is a benzoxazinoid metabolite.

Table 6 Potential Biomarkers of Whole Grain Wheat Intake Identified from the Barley Bread Study

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	113.02 153.02 175.02	[3,5-dihydroxybenzoic acid - H]$^-$ [M-H]$^-$	3,5-DHBA glucuronide[1]
2	329.05	$C_{13}H_{14}O_{10}$	3.96	1.12	109.04 135.06 153.03 329.08	[3,5-dihydroxybenzoic acid - H]$^-$ [3,5-dihydroxybenzene - H]$^-$ [M-H]$^-$	Glucuronyl dihydroxybenzoate[2]
3	232.97	$C_7H_6O_7S$	4.07	1.37	79.96 96.96 109.03 123.05 153.01 215.09	[SO_3]$^-$ [3,5-dihydroxybenzoic acid - H]$^-$ [M-H-SO_3]$^-$ [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	109.03 153.01	[3,5-dihydroxybenzoic acid - H]$^-$ [M-H]$^-$	3,5-DHBA[2]
8	357.08	$C_{15}H_{18}O_{10}$	4.08	1.91	113.03 137.07 181.07 357.08	[M-glucuronide-H]$^-$ [M-H]$^-$	3,5-DHPPA Glucuronide[2]
9	359.09				165.05 183.07 359.09 376.12	[M+H-glucuronide-H_2O]$^+$ [M+ glucuronide H-]$^+$ [M+H]$^+$ [M+NH_4]$^+$	
10	261.00	$C_9H_{10}O_7S$	4.32	2.10	261.00 523.02 777.14	[M-H]$^-$ [2M-H]$^-$ Unknown adduct	3,5-DHPPA sulfate[2]
11	263.02	$C_9H_{12}O_7S$	5.48	3.31	79.96 153.03 168.06 183.08 248.02 263.02	[SO_3]- [M-CH_3O_4S-H] [M-H-SO_3]- [M-CH_3-H]- [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 glucuronide[3]
13	289.03	$C_{11}H_{14}O_7S$	5.18	3.37	---	[M-H]$^-$	3,5-DHPPTA sulfate[3]
14	340.06	$C_{14}H_{15}NO_9$	---	2.96	326.08 340.06	[M-CH_2-H]$^-$ [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. Superscript notations represent the level of identifications.							

5. Discussions

5.1 The Ambiguity of Using the Term Biomarker

The ambiguity of using the term *biomarker* is common in the retrieved scientific articles. 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* (*27*). In the literature search, it was difficult to quickly extract what was meant by biomarkers before reading the full text.

Dragsted (*28*) and Gao (*27*) 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 may relief this problem in the future.

5.2 Potential Biomarkers for Whole Grain Wheat Intake

The literature search results also included some *food compound intake biomarkers (FCIBs)* (*27*) including phenolic compounds (*65*), benzoxazinoids (BXOs) (*66*, *67*), phytoestrogen (*68*), phytosterol and lignan (*69*) and *effect biomarkers* such as microbial metabolites (*67*). 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-F**.

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 (*69*). Because the 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.

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

In the areas where whole grain wheat is not the only source of AR, the combination of total plasma AR and AR homologue ratio C17:0/C21:0 may indicate whole grain wheat intake. AR-homologue ratio C17:0/C21:0 was first reported by cereal scientists in 2004 to distinguish WG rye and wheat grains (*70*). 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. Because AR are present exclusively high concentration in whole grain wheat and rye (*8*), total AR can potentially indicate the total intake of whole grain rye

and wheat, while the homologue ratio of AR C17:0/C21:0 can distinguish whole grain rye and wheat.

In the areas where whole grain wheat is the only AR source, plasma AR (C19:0, C21:0, and C23:0), and AR urinary metabolites may indicate whole grain wheat intake. ARs (C19:0, C21:0, C23:0) constitutes 85% of ARs in whole grain wheat (*60*). Therefore, these three homologue in plasma might indicate whole grain wheat intake. Unlike plasma ARs, urinary AR metabolites cannot be traced back to their precursors, because all homologues undergo the same metabolism pathway. Therefore, AR metabolites cannot differentiate whole grain wheat and rye intake. But in countries where whole grain wheat is the only source of AR (for example, UK and USA), AR metabolites might be good biomarkers to measure whole grain wheat intake. However, this proposal needs to be further validated. Urinary AR metabolites also have the advantages of non-invasiveness, and longer half-life-time compared with plasma AR (*71*).

5.4 Metabolomics Analysis of the Barley Intervention Study

5.4.1 Urinary AR metabolites

From the intervention study, ten of the identified metabolites differentially increased after the consumption of WG wheat breads are AR metabolites, most of which were proposed within AR metabolism pathway (**Figure 16**). The barley bread intervention study used 24-h pooled urine samples, which accumulated a broad range of metabolites after the intake of barley.

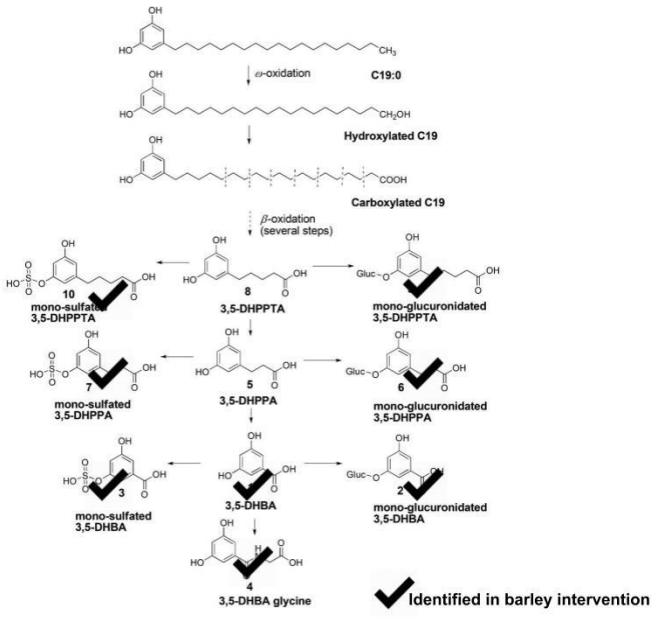


Figure 16 Detected AR Metabolites. Adapted from Zhu et al., J Nutr, 2014, 144(2).

