Untargeted Metabolomics and Novel Data Analysis Strategies to Identify Biomarkers of Diet and Type 2 Diabetes

Lin Shi

Faculty of Natural Resources and Agricultural Sciences

Department of Molecular Sciences

Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2017

Acta Universitatis agriculturae Sueciae
2017:108

Cover: Important words of the thesis
(Photo: Lin Shi)

ISSN 1652-6880
ISBN (print version) 978-91-7760-106-7
ISBN (electronic version) 978-91-7760-107-4

Print: SLU Service/Repo, Uppsala, 2017

Untargeted Metabolomics and Novel Data Analysis Strategies to Identify Biomarkers of Diet and Type 2 Diabetes

Abstract

Type 2 diabetes (T2D) is a major global health problem and prevention could be improved by identifying individuals at risk at an early stage, followed by preventive strategies, *e.g.*, dietary modifications. Untargeted LC-MS metabolomics offers the possibility to identify predictive biomarkers that may improve risk prediction and dietary biomarkers that may facilitate investigation of diet-T2D relationships. However, untargeted metabolomics generates large-scale data, resulting in demanding data processing and statistical analyses preceding meaningful biological interpretation.

The work presented in this thesis sought to develop bioinformatics tools for dealing with large-scale data generated from untargeted LC-MS metabolomics, to apply such tools to identify predictive metabolites of T2D and metabolites related to predefined healthy Nordic dietary indices, and to investigate whether such metabolites are associated with T2D risk in a Swedish population.

Two novel R programming based packages were developed: 'batchCorr', a data-processing strategy to correct for within- and between-batch variability in LC-MS experiments, and 'MUVR', a statistical framework for multivariate analysis with unbiased variable selection. These tools were applied on untargeted LC-MS metabolomics data obtained from plasma samples from a nested case-control study. Overall, 46 predictive metabolites of T2D were identified. Several metabolites showed good long-term reproducibility among healthy participants, reinforcing their potential as predictive biomarkers, while some changed in the disease-associated direction among cases, reflecting disease progression. In total, 38 metabolites were found to be associated with two predefined healthy Nordic dietary indices. No evidence was found to support association between indices and T2D risk. Instead, metabolites related to unhealthy foods not captured in indices were associated with increased risk.

In conclusion, the novel bioinformatics tools developed here can overcome vital data-analytical challenges inherent in large-scale untargeted metabolomics studies. Predictive metabolites have great potential to provide information related to T2D pathophysiology and monitoring of disease progression, though only a limited improvement in disease prediction was achieved when adding them to models based on optimally selected traditional risk factors. Moreover, no evidence was found of an association between healthy Nordic dietary indices and T2D risk. Future studies should investigate how diet/lifestyle risk factors affect pathological pathways of T2D and prevent disease development by integration of multi-omics techniques and traditional methods.

Keywords: Biomarkers, bioinformatics, healthy Nordic dietary index, multivariate analysis, nested case-control study, risk prediction, type 2 diabetes, untargeted LC-MS metabolomics

Author's address: Lin Shi, Department of Molecular Sciences, SLU, P.O. Box7015, 75007 Uppsala, Sweden

Dedication

To my family

“Dear me, one day I'll make you proud.”

Charlotte Eriksson

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Brunius C.*, **Shi L**., Landberg R. (2016). Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* 12(173), 1-13.

II. **Shi L.*,** Rosén J., Westerhuis J.A., Landberg R., Brunius C. (2017). Unbiased variable selection and validation in multivariate modelling (*Submitted*).

III. **Shi L**.*, Brunius C., Lehtonen M., Auriola S., Bergdahl I.A., Rolandsson O., Hanhineva K., Landberg R. (2017). Plasma metabolites associated with type 2 diabetes in a Swedish population – A case-control study nested in a prospective cohort. *Diabetologia* (*In press*).

IV. **Shi L.*,** Brunius C., Johansson J., Bergdahl J.A., Lindahl B., Hanhineva K., Landberg R. (2017) Plasma metabolites associated with healthy Nordic dietary indices and risk of type 2 diabetes– A nested case control study in a Swedish population (*Submitted*).

Papers I, III are reproduced with the permission of the publishers.

*Corresponding author.

The author contributed to the following publications during her PhD studies which were not included in the thesis:

- Brunius C.*#, **Shi L.*#,** Landberg R. (2015) Metabolomics for improved understanding and prediction of cardiometabolic diseases—recent findings from human studies. *Current Nutrition Rep*orts 4, 348-364 (#equal contributions).

- Lee I.*, **Shi L.,** Webb D.L., Hellström P.M., Risérus U., Landberg R. (2016) Effects of whole-grain rye porridge with added inulin and wheat gluten on appetite, gut fermentation and postprandial glucose metabolism: a randomised, cross-over, breakfast study. *British Journal of Nutrition* 116(12), 2139-2149.

- **Shi L.*,** Brunius C., Lindelöf M., Shameh S.A., Wu H.X., Lee I., Landberg R., Moazzami A.A. (2017) Targeted metabolomics reveals differences in the extended postprandial plasma metabolome of healthy subjects after intake of whole-grain rye porridges versus refined wheat bread. *Molecular Nutrition & Food Research* 61(7), 1-12.

- de Mello V.D.*, Paananen J., Lindström J., Lankinen M.A., **Shi L.,** Kuusisto J., Pihlajamäki J., Auriola S., Lehtonen M., Rolandsson O., Bergdahl I.A., Nordin E., Ilanne-Parikka P., Keinänen-Kiukaanniemi S., Landberg R., Eriksson J.G., Tuomilehto J., Hanhineva K.*, Uusitupa M. (2017) Indolepropionic acid and novel lipid metabolites are associated with a lower risk of type 2 diabetes in the Finnish Diabetes Prevention Study. *Scientific Reports* 7, 46337.

- Landberg R., Wierzbicka R., **Shi L.,** Nybacka S., Kamal-Eldin A., Hedblad B., Lindroos A.K., Winkvist A., Forslund H.B. (2017) New alkylresorcinol metabolites in spot urine as biomarkers of whole grain wheat and rye intake in a Swedish middle-aged population, *European Journal of Clinical Nutrition*. (*Accepted*)

- Brunius C., Madawala S.R.P., Dutta P.C., Adoleya A., Tripathi S.B., Hanhineva K., Hajazimi E., **Shi L.,** Dimberg L., Landberg R. (2017) Impact of location on composition of selected phytochemicals in wild sea buckthorn (*Hippophae rhamnoides*). (*Submitted*)

The contribution of Lin Shi to the papers included in this thesis was as follows:

I Participated in planning of the work. Had main responsibility for data acquisition, pre-processing, and evaluation of results.

II Participated in planning the work. Performed data processing and statistics. Contributed to evaluation of results and manuscript writing.

III Participated in data acquisition. Had main responsibility for data processing, statistical analyses, results interpretation, and manuscript writing.

IV Participated in data acquisition. Had main responsibility for data processing, statistical analyses, results interpretation, and manuscript writing.

List of tables

List of figures

Abbreviations

2h-PG	2-hour plasma glucose
AAA	Aromatic amino acids
ADA	American Diabetes Association
ANOVA	Analysis of variance
AROC	Area under the receiver operation characteristics curve
BCAA	Branched chain amino acid
BMI	Body mass index
BSDS	Baltic Sea Diet Score
CMPF	3-carboxy-4-methyl-5-propyl-2-furanpropionic acid
CS	Combined score
DAG	Diglycerides
DHA	Docosahexaenoic acid
EPA	Eicosapentaenoic acid
ESI	Electrospray ionization
FDR	False discovery rate
FFQ	Food frequency questionnaire
FPG	Fasting plasma glucose
FR	Fat ratio

GC	Gas chromatography
HDL	High-density lipoprotein
HILIC	Hydrophilic interaction chromatography
HMDB	Human Metabolome Database
HNFI	Healthy Nordic Food Index
HOMA-B%	Homeostatic model assessment for β-cell function
HOMA-IR	Homeostatic model assessment for insulin resistance
ICC	Intra-class correlation
IDF	International Diabetes Federation
IDI	Integrated discrimination improvement
LC	Liquid chromatography
LDL	Low-density lipoprotein
LPC	Lysophosphatidylcholine
LPE	Lysophosphatidylethanolamine
m/z	Mass to charge ratio
MG	Monoacylglycerols
ML-PLS	Multilevel-Partial Least Squares
MP	Metabolite patterns
MS	Metabolite score or mass spectrometry depending on context
MSI	Metabolomics Standard Initiative
NMR	Nuclear magnetic resonance
NRI	Net reclassification improvement
OC-FA	Odd chain-fatty acid
OGTT	Oral glucose tolerance test
OTU	Operational taxonomic units
PC	Phosphatidylcholine
PCA	Principal component analysis

PE	Phosphatidylethanolamine
PLS	Partial least squares regression
QC	Quality control
QTOF	Quadrupole time-of-flight
rdCV	Repeated double cross validation
RF	Random forest
RMSEP	Root mean square error of prediction
ROC	Receiver operating characteristic
RP	Reverse phase
RT	Retention time
sPLS	Sparse partial least squares regression
T2D	Type 2 diabetes
TCA	Tricarboxylic-acid-cycle
TMAO	Trimethylamine N-oxide
TS	Traditional risk score
UV	Univariate analysis
VIP	Västerbotten Intervention Programme cohort
WHO	World Health Organization

1 Background

Diabetes is a chronic, progressive non-communicable disease and is an important cause of morbidity, mortality, and high healthcare costs worldwide (World Health Organization, 2016; Zhou *et al.*, 2016). According to the latest report by the International Diabetes Federation (IDF, 2015, https://www.idf.org/), approximately 415 million adults have diabetes and by 2040 this number is estimated to rise to 642 million. In an effort to address this growing health challenge, one of the major goals of the 2030 Agenda for Sustainable Development is to reduce premature mortality from non-communicable diseases, including diabetes, by one-third through prevention and treatment by 2030.

Type 2 diabetes (T2D) constitutes about 90-95% of diabetes cases and is a metabolic disorder characterized by insulin resistance in target tissues and deficiency of insulin secretion in the pancreas (Hameed *et al.*, 2015; World Health Organization, 2016). The etiology of T2D involves interactions between genetic predisposition and environmental factors, including diet, which has been suggested as an important factor that has a strong impact on the risk of developing T2D (AlEssa *et al.*, 2017; Schwingshackl *et al.*, 2017). The incidence of T2D could be effectively reduced if at-risk individuals who might benefit from effective and preventive therapies/interventions could be detected early (Zhang *et al.*, 2013; World Health Organization, 2016). Therefore, understanding the biological mechanisms in T2D pathophysiology and identifying populations at risk are critically important.

State-of-the-art analytical technologies and bioinformatics have boosted the development of ‘omics’ technologies, *i.e.*, genomics, transcriptomics, proteomics, and metabolomics (Haring and Wallaschofski, 2012; Fondi and Liò, 2015). Among these techniques, metabolomics is a useful tool for analyzing physiological and disease-induced biological states at the molecular level, taking into account both properties of the organism, such as genetic factors, health status, and other phenotypic traits, and the effects of environmental factors, *e.g.*, diet, physical activity, and medication (Trifonova *et al.*, 2013; Menni *et al.*, 2017).

Applications of metabolomics in T2D research have shown great potential to identify novel factors, *i.e.*, metabolite biomarkers that relate to T2D development, thereby providing complementary information to established risk factors and facilitating personalized medical/nutrition interventions. Moreover, metabolomics is emerging as a key technology for identification of biomarkers that reflect the exposome, including dietary biomarkers, which provide a complement to error-prone self-reported dietary assessment (Scalbert *et al.*, 2014; Cheung *et al.*, 2017; Marushka *et al.*, 2017). The use of such biomarkers may also aid in investigation of the diet-disease relationship (Biskup *et al.*, 2016; Savolainen *et al.*, 2017).

1.1 The human metabolome and metabolomics

The human metabolome can be defined as the complete set of low molecular weight compounds (<2000 Da), known as metabolites, that are present in cells, organs, tissues, or bio-fluids (Wishart *et al.*, 2007). The metabolome represents the level downstream of genetic variation, transcriptional changes, and post-translational modifications of proteins, and is also influenced by environmental and lifestyle factors, *e.g.*, diet, physical activity, microorganisms, and medications (Wishart *et al.*, 2007; Wishart, 2016) (Figure 1). It includes a diverse group of intermediates and products of metabolism, such as lipids, amino acids, peptides, nucleic acids, organic acids, vitamins, and thiols. According to the latest report in the Human Metabolome Database (HMDB), more than 18,000 metabolites have been detected and quantified in human samples and the vast majority of these have been found in the blood (http://www.hmdb.ca/statistics).

Metabolomics is defined as comprehensive analysis of metabolites in biological samples (Dunn *et al.*, 2011; Menni *et al.*, 2017). Different biological samples can be measured in a metabolomics study, with the research question driving the optimal selection of sample types. The commonly used samples in human studies are urine, plasma, serum, saliva, and biopsy samples.

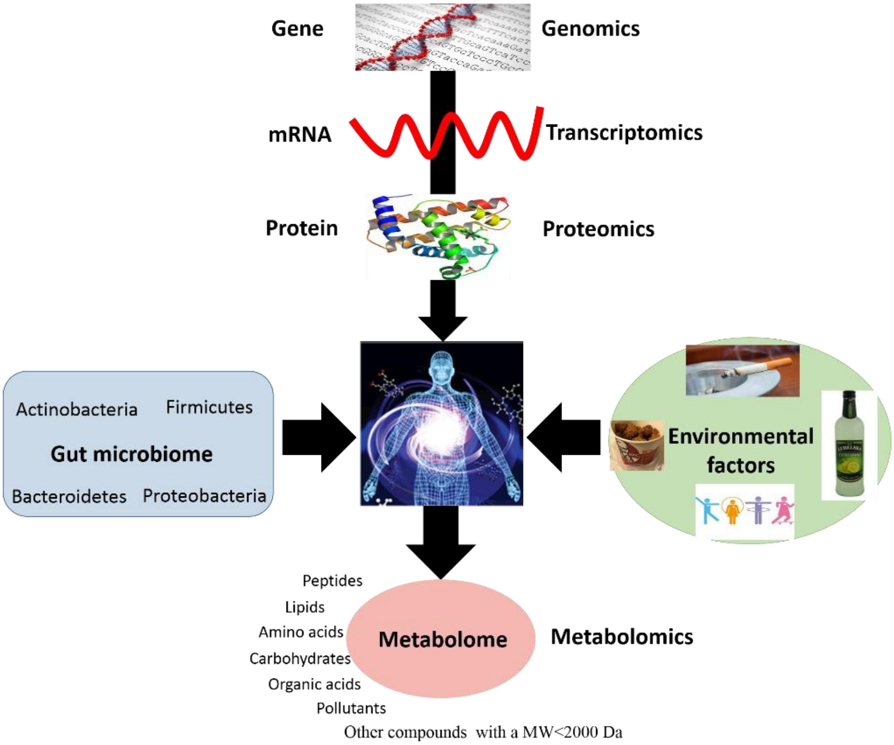


Figure 1. Different 'Omics'. Factors influencing the human metabolome that could be measured by metabolomics. MW: molecular weight.

1.1.1 Targeted and untargeted metabolomics

Metabolomics studies often apply one of two fundamental approaches: targeted or untargeted analysis (Lazar *et al.*, 2015; Gorrochategui *et al.*, 2016). These approaches are complementary. Targeted metabolomics aims to quantify a subset of pre-defined groups of chemically characterized and biochemically annotated metabolites (typically <200) based on *a priori* hypotheses. This approach leads to higher precision and the possibility to use absolute quantification of metabolites in samples. However, targeted metabolomics covers a limited number of metabolites, which may limit the scope to generate novel hypotheses.

Untargeted metabolomics is an intended comprehensive analysis of ‘all’ measurable metabolites with the chosen technology in a biological sample, including unknown compounds (Gorrochategui *et al.*, 2016). It provides a holistic view of the small molecules in the biological sample and has great potential for hypothesis-generating discovery studies. However, untargeted metabolomics often results in a massively large amount of information-rich and complex data, which leads to challenges regarding unbiased and optimal strategies for data acquisition, data processing, and statistical analyses. It therefore requires sophisticated bioinformatics tools for interpretation of the results (Gorrochategui *et al.*, 2016; Schrimpe-Rutledge *et al.*, 2016). Moreover, identification of metabolites, in particular unknowns, is an essential, yet problematic and challenging, aspect of untargeted metabolomics studies (Dunn *et al.*, 2013; Alonso *et al.*, 2015).

1.1.2 Analytical platforms for metabolomics

Regardless of the approach used, the analytical techniques/platforms that have often been applied in metabolomics studies include nuclear magnetic resonance (NMR), and mass spectrometry (MS). Mass spectrometry is normally coupled to different separation techniques, in particular gas chromatography (GC) or liquid chromatography (LC), whereas direct analysis is most often performed with NMR(Dunn *et al.*, 2011; Wishart, 2016). Mass spectrometry and NMR are complementary and each have their advantages and disadvantages. Many reviews have comprehensively discussed how each of these techniques works and how each can be used in metabolomics(Alonso *et al.*, 2015; Wishart, 2016). In brief, NMR has the advantages of simple sample preparation, non-destructive sample analysis, simultaneously detectable molecular species and excellent reproducibility, but has relatively low sensitivity, detecting only the most abundant compounds (typically >1 µmol/L) in a sample, whereas MS-based techniques are able to detect compounds in much lower concentrations (Wishart, 2016).

Liquid chromatography-mass spectrometry (LC-MS) is gaining in popularity in the metabolomics field. High-performance liquid chromatography (HPLC) and ultra-HPLC (UHPLC) are considered more comprehensive than GC-MS, since they permit analysis of a broader range of metabolites, *i.e.*, from peptides to hydrophilic organic acids, even to hydrophobic lipids, without sample derivatization (Wishart, 2016). Coupling LC to high-resolution mass spectrometry, such as time-of flight (TOF), quadrupole time-of-flight (QTOF), Fourier transform ion cyclotron resonance, and orbital ion traps, gives great specificity and sensitivity, thereby improving the quality of metabolome data. These advances allow LC-MS to be used for routine separation and measurement of thousands of discrete chemical features from large-scale sample sets in the untargeted fashion (Zhu *et al.*, 2013; Gorrochategui *et al.*, 2016; Rochat, 2016).

This thesis focuses on untargeted LC-MS metabolomics, since this was the technique of choice for all studies performed (**Papers I-IV**).

1.2 Untargeted LC-MS metabolomics

A typical untargeted metabolomics workflow constitutes several fundamental steps that need careful consideration, *i.e.*, study design, sample preparation, data acquisition and storage, data processing and analysis, metabolite identification, and biological interpretation (Theodoridis *et al.*, 2008; Gorrochategui *et al.*, 2016) (Figure 2). Challenges and detailed analytical aspects have been comprehensively reviewed in the literature, providing informative and valuable guidance for researchers (Dunn *et al.*, 2011; Yin and Xu, 2014; Yi *et al.*, 2016; Dudzik *et al.*, 2017).

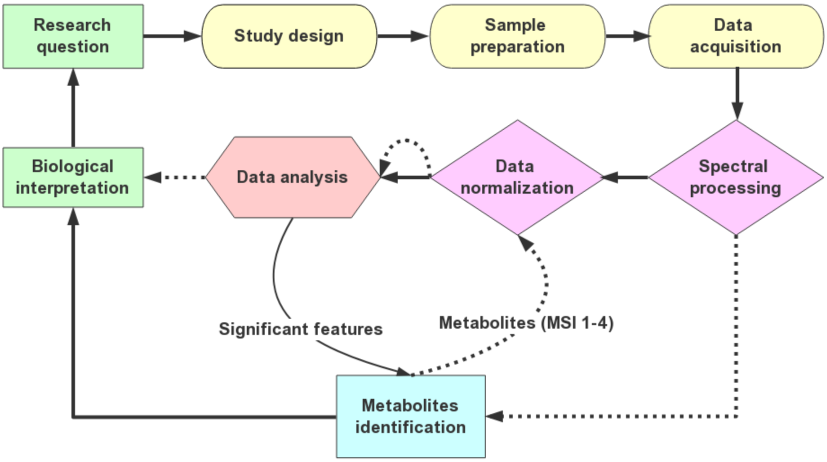


Figure 2. General workflow of untargeted metabolomics. MSI: Metabolomics Standards Initiative. The workflows indicated by dotted lines indicate alternate workflows where metabolites are identified and quantified prior to statistics.

Applying LC-MS in large-scale studies, such as in epidemiological investigations with thousands of samples, generates complex data with an abundance of MS spectra for each observation and at all retention times (RT), which makes data processing and analysis a challenging task. Inappropriate data processing and analysis increases the risk of chance findings and misinterpretation. In addition, metabolite identification is indispensable in conferring biological meaning to the observed features in a MS-based metabolomics study (Dunn *et al.*, 2013; Alonso *et al.*, 2015; Gorrochategui *et al.*, 2016).

Overall, translating metabolomics data into biologically meaningful information involves several steps: spectral processing, data normalization, downstream statistical analysis of processed datasets, and identification of the metabolites underlying the observed metabolic features.

1.2.1 Spectral processing of LC-MS data

Raw LC-MS data contain three dimensions: m/z (mass/charge ratio), RT, and signal intensity. Pre-processing is conducted to identify and quantify features and transform three-dimensional (3D) data into a 2D numerical dataset suitable for downstream analysis (Alonso *et al.*, 2015). A feature refers to a molecular entity with a unique mass to charge ratio (m/z) and RT, with an intensity representing feature height or area. It is noteworthy that each feature can be a molecular ion, adduct, fragment, or isotope of a metabolite, and that one metabolite may be represented by several features.

Raw data acquired by instruments can be processed using vendor software packages, such as Mass Profiler Professional (Agilent Technologies), DataAnalysis (Bruker Daltonics), and MarkerLynx™ (Waters Corporation). A number of open-source software tools have been developed, *e.g.*, XCMS (Smith *et al.*, 2006) and the MZmine packages (Katajamaa *et al.*, 2006; Pluskal *et al.*, 2010), due to their wider and more flexible practicality and feasibility (Myers *et al.*, 2017; Spicer *et al.*, 2017). Regardless of the software used, spectral processing typically includes algorithms for the two key steps, feature detection and alignment (Lazar *et al.*, 2015; Gorrochategui *et al.*, 2016).

Feature detection

Many algorithms have been proposed for feature detection (Tautenhahn *et al.*, 2008; Gorrochategui *et al.*, 2016). Among these, the continuous wavelet transform-based *centWave* algorithm implemented in the modular framework of widely used open source software packages, *i.e.*, XCMS and Mzmine, is often used in the metabolomics community (Myers *et al.*, 2017; Spicer *et al.*, 2017). This algorithm processes the sample spectra individually and the features are identified using multiple detection thresholds applied to different parameters such as the signal-to-noise ratio, ppm according to mass accuracy, and peak width according to the chromatographic feature width range.

These parameters are suggested to be instrument-dependent and study-specific (Eliasson *et al.*, 2012; Libiseller *et al.*, 2015; Ganna *et al.*, 2016). Parameter setting has great impact on the number and quality of features identified (Ganna *et al.*, 2016). Some strategies have been proposed to optimize the relevant parameters, *e.g.*, isotopologue parameter optimization (IPO), which is dedicated to parameters specific to XCMS (Libiseller *et al.*, 2015). However, optimization of parameters is rarely reported in LC-MS based metabolomics studies, especially when the feature detection is performed by vendor-specific software.

Alignment

When analyzing multiple samples, undesired shifts in RT and m/z are often seen due to many factors such as environmental temperature, changes in stationary or mobile phase of the chromatographic system, complex sample matrix, and ion suppression (Smith and Ventura, 2013). Shifts in m/z and RT of features between analytical runs result in different extracted spectrum patterns for metabolites across samples and severely affect subsequent statistical analysis and further metabolite identification. Owing to the continuous mass axis calibration that is commonly applied during data acquisition, shifts in m/z are in general small and the magnitude may be within an acceptable range. In contrast, the differences in RT can be quite large and non-linear (Podwojski *et al.*, 2009; Smith and Ventura, 2013) (Figure 3).

Most importantly, the shifts in RT increase with increasing period of time between experiments and become much more severe between multiple analytical batches than within each batch (Podwojski *et al.*, 2009) (Figure 3). This may increase the risk of misalignment and give rise to critical challenges regarding information loss and data processing, especially in large studies where data are often collected over long periods and analyzed in multiple batches.

Although many alignment methods for LC-MS metabolomics data have been developed, and are summarized in the literature (see *e.g.* Smith and Ventura, 2013; Spicer *et al.*, 2017), to the best of our knowledge there were no methods available to specifically address systematic misalignment across multiple batches before this thesis work. Therefore, we saw that there was an urgent need for improved alignment algorithms.