Two novel AR metabolites were identified from the barley bread study, glucoronyl dihydroxybenzoate and DHMBA sulfate. Glucuronyl dihydroxybenzoate is an isomer of 3,5-

DHBA glucuronide (**Figure 17**), which is formed during phase II metabolism by substituting the phenol group. The retention time of DHMBA sulfate in the barley bread intervention study (3.31 min) was different from the Sysdiet study (0.78 min) (unpublished work). Therefore, they could be isomers.

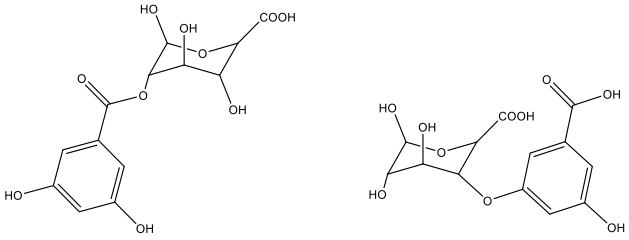


Figure 17 Structure of 3,5-DHBA Glucuronide and Glucuronyl Dihydroxybenzoate

The barley bread intervention was one of the few studies that detected urinary AR metabolites by LC-MS. Most of the studies used GC-MS to detect urinary AR metabolites (*61*, *62*).

5.4.2 Plasma AR

In the barley bread intervention study, intact ARs from fasting plasma were not detected. The half-life-time ($t_{1/2}$) for intact ARs are around 5 h (*72*). Therefore, the intact ARs could be fully metabolized and excreted, or partially metabolized to the concentration below the limit of detection.

5.5 Identification of the Interfering Ion: Androsterone

The interfering ion is identified as androsterone with level II confidence. The MS^2 spectra of androsterone show the similarities with sitostanol, the plant-originated steroid alcohol (**Figure 24**). CFM-ID predicted spectra also show a similar pattern.

Androsterone is an endogenous steroid hormone, neurosteroid, and putative pheromone (*73*). The elevated urinary concentration of androsterone have been associated with several diseases, such as hirsutism (*74*). Androsterone has also been proposed as a potential biomarker to diagnose schizophrenia together with other endogenous steroids (*75*). In doping tests, the ratio of urinary androsterone and etiocholanolone can be used to monitor urine manipulation.

The concentration of androsterone does not seem to vary before and after the intervention, but the concentration seems higher in males than females (1.3-fold, p=0.09), as shown in **Figure 18**Error! Reference source not found..

Androsterone is also detectable in the Fatmed, Sysdiet, Coffee, and Seaweed studies (not covered in this thesis). These studies covers most adult age groups with or without metabolic syndromes

and mostly without any barley intake. However, the distribution in different genders is not investigated due to the lack of information.

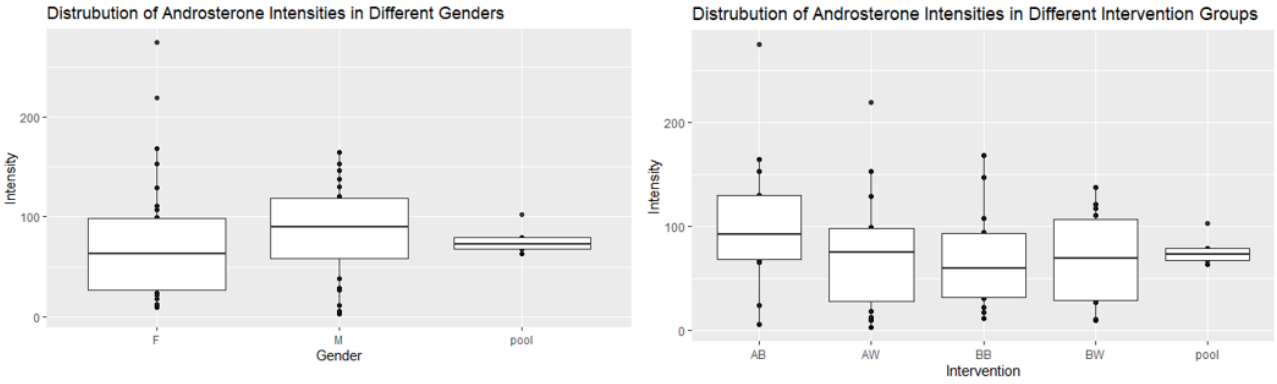


Figure 18 Distribution of Androsterone Intensities in Different Genders (Left, F=Female, M=Male) and Different Intervention Groups (AB=After Barley, AW=After Wheat, BB= Before Barley, BW= Before Wheat)

5.6 Data Processing Workflow

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

5.6.1 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 introduce noise'*. 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 19** 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.

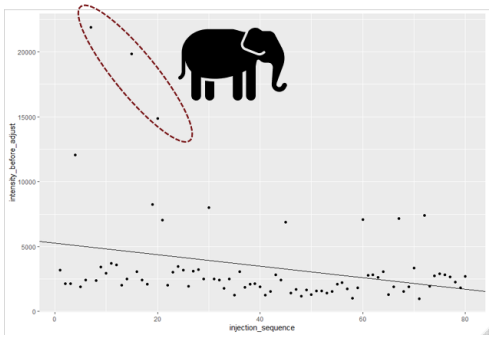


Figure 19 Elephants in the data

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*.

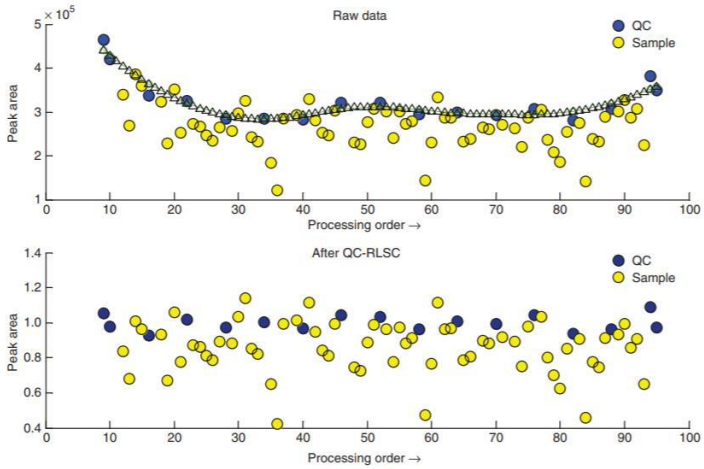


Figure 20 Intra-batch effects calibrated by LOESS model, adapted from Dunn et al., Nat Protoc. 2011

Libra might be a more appropriate calibration tool when the analysis includes too few QC samples, while LOESS might introduce artifacts when smoothing the intensities. The good smooth effect by LOESS can be achieved when 20% of samples are QC samples (*76*) (**Figure 20**).

5.6.2 Annotate_kudb

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

5.6.3 Comparisons of New and Old Workflows

The primary advantage of the new data workflow is that it incorporates all data processing steps into the R programming language. **Table 7** compares the new workflow (**3.3.3 Streamlined Data Processing Workflow**) with the old one (**3.3.2 Data Processing Workflow**). 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 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.) - only be able to only convert waters *.raw* to *.cdf* - Proprietary software - Only available in Windows OS	ProteoWizard MSconvert	- Robust - Able to convert from multiple formats to multiple formats - Open-source and freely-available - Available on all platforms - User friendly interface	---
Data preprocessing	MZmine	- Open-source - Good visualisation functions - Informative manuals and documentations - Easy-to-use with graphic user interface	- High demand of hardware (at least 32 Gb ROM) - Programmed in Java. Difficult to incorporate into the workflows seamlessly	R package XCMS	- Open-source - Easy to incorporate into workflow seamlessly	- Poor visualisation functions - Hard-to-learn
Annotation and identification	Manual Selection	- The capability of identifying novel compounds	- Time-consuming - 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 - Not delivered as a packaged. Difficult to be re-used by other users.
Statistics	PLS Toolbox	- Powerful multivariable data analysis - Good customer support - User friendly interface	- Programmed in proprietary language (MATLAB) - Proprietary software	R package stats R package mixOmics	- Open-source - In active development - Contributions from other peer-users	---

6. Conclusions

For whole grain barley, the systematic literature review showed that no biomarker can indicate the intake of whole grain barley. The metabolomics study identified 4-hydroxybenzoic acid-4-sulphate and some unknown metabolites as potential biomarkers for whole grain barley intake.

For whole grain wheat, the systematic literature review showed that, in the areas where whole grain wheat is not the only source of AR, the combination of total plasma AR and AR homologue ratio C17:0/C21:0 may indicate whole grain wheat intake. While in the areas where whole grain wheat is the only AR source, plasma AR, AR (C19:0, C21:0, and C23:0), and AR urinary metabolites may indicate whole grain wheat intake. The metabolomics study identified 10 AR metabolites (including two novel isomers) and HBOA glucuronide as potential biomarkers for whole grain wheat intake.

The metabolomics data processing workflow was improved by incorporating free-available tools and self-developed functions into a unified programming language, R. This new toolset was validated by being applied to investigate the biomarkers of whole grain barley and wheat intake.

7. Perspectives

The identification of barley intake biomarkers was 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 may 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 may be included as the source for matching.

Batch effect calibration is a critical step to guarantee data qualities (*77*). However, few bioinformatics tools and research have been devoted to this area. *Libra* could be a potentially useful tool. *Libra* should be systematically tested with more data set and the performance should evaluated by comparing with tools.

8. Acknowledgements

Thank you! My dear supervisors Lars Ove Dragsted, and Gözde Gürdeniz. Thanks for your guidance and help during my thesis. Lars impressed me as a good teacher. He can always explain the complex concepts in a vivid and understandable manner. And, Gözde offered her babysitting-styled supervision and helped me in every detail. I also acknowledge Susanne Bügel for offering me the chance to analyze the data, and the examiner Ashfaq Ali for assessing my thesis.

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

The researchers from the Section of Preventive and Clinical Nutrition are also acknowledged. Steen, and Inge shared their knowledge of fibers and cardiovascular disease. I really acknowledge my friend Simon. Besides the help in science, Simon also inspired 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 thank my peer-reviewers, Nina Sondrup, Simon Schacht, and Kristina Pigsborg Jensen, for their comments and suggestions for the thesis.