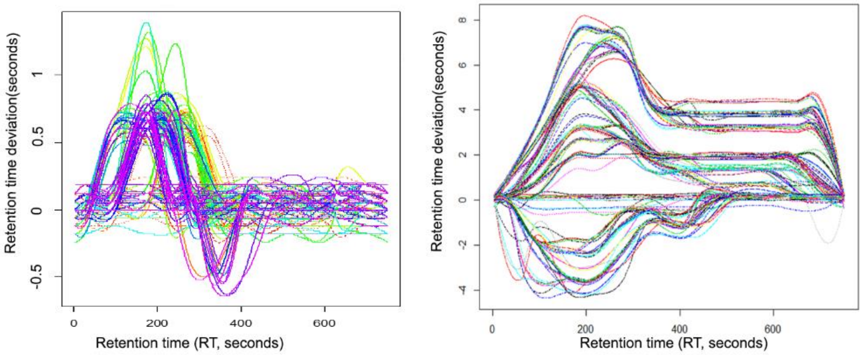


Figure 3. Retention time (RT) deviation for samples within an analytical batch (left) and between multiple analytical batches. The subsets of samples were randomly selected from samples analysed as part of the studies presented in **Papers I-IV**. Each colored line represents a different sample processed. Note that the RT deviation is different for each sample and that it is not linear.

1.2.2 Data normalization

Even when analytical procedures and spectral processing have been appropriately conducted, data will inevitably suffer from systematic and random variability in signal intensity, possibly due to matrix effects and/or variations in MS instrument sensitivity (Warrack *et al.*, 2009; Brown *et al.*, 2011; Gorrochategui *et al.*, 2016). Such variability adversely affects data quality and may hamper investigation of inherent biological variability of interest in the dataset (Mizuno *et al.*, 2017). Therefore, normalization of LC-MS data without artificially altering the biological differences between samples prior to statistics is of great importance (Sysi-Aho *et al.*, 2007; Dunn *et al.*, 2011; Mizuno *et al.*, 2017).

Several normalization approaches have been developed aiming to minimize signal drifts, including methods based on internal standards, total intensity, or intensity of the most stable features, signal intensity distributions (*e.g.*, unit norm, median, and quantile), and quality control (QC) samples. Among these, QC sample strategies are most commonly applied (Dunn *et al.*, 2012; Kirwan *et al.*, 2013; Fernández-Albert *et al.*, 2014; Petersen and Julia, 2015). The importance of QC samples and how they could be effectively applied for normalization of data have been well summarized (Dunn *et al.*, 2011, 2012). In brief, QC samples are in general pooled samples under study, thereby having a similar matrix composition to the samples to be investigated. A sufficient number of QC samples should be processed together with study samples from preparation to data acquisition and data processing. Very often, QCs are injected randomly or regularly within each analytical batch. In this way, QC samples provide a robust measure of the repeatability for each feature detected and allow effective correction for drifts within and between analytical batches (*e.g.*, quality control-based robust LOESS signal correction) (Dunn *et al.*, 2011, 2012) (Figure 4). Following normalization, coefficient of variation (CV) for each single feature (CV <20%) and its frequency of presence in QC samples, *i.e.*, the '80% rule' (Smilde *et al.*, 2005), are often used as criteria for determination of qualified features that are eligible for subsequent statistical analysis.

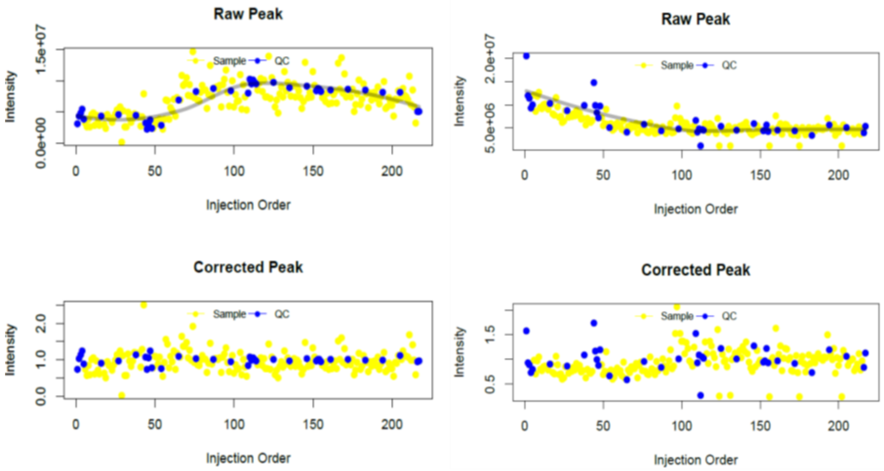


Figure 4. Quality control-based robust LOESS signal correction. A correction procedure was applied for each feature. The shift correction algorithm implemented in R package 'statTarget' was performed on the example data with default setting (Luan H, 2016).

1.2.3 Statistical methods applied in metabolomics

Once a feature matrix that consists of qualified features has been produced, followed by feature-wise normalization (centering, scaling, and/or transformation) (van den Berg *et al.*, 2006), one or multiple statistical analyses are subsequently performed to identify relevant and informative variables with biologically meaningful information (Broadhurst and Kell, 2006; Xi *et al.*, 2014).

Metabolomics studies are often conducted to identify biomarkers of disease or biological processes (Brennan *et al.*, 2015; Klein *et al.*, 2016), to characterize complicated biochemical systems (Schmid and Blank, 2010; Fondi and Liò, 2015), and to reveal insights into the mechanism of pathophysiological processes (Tanaka and Ogishima, 2011; Barallobre-Barreiro *et al.*, 2013). In general, data derived from metabolomics experiments are often subjected to one of three major statistical methods, *i.e.*, regression (where dependent response variable(s) Y is/are continuous), classification, (where dependent response variable(s) Y, is/are categorical), and sometimes the analytical strategy needs to cope with dependent data structures, derived *e.g.*, from repeated measures and/or cross-over experimental designs. Different situations require different univariate and multivariate statistical methods to analyze the data.

Univariate analysis

Univariate methods focus on independent changes in individual features without considering the presence of inter-correlations between features. Each feature is tested by a parametric test, *e.g.*, Student's t test and Analysis of variance (ANOVA), or corresponding non-parametric tests, *e.g.*, Wilcoxon-rank test and Kruskal-Wallis one-way ANOVA, which are preferable when assumptions for parametric methods are not met. Correction for multiple testing is necessary to protect against the probability of false positives, *i.e.*, finding a statistically significant result by chance, due to multiple comparisons. This is a crucial step that should always be considered when dealing with metabolomics data (Vinaixa *et al.*, 2012; Gorrochategui *et al.*, 2016). However, in metabolomics data, many features represent the same metabolite and several features can be artifacts caused by chemical and/or bioinformatics noise. Therefore, traditional multiple test correction methods, *i.e.* Bonferroni correction and false-discovery rate (FDR) (Alonso *et al.*, 2015; Chong *et al.*, 2015) may be overly conservative. This may result in substantial loss of statistical power to detect truly differential features.

Multivariate analysis

Unlike univariate methods, multivariate analysis methods take the relations between variables into consideration, and can be classified into unsupervised and supervised methods. Detailed descriptions and discussions on the working principles of commonly used techniques are available in excellent reviews (Broadhurst and Kell, 2006; Zhang *et al.*, 2014; Gromski *et al.*, 2015). With unsupervised methods, *e.g.*, principal component analysis (PCA) and hierarchical clustering analysis, data patterns can be detected without considering their orchestrated or complementary behavior in relation to biological responses/processes.

Supervised methods have been extensively applied in metabolomics studies to extract relevant information from datasets that involve large amounts of qualified features relevant to specific research questions (Gromski *et al.*, 2015; Ren *et al.*, 2015). The main supervised techniques applied in metabolomics studies are partial least squares regression (PLS) or PLS discrimination analysis (PLS-DA) and random forest (RF) (Gromski *et al.*, 2015; Afanador *et al.*, 2016; Yi *et al.*, 2016). Model over-fitting has emerged as one of the major problems related to application of supervised modeling. An over-fitted model describes random error or noise in addition to the underlying relationship under investigation, thereby substantially increasing the risk of false positive discoveries and consequently leading to erroneous conclusions and hypotheses (Broadhurst and Kell, 2006; Saccenti *et al.*, 2014). Therefore, great efforts and particular attention are required to avoid model over-fitting.

Cross-validation is one of the most popular approaches to optimize models (*e.g.*, the number of components in PLS), to evaluate model performance, and to reduce model over-fitting (Bro *et al.*, 2008; Filzmoser *et al.*, 2009; Arlot and Celisse, 2010; Baumann and Baumann, 2014). The repeated double cross-validation (rdCV) algorithm has been shown to achieve better performance in model optimization and in reducing the risk of over-fitting (Filzmoser *et al.*, 2009) than leave-one-out and K-fold cross-validation.

Variable selection aims to generate a compact data structure with less noisy and redundant predictors, which facilitates production of parsimonious models that exhibit robust explanatory predictive power and minimized over-fitting (Saeys *et al.*, 2007; Barbu *et al.*, 2013; Rudnicki *et al.*, 2015). Most existing variable selection techniques are designed to identify the '*minimal-optimal*' variables set, *i.e*., maximum information density per variable, while optimizing prediction performance (Rudnicki *et al.*, 2015). This strategy is particularly useful for *e.g.*, designing diagnostic, predictive, or prognostic biomarkers (Nilsson *et al.*, 2007; Saeys *et al.*, 2014). However, only a limited number of algorithms have been tailored for identifying the '*all-relevant*' variable set (Kursa and Rudnicki, 2011). This strategy facilitates understanding of complicated biochemical systems or mechanisms of *e.g.*, pathophysiological or metabolic processes, where all changes in the metabolites in relation to research questions are of interest.

Several studies have applied variable selection to construct robust models (Gromski *et al.*, 2014; Bujak *et al.*, 2016; Grissa *et al.*, 2016). However, many of these have suffered from selection bias, which in turn induces over-optimistic estimations and an increasing risk of false positive discoveries due to over-fitting (Ambroise and McLachlan, 2002; Krawczuk and Lukaszuk, 2016). Selection bias occurs when a given dataset is used for both variable selection and assessment of model performance. To address this issue, a few variable selection-within-validation schemes have been introduced. For example, Boulesteix *et al.* (2007) propose implementation of variable selection based on the Wilcoxon test within a Monte Carlo CV scheme to avoid biased estimation, but this algorithm can only be used for classification analysis. Goodacre *et al.* (2011) have developed a generic algorithm based on a Bayesian network approach, which assesses the predictive accuracy on a test set which was held out during model training and variable selection. The variable selection-within-validation schemes are highly recommended and outperform other bias-prone strategies. However, the scarcity of freely available, easy-to-use algorithms is a limiting factor. Therefore, easy-to-use software tools that support a wide range of comprehensive supervised modeling operations with proper validation are necessary for widespread implementation of these tools, which will ultimately have a major impact on the quality of metabolomics data.

1.2.4 Metabolite identification

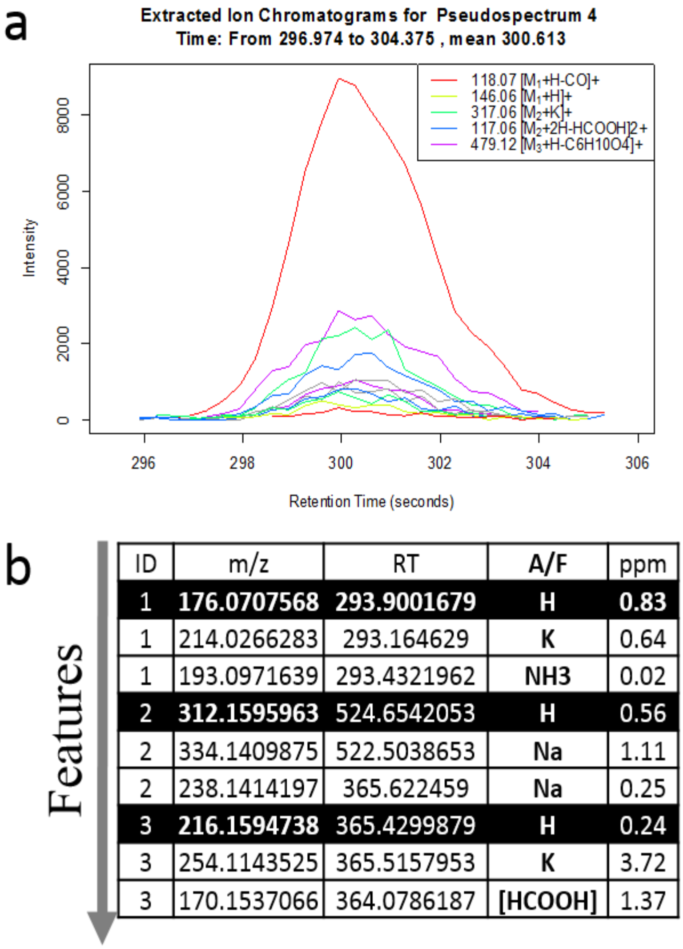
A great potential in untargeted metabolomics is the discovery of novel molecular species, possibly unknown compounds, that are statistically associated with specific biological outcomes. This requires identification of metabolite(s) of interest. The commonly applied strategy for metabolite identification is to match the neutral molecular mass or m/z of a potential metabolite to its fragmentation pattern obtained by a tandem MS technique and compare the outcome against public databases, authentic standards, and/or published literature using a tolerance window for any deviation in mass accuracy (*e.g.* ppm threshold ranges from 5-30). The widely used open databases and their main characteristics are summarized in the literature (*e.g.*, Lazar *et al.*, 2015). Examples of compound-centric databases are *e.g.*, HMDB, lipidMaps, Metlin.

The Metabolomics Standard Initiative (MSI) reporting standards have been used to standardize the confidence level of metabolite identification over a decade (Spicer *et al.*, 2017). The MSI defines four different levels of metabolite identification: Level 1 (highest level of identification) is identified metabolites that must be compared to an authentic chemical standard analyzed in the same laboratory, using the same analytical techniques as the experimental data. Level 2 and 3 are putatively annotated compounds and compound classes, respectively, that require matching to databases rather than authentic chemical standards, while Level 4 is unknown compounds. A more stringent reporting criterion proposed recently that splits MSI reporting standards into 5 levels, though this is not commonly complied with (Schymanski *et al.*, 2014). Great efforts are still required to reassess and update the reporting standards to fit the current community needs and benefit all investigators(Spicer *et al.*, 2017).

Chemical identification of unknown compounds represents a major challenge in untargeted metabolomics (Dunn *et al.*, 2013; Schrimpe-Rutledge *et al.*, 2016). The neutral molecular mass is inferred from the molecular ion m/z value, and depends on *e.g.*, ionization mode (H+ or H-), ionization adducts (*e.g.*, Na^+, K^+, $NH4^+$), and fragments that are common neutral losses, such as H_2O, CO_2, and HCOOH. Querying databases for a single peak m/z value may lead to multiple plausible neutral molecular masses, which increases false positive annotations and the workload in querying precursor ions in databases (Dunn *et al.*, 2013; Alonso *et al.*, 2015).

To address this issue, several tools have been developed, *e.g.*, PUTMEDID-LCMS (Brown *et al.*, 2011) and CAMERA (Kuhl *et al.*, 2014), to provide putative annotations for a panel of features deriving from the same compound, *e.g.*, molecular ions, isotopes, adducts, and in-source fragmentations (Figure 5).

Putative annotation normally relies on accurate m/z, m/z differences, RT similarity, pairwise correlation between measured responses, known adduct lists, and similarity of chromatographic peak shape. Automated annotation reduces the number of features to be considered for identification, increases querying efficiency, and provides an unequivocal molecular formula and/or putative identification. It is noteworthy that such annotation depends on the strength of correlations between measured features within a small RT window. However, the presence of misalignment due to the RT shifting and variability in signal intensity correlation across analytical runs may influence the correlations between features, thereby leading to false positive annotations and/or different annotations.



ID	m/z	RT	A/F	ppm
1	**176.0707568**	**293.9001679**	**H**	**0.83**
1	214.0266283	293.164629	K	0.64
1	193.0971639	293.4321962	NH3	0.02
2	**312.1595963**	**524.6542053**	**H**	**0.56**
2	334.1409875	522.5038653	Na	1.11
2	238.1414197	365.622459	Na	0.25
3	**216.1594738**	**365.4299879**	**H**	**0.24**
3	254.1143525	365.5157953	K	3.72
3	170.1537066	364.0786187	[HCOOH]	1.37

Figure 5. Example of putative annotation of some randomly chosen features using R package ‘CAMERA’ (a) and PUTMEDID-LCMS (b). a) Extracted ion chromatograms (EIC) of labelled, putatively annotated peaks from one pseudospectrum, potentially representing a metabolite, at retention time 300 seconds. b) Key information on annotation using PUTMEDID-LCMS. Peaks with the same ID may originally come from one metabolite. A/F denotes possible annotated adducts or fragments.

Recently, new strategies/pipelines in cooperation with advanced bioinformatics tools for metabolite identification tasks have emerged (Benton *et al.*, 2015; Dührkop *et al.*, 2015; Bingol *et al.*, 2016; van der Hooft *et al.*, 2016; Kouassi Nzoughet *et al.*, 2017). These approaches may enable accurate, reliable, and rapid identification of metabolites in complex mixtures, but need to be validated in further studies.

1.3 Metabolomics for prediction of T2D

Diagnosis of T2D is generally made on the basis of glucose and/or hemoglobin A1c concentrations in the blood. According to the American Diabetes Association (ADA) criteria, the diagnosis for diabetes is based on elevated fasting plasma glucose (FPG) (≥7.0 mmol/L) or elevated 2-hour plasma glucose (2h-PG) level (≥11.1 mmol/L) in an oral glucose tolerance test (OGTT) or elevated HbA1c level (≥6.5%) (American diabetes association, 2010). The current World Health Organization (WHO) diagnostic criterion for diabetes is FPG ≥ 7.0 mmol/L (126 mg/dL) or 2h-PG ≥11.1 mmol/L (200 mg/dL). Diabetes symptoms (*e.g.*, polyuria and polydipsia) also need to be considered. For people who do not show symptoms, diagnosis requires confirmatory plasma venous determination. At least one additional glucose test result on another day with a value in the diabetic range is needed for diagnosis, either fasting, from a random blood sample, or from the 2h-PG (WHO, 2006).

A major shortcoming of the glucose-based diagnosis strategy is the poor sensitivity, resulting in late disease discovery when complications have already arisen, which have detrimental effects on human quality of life and longevity (Gjesing and Pedersen, 2012; Olokoba *et al.*, 2012; World Health Organization, 2016). Given the increasing burden of the condition worldwide, earlier identification of at-risk individuals is of particular importance.

1.3.1 Established risk factors for T2D

T2D is multifactorial disease caused by a complex interplay between genetic and environmental factors (Temelkova-Kurktschiev and Stefanov, 2011; Abdullah *et al.*, 2017; Läll *et al.*, 2017). Several genome-wide association studies have identified hundreds of common variants associated with T2D (Temelkova-Kurktschiev and Stefanov, 2011; Läll *et al.*, 2017). Despite high heritability (26-69%), the effects of genetic determinants may interact with other risk factors, *e.g.*, lifestyle factors (physical inactivity, unhealthy dietary intake, tobacco use), thus providing relatively weak independent prediction power (Franks and McCarthy, 2016).

Known risk factors have been incorporated into risk scores that are suggested for use as a practical screening tool to identify at-risk individuals for further testing among populations, such as the Diabetes Risk Score (Lindström and Tuomilehto, 2003), RuralDiab risk score (Zhou *et al.*, 2017), and Framingham Risk Score. Despite moderate to good prediction in some studies, models with established risk factors still leave room for improvement in risk prediction (Herder *et al.*, 2014). Furthermore, established risk factors provide limited evidence of early mechanisms in relation to development of T2D.

1.3.2 Metabolomics to identify predictive biomarkers

Metabolomics has been successfully applied in a number of epidemiological studies, in particular in prospective study settings, to identify early metabolic alterations associated with a risk of developing T2D (Dunn, 2012; Lu *et al.*, 2013; Zhang *et al.*, 2013; Gonzalez-Franquesa *et al.*, 2016; de Mello *et al.*, 2017).

There is a growing body of evidence, mostly from targeted metabolomics, on various species of metabolites that predict future T2D. Such metabolites include sugars, *e.g.*, glucose, fructose, and hexose; branched-chain and aromatic amino acids (BCAA and AAA); free fatty acids and ketones, *i.e.*, hydroxybutyrate, acetone, and acetoacetate originating from fatty acid oxidation; and lipids and their metabolic by-products, such as diacylglycerides, ceramides, choline-containing phospholipids (Lu, 2013; Gonzalez-Franquesa, 2016; Klein *et al.*, 2016). The causative effect and underlying mechanism have yet to be clearly elucidated, but circulating concentrations have been shown to be altered 3-15 years before clinical diagnosis of T2D (see *e.g.*, Herder *et al.*, 2014; Klein and Shearer, 2016).

Before the research presented in this thesis, few prospective studies had applied untargeted metabolomics for discovery of predictive metabolites, due to the high cost and complexity of robust large-scale metabolomics instrumental and data analyses. During the past four years, and benefiting from the rapid development of analytical techniques and bioinformatics tools, an increasing number of applications of untargeted metabolomics have uncovered novel metabolites associated with future T2D (Menni *et al.*, 2013; Drogan *et al.*, 2014; Nikiforova *et al.*, 2014; Zhao *et al.*, 2015; Fall *et al.*, 2016; Qiu *et al.*, 2016; de Mello *et al.*, 2017).

Typical problems related to studies in the field include small number of incident cases, lack of generalizability, and/or ambiguous metabolite annotation by only matching molecular mass against online databases. It is also worth mentioning that most novel findings are study-specific and thus need to be further validated through independent studies. In addition, few studies assessed long-term reproducibility of identified metabolites (their fluctuations within individual over time). In most observational studies, metabolites have been determined in a single sample at baseline, assumed to reflect the average concentration in an individual over time. However, metabolites are in fact subject to both random and systematic variations over time, particularly those metabolites that are related to disease. Random intra-individual variability of a measured biomarker over time will affect the precision in risk estimates linking biomarker to endpoint and drive associations towards the null (Mcdermott, 1997). Reproducibility is therefore an important characteristic determining the applicability of metabolites in clinical investigations that needs to be investigated. The intraclass correlation coefficients (ICC), ranging between 0 and 1, is a typically used index of reproducibility. In general, ICC value 0.4–0.75 indicates moderate to good reproducibility (Rosner, 2005; Carayol *et al.*, 2015).

1.3.3 Incremental value of predictive metabolites in T2D risk prediction

Several studies have investigated whether incorporation of metabolites together with established risk factors can improve T2D risk prediction (Herder *et al.*, 2014; Yao *et al.*, 2015; Zhao *et al.*, 2015; Lu *et al.*, 2016; Yengo *et al.*, 2016; Peddinti *et al.*, 2017).

The most common method used to assess prediction power is the receiver operating characteristic (ROC) curves approach (Figure 6). The ROC is created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) for measuring discrimination of a binary variable (case compared with control). For comparison between models, the area under the curve (AROC) is calculated. AROC varies between 0 and 1, where 1 indicates perfect discrimination and 0.5 represents no discrimination. An AROC of 0.8 corresponds to 80% likelihood that a randomly selected case (*i.e.*, a person who will develop disease) will be assigned a higher estimated diabetes risk than a randomly selected control (*i.e.*, a person who will remain disease-free).

The net reclassification improvement (NRI) is more sensitive for assessing the incremental predictive power of a new marker than AROC (Herder *et al.*, 2014), where a large NRI value indicates good predictive ability. The integrated discrimination improvement (IDI) is another useful measure, with an IDI value of 0 meaning no predictive improvement. To assess the incremental value of a new predictor marker, use of all three measures is recommended to enhance the certainty of assessments(Herder *et al.*, 2014).

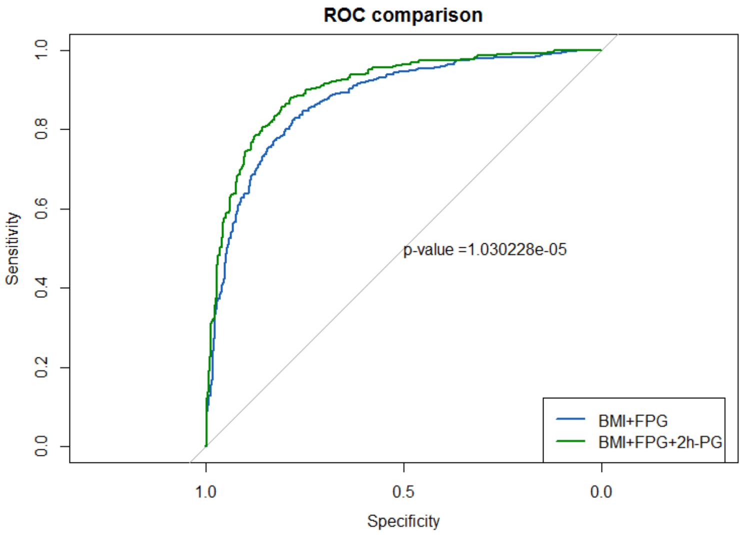


Figure 6. Receiver operating characteristic (ROC) curves and area under the curves (AROC) from conditional logistic regression of T2D status including body mass index (BMI) and fasting plasma glucose (FPG; AROC=0.87) or additionally including 2h-plasma glucose (2h-PG; AROC=0.91). Data on BMI, fasting glucose and 2h-PG from the human studies in Papers I-IV were used. Perfect prediction would be a point in the upper left corner with coordinates (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). A random classification would give a point along the diagonal line of no-discrimination. The p value relates to the difference in AROC between the two models. The addition of 2h-PG resulted in NRI (category-free) [95% CI]: 0.72 [0.61-0.84] and IDI [95% CI]: 0.1 [0.08-0.11] (p<0.05 for both measurements), indicating improvement in risk prediction. In this example, the same data were used to train and predict the model.