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

9. Reference

1. S. E. Ullrich, *Barley: Production, improvement, and uses* (John Wiley & Sons, 2010), vol. 12.
2. M. F. Piepoli, A. W. Hoes, S. Agewall, C. Albus, C. Brotons, A. L. Catapano, M.-T. Cooney, U. Corrà, B. Cosyns, C. Deaton, I. Graham, M. S. Hall, F. D. R. Hobbs, M.-L. Løchen, H. Löllgen, P. Marques-Vidal, J. Perk, E. Prescott, J. Redon, D. J. Richter, N. Sattar, Y. Smulders, M. Tiberi, H. Bart van der Worp, I. van Dis, W. M. M. Verschuren, W. M. M. Verschuren, 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Atherosclerosis*. **252**, 207–274 (2016).
3. Nordic Council of Ministers., *Nordic nutrition recommendations 2012 : integrating nutrition and physical activity*.
4. FAO, *FAO Statistics* (2016).
5. A. C. Newton, A. J. Flavell, T. S. George, P. Leat, B. Mullholland, L. Ramsay, C. Revoredo-Giha, J. Russell, B. J. Steffenson, J. S. Swanston, W. T. B. Thomas, R. Waugh, P. J. White, I. J. Bingham, 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).
6. B. Baik, S. E. Ullrich, Barley for food : Characteristics , improvement , and renewed interest. *J. Cereal Sci.* **48**, 233–242 (2008).
7. S. R. Baerson, J. Schröder, D. Cook, A. M. Rimando, Z. Pan, F. E. Dayan, B. P. Noonan, S. O. Duke, Alkylresorcinol biosynthesis in plants New insights from an ancient enzyme family? *Plant Signal. Behav.* **5**, 1286–1289 (2010).
8. Y. Suzuki, Y. Esumi, I. Yamaguchi, Structures of 5-alkylresorcinol-related analogues in rye. *Phytochemistry*. **52**, 281–289 (1999).
9. A. B. Ross, M. J. Shepherd, M. Schüpphaus, V. Sinclair, B. Alfaro, A. Kamal-Eldin, P. Åman, Alkylresorcinols in cereals and cereal products. *J. Agric. Food Chem.* **51**, 4111–4118 (2003).
10. J. Slavin, Whole grains and human health. *Nutr. Res. Rev.* **17**, 99–110 (2004).
11. J. Slavin, Why whole grains are protective: biological mechanisms. *Proc. Nutr. Soc.* **62**, 129–134 (2003).
12. 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).
13. R. M. van Dam, F. B. Hu, Are alkylresorcinols accurate biomarkers for whole grain intake? *Am. J. Clin. Nutr.* **87**, 797–798 (2008).
14. S. S. Jonnalagadda, L. Harnack, R. H. Liu, N. McKeown, C. Seal, S. Liu, G. C. Fahey, 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).
15. O. Fiehn, Metabolomics - The link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
16. Anonymous, Signs of a long life. *Economist*. **387** (2008).
17. C. B. Clish, Metabolomics: an emerging but powerful tool for precision medicine. *Mol. Case Stud.* **1**, a000588 (2015).
18. 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).
19. 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).
20. 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).
21. T. Cajka, O. Fiehn, Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal. Chem.* **88** (2016), pp. 524–545.
22. M. Witting, P. Schmitt-Kopplin, in *Comprehensive Analytical Chemistry* (Elsevier, 2014; https://www.sciencedirect.com/science/article/pii/B9780444626516000180), vol. 63, pp. 421–442.
23. J. L. Markley, R. Brüschweiler, A. S. Edison, H. R. Eghbalnia, R. Powers, D. Raftery, D. S. Wishart, The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **43** (2017), pp. 34–40.
24. K. Bingol, L. Bruschweiler-Li, D. Li, B. Zhang, M. Xie, R. Brüschweiler, Emerging new strategies for successful metabolite identification in metabolomics. *Bioanalysis*. **8**, 557–573 (2016).
25. S. Beisken, M. Eiden, R. M. Salek, Getting the right answers: understanding metabolomics challenges. *Expert Rev. Mol. Diagn.* **15**, 97–109 (2015).
26. R. Spicer, R. M. Salek, P. Moreno, D. Cañueto, C. Steinbeck, Navigating freely-available software tools for metabolomics analysis. *Metabolomics*. **13**, 1–16 (2017).
27. Q. Gao, G. Praticò, A. Scalbert, G. Vergères, M. Kolehmainen, C. Manach, L. Brennan, L. A. Afman, D. S. Wishart, C. Andres-Lacueva, M. Garcia-Aloy, H. Verhagen, E. J. M. Feskens, L. O. Dragsted, A scheme for a flexible classification of dietary and health biomarkers. *Genes Nutr.* **12**, 34 (2017).
28. L. O. Dragsted, Q. Gao, G. Praticò, C. Manach, D. S. Wishart, A. Scalbert, E. J. M. Feskens, Dietary and health biomarkers—time for an update. *Genes Nutr.* **12**, 1–7 (2017).
29. G. Praticò, Q. Gao, A. Scalbert, G. Vergères, M. Kolehmainen, C. Manach, L. Brennan, S. H. Pedapati, L. A. Afman, D. S. Wishart, R. Vázquez-Fresno, C. A. Lacueva, M. Garcia-Aloy, H. Verhagen, E. J. M. Feskens, L. O. Dragsted, 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.
30. D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vazquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
31. FooDB, (available at http://foodb.ca/).
32. A. Medina-Remón, C. Manach, C. Knox, D. S. Wishart, J. Perez-Jimenez, J. A. Rothwell, N. M'Hiri, P. García-Lobato, R. Eisner, V. Neveu, A. Scalbert, 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.
33. Dictionary of food compounds with CD-ROM. *Choice Rev. Online* (2013), doi:10.5860/choice.51-1824.
34. V. Popovici, thesis, University of Copenhagen (2016).
35. M.-B. S. Andersen, H. C. Reinbach, Å. Rinnan, T. Barri, C. Mithril, L. O. Dragsted, 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).
36. 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.
37. 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.
38. 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.
39. R core team, R: A language and environment for statistical computing (2013).
40. M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, P. Mallick, A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
41. R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*. **9**, 504 (2008).
42. 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).
43. 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).
44. 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).
45. J. Stanstrup, S. Neumann, U. Vrhovšek, PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal. Chem.* **87**, 9421–9428 (2015).
46. H. Wickham, Tidy Data. *J. Stat. Software; Vol 1, Issue 10* (2014) (available at https://www.jstatsoft.org/v059/i10).
47. 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).
48. E. L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H. P. Singer, J. Hollender, Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
49. Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, D. S. Wishart, CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. *Metabolites*. **9** (2019), doi:10.3390/metabo9040072.
50. Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, D. S. Wishart, BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminform.* **11**, 2 (2019).
51. A. R. Bird, M. S. Vuaran, R. A. King, M. Noakes, J. Keogh, M. K. Morell, D. L. Topping, 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).
52. N. Ames, H. Blewett, J. Storsley, S. J. Thandapilly, P. Zahradka, C. Taylor, 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).
53. 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).
54. 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).
55. A. P. Foster, T. G. Knowles, A. H. Moore, P. D. G. Cousins, M. J. Day, E. J. Hall, 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).
56. C. Kyro, A. Olsen, R. Landberg, G. Skeie, S. Loft, P. Aman, M. Leenders, V. K. Dik, P. D. Siersema, T. Pischon, J. Christensen, K. Overvad, M.-C. Boutron-Ruault, G. Fagherazzi, V. Cottet, T. Kuehn, J. Chang-Claude, H. Boeing, A. Trichopoulou, C. Bamia, D. Trichopoulos, D. Palli, V. Krogh, R. Tumino, P. Vineis, S. Panico, P. H. Peeters, E. Weiderpass, T. Bakken, L. A. Asli, M. Argueelles, P. Jakszyn, M.-J. Sanchez, P. Amiano, J. M. Huerta, A. Barricarte, I. Ljuslinder, R. Palmqvist, K.-T. Khaw, N. Wareham, T. J. Key, R. C. Travis, P. Ferrari, H. Freisling, M. Jenab, M. J. Gunter, N. Murphy, E. Riboli, A. Tjonneland, H. B. Bueno-de-Mesquita, 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.
57. 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).
58. K. J. Raninen, J. E. Lappi, M. L. Mukkala, T.-P. Tuomainen, H. M. Mykkanen, K. S. Poutanen, O. J. Raatikainen, Fiber content of diet affects exhaled breath volatiles in fasting and postprandial state in a pilot crossover study. *Nutr. Res.* **36**, 612–619 (2016).
59. 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).
60. N. M. McKeown, M. Marklund, J. Ma, A. B. Ross, A. H. Lichtenstein, K. A. Livingston, P. F. Jacques, H. M. Rasmussen, J. B. Blumberg, C.-Y. O. Chen, 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).
61. R. Landberg, R. Wierzbicka, L. Shi, S. Nybacka, A. Kamal-Eldin, B. Hedblad, A. K. Lindroos, A. Winkvist, H. B. Forslund, 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).
62. R. Landberg, M. K. Townsend, N. Neelakantan, Q. Sun, L. Sampson, D. Spiegelman, R. M. van Dam, 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).
63. 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).
64. K. Levsen, H.-M. Schiebel, B. Behnke, R. Dötzer, W. Dreher, M. Elend, H. Thiele, Structure elucidation of phase II metabolites by tandem mass spectrometry: an overview. *J. Chromatogr. A.* **1067**, 55–72 (2005).
65. L. Bresciani, F. Scazzina, R. Leonardi, E. Dall'Aglio, M. Newell, M. Dall'Asta, C. Melegari, S. Ray, F. Brighenti, D. Del Rio, Bioavailability and metabolism of phenolic compounds from wholegrain wheat and aleurone-rich wheat bread. *Mol. Nutr. Food Res.* **60**, 2343–2354 (2016).
66. B. M. Jensen, K. B. Adhikari, H. J. Schnoor, N. Juel-Berg, I. S. Fomsgaard, L. K. Poulsen, 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).
67. M. Garcia-Aloy, R. Llorach, M. Urpi-Sarda, S. Tulipani, J. Salas-Salvado, M. Angel

Martinez-Gonzalez, D. Corella, M. Fito, R. Estruch, L. Serra-Majem, C. Andres-Lacueva, Nutrimetabolomics fingerprinting to identify biomarkers of bread exposure in a free-living population from the PREDIMED study cohort. *METABOLOMICS*. **11**, 155–165 (2015).
68. 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.
69. 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).
70. 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).
71. 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** (2016), doi:10.1038/srep36278.
72. R. Landberg, A.-M. Linko, A. Kamal-Eldin, B. Vessby, H. Adlercreutz, P. Åman, Human Plasma Kinetics and Relative Bioavailability of Alkylresorcinols after Intake of Rye Bran. *J. Nutr.* **136**, 2760–2765 (2006).
73. Wikipedia, Androsterone, (available at https://en.wikipedia.org/wiki/Androsterone).
74. 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).
75. 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).
76. W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. D. Knowles, A. Halsall, J. N. Haselden, A. W. Nicholls, I. D. Wilson, D. B. Kell, R. Goodacre, T. H. S. M. (HUSERMET) Consortium, 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).
77. R. Wehrens, J. A. Hageman, F. van Eeuwijk, R. Kooke, P. J. Flood, E. Wijnker, J. J. B. Keurentjes, A. Lommen, H. D. L. M. van Eekelen, R. D. Hall, R. Mumm, R. C. H. de Vos, Improved batch correction in untargeted MS-based metabolomics. *Metabolomics*. **12**, 88 (2016).
78. E. M. Keaveney, R. K. Price, L. L. Hamill, J. M. W. Wallace, H. McNulty, M. Ward, J. J. Strain, P. M. Ueland, A. M. Molloy, V. Piironen, W. von Reding, P. R. Shewry, J. L. Ward, R. W. Welch, 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).

10. Appendix

A. 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).

B. Streamlined Data Processing Workflow

Table 8 XCMS parameters for data pre-processing

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 for annotation

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

C. 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
```

D. Source Code of R Function m2r

Available on GitHub: github.com/tuhulab/bfi-wholegrain/blob/master/r/m2r.R

E. 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)
}
```

F. Biomarkers of Whole Grain Barley and Wheat Intake

Table 10 Putative Biomarkers for Whole Grain Wheat Intake from Systematic Literature Review, but Not Specific

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)	(*78*)
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	(*67*)

G. 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

H. 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

I. Retention time and mass-to-chage ratio of of alkylresorcinol metabolites

Table 11 Retention time and mass-to-chage ratio 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.