1.4 Metabolomics to identify dietary biomarkers linking diet to T2D

Diet has been associated with the risk of developing T2D (Alhazmi *et al.*, 2014; Schwingshackl *et al.*, 2017). To improve understanding of the diet-disease relationship, objective assessment of dietary intake of various types of foods in free-living populations is imperative, but is still a major challenge in nutritional science (Garcia-Perez *et al.*, 2017). Metabolomics offers great possibilities to objectively reflect dietary exposures, and could also link such exposures to biological effects of relevance for disease development (O'Gorman and Brennan, 2015; Playdon, Ziegler, *et al.*, 2017; Savolainen *et al.*, 2017).

1.4.1 Epidemiological evidence for the association between diet and T2D

Epidemiological studies have shown a protective role of individual foods such as whole grains, fish, vegetables, fruits, and coffee in developing T2D (Schwingshackl *et al.*, 2017). In contrast, a poor-quality diet, characterized by high consumption of *e.g.*, red/processed meat, high-sugar snacks, and soft drinks, has been associated with an increased risk (Pan *et al.*, 2011; Schwingshackl *et al.*, 2017).

Humans consume combinations of foods, *i.e.*, a diet. Analysis of individual foods however does not take the complexity of the entire diet, including interactions into consideration. Dietary pattern analysis has been suggested as an approach to provide a snapshot of the entire diet, considering extensive interaction between foods, which may be more informative in investigating diet-disease relationships than single food item analysis (Alhazmi *et al.*, 2014; Osonoi *et al.*, 2015). Two main approaches have been used to determine dietary patterns: dietary indices that assess compliance with *a priori*-defined dietary patterns, such as Mediterranean Diet Score (MDS) (Panagiotakos *et al.*, 2006), Dietary Approaches to Stop Hypertension (DASH) score (Sacks *et al.*, 1995), Healthy Eating Index (HEI)(Kennedy *et al.*, 1995); and data-driven approaches, *e.g.*, principal component analysis (PCA) and cluster analysis, which describe actual intake patterns in the population (Cespedes and Hu, 2015).

Many studies have found that higher adherence to healthy dietary patterns, *e.g.*, MDS and DASH, is associated with a reduced risk of developing T2D (De Koning *et al.*, 2011; Abiemo *et al.*, 2012; Jacobs *et al.*, 2017). In recent randomized controlled trials, adherence to a healthy Nordic dietary pattern was shown to have beneficial effects on cardio-metabolic risk factors (Adamsson *et al.*, 2011, 2014; Damsgaard *et al.*, 2012; Uusitupa *et al.*, 2013; Poulsen *et al.*, 2014). A healthy Nordic dietary pattern is characterized by traditional Nordic food items with potential health effects, such as vegetables, fish, fruits, berries, whole-grain rye and oat products, and various seafood products. However, epidemiological evidence regarding the role of a healthy Nordic dietary pattern for prevention of T2D is limited (Kanerva *et al.*, 2014; Lacoppidan *et al.*, 2015).

1.4.2 Metabolomics to identify dietary biomarkers

Most evidence on the role of diet in relation to T2D comes from epidemiological studies where food intake is typically based on selfreporting measurements by food frequency questionnaires (FFQ). However, such studies are subject to systematic and random measurement errors, *e.g.*, due to misreporting and poor information from databases (Paeratakul *et al.*, 1998; Streppel *et al.*, 2013). These errors contribute to data inaccuracy and may thus hamper the discovery of existing diet-disease relationships or lead to false positive associations (Paeratakul *et al.*, 1998).

Measuring dietary biomarkers in biological samples has been suggested as a complementary approach to self-reported methods to improve nutritional epidemiological studies (Scalbert *et al.*, 2014; Zheng *et al.*, 2014; Brennan, 2015; O'Gorman, 2015; Gibbons *et al.*, 2017). Dietary biomarkers should objectively reflect the intake of a specific dietary exposure and can be used to validate self-reported measurements and/or be linked to disease endpoints (Hedrick *et al.*, 2012; Scalbert *et al.*, 2014; Bordoni and Capozzi, 2015). For validation and calibration of self-reported methods, recovery biomarkers are most useful. Such biomarkers reflect the intake, quantitatively, on the same scale and are not affected by non-dietary determinants to a large extent. Unfortunately, only a few recovery biomarkers exist (Freedman *et al.*, 2010; Van Dam and Hunter, 2012). Most dietary biomarkers fall into the category of concentration biomarkers (Freedman *et al.*, 2010). Such biomarkers represent a concentration in a biological tissue or fluid that is correlated with intake, but also affected by other determinants (Van Dam and Hunter, 2012).

Metabolomics has emerged as a key technology for discovery and validation of dietary biomarkers that could objectively reflect food consumption (Gibbons and Brennan, 2017). Such biomarkers belong to the concentration biomarkers. Biomarker discovery studies can be made in controlled intervention settings or in cross-sectional observational studies (Scalbert *et al.*, 2014). Of note, dietary biomarkers that were discovered in intervention studies may not be specific for the food under study at population level, because habitual diets may include other foods containing precursors of the same biomarkers. Thus, identifying diet-metabolite associations in observational studies is a complementary strategy to overcome this problem (Scalbert, 2014; Garcia-Perez, 2017; Gibbons and Brennan, 2017; Playdon *et al.*, 2017).

Metabolomics may also be used to provide objective assessments of overall dietary patterns in free-living populations (Garcia-Perez *et al.*, 2017; Playdon *et al.*, 2017). However, few biomarkers reflecting adherence to dietary patterns derived either by data-driven methods or *a priori* defined pattern indices exist (Floegel *et al.*, 2013; Esko *et al.*, 2017; Garcia-Perez *et al.*, 2017; Playdon *et al.*, 2017).

1.4.3 Dietary biomarkers linking diet with risk of T2D

Dietary biomarkers could be used for investigation of associations between diet and T2D in large population-based cohort studies. Plasma alkylresorcinols represent a good example of dietary biomarkers that specifically reflect whole-grain wheat and rye intake (Landberg *et al.*, 2008, 2014; Kyrø *et al.*, 2014; Biskup *et al.*, 2016). The alkylresorcinol C17:0/ C21:0 ratio could reflect relative whole-grain rye to whole-grain wheat intake, and has been associated with improved insulin sensitivity and reduced risk of developing T2D (Biskup *et al.*, 2016).

Very recently, Savolainen *et al*. (2017) analyzed and linked previously known dietary biomarkers with future T2D in a nested case-control study where self-reported dietary intake data were lacking. The multiple biomarkers captured different dietary components and suggested a lowered risk of developing T2D related to high concentrations of biomarkers previously shown to reflect fish, whole grains, vegetable oils, and good vitamin E levels, while biomarkers that reflected red meat and saturated fatty acids were associated with increased risk (Savolainen *et al.*, 2017).

However, no previous study has examined whether metabolites related to a specific dietary pattern are associated with incident T2D.

2 Aims of the thesis

The overall aims of this thesis were to develop bioinformatics tools for data processing and analysis of large-scale untargeted LC-MS metabolomics and to apply such tools to identify predictive metabolites of T2D, to identify metabolites associated with *a priori* defined healthy Nordic dietary indices, and to investigate whether such metabolites are associated with risk of developing T2D in a Swedish nested case-control study.

The specific objectives addressed through the thesis:

(1) To develop a novel data-processing strategy to correct for within- and between-batch variability in processing multiple-batch untargeted LC-MS metabolomics data and to provide unbiased measures of improved data quality (**Paper I**).

(2) To develop a novel statistical framework for multivariate modeling with unbiased variable selection aiming to identify informative variables from large-scale and high-dimensional datasets without compromising predictive accuracy and/or increasing the risk of over-fitting (**Paper II**).

(3) To identify plasma metabolites that predict future T2D, and assess the extent to which inclusion of predictive metabolites beyond traditional risk factors could improve risk prediction. Further, to investigate the changes over time in identified metabolites among healthy individuals (*i.e.*, reproducibility) and individuals who later developed T2D (*i.e.*, systematic change reflecting disease progression) (**Paper III**).

(4) To identify plasma metabolites associated with adherence to a healthy Nordic dietary pattern by two *a priori*-defined indices, the Baltic Sea Diet (BSDS) and the Healthy Nordic Food Index (HNFI), and to investigate associations of identified metabolites and the risk of T2D (**Paper IV**).

3 Materials and methods

3.1 Study design and samples

3.1.1 Participants and samples from the Västerbotten Intervention Programme cohort (Papers I, III, IV)

Data and fasting plasma samples were obtained from the Västerbotten Intervention Programme (VIP) cohort, which is one of the sub-cohorts of the Northern Sweden Health and Disease Study cohort (Norberg *et al.*, 2010). The cohort included 141,000 sampling occasions involving 98,300 individuals by March 2015, of which 36,100 individuals have provided repeated samples. Recruitment for the VIP cohort started in 1985. Individuals at age 30, 40, 50, and 60 years were invited to participate in a systematic risk factor screening program, including individual counseling about lifestyle habits, at their local healthcare center. Blood samples were drawn and stored at the Northern Sweden Medical Biobank. In all assessments, all participants were asked to complete a questionnaire that captured information on socio-economic conditions, self-rated health, personal health history, and family history of diabetes, tobacco use, and physical activity.

Incident T2D cases were identified according to the diabetes registry DiabNorth (Rolandsson *et al.*, 2012). Among cases, participants that had an unthawed fasting plasma sample in the biobank and/or had follow-up samples available were considered eligible for this thesis work. Among eligible cases, 503 participants who developed T2D after a median time of seven years after baseline were selected (Figure 7a). Each case was individually matched to one non-diabetic individual according to age (+-2 years), gender, ethnic group, and season of blood draw. Among the 503 pairs of selected participants, 187 case-control pairs had a second follow-up sample drawn and data collected 10 years after baseline (Figure 7b). The corresponding characteristics of the subgroup of 187 case-control pairs with follow-up data were similar to those of the 503 case-control pairs and those without available repeated samples. The study protocol was approved by the Regional Ethics Committee in Uppsala, Sweden (registration number 2014/011).

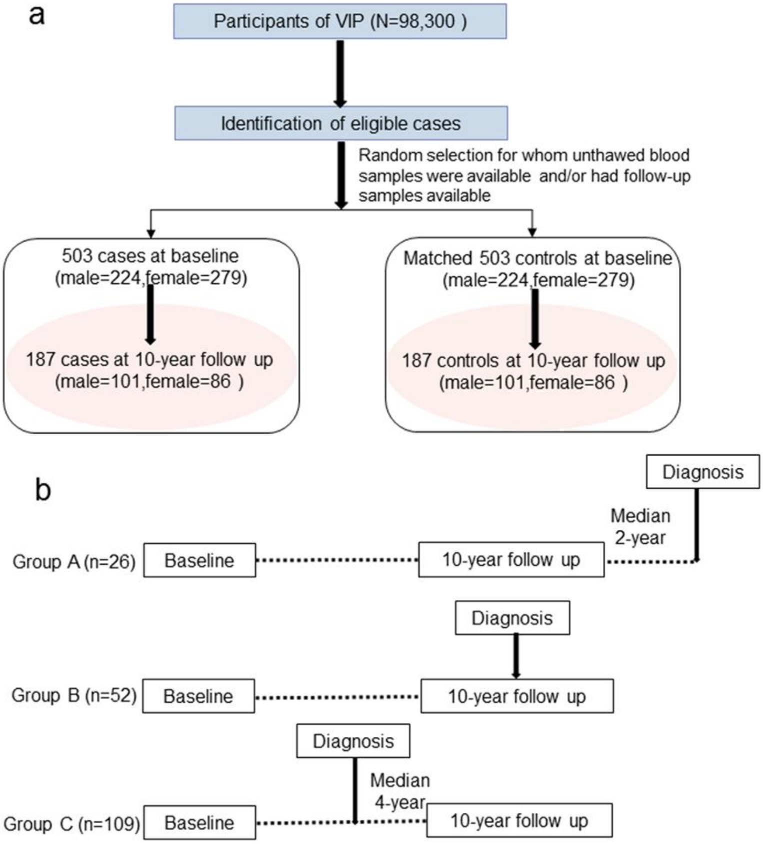

a
Participants of VIP (N=98,300)
Identification of eligible cases
Random selection for whom unthawed blood samples were available and/or had follow-up samples available
503 cases at baseline (male=224,female=279)
187 cases at 10-year follow up (male=101,female=86)
Matched 503 controls at baseline (male=224,female=279)
187 controls at 10-year follow up (male=101,female=86)
b
Diagnosis
Median 2-year
Group A (n=26)
Baseline
10-year follow up
Diagnosis
Group B (n=52)
Baseline
10-year follow up
Diagnosis
Median 4-year
Group C (n=109)
Baseline
10-year follow up

Figure 7. Study design and population selection for the human studies reported in **Papers III-IV**. a) Flowchart of participant selection from the VIP cohort. b) Information on baseline and 10-year follow-up sampling among 187 type 2 diabetes cases.

3.1.2 Diet data and healthy Nordic dietary pattern indices

Two modified and validated versions of the Northern Sweden Food Frequency Questionnaire (FFQ) were used in the VIP cohort: one with 84 food items and another with 64 items (Johansson *et al.*, 2002, 2010). Food items were energy-adjusted by the density method for dietary pattern indices (Playdon, Moore, *et al.*, 2017). Two participants with an implausible caloric intake (<600 or >5000 kcal/day) and/or incomplete food frequency questionnaire (≥10% missing values) were excluded.

Baltic Sea Diet Score (BSDS) was derived based on the approach described by Kanerva *et al.* (2014). The BSDS included nine food component categories (Table 1). All components, except alcohol, were scored according to sex-specific population consumption quartiles (Q1-Q4). For fruits, vegetables, whole-grain cereals, low-fat milk products, fish and FR, a score of 0-3 was given for Q1-Q4, while for meat products and total fat, the score was given in the reversed order. Alcohol scored 1 if intake was less than 20 g/day for men or 10 g/day for women, or otherwise a score of 0 was given. The resulting BSDS ranged from 0 to 25 points, with higher scores representing stronger adherence to a healthy diet.

The Healthy Nordic Food Index (HNFI) was calculated according to Olsen *et al*. (2011), and was initially based on six food items: fish, cabbages, rye bread, oatmeal, apples/pears, and root vegetables. However, since consumption of ‘rye bread’, ‘apples/pears’, and ‘oatmeal’ was not assessed specifically by these FFQs, ‘whole-grain bread’, ‘fruits (apples, pears, oranges, mandarins, grapefruits)’, and ‘oat/rye/barley porridge’ were used instead as respective proxy measures (Table 1). Adherence to the index was calculated by summarizing indices, where each food item was scored 0 or 1 when intake was less than or greater than the sex-specific median intake of participants. The HNFI ranged from 0-6 and a higher score was indicative of stronger adherence to a healthy diet.

Table 1. *Food components involved in* a priori-*defined healthy Nordic dietary pattern indices*

Food components of dietary index	Food items used to define the food component
Baltic Sea Diet Score(BSDS)	
Nordic fruits	Apples, pears, berries, oranges, mandarins, grapefruits
Nordic vegetables	Tomatoes, cucumbers, brassicas, legumes, carrots
Nordic whole-grain cereals	Whole-grain rye/wheat bread, rye, oat, barley porridge
Fish	Perch, cod, herring, salmon, white fish, shellfish
Red and processed meat	Beef, pork, processed meat and sausages
Low-fat milk products	Milk with 0.5% fat, yogurt with ≤3% fat
Fat ratio (FR)	Ratio of polyunsaturated fatty acids to sum of saturated fatty acid and trans-fatty acids
Alcohol	Alcohol

Food components of dietary index	Food items used to define the food component
Total fat	Total fat
Healthy Nordic Food Index(HNFI)	
Fruits	Apples, pears, oranges
Whole-grain bread	Whole-grain rye/wheat bread
Oat/rye/ barley porridge	Whole-grain rye, oat, barley bread, porridge
Fish	Perch, cod, Baltic herring, salmon, white fish, shell fish
Cabbages	Cabbages
Root vegetables	Carrots

3.1.3 Datasets used for developing a statistical framework for multivariate modeling

‘Freelive’*:* Detailed information on study design and metabolomics data acquisition has been described elsewhere (Hanhineva, Brunius, *et al.*, 2015). In brief, free-living participants with no diagnosed or perceived gastrointestinal diseases or symptoms were invited to participate and instructed to adhere to their habitual diet. Untargeted LC-qTOF-MS metabolomics was performed on urine samples from 58 participants who completed three-day weighed food records, morning spot urine tests, and 24 h urine collection on two separate occasions (periods A and B) approximately 2-3 months apart. The dataset consisted of reported whole-grain rye consumption (continuous Y variable) from 112 observations (58 unique participants; 2 individuals had sample from one occasion available), codes for individual (numeric ID variable) and 16,392 features as X matrix (a molecular entity with a unique m/z and retention time as measured by an LC-MS instrument). This study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures involving human subjects were approved by the Regional Ethics Review Board in Uppsala (log no. 2008:040). Written informed consent was obtained from all subjects.

‘Mosquito’*:* This dataset has been described in detail by Buck *et al*. (2016). *Anopheles gambiae* mosquitoes were collected from three villages in western Burkina Faso and whole-body bacterial flora was analyzed by 16S amplicon sequencing. In total, 29 observations were available for village of capture (categorical Y variable; three levels) and 1678 16S operational taxonomic units (OTU, X matrix). However, owing to the non-continuous nature of 16S data, leading to a high degree of data scarcity, 940 OTUs showed near-zero variance. PLS was therefore performed on a subset with 738 OTUs only.

‘Crisp’: The study design is described elsewhere (Zamaratskaia *et al*., 2017, accepted). In brief, rye and wheat crisp breads were consumed as part of isocaloric breakfast interventions in a cross-over design. Untargeted UHPLC-qTOF-MS metabolomics was performed on plasma samples from 20 randomly selected individuals and six time points. Feature signals were numerically integrated using the trapezoidal rule to obtain area-under-the-curve values (AUCs) for all features. This dataset contained 20 subjects (Y, ID) and AUCs of 1587 features as X matrix.

3.2 Untargeted LC-MS metabolomics

3.2.1 Analytical platform and data acquisition

Untargeted LC-MS metabolomics was performed on the plasma samples of participants selected from the VIP cohort, as described in section 3.1.1.

Sample preparation: Fasting heparin-plasma samples (90 µL) were mixed with 360 µL acetonitrile, incubated in a 96 deep well polypropylene plate in an ice bath for 15 min and then centrifuged at 1200 g for 5 min through a 0.2 µm polytetrafluoroethylene filter to collect the filtrate. The filtrate was kept refrigerated at 4 ℃ until further analysis.

Quality controls: Two types of independent biological samples (QC-A: batch-specific quality control sample; QC-B: long-term reference sample) were used to monitor the stability and functionality of the system throughout the instrumental analyses and together constituted approximately 16% of analytical samples.

Data acquisition: For hydrophilic interaction (HILIC) chromatography, 3 µL plasma were injected for analyses. An Acquity UPLC BEH Amide column (2.1 x 100 mm, 1.7 mm; Waters Corporation) was used and maintained at 45 ℃ for separation. The mobile phase was delivered at 600 µL/min and consisted of 50% acetonitrile (vol:vol; eluent A) and 90% acetonitrile (vol:vol; eluent B), both containing 20 mmol/L ammonium formate (pH 3), delivered in a gradient profile: 0-2.5 min: 100% B, 2.5-10 min: 100 to 0% B, 10-10.01 min: 0 to 100% B, 10.01-12.5 min: 100% B. For reversed phase (RP) chromatographic analyses, 4 µL plasma were injected. Separation was performed using a Zorbax Eclipse XDB-C18 column (2.1 x 100 mm, 1.8 mm; Agilent Technologies) at 50 ℃. The mobile phase was delivered at 400 µL/min and consisted of eluent A (water, Milli-Q purified) and eluent B (methanol), both containing 0.1% (vol:vol) of formic acid, delivered in a gradient profile: 0-10 min 2 to 100% B, 10-14.5 min: 100% B, 14.5- 14.51 min: 100 to 2% B, 14.51-16.5: 2% B. The electrospray ionization (ESI) source was operated for both positive mode (+) and negative (-) ionization. The approach used in MS/MS analyses has been described elsewhere (Hanhineva *et al.*, 2015).

In total, samples including QCs were analyzed in eight batches, with randomization being constrained to having sample pairs and repeated samples within the same batch, and otherwise full randomization within batch (Figure 8). Instrumental analyses were performed with approximately 250 injections per batch in eight analytical batches over six months.

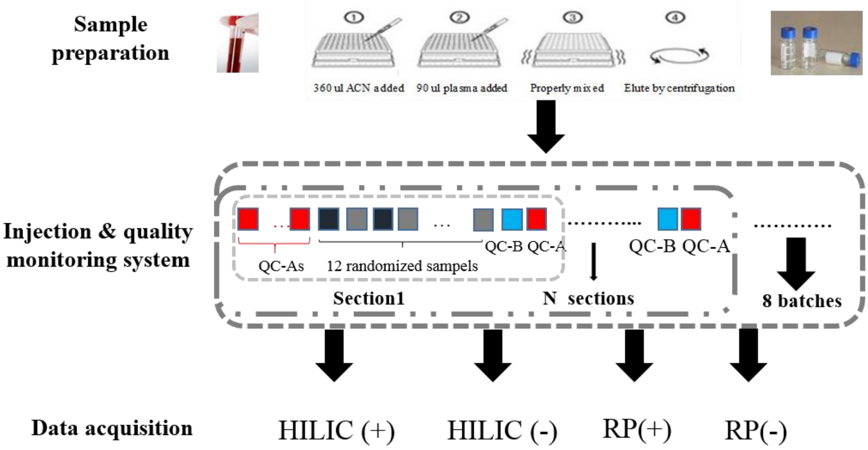


Figure 8. Analytical protocol for untargeted LC-MS metabolomics.

In total, for each sample, four spectra were obtained from reversed phase (RP, ESI+), RP (ESI-), hydrophilic interaction chromatography (HILIC ESI+), HILIC (ESI-) and were subjected to data processing.

3.2.2 Data processing

Data processing comprised three steps: spectral processing, data normalization, and putative feature annotation and aggregation (Figure 9).

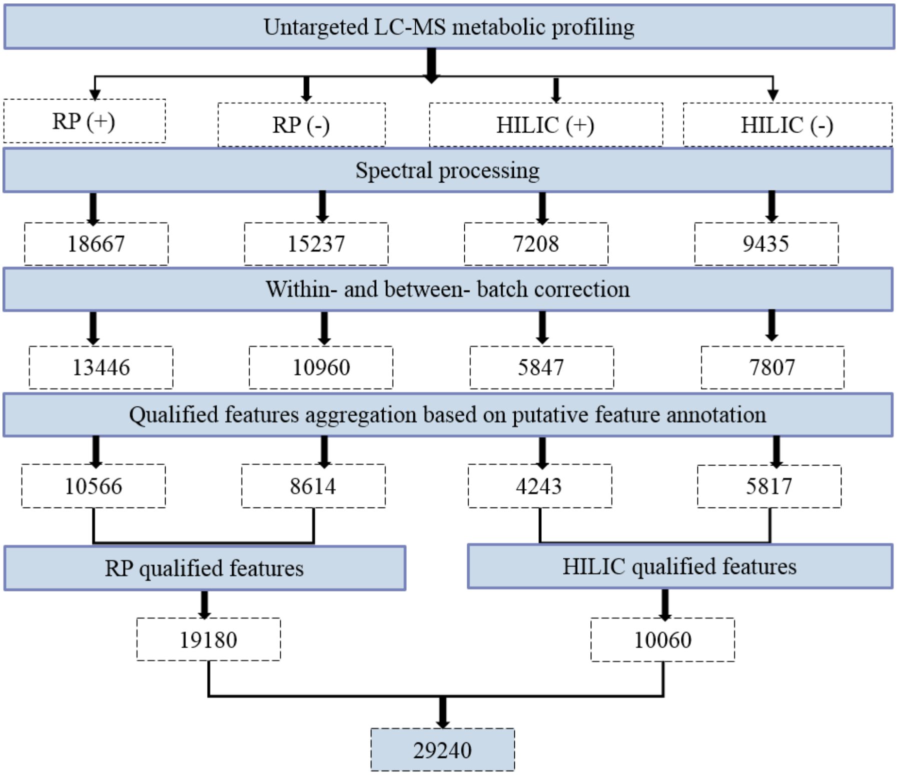


Figure 9. Overall workflow for untargeted LC-MS metabolomics data processing used in this thesis. The number of features retained after each step of processing is presented. HILIC: hydrophilic interaction (HILIC) chromatographic analysis; RP: reversed phase chromatographic analysis.