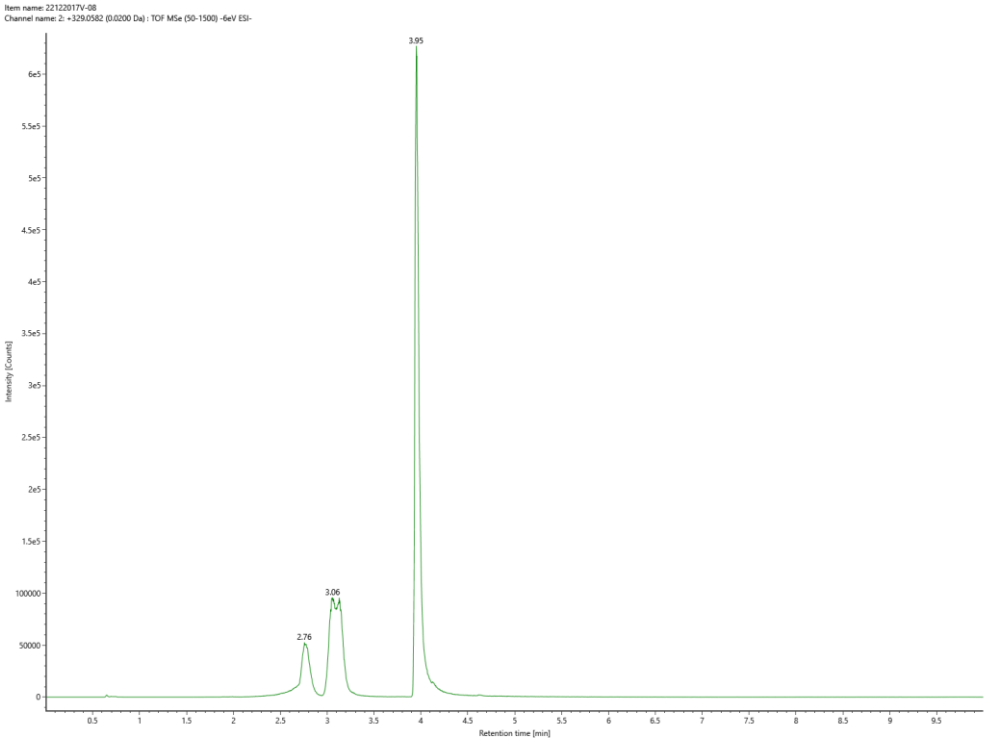
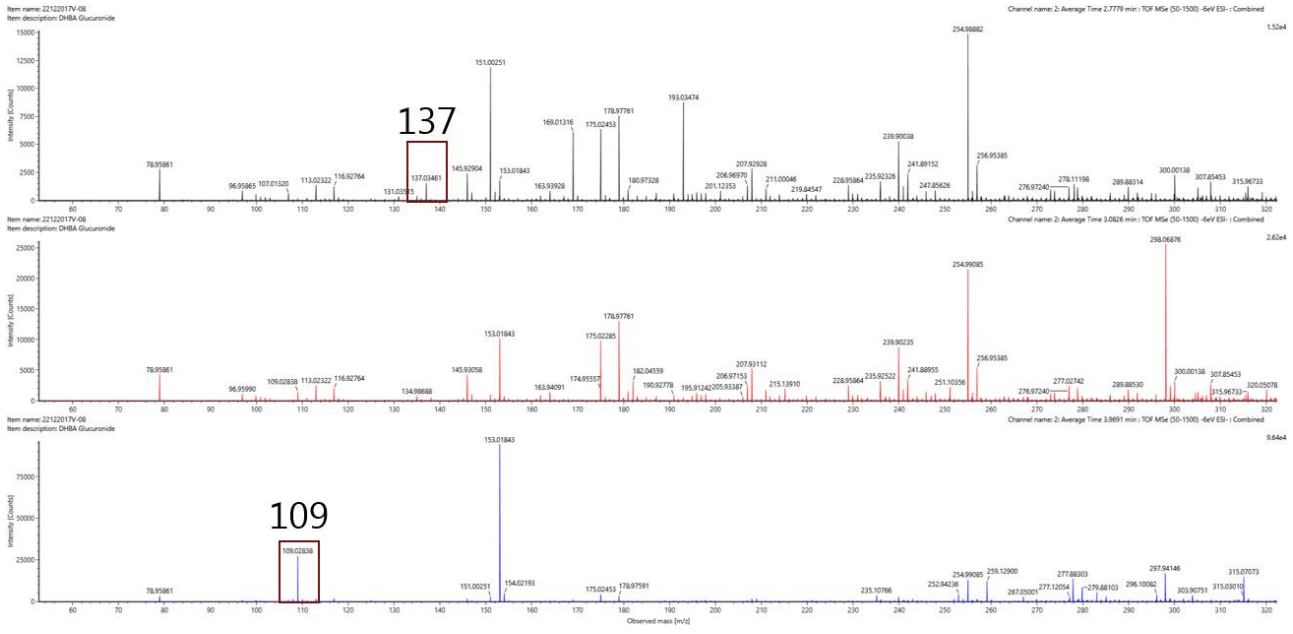


Figure 21 Extracted Ion Chromatogram of Glucuronidation Products of 3,5-DHBA



J. MS^2 spectra

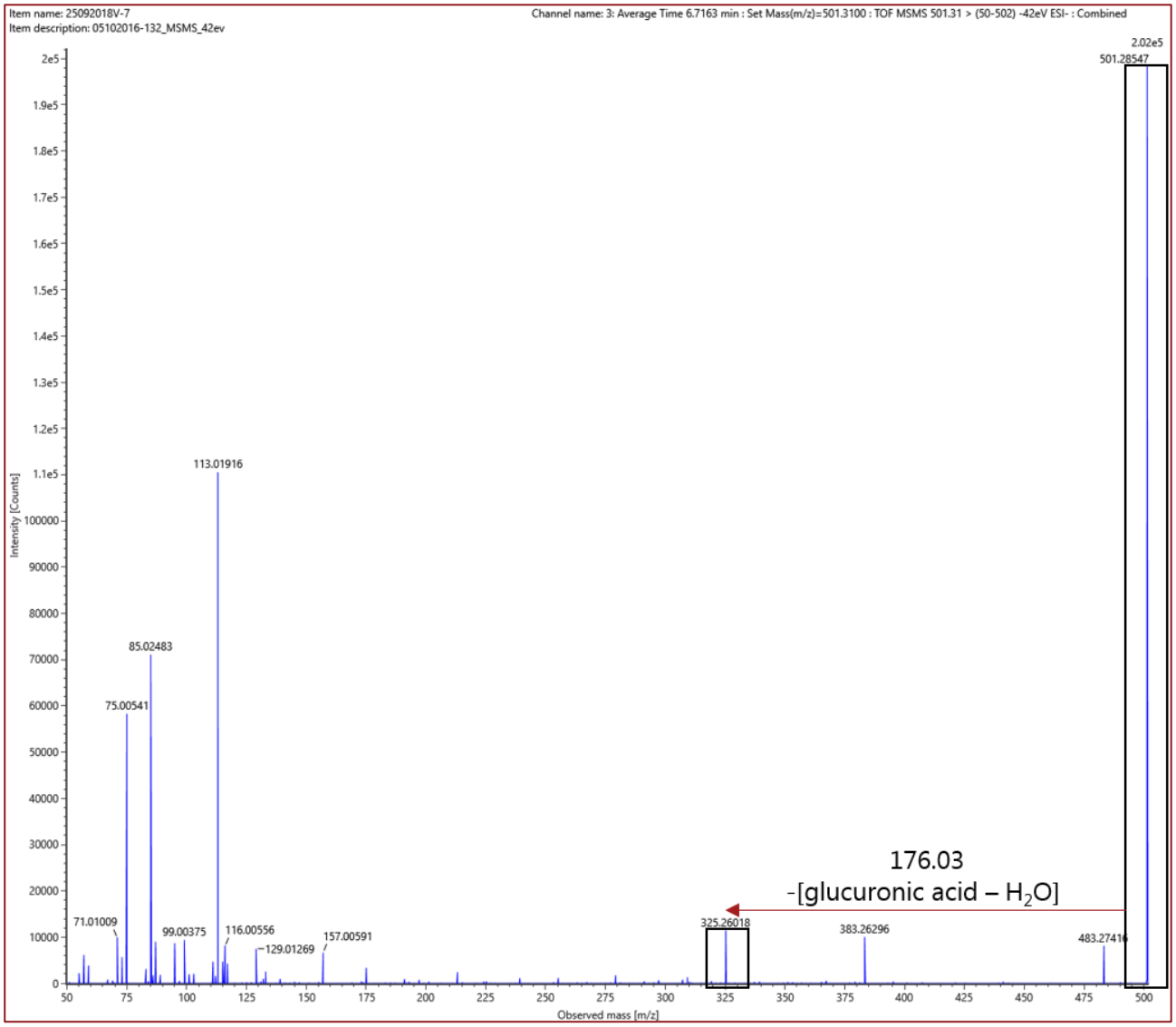


Figure 22 MS/MS spectra of glucuronide ion 501

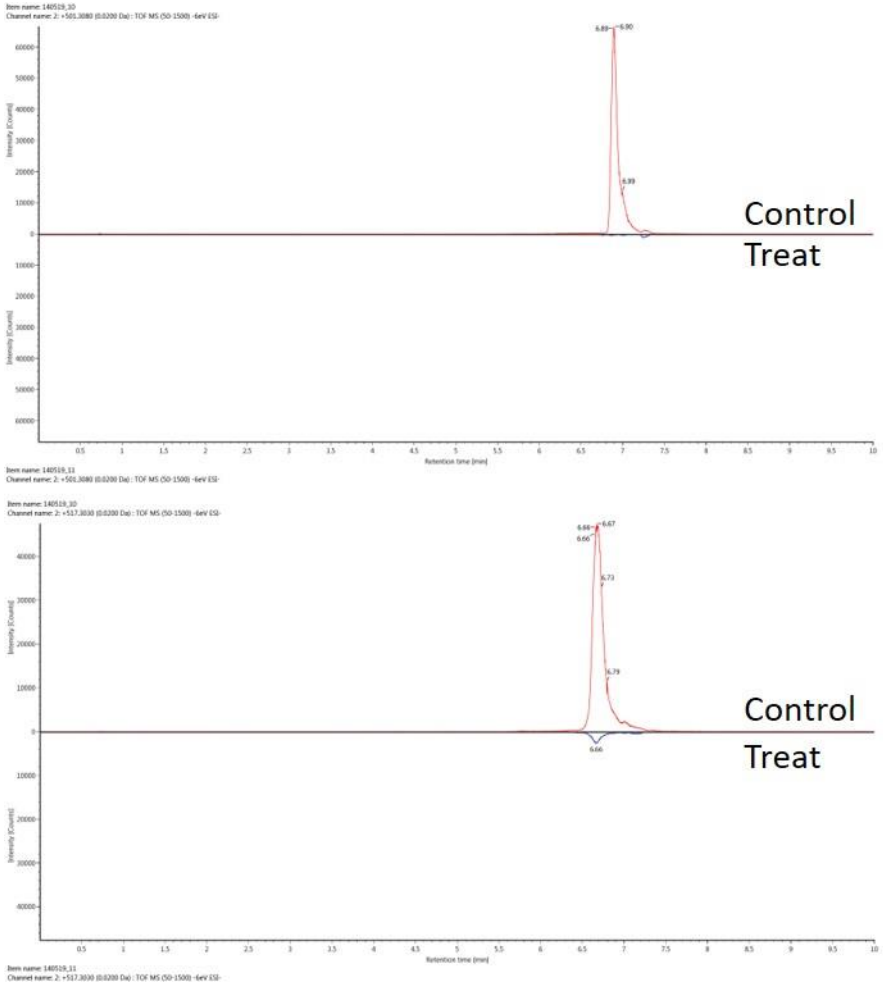


Figure 23 Extracted Ion Chromatogram of Glucuronide Ions (Top: 501, Bottom: 517)