Specifically, spectral processing was conducted using feature detection and alignment algorithms with optimized parameters implemented in XCMS (Scripps Center for Metabolomics, La Jolla, CA). Raw data files were converted to mzXML format using MassHunter Qualitative Analysis B.06.00 (Agilent Technologies) and were subjected to XCMS following standardized procedures (Smith *et al.*, 2006). The *centWave* algorithm was used for peak detection. Parameters were optimized by iterative testing of different settings of parameter configurations and IPO packages (Libiseller *et al.*, 2015). The quality of algorithm performance was evaluated by the number of detected features and visualizing plot of detected peaks. Optimization of main parameters was conducted for each mode separately (Table 2).

Table 2. *Optimized parameters applied for untargeted LC-MS spectral processing*

Parameter	Chromatography and ionization mode [1]			
	HILIC (+)	HILIC (-)	RP(+)	RP (-)
XCMS [2]				
snthresh	6	6	6	6
ppm	15	15	15	15
prefilter	(3,1000)	(3,1000)	(3,1000)	(3,1000)
Peakwidth	(5,20)	(5,20)	(5,76)	(5,84)
mzdiff	0.0056	0.005	0.0045	0.0045
gapInit	0.928	0.296		0.58
gapExtend	2.688	2.4		2.19
bw	1.5	1.5	30, 10, 5 *	2
batchCorr [3]				
mzdiff	0.005	0.005	0.005	0.005
rtdiff	10	15	10	15
Fold change limit	5	5	7	5
PUTMEDID-LCMS [4]				
ppm mass error	10	10	10	10
Retention time shift	2.5	2.5	2.5	2.5
Retention time range	30 - 720 seconds		70-980 seconds	

[1]RP: Reverse phase chromatography; HILIC: hydrophilic interaction chromatography. '+' or ' –' denotes the electrospray ionization, positive or negative, respectively. [2]Freely available software implemented in R was applied for data deconvolution. Retention time correction was achieved using the *Obiwarp* function for HILIC (ESI+), HILIC (ESI-), and RP (ESI-) and the *LOESS* fitting method was used for RP (ESI+). *snthresh*: signal to noise ratio; ppm: measurement error; *prefilter*: prefiltering step for the first analysis step (ROI detection), mass traces are only retained if they contain at least 3 peaks with intensity >= 1000; *peakwidth*: the expected approximate peak width in chromatographic space, given as a range (min, max) in seconds; *mzdiff*: the minimum difference in m/z dimension for peaks with overlapping retention times; *gapInit* and *gapExtend* : penalty for gap opening and gap enlargement, respectively; bw: the bandwidth (standard deviation of the smoothing kernel) for alignment. [3]The developed 'batchCorr' (**Paper I**) was used for within- and between-batch correction. *rtdiff* is equal to retention time deviation of samples in 8 batches. [4]A freely available workflow for putative annotation of metabolites or metabolite groups (http://omictools.com/putmedid-lcms-tool). *Peak group retention time correction and alignment were performed with different bandwidth setting iteratively.

The data processing procedures developed in **Paper I** were applied for batch misalignment correction, within-batch cluster-based drift correction, and between-batch signal intensity normalization. The algorithms of the procedures are available as an R package 'batchCorr' (see chapter 4.1 of this thesis for detailed description of algorithms). The parameters used are given in Table 2.

PUTMEDID-LCMS Workflow was applied to group features potentially resulting from one metabolite (Table 2). This procedure was performed batch-wise for each analytical mode. For each feature, qualified putative annotation was determined only if the same annotation was given for at least in five batches. For features that were determined originally from one metabolite, intensities were then aggregated.

In total, 29,240 features were considered qualified (Figure 9) and thus were subjected to further statistical analyses in two endpoint studies (**Papers III and IV**).

3.3 Statistical analysis

All analyses were carried out using R v.3.4.0, except for intra-class correlation (ICC), which was calculated using a SAS macro '%ICC9' freely available at https://www.hsph.harvard.edu/donna-spiegelman/software/icc9/ (SAS Institute, USA).

3.3.1 Paper I

Clustering of variables in the observation space was analyzed by the 'mclust' algorithm (Fraley *et al.*, 2012). Principal component analysis (PCA) was performed on the data to assess the within- and between-batch drift correction.

3.3.2 Paper II

The R packages used included 'mixOmics' for PLS core modeling (Rohart *et al.*, 2017), 'randomForest' for RF core modeling (Liaw and Wiener, 2002), 'Boruta' (Kursa, 2010) and 'VSURF' (Genuer *et al.*, 2015). For fitness estimation, Q^2 was used for regression analysis, while number of misclassifications was used for classification and multilevel analysis. Permutation-based p-values were used for assessing modeling performance (Lindgren *et al.*, 1996) by calculating the cumulative (1-tailed) probability of actual model prediction performance versus distribution of random permutations using standard Student's t distribution or Student's t test on rank-transformed data for nonparametric tests.

3.3.3 Paper III

The overall workflow for discovery of predictive metabolites of developing type 2 diabetes is shown in Figure 10. To select informative features with minimal risk of model over-fitting, the novel statistical framework for multivariate modeling developed in **Paper II** was applied on the plasma metabolome. These algorithms are available as an R package ('MUVR'). Development of these algorithms is described in detail in Chapter 4.2 of this thesis.

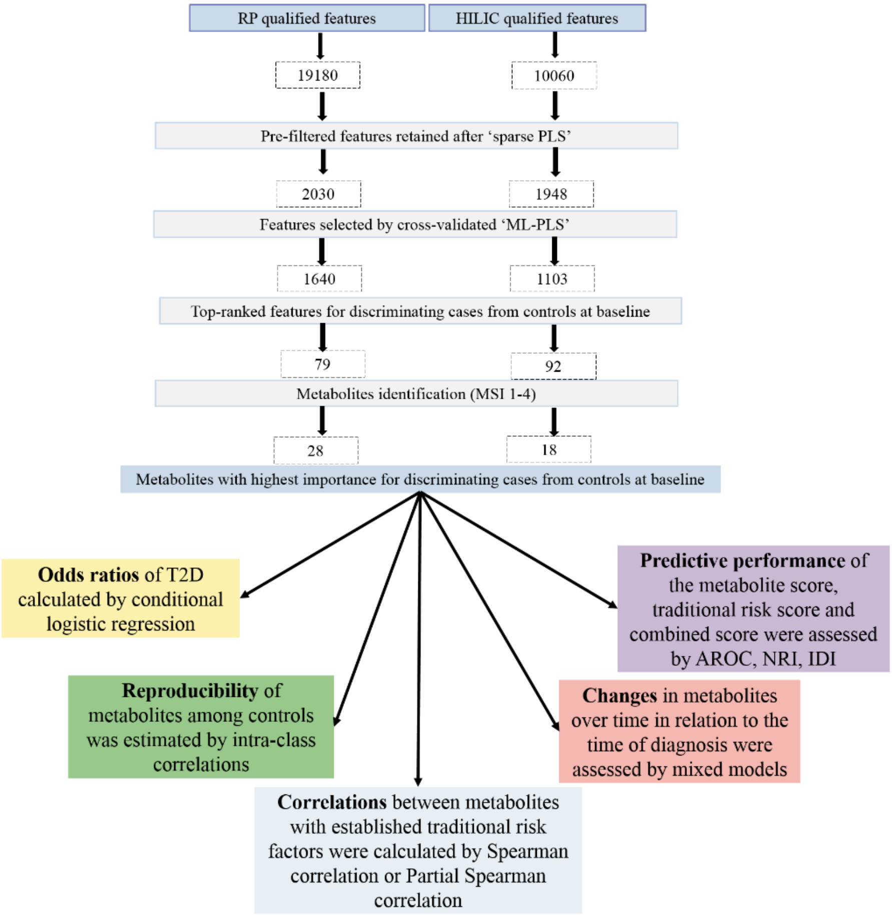

RP qualified features
HILIC qualified features
19180
10060
Pre-filtered features retained after 'sparse PLS'
2030
1948
Features selected by cross-validated 'ML-PLS'
1640
1103
Top-ranked features for discriminating cases from controls at baseline
79
92
Metabolites identification (MSI 1-4)
28
18
Metabolites with highest importance for discriminating cases from controls at baseline
Odds ratios of T2D calculated by conditional logistic regression
Predictive performance of the metabolite score, traditional risk score and combined score were assessed by AROC, NRI, IDI
Reproducibility of metabolites among controls was estimated by intra-class correlations
Changes in metabolites over time in relation to the time of diagnosis were assessed by mixed models
Correlations between metabolites with established traditional risk factors were calculated by Spearman correlation or Partial Spearman correlation

Figure 10. Overall workflow for statistics used in **Paper III**. Number of features retained at each step is presented.

Specifically, sparse PLS was performed as a pre-filter on the entire data to remove the majority of redundant features and/or features that were unlikely to contribute to discrimination between cases and controls. The algorithm Multilevel PLS (ML-PLS) implemented in R package ‘MUVR’ (developed in **Paper II**) was performed on pre-filtered data to determine a parsimonious set of discriminative features ranked according to their importance. Top-ranked features, *i.e.* where the variable importance ranking score <100 (lower score indicates higher importance of a given feature to discriminate between cases and controls) were subjective to identification (see details in section ‘Metabolite identification’) and further analysis.

Conditional logistic regression (R package ‘survival’) (Therneau and Lumley, 2017) was applied to investigate associations between metabolites at baseline and likelihood of developing T2D. Odds ratios (OR) were calculated for quartiles and per SD) increment based on a crude model and multivariate adjusted models controlled for traditional risk factors and/or lifestyle-related factors. To compensate for multiple testing, false discovery rate (FDR)-adjusted p-values were calculated and significance threshold was set at $p<0.05$.

Reproducibility of metabolites was estimated by ICC over the 10-year period between the two sampling occasions among the subset of healthy participants (n=187). Systematic changes in metabolites over the 10-year period in relation to the time of diagnosis among the subset of T2D cases (n=187) were investigated by mixed model.

Incremental predictive ability of metabolites was assessed using two approaches: i) by adding predictive metabolites to predefined traditional risk factors (covariates used in conditional logistic regression); or ii) through selection of optimal variables from traditional risk factors and/or metabolites using a validated random forest algorithm (Buck *et al.*, 2016). For models based on approach ii), the *metabolite score (MS)* was based only on selected variables from the annotated predictive metabolites (MSI 1-2 level), *the traditional risk score (TS)* was based on 14 known traditional risk factors of type 2 diabetes, and *the combined score (CS)* was based on optimal variable selection among both metabolites and traditional risk factors. The area under the receiver operation characteristics curve (AROC, R ‘pROC’) was computed to evaluate the prediction performance of different models. Incremental prediction performance of MS compared with traditional risk factors was further assessed by net reclassification improvement (NRI) and integrated discrimination improvement (IDI) tests (R package ‘PredictABEL’ (Kundu *et al.*, 2011).

Spearman correlation coefficients were calculated to explore the association of metabolites with traditional risk factors at baseline among healthy participants (n=503). Partial Spearman correlations were calculated to investigate independent associations between each of the metabolites with HOMA-IR and HOMA beta cell function (HOMA B%), adjusted for BMI, age, sex, and case/control status among 187 case/control pairs with follow-up samples.

3.3.4 Paper IV

The overall workflow for identification of metabolites related to pre-defined indices and their associations with likelihood of developing T2D are shown in Figure 11.

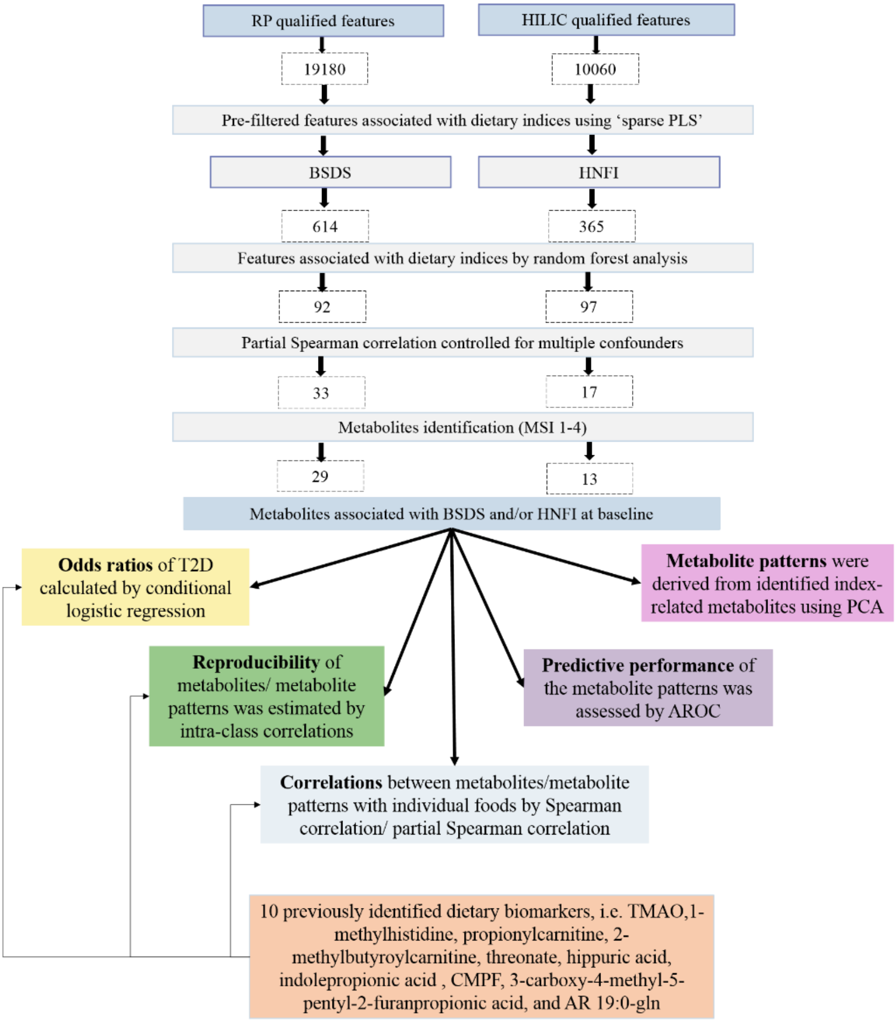

RP qualified features
HILIC qualified features
19180
10060
Pre-filtered features associated with dietary indices using 'sparse PLS'
BSDS
HNFI
614
365
Features associated with dietary indices by random forest analysis
92
97
Partial Spearman correlation controlled for multiple confounders
33
17
Metabolites identification (MSI 1-4)
29
13
Metabolites associated with BSDS and/or HNFI at baseline
Odds ratios of T2D calculated by conditional logistic regression
Metabolite patterns were derived from identified index-related metabolites using PCA
Reproducibility of metabolites/ metabolite patterns was estimated by intra-class correlations
Predictive performance of the metabolite patterns was assessed by AROC
Correlations between metabolites/metabolite patterns with individual foods by Spearman correlation/ partial Spearman correlation
10 previously identified dietary biomarkers, i.e. TMAO,1-methylhistidine, propionylcarnitine, 2-methylbutyroylcarnitine, threonate, hippuric acid, indolepropionic acid , CMPF, 3-carboxy-4-methyl-5-pentyl-2-furanpropionic acid, and AR 19:0-gln

Figure 11. Overall workflow for statistical analysis in **Paper IV**. Number of features retained at each step is presented. AR 19:0-Gln, nonadecyl-benzenediol glucuronide; CMPF, 3-carboxy-4-methyl-5-propyl-2-furanpropionic acid; TMAO, trimethylamine N-oxide.

The RF core modelling implemented in R package 'MUVR' (**Paper II**) was performed on sparse PLS pre-filtered data to discover index-related features. Direct associations between features retained from random forest and indices were assessed using partial Spearman correlation analysis controlled for case-control status, age at blood draw, gender, BMI (kg/m^2), smoking status, education, and physical activity.

Odds ratios and reproducibility of index-related metabolites were assessed as in Paper III. Metabolite patterns were derived from identified index-related metabolites using PCA (R package 'psych') (Revelle, 2017). Correlations between scores and dietary variables, *i.e.*, indices and individual foods reported in both FFQs and with <0.5% missing values, were calculated. Prediction performance of risk modeling for PCA score was assessed using AROC.

Associations with T2D risk were also assessed for 10 *a priori*-defined metabolites previously associated with either a healthy Nordic dietary pattern or with foods captured by Nordic dietary indices.

3.4 Metabolite identification

Metabolites were identified based on accurate mass and MS/MS fragmentation manually matched against an in-house library of authentic standards, online databases, and the literature. The confidence level of annotation was categorized according to the Metabolomics Standard Initiative (MSI) (Sumner *et al.*, 2007).

4 Results and discussions

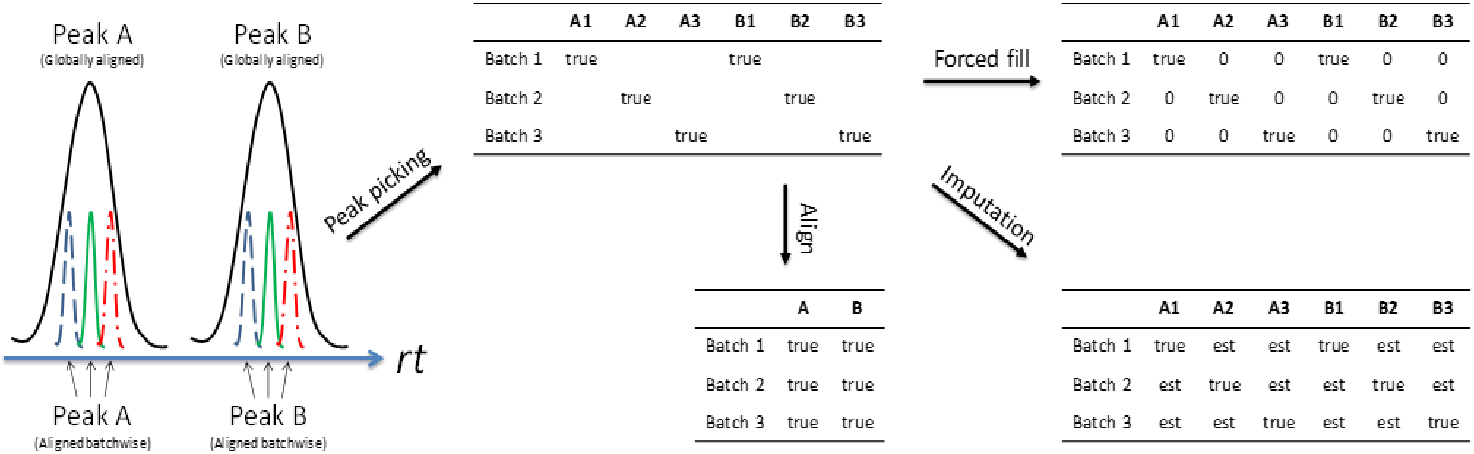
4.1 ‘batchCorr’ for data processing

The data-processing pipeline for untargeted LC-MS metabolomics developed in **Paper I** is available as an open source R package ‘batchCorr’ (https://gitlab.com/CarlBrunius/batchCorr). It includes multiple algorithms to overcome the measurement errors inherent to variations in signal intensity and shifts in m/z and RT between samples in untargeted LC-MS metabolomics data.

4.1.1 Feature alignment between multiple batches

When applying XCMS for grouping and alignment of multiple-batch data, low bandwidth may capture batch-specific features but result in missing data. As one of the standardized procedures in XCMS, forced integration data filling, *i.e.*, ‘fillPeaks’, is often applied following feature alignment. However, there is an obvious risk of splitting one true feature into two or more, thereby artificially creating batch-specific features and increasing noise in the variable set (Figure 12). On the other hand, increasing bandwidth for alignment may introduce the risk of forcing alignment of unrelated features, consequently increasing the information loss.

To overcome this problem, an algorithm for misalignment correction was developed (Figure 13). Datasets should be first examined for systematic feature misalignment between batches to avoid loss of data integrity.



	A1	A2	A3	B1	B2	B3
Batch 1	true			true		
Batch 2		true			true	
Batch 3			true			true

	A1	A2	A3	B1	B2	B3
Batch 1	true	0	0	true	0	0
Batch 2	0	true	0	0	true	0
Batch 3	0	0	true	0	0	true

	A	B
Batch 1	true	true
Batch 2	true	true
Batch 3	true	true

	A1	A2	A3	B1	B2	B3
Batch 1	true	est	est	true	est	est
Batch 2	est	true	est	est	true	est
Batch 3	est	est	true	est	est	true

Figure 12. Strategies to deal with missing data resulting from systematic misalignment across batches exemplified in the retention time (rt) domain. Increasing bandwidth setting to capture all batches (black lines) introduces the risk of forcing alignment of unrelated features. Lower bandwidth settings can accurately capture batch-specific features (dashed blue, whole green or dot-dashed red), but may result in missing data after peak picking. Forced filling will insert mostly baseline noise (approx. zero areas) into the peak table, whereas imputation will provide estimates of the true values. Batch alignment will provide the possibility to aggregate batch-specific features to their global counterparts.

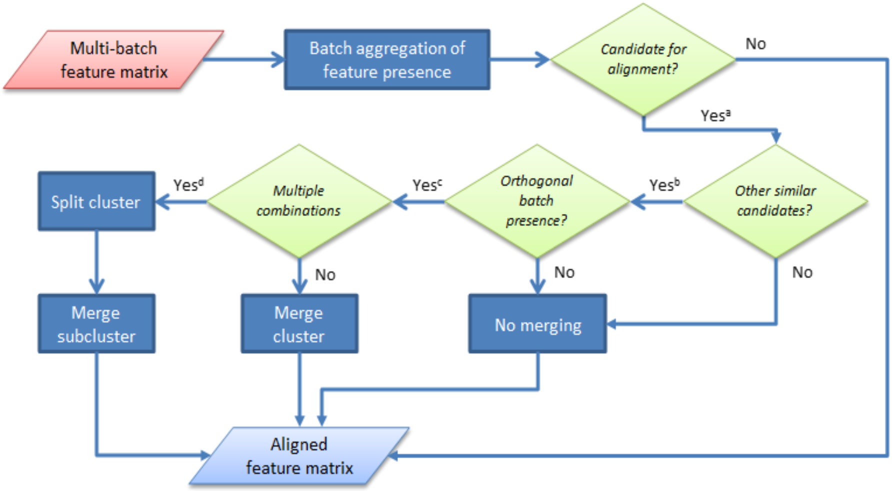


Figure 13. Algorithm developed for between-batch feature alignment involved in 'batchCorr'. Flowchart for alignment of features systematically misaligned between batches. [a]A feature is considered a potential candidate for alignment if 0<total batch presence<number of batches. [b]Potential candidates for alignment are considered similar if m/z and retention time (RT) are within user-defined tolerance. [c]Candidates are considered for alignment and subsequently clustered if not mutually present in the same batch (*i.e.*, presence vector orthogonality). [d]Clusters containing multiple possible alignments are recursively subdivided into sub-clusters.

In brief, presence/missingness of each feature is first aggregated per batch to filter out spurious random noise or erroneous misalignment of individual samples. Potential alignment candidates are features present in at least one, but not all, batches.

Second, for each potential candidate, correspondence with other candidates ('events') is determined if features are sufficiently close in the m/z and RT domains (*i.e.*, within user-defined borders) and under the constraint of batch presence being orthogonal between features, *i.e.*, ensuring that two features present in the same batch cannot be aligned. Key parameters are *mzdiff* and *rtdiff*, representing a user-defined box bounded by largest allowed absolute m/z and RT differences, where *mzdiff* is set according to instrument resolution and *rtdiff* was chosen as the maximum retention drift between batches obtained from XCMS.

Third, all distinct events thus consist of two alignment candidates. Events which share common alignment candidates are then clustered. If all cluster candidates are mutually orthogonal, correspondence is assumed and alignment candidates are then merged. When multiple alignment combinations occur, multiple alignment candidates are disentangled into their respective correspondences through a recursive sub-clustering algorithm.

4.1.2 Within- and between-batch feature intensity normalization

The algorithm for within-batch drift correction consists of four essential parts: clustering of features; drift modelling per cluster; drift correction per cluster; and removal of individual features with poor reproducibility (Figure 14a). The cluster-based strategy can increase within-batch data quality (Figure 15), while minimizing the risk of over-fitting (*e.g.*, modeling of noise in individual features, *e.g.*, QC-LOESS (Dunn *et al.*, 2011)) by adding statistical strength of multiple features to the individual cluster regressions. Note that the availability of two distinct QC samples, *i.e.* QC-A and QC-B, analyzed in the present thesis (see section 3.2.1) allowed unbiased measures to be applied for quality improvement using data correction techniques to avoid over-fitting and introducing bias. In the algorithm, drift correction per cluster was therefore performed only if providing an unbiased measure of increased quality of data measured on QC samples not used for drift modeling.

Between-batch normalization is performed through an iterative process (Figure 14b). Batch-wise qualified data after alignment and within-batch drift correction are first limited to common features, *i.e.*, those features common to all batches. Normalization is then achieved using either of two standard approaches: reference sample intensity or population-based (median) normalization. Normalization by average reference sample intensity per batch is performed only if the precision (as indicated by CV) and accuracy (deviation from general intensity ratio between batches) of the reference samples are within user-defined tolerances (*e.g.*, CV<30% and fold change <5). Regardless of chromatography type and ionization modes, the severe batch effect among multiple batches clearly observed as the main determinant of variance prior to normalization was drastically reduced after applying the between-batch normalization procedure (Figure 16).

It is noteworthy that the algorithms developed can be used either alone or in combination to suit any particular analytical situation. For example, correction of misalignment between batches is easily integrated with available sample-based alignment methods, *e.g.*, *obiwarp* implemented in XCMS. The within-batch correction without alignment or normalization can be applied if samples are analyzed within only one batch. Moreover, in the case of multiple batches, these algorithms can easily be chosen at will, combined with other drift correction and/or normalization procedures, and incorporated into a customized workflow.

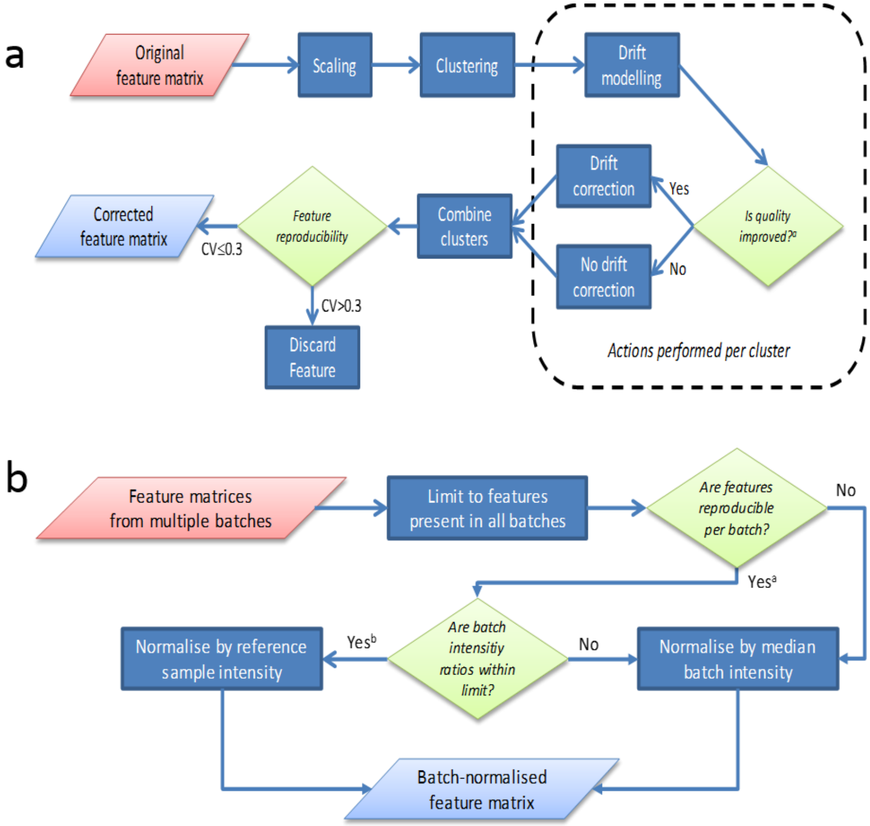

a
Original feature matrix
Scaling
Clustering
Drift modelling
Is quality improved?[a]
Yes
No
Drift correction
No drift correction
Actions performed per cluster
Combine clusters
Feature reproducibility
CV≤0.3
CV>0.3
Corrected feature matrix
Discard Feature
b
Feature matrices from multiple batches
Limit to features present in all batches
Are features reproducible per batch?
No
Yes[a]
Are batch intensitiy ratios within limit?
Yes[b]
No
Normalise by reference sample intensity
Normalise by median batch intensity
Batch-normalised feature matrix

Figure 14. Algorithms developed for within- and between-batch signal intensity drift corrections. a) Flowchart of cluster-based within-batch intensity drift correction. [a]Cluster quality is considered to be improved if $\text{rmsd(QC-B)}_{\text{with correction}} < \text{rmsd (QC-B)}_{\text{without correction}}$, where ‘rmsd’ denotes root mean squared distance from the cluster center point. In this algorithm, QC-A is used for drift modeling while QC-B is used for assessment of data quality improvement. b) Flowchart of between-batch intensity normalization algorithm. [a]Features were considered reproducible if reference sample intensity per batch CV < *e.g.*, 30%. [b]Reference sample average feature intensity ratios between batches were considered within limit if not deviating from corresponding average feature intensity ratios by more than a fold change of five. For features meeting both criteria, batches were normalized by average reference sample intensity. For other features, reference samples were not considered sufficiently representative of the sample population and features were thus normalized by median batch intensity.

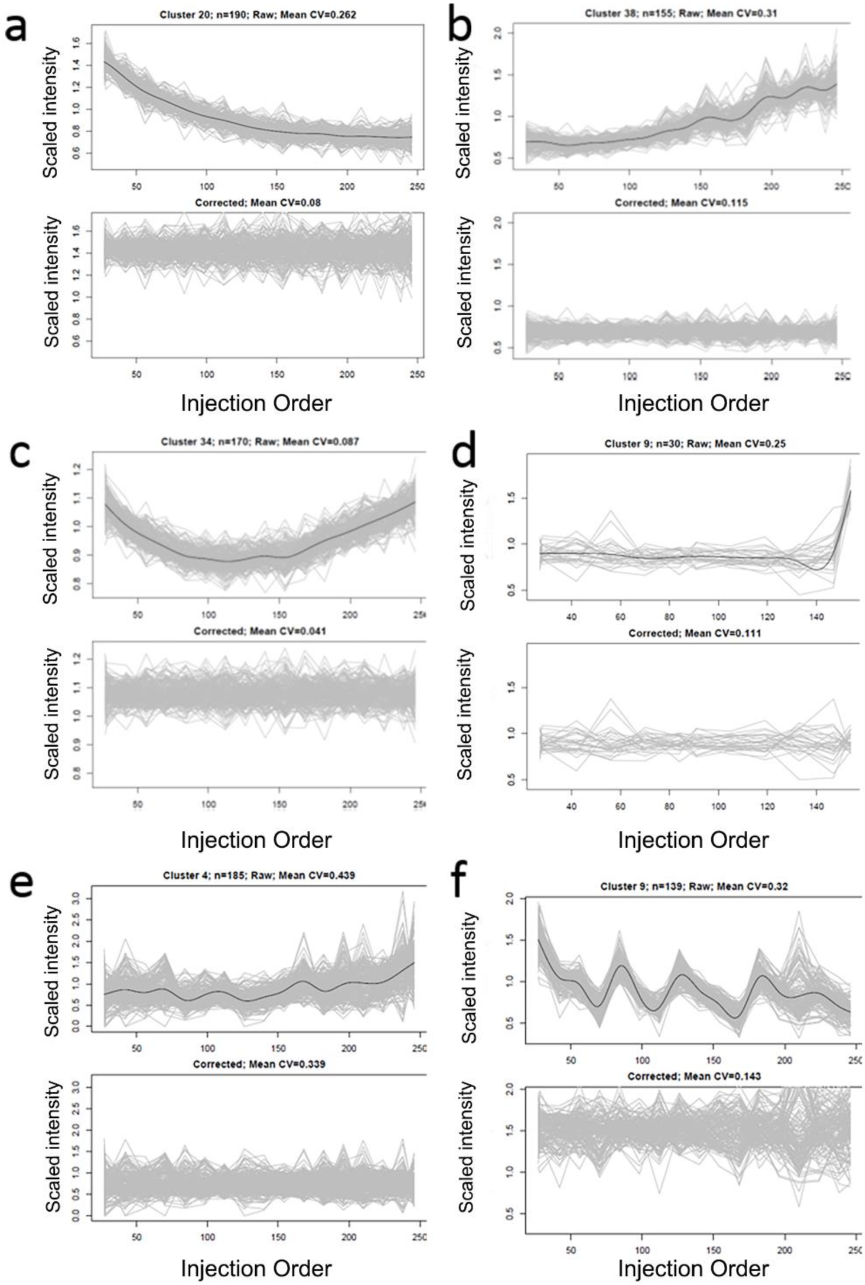


Figure 15. Feature intensity drift before and after within-batch correction. Authentic QC features separated into different clusters (number of features ranges from 30 to 185). (a-f**)** depict different intensity drift behaviors. The clusters presented in a) and b) closely follow the general within-batch intensity drift, whereas the other clusters (c–f) represent several distinctly diverse drift patterns. For each cluster, the upper graph shows the scaled features in grey and the cluster drift function in

black. The lower half shows the same features on the same y scale after application of cluster-based drift correction. Signal variations as indicated by coefficient of variation were considerably reduced in clusters presented after with-batch normalization.

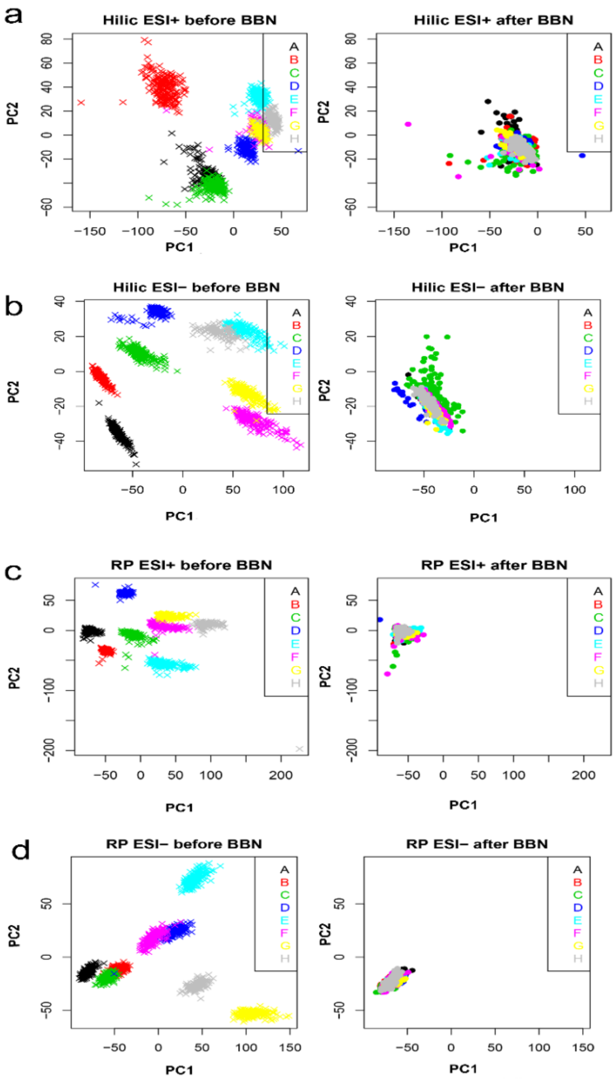


Figure 16. PCA score plots for visualization of performance of between-batch normalization (BBN). Data from eight analytical batches of human plasma samples used for **Papers III and IV** are shown. Data obtained by hydrophilic interaction chromatography (HILIC) and reverse phase chromatography (RP) using both positive (ESI+) and negative (ESI-) ionization. Fold change of the ratio of average reference sample intensity for a specific feature measured in batches and the ratio for the average intensity of all features within the batches was 7 for RP (ESI+) and 5 for others. Batches A-H are presented in different colors. PCA scores of qualified feature intensities before and after BBN are shown as crosses and circles, respectively.

4.2 ‘MUVR’ for multivariate modelling

A statistical validation framework for multivariate modelling was developed and is available as an open source R package ‘MUVR’ (https://gitlab.com/CarlBrunius/MUVR) (**Paper II**). MUVR incorporates an effectively unbiased variable selection process within an rdCV scheme and enables simultaneous identification of all-relevant variable sets and the minimal-optimal variable set (Figure 17). The algorithm developed allows for both PLS and RF core modeling and supports regression, classification, and multilevel modeling.

4.2.1 MUVR principle

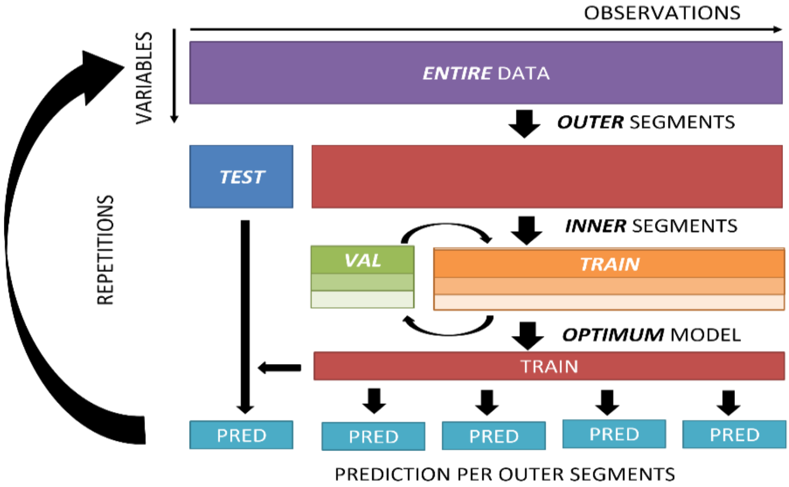


Figure 17. Graphical representation of unbiased variable selection in the MUVR algorithm.

The original data are randomly subdivided into OUTER segments. For each OUTER segment, the remaining (INNER) data are used for training and tuning of model parameters, including recursive ranking and backwards elimination of variables. In each of the INNER training models, variables are ranked by variable importance in projection for PLS analysis, whereas for RF analysis ranking is based on mean decrease in Gini index (classification) or mean decrease in accuracy (regression). For each iteration of the variable tuning, variable ranks are averaged between the inner models. After averaging, a user-specified proportion of the variables is removed from the data matrix before the next iteration (*varRatio*), where INNER segments are again randomly sampled to decrease bias to individual segments. Model performance is then estimated from predictions of the untouched test segments, using the number of selected variables (and optimal number of components for PLS modeling) determined by a consensus model from all INNER observations. Arbitration of model performance in variable tuning within the inner validation loop is performed using different fitness functions specifically adapted to the problem type: Root mean square error of prediction for regression and number of misclassifications for multilevel or general classification analysis (two or more classes). The area under the receiver operation characteristics curve is also supported as an optional fitness metric for two-class discriminant analysis (Szymańska *et al.*, 2012).

The procedure is then repeated for improved modeling performance. These models are created *nRep* (number of repetitions) × *nOuter* (number of outer segments) times for prediction of test segment observations ensuring that test segment observations are never used for model training or tuning. For key parameters of MUVR, results of stability testing suggest having $6 \leq nOuter \leq 8$, $nInner = nOuter - 1$, nRep≈15 and 0.5≤varRatio≤0.75 for initial preliminary analysis. Users could then increase nRep≥50 for reproducible results.

4.2.2 MUVR performance

The MUVR algorithm was performed on three -omics datasets with different characteristics, *i.e.*, 'Freelive' data used for regression analysis, 'Mosquito' data for classification analysis, and 'Crisp' data for multilevel analysis. The predictive results are shown in Figure 18(a-c). Regardless of the problem type (*i.e.*, regression, classification, multilevel) and core modeling applied (*i.e.*, PLS and RF), it was observed that variable reduction from the entire data to the 'max' model effectively resulted in removal of noise until an optimal validation performance as measured by the fitness function was obtained (Figure 18 d-f). The 'max' model would be an all-relevant solution, corresponding to a maximum of biologically relevant information and containing both optimal and redundant variables, and could be used to get an overview of biological perturbations or metabolic regulations associated with the research question. Further variable reduction from the 'max' model maintained validation performance down to the 'min' model, indicating dismissal of informative but redundant variables. Thus, the 'min' model corresponds to the minimal-optimal variable set, useful *e.g.*, for biomarker discovery. Further variable reduction increased model prediction errors, however, probably due to elimination of genuinely relevant, non-redundant information.

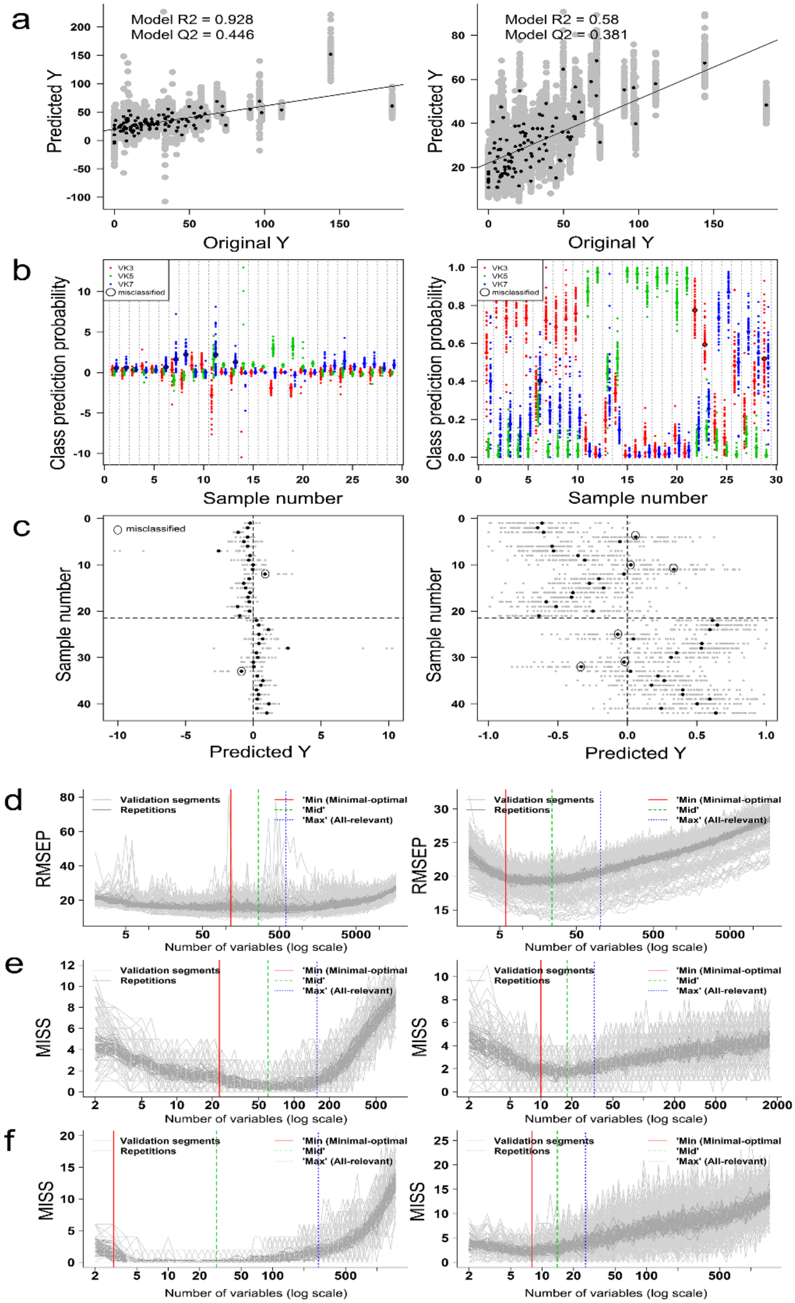


Figure 18. Predictive results and validation of models conducted for three datasets using MUVR PLS (left) and random forest (right) core modeling: (a-c) Predictive results of 'mid' model. a)

Regression using ‘Freelive’ data. Prediction estimates are shown in grey for each repetition, and in black for the prediction estimates averaged over all repetitions. b) Classification using ‘Mosquito’ data. Each swimlane represents one observation. Class probabilities are color-coded by class and presented per repetition (smaller dots) and averaged over all repetitions (larger dots), with misclassified samples circled. c) Multilevel analysis using ‘Crisp’ data. Samples are matched row-wise between upper and lower half for the positive and negative effect matrix, respectively. All upper half predictions should be negative and all lower half predictions positive for ideal classification. Prediction estimates are shown in grey for each repetition, and in black for the prediction estimates averaged over all repetitions and misclassified samples circled. (d-f) MUVR validation plots for identification of the all-relevant (‘max’ model) and minimal-optimal (‘min’ model) variables on three datasets. d) ‘Freelive’, regression; e) ‘Mosquito’, classification; f) ‘Crisp’, multilevel.

Benefiting from the rdCV scheme, MUVR minimizes selection bias by performing variable selection and tuning of model parameters in the inner segments, followed by assessment of modeling performance using outer loop data that are held out of model construction and variable reduction. MUVR yielded parsimonious models with minimal over-fitting and improved model performance with rdCV, particularly in cases where PLS core modelling was applied (Figure 19).

MUVR outperformed other RF-based variable selection techniques tailored for identification of all-relevant variables, *i.e.*, Boruta and VSURF, in various aspects, *e.g.*, less-stringent variable inclusion criteria, feasible computation time, minimum general over-fitting. It is also worth noting that, although Boruta and VSURF enable variable selection, some form of cross-validation is still required to avoid selection bias and to assess prediction performance (Kursa, 2010) (https://m2.icm.edu.pl/boruta/). In this respect, MUVR is more easy and efficient to use, allowing for simultaneous variable selection and validation with minimized selection bias.

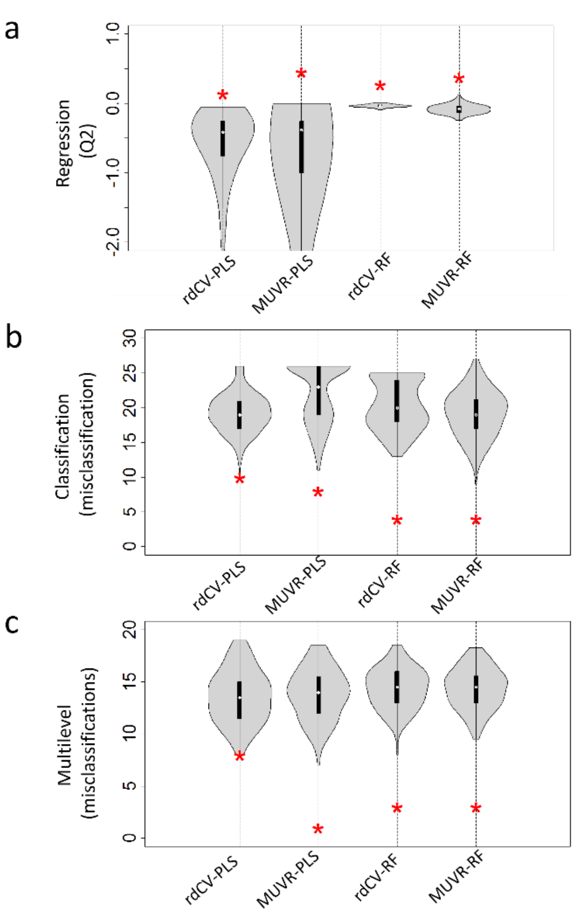


Figure 19. Performance of repeated double cross-validation models without (rdCV) or with unbiased variable selection (MUVR) built from actual data and random permutations for three data sets: a) 'Freelive', regression; b) 'Mosquito', classification; c) 'Crisp', multilevel. The performance distributions of random permutations are represented as violin plots, with the asterisks representing actual model performance (Q^2 for regression, number of misclassifications for classification and multilevel analysis).

4.3 Plasma metabolites associated with risk of T2D

In **Paper III**, several predictive metabolites of T2D were identified. These metabolites had limited incremental value in risk prediction beyond optimal utilization of traditional risk factors. However, they may provide other advantages over traditional risk factors, through higher long-term stability and their potential in providing information related to disease pathophysiology and progression.

4.3.1 Plasma metabolites associated with risk of developing T2D

Forty-six predictive metabolites of T2D were identified, including novel findings, *i.e.*, phosphatidylcholines (PCs) containing odd-chain fatty acids (OC-FA) and 2-hydroxyethanesulfonate, and previously identified predictive biomarkers. Of these, 41 were associated with insulin resistance and/or β-cell dysfunction at baseline. Among the 46 metabolites, 26 showed intermediate to good reproducibility among healthy controls ($0.4 \leq ICC \leq 0.75$), reinforcing their potential as predictive biomarkers.

Novel findings regarding PCs containing OC-FAs are encouraging enough to merit further investigation. It was found that lysoPC(19:1), lysoPC(15:1) and PC(15:1/18:2) were inversely associated with risk of developing T2D and exhibited moderate to high long-term reproducibility ($ICC>0.4$). Moreover, the three lipids changed in the disease-associated direction at follow-up measurement, thus most likely representing disease progression.

Paper III also replicated findings from several previous studies and supports existing hypotheses in relation to T2D pathogenesis, *e.g.*, dysregulated lipid metabolism (Markgraf *et al.*, 2016; de Mello *et al.*, 2017; Meikle and Summers, 2017), impaired BCAA metabolism (Chen *et al.*, 2016; Lotta *et al.*, 2016), and abnormal DAG (diglyceride) accumulation (Markgraf *et al.*, 2016). However, Paper III also provides information on the reproducibility of previously reported metabolites, an issue which has rarely been investigated previously. For instance, the remaining four PCs, *i.e.*, lysoPC(18:2), lysoPC(17:0), lysoPC(20:1), and PC(16:0/16:1), had only weak to moderate long-term reproducibility ($ICC<0.4$), which limits their potential as predictive biomarkers. BCAAs and 3-methyl-2-oxovaleric acid had high long-term (10 years) reproducibility ($ICC >0.6$), similar to reported two-year ICCs (Carayol *et al.*, 2015) but weaker shorter-term reproducibility (Floegel *et al.*, 2011; Breier *et al.*, 2014).