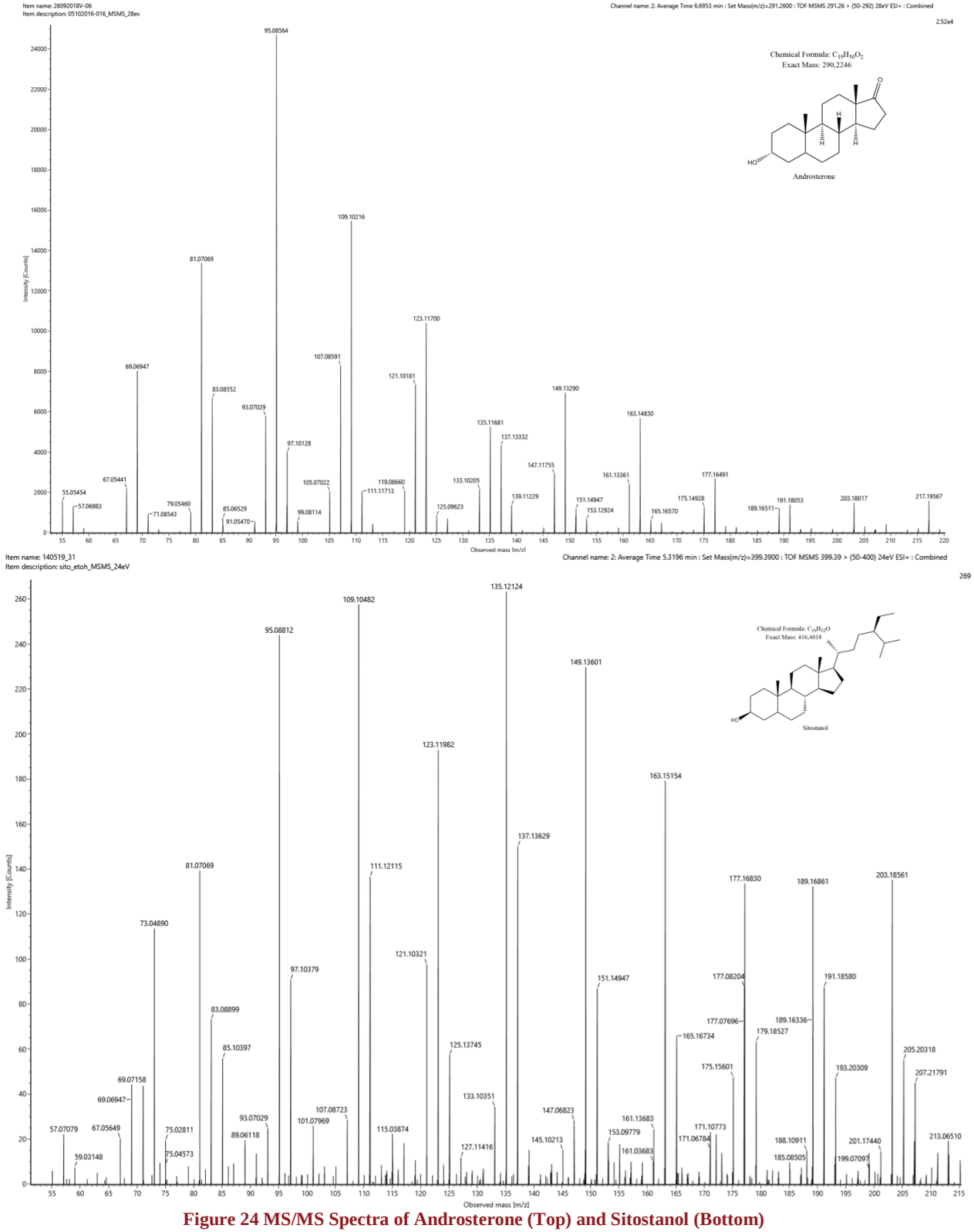


Figure 24 MS/MS Spectra of Androsterone (Top) and Sitostanol (Bottom)

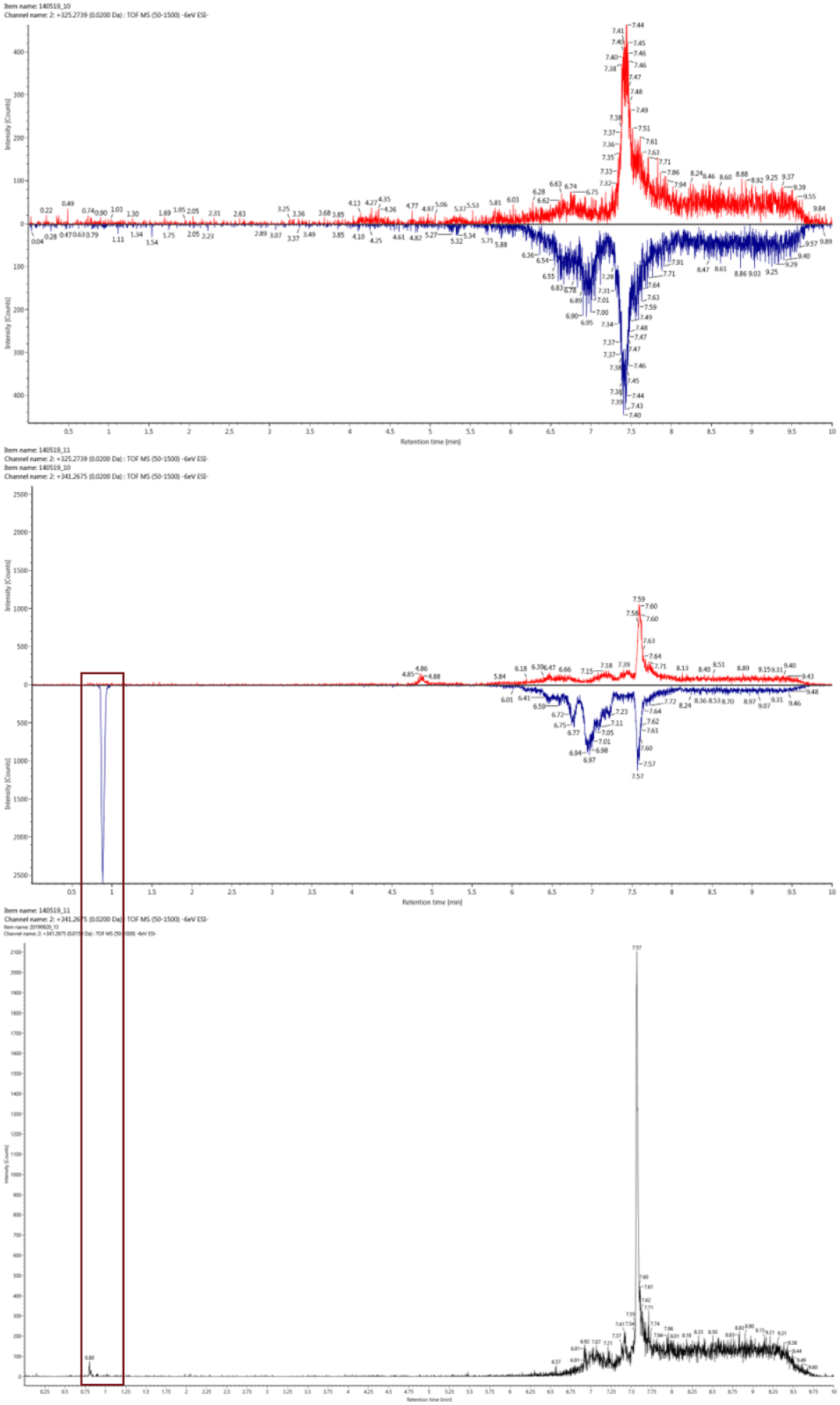


Figure 25 EIC of expected ions