4.3.2 Improved risk prediction with optimal variable selection

For the first time, by adopting an unbiased variable selection approach among study-specific traditional risk factors and/or metabolites (Figure 20a), results from the comprehensive predictive analyses illustrate that predictive metabolites can provide complementary information, albeit with limited predictive improvement beyond the optimal utilization of traditional risk factors (Figure 20b).

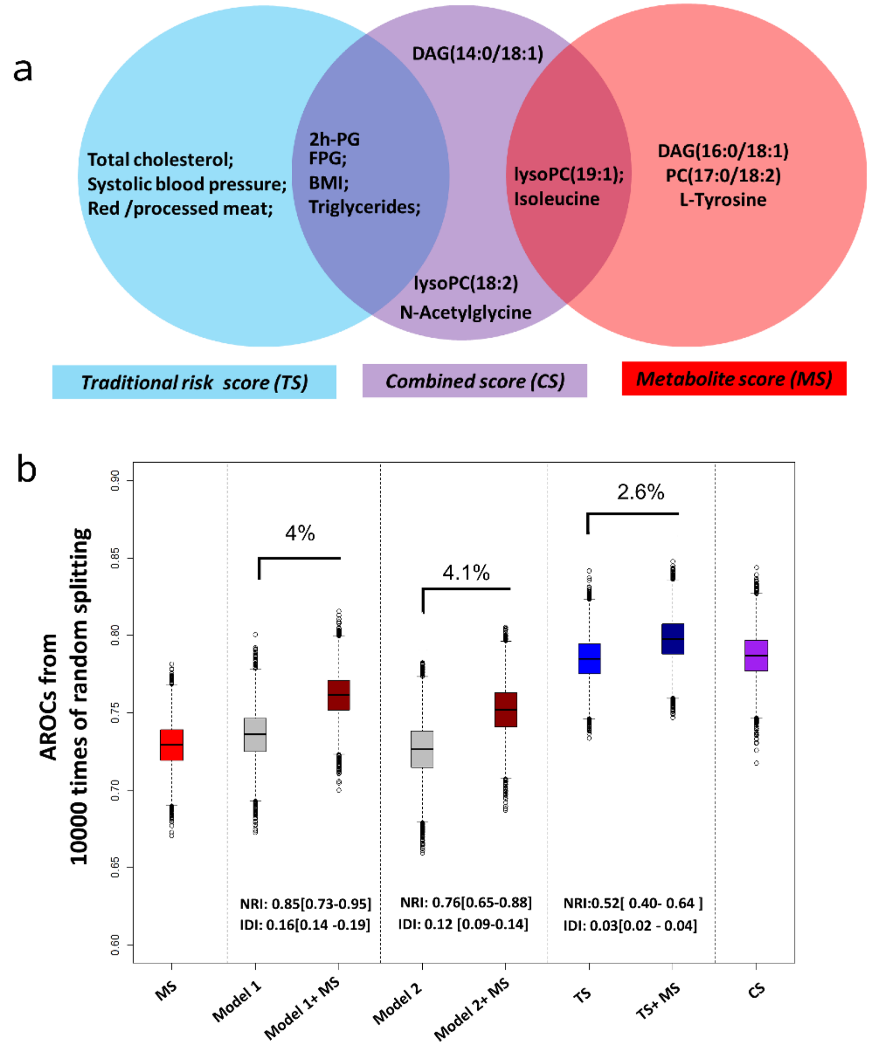


Figure 20. Comparison of the prediction performance of clinical risk factors, metabolites, and their combinations for risk of type 2 diabetes. a) Optimally selected subset of predictors, employing a validated random forest algorithm, for metabolite score (MS), traditional risk score (TS), and combined score (CS). b) Prediction performance of different models trained from metabolites,

traditional risk factors, and their combinations. The area under the receiver operation characteristics curves (AROC) was obtained from 10,000 models where the samples were randomly split into training (60%) and test sets (40%) for prediction and validation. Model 1 included fasting plasma glucose (mmol/l) and BMI (kg/m2) as predictors; Model 2 further included physical activity, education, smoking, alcohol, dietary fibre, red and processed meat consumption (g/day), and coffee intake as predictors; 2h-PG: 2h plasma glucose. DAG: diglycerides. FPG: fasting plasma glucose. PC: Phosphatidylcholine.

Several prospective studies have assessed the incremental predictive utility of adding metabolite biomarkers to traditional risk factors (Figure 21). These studies all used either an established risk score as reference or pre-defined subset of traditional risk factors as covariates adjusted in prediction models. Adding metabolites improved AROCs by 0.4-16% (Figure 21, blue bars).

In a majority of studies, traditional risk factors have not been optimally selected, which may under-estimate the predictive power of such risk factors in cohorts. This may give an overly optimistic impression of the benefits of adding metabolite data in risk prediction models, as clearly supported by the finding that adding best predictive metabolites (***MS***) into models constructed using pre-defined risk factors improved risk prediction by about 4% (Figure 21, red bar). However, compared with a model based on optimal selection of available cohort-specific traditional risk factors, only about 2.5% improvement was achieved (Figure 21, black bar). A larger improvement (6-8%) came, in fact, from optimal selection of traditional risk factors compared with pre-defined risk factors (Figure 21, yellow bars).

Many of the risk predictors typically used in cohort studies, *e.g.*, single measurements of blood lipids, blood pressure, and dietary data from FFQ, suffer from large systematic and random errors, which in turn may lead to inaccurate risk estimates (Tirosh *et al.*, 2008). Combined scores (***CS***) represented the optimal integration of metabolite predictors and traditional risk factors, which showed comparable predictive ability to Traditional risk scores (***TS***). However, the selected metabolite predictors in CS showed higher long-term reproducibility (0.37≤ICC≤0.68) than the total cholesterol, systolic blood pressure, and red/processed meat (0.3≤ICC≤0.41) involved in TS. This finding suggests that, although additional metabolite biomarkers may provide little improvement in prediction performance, reproducible metabolites may under some conditions serve as a complement or alternative to established risk factors, most likely providing more accurate risk estimates.

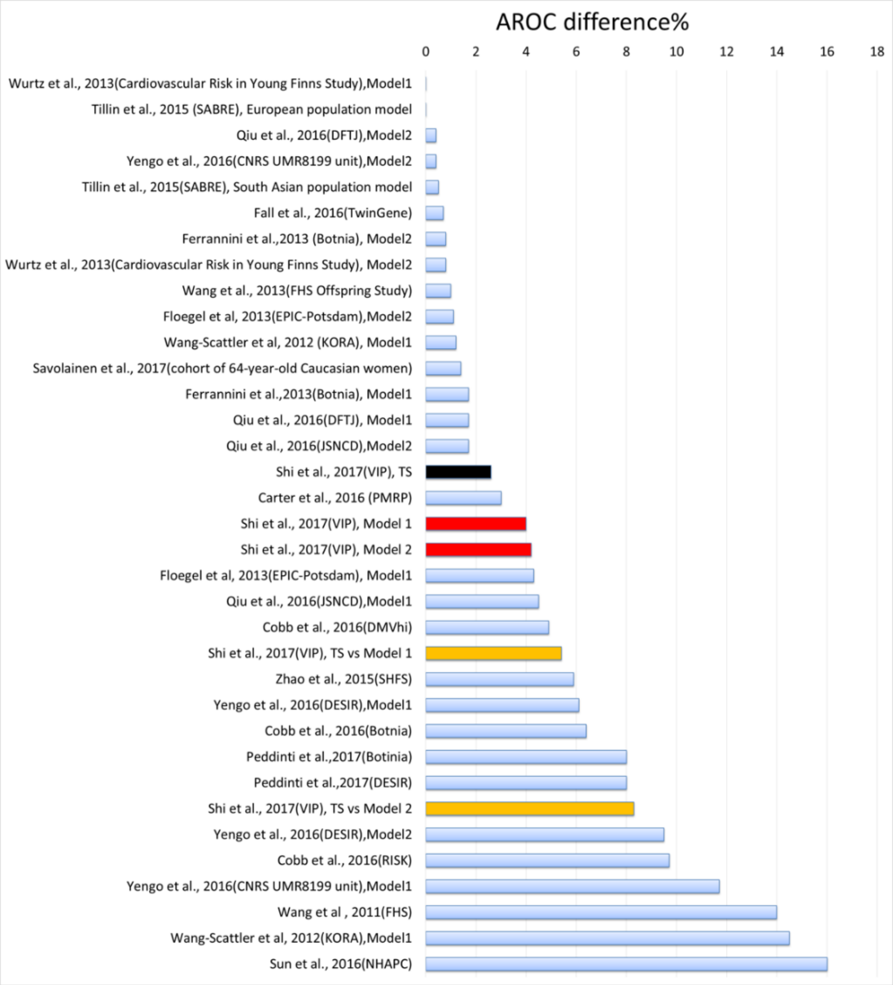


Figure 21. Incremental predictive ability of metabolites added to traditional risk factors. Area under the receiver operation characteristics curve (AROC) difference% was calculated by (AROC of model$_{\text{traditional risk factors+ metabolites}}$ – AROC of model$_{\text{traditional risk factors}}$)/AROC of model$_{\text{traditional risk factors}}$ ×100%. Model 1 and Model 2 are models conducted using *a priori*-selected traditional risk factors. TS: Traditional risk factor score. Black bar denotes the additional predictive ability of metabolites beyond TS. Yellow bars denote the AROC difference% between TS and Model 1/2.

4.4 Metabolites related to healthy Nordic dietary indices and risk of T2D

In **Paper IV,** metabolites related to predefined healthy Nordic dietary indices were identified and their associations with risk of T2D were investigated. The results did not support any association between indices and risk of T2D. Rather, unhealthy foods not included in the indices, such as pizzas, snack, liquor, and hamburgers, were associated with increased risk in the VIP population.

4.4.1 Metabolites related to BSDS and HNFI

In total, 29 metabolites were associated with Baltic Sea Diet Score (BSDS) ($-0.22<r<0.21$) and 13 with Healthy Nordic Food Index (HNFI) ($-0.14<r<0.17$). Correlations were in the same range as previously reported for BSDS and other established healthy dietary pattern indices, *e.g.*, the Healthy Eating Index and the WHO Healthy Diet Indicator(Playdon *et al.*, 2017). Among these metabolites, only four were associated with both indices, which is most likely attributable to the different food items used to score adherence (Playdon *et al.*, 2017).

The associations between metabolites and indices are not easily interpreted: some of the index-related metabolites could be biomarkers of particular food items that were key contributors to the indices investigated, *e.g.,* fish (eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA)) (Marckmann and Lassen, 1995; Silva *et al.*, 2014) and whole-grain products (pipecolic acid betaine) (Hanhineva, Lankinen, *et al.*, 2015; Pekkinen *et al.*, 2015). However, the majority of metabolites were associated with multiple foods captured by the indices as well as foods that were not (Figure 22). Thus, index-related metabolites may be related to general lifestyle patterns, which may possibly reflect the co-consumption of different foods (Lloyd *et al.*, 2013; Guertin *et al.*, 2014; Playdon *et al.*, 2017), or metabolites formed endogenously in response to dietary/lifestyle exposures (Esko *et al.*, 2017).

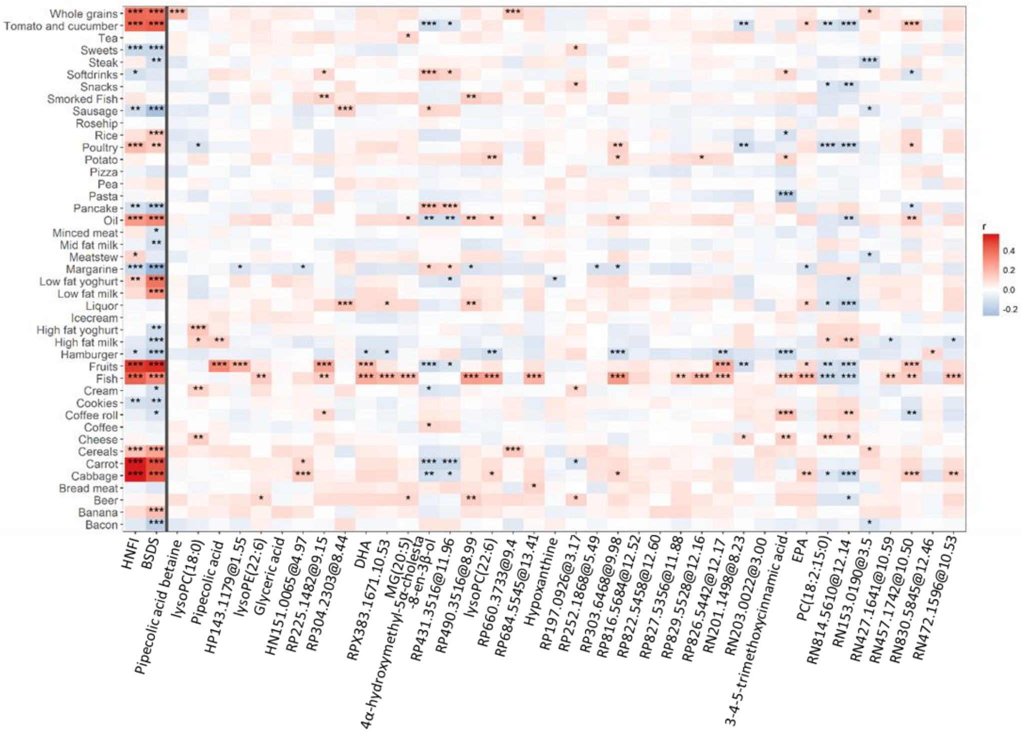


Figure 22. Correlations between index-related metabolites, healthy Nordic dietary indices, and food categories. *** p<0.00001, ** p<0.0001, * p<0.001,.

4.4.2 Healthy Nordic dietary indices and risk of T2D

No association was found between *a priori*-defined healthy Nordic dietary indices and T2D risk. This was surprising, given the beneficial cardiometabolic effects shown for healthy Nordic dietary patterns and/or food items in dietary intervention studies and previous observational studies (Adamsson *et al.*, 2011; Uusitupa *et al.*, 2013; Khakimov *et al.*, 2014; Lankinen *et al.*, 2014; Roswall *et al.*, 2015).

Fish is a central food item in the indices investigated. However, it was found that established biomarkers of fish consumption, *i.e.*, EPA and DHA, were positively associated with T2D risk, although attenuated on adjustment for BMI and blood lipids. This association may be due to the presence of organic pollutants known to occur in fish from the Baltic Sea (Mackenzie *et al.*, 2004; Järv *et al.*, 2017), which have been positively associated with T2D (Marushka *et al.*, 2017). The pollutants may counteract health effects of DHA and EPA on T2D risk (Marushka *et al.*, 2017). In contrast, several metabolites that were related with whole grains and fruits were inversely related with T2D (Parker *et al.*, 2013; Marushka *et al.*, 2017; Schwingshackl *et al.*, 2017).

Index-related metabolites were analyzed by PCA and the self-aggregated metabolites (loadings) were co-visualized with dietary variables and ORs of T2D in a ‘Triplot’, to aid in the interpretation of dietary patterns and risk of developing T2D (Figure 23a). The results clearly showed that indices and most food components of indices, with the exception of whole grains, were not associated with T2D risk in this case. Specifically, PC1 was indicative of adherence to healthy Nordic dietary indices, but was not associated with T2D (Figure 23b). In contrast, index-related metabolites with higher loading in PC2 were primarily associated with foods not included in the indices, *i.e.*, pizza, hamburgers, snacks. The observation that risk was predominantly associated with an unhealthy, rather than healthy, diet is supported by several other studies that have applied data-driven methods (*e.g.*, PCA, cluster analysis and reduced rank regression) (Jannasch *et al.*, 2017).

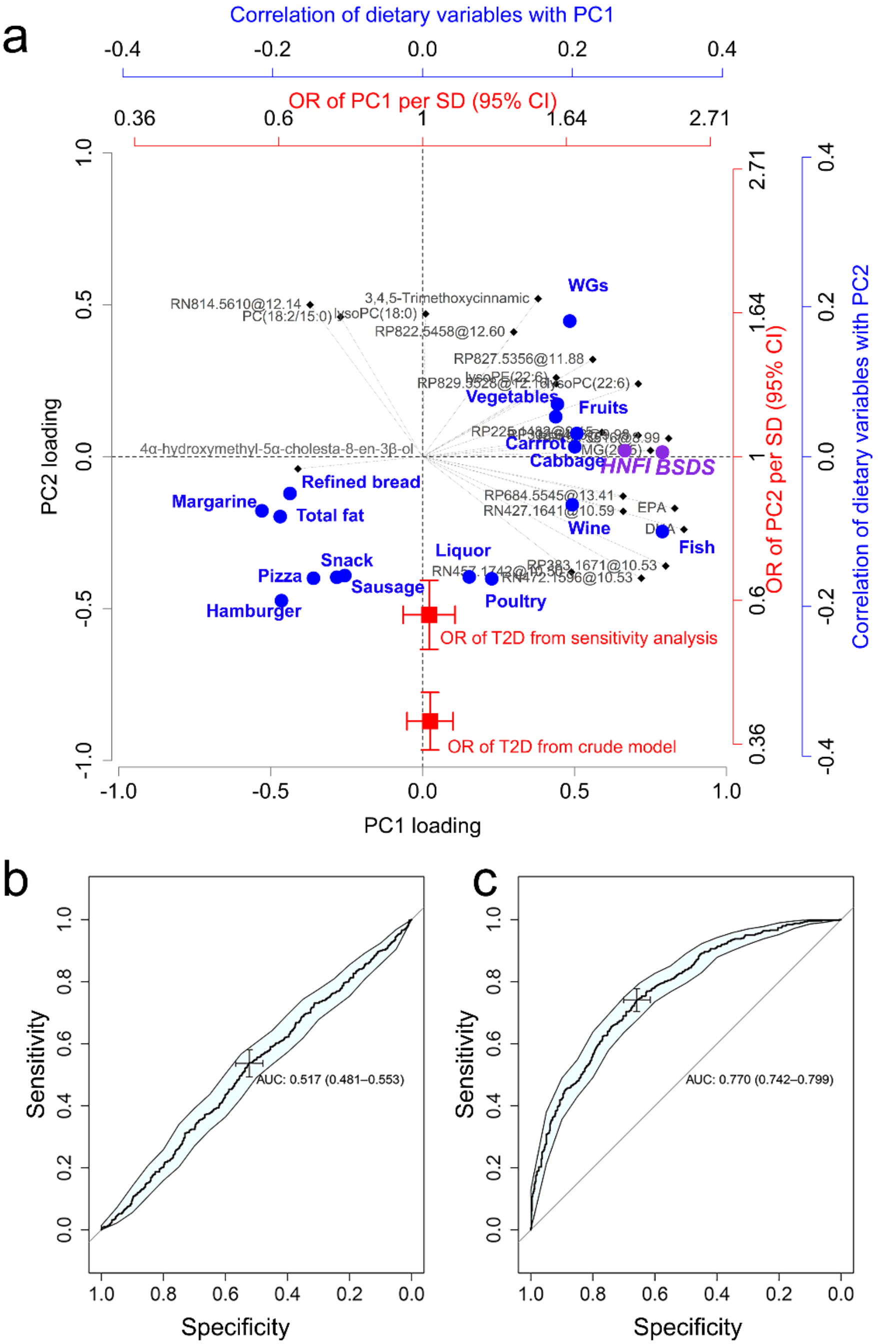


Figure 23. Associations between diet, index-related metabolites, and risk of developing type 2 diabetes (T2D). a) 'Triplot' representing index-related metabolites analyzed by PCA and correlations between principal component (PC) scores with dietary variables (individual foods in

blue circles and dietary indices in purple circles) and odds ratios (ORs) of T2D (per standard deviation; 95% confidence interval; red squares). Metabolites were visualized if |loading| ≥0.4. Correlations between PC scores and dietary variables were visualized if |correlation coefficient| ≥0.15 at a Bonferroni-adjusted p<0.05. ORs were obtained from crude model and model adjusting for lifestyle-related factors, fasting plasma glucose (mmol/L), BMI (kg/m^2), total cholesterol (mmol/L), triglycerides (mmol/L), and systolic and diastolic blood pressure (mmHg). For simple interpretation, metabolites, individual foods, and ORs along the same axis in the plot are associated. Dietary indices are orthogonal to ORs and therefore not associated with T2D risk. Receiver operating characteristics curve of b) PC1 and c) PC2 with the AUC (95% CI) for predicting incident T2D.

4.4.3 Replication of previously reported metabolites and T2D risk

Associations between 10 metabolites that were identified in recent literature to be associated with exposure to healthy Nordic foods and T2D risk in this study population were also investigated.

The findings reinforce the potential of using CMPF and 3-carboxy-4-methyl-5-pentyl-2-furanpropionic acid as biomarkers for fish intake (Hanhineva *et al.*, 2015; Koppe and Poitout, 2016; Savolainen *et al.*, 2017), AR 19:0-Gln for whole grains (Hanhineva *et al.*, 2015), and threonate for vegetable and fruits (Table 3). However, none of these was associated with T2D risk.

Moreover, no evidence was found to support claims for 1-methylhistidine and the two carnitines as generic biomarkers of fish and meat intake (Cheung *et al.*, 2017). However, these two acyl-carnitines with good long-tern reproducibility were strongly associated with increased risk of T2D. Although the mechanisms remain unclear, increased concentration of acyl-carnitines may be attributed to perturbed endogenous metabolism, *e.g.*, abnormal mitochondrial fatty acid oxidation, which may interfere with insulin sensitivity, contributing to the development of T2D (Schooneman *et al.*, 2013).

Table 3. *A priori-selected plasma metabolites and their associations with dietary indices in a nested case-control study in the Västerbotten Intervention Programme (VIP) cohort*

Metabolites[1]	r[2]	r[3]	OR per SD[4]	ICC[5]
TMAO	-	-	0.97(0.86,1.1)	0.18 (0.07, 0.36)
1-methylhistidine	-	-	1.17(1.01,1.35)	0.38 (0.27, 0.51)
Propionylcarnitine	-	-	1.44(1.24,1.66)	0.59 (0.49, 0.68)
2-methylbutyroylcarnitine	-	-	1.03(0.9,1.18)	0.52 (0.41, 0.62)
Threonate	-	Vegetables (0.13)	1(0.88,1.15)	0.21 (0.11, 0.38)

Hipputic acid	-	-	0.88(0.77,1)	0.33 (0.21, 0.47)
Indolepropionic acid	0.15	Cabbage(0.13), carrot(0.12)	0.81(0.7,0.93)	0.23 (0.12, 0.39)
CMPF	-	Fish (0.19)	1.04(0.9,1.2)	0.46 (0.35, 0.57)
3-carboxy-4-methyl-5-pentyl-2-furanpropionic acid	-	Fish (0.11)	1.09(0.95,1.24)	0.43 (0.32, 0.55)
AR 19:0-Gln	-	WG bread (0.14)	0.99(0.87,1.12)	0.51 (0.40, 0.61)

[1] Metabolites previously associated with either healthy Nordic diet or relevant individual food components, detected with MSI 1-2 in the present study. AR 19:0-Gln, nonadecyl-benzenediol glucuronide; CMPF, 3-carboxy-4-methyl-5-propyl-2-furanpropionic acid; TMAO, trimethylamine N-oxide.

[2] Partial Spearman correlation coefficients between metabolites and dietary indices or individual food components, controlling for age, gender, case/control status, BMI (kg/m2), smoking status, education, and physical activity.

[3] Partial Spearman correlation coefficients between metabolites and individual food components.

[4] Odds ratios (OR) per SD increment (95% confidence interval) of metabolites based on results from crude conditional logistic regression models.

[5] Intraclass correlation coefficient, representing long-term reproducibility of metabolites among healthy controls (N=187) over 10 years. ICC ≥0.4 denotes moderate to high reproducibility.

5 General discussion

Untargeted metabolomics has emerged as a promising approach to address various research questions (Rochat, 2016; Dudzik *et al.*, 2017; Han *et al.*, 2017). Although a wealth of analytical techniques and computational tools exist within the community, intuitive interpretation of complex data from untargeted metabolomics studies remains a major challenge, limiting translation of instrument-based results into biologically relevant conclusions (Alonso *et al.*, 2015; Ren *et al.*, 2015; Dudzik *et al.*, 2017). In an effort to improve the overall quality of the untargeted metabolomics data, to reduce false positive discoveries and to promote accurate interpretation, two freely available R packages, *i.e.*, 'batchCorr' (data-processing strategy) and 'MUVR' (multivariate statistical framework) were developed (**Papers I-II**). These packages allow for ease of implementation and use in large-scale applications, including over 2000 samples and spectra from RP (+), RP (-), HILIC (+), and HILIC (-) per sample in the studies on which this thesis is based (**Papers III-IV**).

Prospective cohort studies are particularly suitable for identification of early metabolic alterations before the onset of disease, since samples have been collected before diagnosis (Euser *et al.*, 2009; Thiese, 2014). However untargeted metabolomics had been applied in very few prospective studies before the research in this thesis (**Paper III;** identification of predictive metabolites of T2D). In a large nested case-control study, several novel and previously identified predictive metabolites of T2D were found. Although predictive metabolites resulted in little improvement in prediction performance beyond optimal utilization of traditional risk factors, they provided useful information related to disease pathophysiology, *e.g.* dysregulated lipid metabolism, impaired BCAA metabolism, and abnormal DAG accumulation. Moreover, several metabolites have shown potential to reflect disease progression. In addition, metabolites with good long-term reproducibility may serve as a complement or alternative to established risk factors, most likely resulting in more accurate risk estimates.

There is a large body of evidence from epidemiological studies, based on self-reported data, supporting an association between diet and the development of T2D. Associations have been found for both specific food items and overall diet quality reflected by established healthy dietary indices. However, in this thesis metabolites that were identified as prominent predictive biomarkers of T2D (**Paper III)** were only weakly correlated with self-reported intake of individual foods (-0.1<r<0.13 among healthy participants after adjusting for confounders). In addition, although beneficial cardiometabolic effects have been shown for a healthy Nordic diet in several studies (Kyrø *et al.*, 2013; Uusitupa, 2013; Kanerva *et al.*, 2014), none of the top-ranking predictive metabolites was correlated with the two pre-defined healthy Nordic dietary indices, *i.e.*, BSDS and HNFI. Moreover, the identified metabolites that were related to indices (**Paper IV**) showed remarkably weaker associations with T2D risk in the same population compared with the top-ranking predictive metabolites of T2D (**Paper III**).

These findings imply that endogenous metabolic perturbations, *e.g.*, dysregulated lipid metabolism, impaired amino acids metabolism and fatty acid metabolism might dominate the measurable metabolome, effectively drowning out systematic, but weaker, signals from dietary and lifestyle-related metabolites in relation to disease. Predictive metabolites of T2D, such as BCAAs, carnitines, and phosphatidylcholines, that are repeatedly mentioned in the literature may be related to food intake to some extent, but their perturbations in relation to T2D may not be due to dietary exposures *per se*. Apart from genetic risk factors, the modifiable risk factors, in particular habitual unhealthy lifestyle, including poor diet quality, high energy intake, physical inactivity, smoking, and sleep disorders, are all contributors to dysregulated endogenous metabolism (Ley *et al.*, 2016; Kolb and Martin, 2017).

Although public health recommendations for overall lifestyle modification exist, precision lifestyle guidelines may be more effective for the prevention and treatment of T2D (Franks and Poveda, 2017; Mutie *et al.*, 2017). In this respect, more studies are required to disentangle the contributions of risk factors in pathological pathways of T2D. State-of-the-art multi-omics techniques and traditional methods should be integrated in order to provide comprehensive information, thereby facilitating meaningful optimization of lifestyle modifications.

In conclusion, the novel bioinformatics tools developed proved to be robust and effective to overcome vital data-analytical challenges inherent in large-scale untargeted metabolomics studies. The application of these tools assisted in discovering predictive biomarkers of T2D and biomarkers of healthy Nordic dietary indices in a Swedish population. Findings underline the potential of untargeted metabolomics for discovery of biomarkers, which may provide useful information related to disease pathophysiology and monitoring of disease progression, as well as facilitate investigations into the relationship between diet and risk of developing T2D.

6 Concluding remarks

- A novel data-processing strategy for large-scale untargeted LC-MS metabolomics data correction was developed (**Paper I**) and applied in the presented untargeted metabolomics studies (**Papers III and IV**).
- A novel multivariate statistical framework for improving prediction performance and minimizing the risk of false positive discoveries was developed (**Paper II**) and applied to identify biomarkers of T2D and dietary indices (**Papers III and IV**).
- Several fasting plasma metabolites were associated with risk of developing T2D, with moderate to excellent long-term reproducibility, reinforcing their potential as predictive biomarkers of T2D (**Paper III**).
- Predictive metabolites can only provide limited improvement in T2D risk prediction beyond the optimal utilization of traditional risk factors in the Swedish population investigated (**Paper III**).
- Heathy Nordic dietary indices, *i.e.*, Baltic Sea Diet Score and Healthy Nordic Food Index, were not associated with T2D risk. Dietary patterns characterized by consumption of potentially unhealthy foods not included in these indices were derived from index-related metabolites and appeared to be more important for development of T2D in the study population (**Paper IV**).

7 Future perspectives

- It is worth conducting a more thorough comparison of the two algorithms developed in this thesis, *i.e.*, ‘batchCorr’ and ‘MUVR’, with other existing software tools that have been proposed to address similar tasks.
- MUVR can be further extended by implementing other core modeling techniques (*e.g.*, support vector machine), variable ranking techniques (*e.g.*, rank products), and/or different fitness functions (*e.g.*, balanced error rate for classification problems with unbalanced design).
- There is a pressing need for high-throughput bioinformatics tools to speed up and automate identification of metabolites from untargeted metabolomics experiments, thereby effectively increasing the number of certainly identified metabolites in relation to the total number of features captured.
- On the basis of potential to improve identification strategy, all-relevant metabolites should be identified, as they may provide a holistic review of early metabolite alteration before onset of T2D and facilitate mechanism investigation.
- Other independent cohorts should be further investigated to generalize the conclusion that the incremental predictive ability of metabolites was limited in relation to optimal utilization of available traditional risk factors.
- The association between healthy Nordic dietary indices and T2D risk merits further investigation in the entire VIP cohort or other independent population with high intake of Baltic fish. In this

respect, it is worth investigating the possible interactive effect of pollutants and fish consumption on T2D development.

- Future studies should focus on optimizing recommendations for individuals with particular health needs and disease prevention, *i.e.* personalized lifestyle recommendations. To achieve this purpose, a proper integration of state-of-the-art multi-omics techniques and traditional methods should be conducted to investigate how lifestyle risk factors affect pathological pathways of T2D and prevent disease development, and to improve understanding of variability in individuals in responding to such lifestyle modifications.

References

Abdullah, N., Abdul Murad, N. A., Mohd Haniff, E. A., Syafruddin, S. E., Attia, J., Oldmeadow, C., Kamaruddin, M. A., Abd Jalal, N., Ismail, N., Ishak, M., Jamal, R., Scott, R. J. and Holliday, E. G. (2017) Predicting type 2 diabetes using genetic and environmental risk factors in a multi-ethnic Malaysian cohort, *Public Health*, 149, pp. 31–38.

Abiemo, E. E., Alonso, A., Nettleton, J. a., Steffen, L. M., Bertoni, A. G., Jain, A. and Lutsey, P. L. (2012) Relationships of the Mediterranean dietary pattern with insulin resistance and diabetes incidence in the Multi-Ethnic Study of Atherosclerosis (MESA), *British Journal of Nutrition*, 109, pp. 1490–1497.

Adamsson, V., Cederholm, T., Vessby, B. and Risérus, U. (2014) Influence of a healthy Nordic diet on serum fatty acid composition and associations with blood lipoproteins - Results from the NORDIET study, *Food and Nutrition Research*, 58, pp. 1–6.

Adamsson, V., Reumark, A., Fredriksson, I. B., Hammarström, E., Vessby, B., Johansson, G. and Risérus, U. (2011) Effects of a healthy Nordic diet on cardiovascular risk factors in hypercholesterolaemic subjects: A randomized controlled trial (NORDIET), *Journal of Internal Medicine*, 269, pp. 150–159.

Afanador, N. L., Smolinska, A., Tran, T. N. and Blanchet, L. (2016) Unsupervised random forest: A tutorial with case studies, *Journal of Chemometrics*, 30, pp. 232–241.

AlEssa, H. B., Malik, V. S., Yuan, C., Willett, W. C., Huang, T., Hu, F. B. and Tobias, D. K. (2017) Dietary patterns and cardiometabolic and endocrine plasma biomarkers in US women, *The American Journal of Clinical Nutrition*, 105, pp. 432–441.

Alhazmi, A., Stojanovski, E., Mcevoy, M. and Garg, M. L. (2014) The association between dietary patterns and type 2 diabetes: A systematic review and meta-analysis of cohort studies, *Journal of Human Nutrition and Dietetics*, 27, pp. 251–260.

Alonso, A., Marsal, S. and JuliÃ, A. (2015) Analytical Methods in Untargeted Metabolomics: State of the Art in 2015, *Frontiers in Bioengineering and Biotechnology*, 3, pp. 1–20.

Ambroise, C. and McLachlan, G. J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data, *Proceedings of the National Academy of Sciences*, 99, pp. 6562–6566.

American diabetes association (2010) Diagnosis and Classification of Diabetes Mellitus, *Diabetes Care*, 33, pp. 62–69.

Arlot, S. and Celisse, A. (2010) A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, pp. 40–79.

Barallobre-Barreiro, J., Chung, Y.-L. and Mayr, M. (2013) Proteomics and metabolomics for mechanistic insights and biomarker discovery in cardiovascular disease., *Revista española de cardiología (English ed.)*, 66, pp. 657–61.

Barbu, A., She, Y., Ding, L. and Gramajo, G. (2013) Feature Selection with Annealing for Big Data Learning, *arXiv preprint arXiv:1310.2880*, pp. 1–18.

Baumann, D. and Baumann, K. (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation, *Journal of Cheminformatics*, 6, pp.1-19.

Benton, H. P., Ivanisevic, J., Mahieu, N. G., Kurczy, M. E., Johnson, C. H., Franco, L., Rinehart, D., Valentine, E., Gowda, H., Ubhi, B. K., Tautenhahn, R., Gieschen, A., Fields, M. W., Patti, G. J. and Siuzdak, G. (2015) Autonomous metabolomics for rapid metabolite identification in global profiling, *Analytical Chemistry*, 87, pp. 884–891.

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. and van der Werf, M. J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data., *BMC Genomics*, 7, p. 142.

Bingol, K., Bruschweiler-Li, L., Li, D., Zhang, B., Xie, M. and Brüschweiler, R. (2016) Emerging new strategies for successful metabolite identification in metabolomics, *Bioanalysis*, 8, pp. 557–573.

Biskup, I., KyrO, C., Marklund, M., Olsen, A., Van Dam, R. M., Tjonneland, A., Overvad, K., Lindahl, B., Johansson, I. and Landberg, R. (2016) Plasma alkylresorcinols, biomarkers of whole-grain wheat and rye intake, and risk of type 2 diabetes in Scandinavian men and women, *The American Journal of Clinical Nutrition*, 104, pp. 88–96.

Bordoni, A. and Capozzi, F. (2015) The foodomics approach for discovering biomarkers of food consumption in nutrition studies, *Current Opinion in Food Science*, 4, pp. 124–128.

Boulesteix, A. L. (2007) WilcoxCV: An R package for fast variable selection in cross-validation, *Bioinformatics*, 23, pp. 1702–1704.

Breier, M., Wahl, S., Prehn, C., Fugmann, M., Ferrari, U., Weise, M., Banning, F., Seissler, J., Grallert, H., Adamski, J. and Lechner, A. (2014) Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples, *PLoS ONE*, 9, pp. 1–11.

Brennan, L., Gibbons, H. and O'Gorman, A. (2015) An Overview of the Role of Metabolomics in the Identification of Dietary Biomarkers, *Current Nutrition Reports*, 4, pp. 304–312.

Bro, R., Kjeldahl, K., Smilde, a K. and Kiers, H. a L. (2008) Cross-validation of component models: a critical look at current methods., *Analytical and Bioanalytical Chemistry*, 390, pp. 1241–51.

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

Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., Mamas, M. A., Neyses, L. and Dunn, W. B. (2011) Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets, *Bioinformatics*, 27, pp. 1108–1112.

Buck, M., Nilsson, L. K. J., Brunius, C., Dabiré, R. K., Hopkins, R. and Terenius, O. (2016) Bacterial associations reveal spatial population dynamics in Anopheles gambiae mosquitoes, *Scientific Reports*, 6, p. 22806.

Bujak, R., Daghir-Wojtkowiak, E., Kaliszan, R. and Markuszewski, M. J. (2016) PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-targeted Metabolomics Data, *Frontiers in Molecular Biosciences*, 3, pp.1-10.

Carayol, M., Licaj, I., Achaintre, D., Sacerdote, C., Vineis, P., Key, T. J., Moret, N. C. O., Scalbert, A., Rinaldi, S. and Ferrari, P. (2015) Reliability of serum metabolites over a two-year period: A targeted metabolomic approach in fasting and non-fasting samples from EPIC, *PLoS ONE*, 10, pp. 1–10.

Cespedes, E. M. and Hu, F. B. (2015) Dietary patterns : from nutritional epidemiologic analysis to national guidelines, *The American Journal of Clinical Nutrition*, 101, pp. 899–900.

Chen, T., Ni, Y., Ma, X., Bao, Y., Liu, J., Huang, F., Hu, C., Xie, G., Zhao, A., Jia, W. and Jia, W. (2016) Branched-chain and aromatic amino acid profiles and diabetes risk in Chinese populations., *Scientific reports*, 6, p. 20594.

Cheung, W., Keski-Rahkonen, P., Assi, N., Ferrari, P., Freisling, H., Rinaldi, S., Slimani, N., Zamora-Ros, R., Rundle, M., Frost, G., Gibbons, H., Carr, E., Brennan, L., Cross, A. J., Pala, V., *et al.* (2017) A metabolomic study of biomarkers of meat and fish intake, *The American Journal of Clinical Nutrition*, 105, pp. 600–608.

Chong, E. Y., Huang, Y., Wu, H., Ghasemzadeh, N., Uppal, K., Quyyumi, A. A., Jones, D. P. and Yu, T. (2015) Local false discovery rate estimation using feature reliability in LC/MS metabolomics data, *Scientific Reports*, 5, p. 17221.

Correa, E. and Goodacre, R. (2011) A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: application to the rapid detection of Bacillus spores and identification of Bacillus species, *BMC Bioinformatics*, 12, pp.1-17.

Damsgaard, C. T., Dalskov, S.-M., Petersen, R. a, Sørensen, L. B., Mølgaard, C., Biltoft-Jensen, A., Andersen, R., Thorsen, A. V, Tetens, I., Sjödin, A., Hjorth, M. F., Vassard, D., Jensen, J. D., Egelund, N., Dyssegaard, C. B., *et al.* (2012) Design of the OPUS School Meal Study: a randomised controlled trial assessing the impact of serving school meals based on the New Nordic Diet., *Scandinavian Journal of Public Health*, 40, pp. 693–703.

Drogan, D., Dunn, W. B., Lin, W., Buijsse, B., Schulze, M. B., Langenberg, C., Brown, M., Floegel, A., Dietrich, S., Rolandsson, O., Wedge, D. C., Goodacre, R., Forouhi, N. G., Sharp, S. J., Spranger, J., *et al*. (2014) Untargeted Metabolic Profiling Identifies Altered Serum Metabolites of Type 2 Diabetes Mellitus in a Prospective, Nested Case-Control Study, *Clinical Chemistry*, 497, pp. 487–497.

Dudzik, D., Barbas-Bernardos, C., García, A. and Barbas, C. (2017) Quality assurance procedures for mass spectrometry untargeted metabolomics. a review, *Journal of Pharmaceutical and Biomedical Analysis*, pp. 1–25.

Dunn, W. B. (2012) Diabetes - the Role of Metabolomics in the Discovery of New Mechanisms and Novel Biomarkers, *Current Cardiovascular Risk Reports*, 7, pp. 25–32.

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J. D., Halsall, A., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Kell, D. B. and Goodacre, R. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry., *Nature Protocols*, 6, pp. 1060–1083.

Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J. and Viant, M. R. (2013) Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics, *Metabolomics*, 9(SUPPL.1), pp. 44–66.

Dunn, W. B., Wilson, I. D., Nicholls, A. W. and Broadhurst, D. (2012) The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, *Bioanalysis*, 4, pp. 2249–2264.

Dührkop, K., Shen, H., Meusel, M., Rousu, J. and Böcker, S. (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID, *Proceedings of the National Academy of Sciences*, 112, pp. 12580–12585.

Eliasson, M., Rännar, S., Madsen, R., Donten, M. A., Marsden-Edwards, E., Moritz, T., Shockcor, J. P., Johansson, E. and Trygg, J. (2012) Strategy for optimizing LC-MS data processing in metabolomics: A design of experiments approach, *Analytical Chemistry*, 84, 6869−6876.

Esko, T., Hirschhorn, J. N., Feldman, H. A., Hsu, Y. H. H., Deik, A. A., Clish, C. B., Ebbeling, C. B. and Ludwig, D. S. (2017) Metabolomic profiles as reliable biomarkers of dietary composition, *The American Journal of Clinical Nutrition*, 105(3), pp. 547–554.

Euser, A. M., Zoccali, C., Jager, K. J. and Dekker, F. W. (2009) Cohort studies: Prospective versus retrospective, *Nephron - Clinical Practice*, 113, 3.

Fall, T., Salihovic, S., Brandmaier, S., Nowak, C., Ganna, A., Gustafsson, S., Broeckling, C. D., Prenni, J. E., Kastenmüller, G., Peters, A., Magnusson, P. K., Wang-Sattler, R., Giedraitis, V., Berne, C., Gieger, C., *et al.* (2016) Non-targeted metabolomics combined with genetic analyses identifies bile acid synthesis and phospholipid metabolism as being associated with incident type 2 diabetes, *Diabetologia*, 59, pp. 2114–2124.

Fernández-Albert, F., Llorach, R., Garcia-Aloy, M., Ziyatdinov, A., Andrés-Lacueva, C. and Perera, A. (2014) Intensity drift removal in LC/MS metabolomics by Common Variance Compensation., *Bioinformatics*, 30, pp. 2899–2905.

Filzmoser, P., Liebmann, B. and Varmuza, K. (2009) Repeated double cross validation, *Journal of Chemometrics*, 23, pp. 160–71.

Floegel, A., Drogan, D., Wang-Sattler, R., Prehn, C., Illig, T., Adamski, J., Joost, H. G., Boeing, H. and Pischon, T. (2011) Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach, *PLoS ONE*, 6, e21103.

Floegel, a, von Ruesten, A., Drogan, D., Schulze, M. B., Prehn, C., Adamski, J., Pischon, T. and Boeing, H. (2013) Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam., *European Journal of Clinical Nutrition*, 67, pp. 1100–8.

Fondi, M. and Liò, P. (2015) Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology, *Microbiological Research*, 171, pp. 52–64.

Fraley, C., Raftery, A., Murphy, T. and Scrucca, L. (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation, *Technical Report No. 597*.

Franks, P. W. and McCarthy, M. I. (2016) Exposing the exposures responsible for type 2 diabetes and obesity, *Science*, 354, pp. 69–73.

Franks, P. W. and Poveda, A. (2017) Lifestyle and precision diabetes medicine: will genomics help optimise the prediction, prevention and treatment of type 2 diabetes through lifestyle therapy?, *Diabetologia*, 60, pp. 784–792.

Freedman, L. S., Kipnis, V., Schatzkin, A., Tasevska, N. and Potischman, N. (2010) Can we use biomarkers in combination with self-reports to strengthen the analysis of nutritional epidemiologic studies?, *Epidemiologic Perspectives & Innovations*, 7, pp. 1–9.

Zamaratskaia G., Johansson D.P., Junqueira M.A., Deissler L., Langton M, Hellström P.M., Landberg R. (2017) Impact of sourdough fermentation on appetite and postprandial metabolic responses – a randomised cross-over trial with whole grain rye crispbread, *British Journal of Nutrition*, Accepted.

Ganna, A., Fall, T., Salihovic, S., Lee, W., Broeckling, C. D., Kumar, J., Hägg, S., Stenemo, M., Magnusson, P. K. E., Prenni, J. E., Lind, L., Pawitan, Y. and Ingelsson, E. (2016) Large-scale non-targeted metabolomic profiling in three human population-based studies, *Metabolomics*, 12, p. 1-13.

Garcia-Perez, I., Gibson Bsc, R., Chambers, E. S., Frost, G., Posma, J. M., Nicholson, J. K., Holmes, E., Elliott, P., Hansen, T. H., Vestergaard, H., Hansen, T., Pedersen, O., Beckmann, M., Draper, J., Garcia-Perez, I., *et al.* (2017) Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial, *Lancet Diabetes & Endocrinology*, 5, pp. 184–195.

Genuer, R., Poggi, J. and Tuleau-malot, C. (2015) VSURF : An R Package for Variable Selection Using Random Forests, *The R Journal* , 7, pp. 19–33.

Gibbons, H. and Brennan, L. (2017) Metabolomics as a tool in the identification of dietary biomarkers, *Proceedings of the Nutrition Society*, 76, pp. 42–53.

Gjesing, A. P. and Pedersen, O. (2012) "Omics"-driven discoveries in prevention and treatment of type 2 diabetes, *European Journal of Clinical Investigation*, 42, pp. 579–588.

Gonzalez-franquesa, A., Burkart, A. M. and Isganaitis, E. (2016) What Have Metabolomics Approaches Taught Us About Type 2 Diabetes ?, *Current Diabetes Reports*, 16(74). pp.1-10.

Gorrochategui, E., Jaumot, J., Lacorte, S. and Tauler, R. (2016) Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC Trends in Analytical Chemistry*, 82, pp. 425–442.

Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B. and Pujos-Guillot, E. (2016) Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data, *Frontiers in Molecular Biosciences*, 3, pp. 1–15.

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L. and Goodacre, R. (2015) A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding, *Analytica Chimica Acta*, 879, pp. 10–23.

Gromski, P. S., Xu, Y., Correa, E., Ellis, D. I., Turner, M. L. and Goodacre, R. (2014) A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data, *Analytica Chimica Acta*, 829, pp. 1–8.

Guertin, K. a., Moore, S. C., Sampson, J. N., Huang, W. Y., Xiao, Q., Stolzenberg-Solomon, R. Z., Sinha, R. and Cross, A. J. (2014) Metabolomics in nutritional epidemiology: Identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations, *The American Journal of Clinical Nutrition*, 100, pp. 208–217.

Hameed, I., Masoodi, S. R., Mir, S. A., Nabi, M., Ghazanfar, K. and Ganai, B. A. (2015) Type 2 diabetes mellitus: From a metabolic disorder to an inflammatory condition., *World journal of diabetes*, 6, pp. 598–612.

Han, T.-L., Yang, Y., Zhang, H. and Law, K. P. (2017) Analytical challenges of untargeted GC-MS-based metabolomics and the critical issues in selecting the data processing strategy, *F1000Research*, 6:967.

Hanhineva, K., Brunius, C., Andersson, A., Marklund, M., Juvonen, R., Keski-Rahkonen, P., Auriola, S. and Landberg, R. (2015) Discovery of urinary biomarkers of whole grain rye intake in free-living subjects using nontargeted LC-MS metabolite profiling, *Molecular Nutrition & Food Research*, 59, pp. 2315–2325.

Hanhineva, K., Lankinen, M. A., Pedret, A., Schwab, U., Kolehmainen, M., Paananen, J., Mello, V. De, Sola, R., Lehtonen, M., Poutanen, K. and Uusitupa, M. (2015) Nontargeted metabolite profiling discriminates diet-specific biomarkers for consumption of whole grains , fatty fish , and bilberries in a randomized controlled trial, *The Journal of Nutrition*, 145, pp. 7–17.

Haring, R. and Wallaschofski, H. (2012) Diving Through the "-Omics": The Case for Deep Phenotyping and Systems Epidemiology, *OMICS: A Journal of Integrative Biology*, 16, pp. 231–234.

Hedrick, V. E., Dietrich, A. M., Estabrooks, P. a, Savla, J., Serrano, E. and Davy, B. M. (2012) Dietary biomarkers: advances, limitations and future directions., *Nutrition Journal*, 11, p. 109.

Herder, C., Kowall, B., Tabak, A. G. and Rathmann, W. (2014) The potential of novel biomarkers to improve risk prediction of type 2 diabetes, *Diabetologia*, pp. 16–29.

van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. and Rogers, S. (2016) Topic modeling for untargeted substructure exploration in metabolomics, *Proceedings of the National Academy of Sciences*, 113, pp. 13738–13743.

Jacobs, S., Boushey, C. J., Franke, A. A., Shvetsov, Y. B., Monroe, K. R., Haiman, C. A., Kolonel, L. N., Le Marchand, L. and Maskarinec, G. (2017) A priori-defined diet quality indices, biomarkers and risk for type 2 diabetes in five ethnic groups: the Multiethnic Cohort, *British Journal of Nutrition*, 118, pp. 312–320.

Jannasch, F., Kröger, J. and Schulze, M. B. (2017) Dietary Patterns and Type 2 Diabetes: A Systematic Literature Review and Meta-Analysis of Prospective Studies, *The Journal of Nutrition*, 147, pp. 1174–1182.

Johansson, I., Van Guelpen, B., Hultdin, J., Johansson, M., Hallmans, G. and Stattin, P. (2010) Validity of food frequency questionnaire estimated intakes of folate and other B vitamins in a region without folic acid fortification., *European Journal of Clinical Nutrition*, 64, pp. 905–913.

Johansson, I., Hallmans, G., Wikman, A., Biessy, C., Riboli, E. and Kaaks, R. (2002) Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort., *Public Health Nutrition*, 5, pp. 487–496.

Järv, L., Kiviranta, H., Koponen, J., Rantakokko, P., Ruokojärvi, P., Radin, M., Raid, T., Roots, O. and Simm, M. (2017) Persistent organic pollutants in selected fishes of the Gulf of Finland, *Journal of Marine Systems*, 171, pp. 129–133.

Kanerva, N., Kaartinen, E., Rissanen, H., Knekt, P., Eriksson, J. G., Sääksjärvi, K., Sundvall, J. and Männistö, S. (2014) Associations of the Baltic Sea diet with cardiometabolic risk factors – a meta-analysis of three Finnish studies, *British Journal of Nutrition*, 112, pp. 616–626.

Kanerva, N., Kaartinen, N. E., Schwab, U., Lahti-Koski, M. and Männistö, S. (2014) The Baltic Sea Diet Score: a tool for assessing healthy eating in Nordic countries., *Public Health Nutrition*, 17, pp. 1697–705.

Kanerva, N., Rissanen, H., Knekt, P., Havulinna, A. S., Eriksson, J. G. and Männistö, S. (2014) The healthy Nordic diet and incidence of Type 2 Diabetes - 10-year follow-up, *Diabetes Research and Clinical Practice*, 106, pp. e34–e37.

Katajamaa, M., Miettinen, J. and Orešič, M. (2006) MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data, *Bioinformatics*, 22, pp. 634–636.

Kennedy E, James, O., Steven, C. and Kathryn, F. (1995) The Healthy Eating Index, *Journal of the American Dietetic Association*, 95, pp. 1103–1108.

Khakimov, B., Poulsen, S. K., Savorani, F., Acar, E., Larsen, T. M., Astrup, A., Dragsted, L. O., Engelsen, S. B., Lacoppidan, S. A., Kyrø, C., Loft, S., Helnæs, A., Christensen, J., Hansen, C. P., Dahm, C. C., *et al*. (2014) Associations of adherence to the New Nordic Diet with risk of preeclampsia and preterm delivery in the Norwegian Mother and Child Cohort Study (MoBa), *Public Health Nutrition*, 17, pp. 920–7.

Kirwan, J. a., Broadhurst, D. I., Davidson, R. L. and Viant, M. R. (2013) Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow, *Analytical and Bioanalytical Chemistry*, 405, pp. 5147–5157.

Klein, M. S. and Shearer, J. (2016) Metabolomics and Type 2 Diabetes: Translating Basic Research into Clinical Application, *Journal of Diabetes Research*, Article ID 824814, pp. 1–10.

Kolb, H. and Martin, S. (2017) Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes, *BMC Medicine*, 15, p. 131.

De Koning, L., Chiuve, S. E., Fung, T. T., Willett, W. C., Rimm, E. B. and Hu, F. B. (2011) Diet-quality scores and the risk of type 2 diabetes in men, *Diabetes Care*, 34(5), pp. 1150–1156.

Koppe, L. and Poitout, V. (2016) CMPF: A Biomarker for Type 2 Diabetes Mellitus Progression?, *Trends in Endocrinology and Metabolism*, 27, pp. 439–440.

Kouassi Nzoughet, J., Bocca, C., Simard, G., Prunier-Mirebeau, D., Chao De La Barca, J. M., Bonneau, D., Procaccio, V., Prunier, F., Lenaers, G. and Reynier, P. (2017) A Nontargeted UHPLC-HRMS Metabolomics Pipeline for Metabolite Identification: Application to Cardiac Remote Ischemic Preconditioning, *Analytical Chemistry*, 89(3), pp. 2138–2146.

Krawczuk, J. and Lukaszuk, T. (2016) The feature selection bias problem in relation to high-dimensional gene data, *Artificial Intelligence in Medicine*, 66, pp. 63–71.

Kuhl, C., Tautenhahn, R. and Neumann, S. (2014) LC-MS Peak Annotation and Identification with CAMERA, pp. 1–14.

Kundu, S., Aulchenko, Y. S., Van Duijn, C. M. and Janssens, A. C. J. W. (2011) PredictABEL: An R package for the assessment of risk prediction models, *European Journal of Epidemiology*, 26, pp. 261–264.

Kursa, M. B. (2010) Feature Selection with the Boruta Package, *Journal of Statistical Software,* 36.

Kursa, M. B. and Rudnicki, W. R. (2011) The All Relevant Feature Selection using Random Forest.

Kyrø, C., Olsen, A., Landberg, R., Skeie, G., Loft, S., Aman, P., Leenders, M., Dik, V. K., Siersema, P. D., Pischon, T., Christensen, J., Overvad, K., Boutron-Ruault, M.-C., Fagherazzi, G., Cottet, V., *et al.* (2014) Plasma alkylresorcinols, biomarkers of whole-grain wheat and rye intake, and incidence of colorectal cancer., *Journal of the National Cancer Institute*, 106, pp.1-9.

Kyrø, C., Skeie, G., Loft, S., Overvad, K., Christensen, J., Tjønneland, A. and Olsen, A. (2013) Adherence to a healthy Nordic food index is associated with a lower incidence of colorectal cancer in women: The Diet, Cancer and Health cohort study, *British Journal of Nutrition*, 109, pp. 920–927.

Lacoppidan, S., Kyrø, C., Loft, S., Helnæs, A., Christensen, J., Hansen, C., Dahm, C., Overvad, K., Tjønneland, A. and Olsen, A. (2015) Adherence to a Healthy Nordic Food Index Is Associated with a Lower Risk of Type-2 Diabetes—The Danish Diet, Cancer and Health Cohort Study, *Nutrients*, 7, pp. 8633–8644.

Landberg, R., Kamal-Eldin, A., Andersson, A., Vessby, B. and Aman, P. (2008) Alkylresorcinols as biomarkers of whole-grain wheat and rye intake: plasma concentration and intake estimated from dietary records., *The American Journal of Clinical Nutrition*, 87, pp. 832–8.

Landberg, R., Marklund, M., Kamal-Eldin, A. and Åman, P. (2014) An update on alkylresorcinols - Occurrence, bioavailability, bioactivity and utility as biomarkers, *Journal of Functional Foods*, 7, pp. 77–89.

Lankinen, M., Kolehmainen, M., Jääskeläinen, T., Paananen, J., Joukamo, L., Kangas, A. J., Soininen, P., Poutanen, K., Mykkänen, H., Gylling, H., Oreŝiĉ, M., Jauhiainen, M., Ala-Korpela, M., Uusitupa, M. and Schwab, U. (2014) Effects of whole grain, fish and bilberries on serum metabolic profile and lipid transfer protein activities: A randomized trial (Sysdimet), *PLoS ONE*, 9, pp. 1–12.

Lazar, A. G., Romanciuc, F., Socaciu, M. A. and Socaciu, C. (2015) Bioinformatics Tools for Metabolomic Data Processing and Analysis Using Untargeted Liquid Chromatography Coupled With Mass Spectrometry, *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Animal Science and Biotechnologies*, 72, pp. 250–255.

Ley, S. H., Korat, A. V. A., Sun, Q., Tobias, D. K., Zhang, C., Qi, L., Willett, W. C., Manson, J. A. E. and Hu, F. B. (2016) Contribution of the nurses' health studies to uncovering risk factors for type 2 diabetes: diet, lifestyle, biomarkers, and genetics, *American Journal of Public Health*, 106, pp. 1624–1630.

Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest, *R News*, 2, pp. 18–22.

Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., Neumann, S., Trausinger, G., Sinner, F., Pieber, T. and Magnes, C. (2015) IPO: a tool for automated optimization of XCMS parameters, *BMC Bioinformatics*, 16, pp. 1–10.

Lindgren, F., Hansen, B. and Karcher, W. (1996) Model Validation By Permutation Tests :, *Journal of Chemometrics*, 10, pp. 521–532.

Lindström, J. and Tuomilehto, J. (2003) The diabetes risk score: A practical tool to predict type 2 diabetes risk, *Diabetes Care*, 26, pp. 725–731.

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

Loca, L. A., Scott, R. A., Sharp, S. J., Burgess, S., Luan, J., Tillin, T., Schmidt, A. F., Imamura, F., Stewart, I. D., Perry, J. R. B., Marney, L., Koulman, A., Karoly, E. D., Forouhi, N. G., Sjögren, R. J. O., *et al.* (2016) Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis, *PLoS Medicine*, 13, pp. 1–22.

Lu, J., Xie, G., Jia, W. and Jia, W. (2013) Metabolomics in human type 2 diabetes research., *Frontiers of medicine*, 7, pp. 4–13.

Lu, Y., Wang, Y., Ong, C. N., Subramaniam, T., Choi, H. W., Yuan, J. M., Koh, W. P. and Pan, A. (2016) Metabolic signatures and risk of type 2 diabetes in a Chinese population: an untargeted metabolomics study using both LC-MS and GC-MS, *Diabetologia*, 59, pp. 2349–2359.

Luan H (2016) statTarget: Statistical Analysis of Metabolite Profile. R package version 1.4.0.

Läll, K., Mägi, R., Morris, A., Metspalu, A. and Fischer, K. (2017) Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores, *Genetics in Medicine*, 19, pp. 322–329.

Mackenzie, B. R., Almesjö, L. and Hansson, S. (2004) Fish, Fishing, and Pollutant Reduction in the Baltic Sea, *Environmental Science and Technology*, 38, pp. 1970–1976.

Marckmann, P. and Lassen, A. (1995) Biomarkers fish intake, *The American Journal of Clinical Nutrition*, 62, 956-959.

Markgraf, D., Al-Hasani, H. and Lehr, S. (2016) Lipidomics—Reshaping the Analysis and Perception of Type 2 Diabetes, *International Journal of Molecular Sciences*, 17, p. 1841.

Marushka, L., Batal, M., David, W., Schwartz, H., Ing, A., Fediuk, K., Sharp, D., Black, A., Tikhonov, C. and Chan, H. M. (2017) Association between fish consumption, dietary omega-3 fatty acids and persistent organic pollutants intake, and type 2 diabetes in 18 First Nations in Ontario, Canada., *Environmental Research*, 156, pp. 725–737.

Marushka, L., Batal, M., David, W., Schwartz, H., Ing, A., Fediuk, K., Sharp, D., Black, A., Tikhonov, C., Chan, H. M., Schwingshackl, L., Hoffmann, G., Lampousi, A.-M., Knüppel, S., Iqbal, K., *et al.* (2017) The association of whole grain consumption with incident type 2 diabetes: The Women's Health Initiative Observational Study, *BMJ Open*, 104(Xl), p. e005497.

Mcdermott, A. (1997) Regression calibration methodfor correcting measurement error bias in nutritionalepidemiology, *Public Health*, 65.

Meikle, P. J. and Summers, S. A. (2017) Sphingolipids and phospholipids in insulin resistance and related metabolic disorders, *Nature Reviews Endocrinology*, 13, pp. 79–91.

de Mello, V. D., Paananen, J., Lindström, J., Lankinen, M. A., Shi, L., Kuusisto, J., Pihlajamäki, J., Auriola, S., Lehtonen, M., Rolandsson, O., Bergdahl, I. A., Nordin, E., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Landberg, R., *et al.* (2017) Indolepropionic acid and novel lipid metabolites are associated with a lower risk of type 2 diabetes in the Finnish Diabetes Prevention Study, *Scientific Reports*, 7, p. 46337.

Menni, C., Fauman, E., Erte, I., Perry, J. R. B., Kastenmüller, G., Shin, S. Y., Petersen, A. K., Hyde, C., Psatha, M., Ward, K. J., Yuan, W., Milburn, M., Palmer, C. N. a, Frayling, T. M., Trimmer, J., *et al.* (2013) Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach, *Diabetes*, 62, pp. 4270–4276.

Menni, C., Zierer, J., Valdes, A. M. and Spector, T. D. (2017) Mixing omics: combining genetics and metabolomics to study rheumatic diseases, *Nature Reviews Rheumatology*, 13(3), pp. 174–181.

Mizuno, H., Ueda, K., Kobayashi, Y., Tsuyama, N., Todoroki, K., Min, J. Z. and Toyo'oka, T. (2017) The great importance of normalization of LC–MS data for highly-accurate non-targeted metabolomics, *Biomedical Chromatography*, 31: e3864.

Mutie, P. M., Giordano, G. N. and Franks, P. W. (2017) Lifestyle precision medicine: the next generation in type 2 diabetes prevention?, *BMC Medicine*, 15, p. 171.

Myers, O. D., Sumner, S. J., Li, S., Barnes, S. and Du, X. (2017) Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data, *Analytical Chemistry*, 89, pp. 8689–8695.

Nikiforova, V. J., Giesbertz, P., Wiemer, J., Bethan, B., Looser, R., Liebenberg, V., Ruiz Noppinger, P., Daniel, H. and Rein, D. (2014) Glyoxylate, a new marker metabolite of type 2 diabetes, *Journal of Diabetes Research*, 2014.

Nilsson, R., Peña, J. M., Björkegren, J. and Tegnér, J. (2007) Consistent Feature Selection for Pattern Recognition in Polynomial Time, *The Journal of Machine Learning Research*, 8, pp. 589–612.

Norberg, M., Wall, S., Boman, K. and Weinehall, L. (2010) The Västerbotten Intervention Programme: background, design and implications., *Global Health Action*, 3, pp. 1–15.

O'Gorman, A. and Brennan, L. (2015) Metabolomic applications in nutritional research: A perspective, *Journal of the Science of Food and Agriculture*, 95(13), pp. 2567–2570.

Olokoba, A. B., Obateru, O. A. and Olokoba, L. B. (2012) Type 2 diabetes mellitus: A review of current trends, *Oman Medical Journal*, 27, pp. 269–273.

Olsen, A., Egeberg, R., Halkjær, J., Christensen, J., Overvad, K. and Tjønneland, A. (2011) Healthy Aspects of the Nordic Diet Are Related to Lower Total Mortality, *The Journal of Nutrition* , 141, PP. 639–644.

Osonoi, Y., Mita, T., Osonoi, T., Saito, M., Tamasawa, A., Nakayama, S., Someya, Y., Ishida, H., Kanazawa, A., Gosho, M., Fujitani, Y. and Watada, H. (2015) Relationship between dietary patterns and risk factors for cardiovascular disease in patients with type 2 diabetes mellitus: a cross-sectional study, *Nutrition Journal*, 15, p. 15.

Paeratakul, S., Popkin, B. M., Kohlmeier, L., Hertz-Picciotto, I., Guo, X. and Edwards, L. J. (1998) Measurement error in dietary data: implications for the epidemiologic study of the diet-disease relationship., *European Journal of Clinical Nutrition*, 52, pp. 722–727.

Pan, A., Sun, Q. and Bernstein, A. (2011) Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis, *The American Journal of Clinical Nutrition, 94, 1088-1096.*

Panagiotakos, D. B., Pitsavos, C. and Stefanadis, C. (2006) Dietary patterns: A Mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk, *Nutrition, Metabolism and Cardiovascular Diseases*, 16, pp. 559–568.

Parker, E. D., Liu, S., Van Horn, L., Tinker, L. F., Shikany, J. M., Eaton, C. B. and Margolis, K. L. (2013) The association of whole grain consumption with incident type 2 diabetes: The Women's Health Initiative Observational Study, *Annals of Epidemiology*, 23, pp. 321–327.

Peddinti, G., Cobb, J., Yengo, L., Froguel, P., Kravi?, J., Balkau, B., Tuomi, T., Aittokallio, T. and Groop, L. (2017) Early metabolic markers identify potential targets for the prevention of type 2 diabetes, *Diabetologia*, 60:1740–1750.

Pekkinen, J., Rosa-Sibakov, N., Micard, V., Keski-Rahkonen, P., Lehtonen, M., Poutanen, K., Mykkänen, H. and Hanhineva, K. (2015) Amino acid-derived betaines dominate as urinary markers for rye bran intake in mice fed high-fat diet-A nontargeted metabolomics study, *Molecular Nutrition and Food Research*, 59, pp. 1550–1562.

Petersen, B. and Julia, K. (2015) Validation of a two-step quality control approach for a large-scale human urine metabolomic study conducted in seven experimental batches with LC / QTOF-MS, *Bioanalysis*, 7, pp. 103–112.

Playdon, M. C., Moore, S. C., Derkach, A., Reedy, J., Subar, A. F., Sampson, J. N., Albanes, D., Gu, F., Kontto, J., Lassale, C., Liao, L. M., Männistö, S., Mondul, A. M., Weinstein, S. J., Irwin, M. L., *et al.* (2017) Identifying biomarkers of dietary patterns by using metabolomics, *The American Journal of Clinical Nutrition*, 105, pp. 450–465.

Playdon, M. C., Ziegler, R. G., Sampson, J. N., Stolzenberg-Solomon, R., Thompson, H. J., Irwin, M. L., Mayne, S. T., Hoover, R. N. and Moore, S. C. (2017) Nutritional metabolomics and breast cancer risk in a prospective study, *The American Journal of Clinical Nutrition*, 106, pp. 637–649.

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

Podwojski, K., Fritsch, A., Chamrad, D. C., Paul, W., Sitek, B., Stühler, K., Mutzel, P., Stephan, C., Meyer, H. E., Urfer, W., Ickstadt, K. and Rahnenführer, J. (2009) Retention time alignment algorithms for LC/MS data must consider non-linear shifts, *Bioinformatics*, 25, pp. 758–764.

Poulsen, S. K., Due, A., Jordy, A. B., Kiens, B., Stark, K. D., Stender, S., Holst, C., Astrup, A. and Larsen, T. M. (2014) Health effect of the new nordic diet in adults with increased waist circumference: A 6-mo randomized controlled trial, *The American Journal of Clinical Nutrition*, 99, pp. 35–45.

Qiu, G., Zheng, Y., Wang, H., Sun, J., Ma, H., Xiao, Y., Li, Y., Yuan, Y., Yang, H., Li, X., Min, X., Zhang, C., Xu, C., Jiang, Y., Zhang, X., *et al.* (2016) Plasma metabolomics identified novel metabolites associated with risk of type 2 diabetes in two prospective cohorts of Chinese adults., *International Journal of Epidemiology*, 1507-1516.

Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D. and Lu, L. J. (2015) Computational and statistical analysis of metabolomics data, *Metabolomics*, 11, pp. 1492–1513.

Revelle, W. (2017) psych: Procedures for Personality and Psychological Research.

Rochat, B. (2016) From targeted quantification to untargeted metabolomics: Why LC-high-resolution-MS will become a key instrument in clinical labs, *TrAC - Trends in Analytical Chemistry*, 84, pp. 151–164.

Rohart, F., Gautier, B., Singh, A. and Cao, K.-A. Le (2017) mixOmics: an R package for ’omics feature selection and multiple data integration, *bioRxiv*, p. 108597.

Rolandsson, O., Norberg, M., Nyström, L., Söderberg, S., Svensson, M., Lindahl, B. and Weinehall, L. (2012) How to diagnose and classify diabetes in primary health care: Lessons learned from the Diabetes Register in Northern Sweden (DiabNorth), *Scandinavian Journal of Primary Health Care*, 30, pp. 81–87

Rosner B (2005) *Fundamentals of biostatistics*. 6th edn.

Roswall, N., Sandin, S., Löf, M., Skeie, G., Olsen, A., Adami, H. O. and Weiderpass, E. (2015) Adherence to the healthy Nordic food index and total and cause-specific mortality among Swedish women, *European Journal of Epidemiology*, 30(6), pp. 509–517.

Rudnicki, W. R., Wrzesie, M. and Paja, W. (2015) Feature Selection for Data and Pattern Recognition, *Studies in Computational Intelligence* 584.

Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. a. and Hendriks, M. M. W. B. (2014) Reflections on univariate and multivariate analysis of metabolomics data, *Metabolomics*, 10, pp. 361–374.

Sacks, F. M., Obarzanek, E., Windhauser, M. M., Svetkey, L. P., Vollmer, W. M., McCullough, M., Karanja, N., Lin, P. H., Steele, P., Proschan, M. A., Evans, M. A., Appel, L. J., Bray, G. A., Vogt, T. M., Moore, T. J., *et al.* (1995) Rationale and design of the Dietary Approaches to Stop Hypertension trial (DASH). A multicenter controlled-feeding study of dietary patterns to lower blood pressure, *Annals of Epidemiology*, 5, pp. 108–118.

Saeys, Y., Inza, I. and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23, pp. 2507–2517.

Saeys, Y., Inza, I., Larranaga, P., Larrañaga, P., Nielsen, R., Peña, J., Björkegren, J., Tegnér, J., Touw, W., Bayjanov, J., Overmars, L., Backus, L., Boekhorst, J., Wels, M., Hijum, S. van, *et al.* (2014) Robustness of Random Forest-based gene selection methods, *Bioinformatics*, 23, pp. 2507–2517.

Savolainen, O., Lind, M. V., Bergström, G., Fagerberg, B., Sandberg, A.-S. and Ross, A. (2017) Biomarkers of food intake and nutrient status are associated with glucose tolerance status and development of type 2 diabetes in older Swedish women, *The American Journal of Clinical Nutrition*, 31, p. ajcn152850.

Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L. O., Draper, J., Rappaport, S. M., van der Hooft, J. J. and Wishart, D. S. (2014) The food metabolome: a window over dietary exposure, *The American Journal of Clinical Nutrition*, 99, pp. 1286–1308.

Schmid, A. and Blank, L. M. (2010) Systems biology: Hypothesis-driven omics integration., *Nature Chemical Biology*, pp. 485–487.

Schooneman, M. G., Vaz, F. M., Houten, S. M. and Soeters, M. R. (2013) Acylcarnitines: Reflecting or inflicting insulin resistance?, *Diabetes*, 62, pp. 1–8.

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. and McLean, J. A. (2016) Untargeted Metabolomics Strategies—Challenges and Emerging Directions, *Journal of the American Society for Mass Spectrometry*, 27(12), pp. 1897–1905.

Schwingshackl, L., Hoffmann, G., Lampousi, A.-M., Knüppel, S., Iqbal, K., Schwedhelm, C., Bechthold, A., Schlesinger, S. and Boeing, H. (2017) Food groups and risk of type 2 diabetes mellitus: a systematic review and meta-analysis of prospective studies, *European Journal of Epidemiology*, DOI 10.1007/s10654-017-0246-y

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P. and Hollender, J. (2014) Identifying small molecules via high resolution mass spectrometry: Communicating confidence, *Environmental Science and Technology*, 48, pp. 2097–2098.

Silva, V., Barazzoni, R. and Singer, P. (2014) Biomarkers of Fish Oil Omega-3 Polyunsaturated Fatty Acids Intake in Humans, *Nutrition in Clinical Practice*, 29(1), pp. 63–72.

Smilde, A. K., Van Der Werf, M. J., Bijlsma, S., Van Der Werff-Van Der Vat, B. J. C. and Jellema, R. H. (2005) Fusion of mass spectrometry-based metabolomics data, *Analytical Chemistry*, 77, pp. 6729–6736.

Smith, C. a, Want, E. J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification., *Analytical Chemistry*, 78, pp. 779–787.

Smith, R. and Ventura, D. (2013) LC-MS alignment in theory and practice : a comprehensive algorithmic review, *Briefings in Bioinformatics*, 16, 104-117

Spicer, R. A., Salek, R. and Steinbeck, C. (2017) A decade after the metabolomics standards initiative it's time for a revision, *Scientific Data*, 4:170138.

Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. and Steinbeck, C. (2017) Navigating freely-available software tools for metabolomics analysis, *Metabolomics*, 13:106, 1-16.

Streppel, M. T., de Vries, J. H. M., Meijboom, S., Beekman, M., de Craen, A. J. M., Slagboom, P. E. and Feskens, E. J. M. (2013) Relative validity of the food frequency questionnaire used to assess dietary intake in the Leiden Longevity Study., *Nutrition Journal*, 12, pp.1-8.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., *et al.* (2007) Proposed minimum reporting standards for chemical analysis, *Metabolomics*, 3, pp. 211–221.

Sysi-Aho, M., Katajamaa, M., Yetukuri, L. and Orešič, M. (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards, *BMC Bioinformatics*, 8, pp. 1-17.

Szymańska, E., Saccenti, E., Smilde, A. K. and Westerhuis, J. A. (2012) Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies., *Metabolomics : Official journal of the Metabolomic Society*, 8(Suppl 1), pp. 3–16.

Tanaka, H. and Ogishima, S. (2011) Omics-based identification of pathophysiological processes., *Methods in Molecular Biology*, 719, pp. 499–509.

Tautenhahn, R., Böttcher, C. and Neumann, S. (2008) Highly sensitive feature detection for high resolution LC/MS., *BMC bioinformatics*, 9, pp.1-16.

Temelkova-Kurktschiev, T. and Stefanov, T. (2012) Lifestyle and Genetics in Obesity and type 2 Diabetes, *Exp.Clin Endocrinol Diabetes*, 120, pp. 1–6.

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

Therneau, T. M. and Lumley, T. (2017) Survival Analysis Guide.

Thiese, M. S. (2014) Observational and interventional study design types; an overview, *Biochemia Medica*, 24, pp. 199–210.

Tirosh, A., Shai, I. and Bitzur, R. (2008) Changes in triglyceride levels over time and risk of type 2 diabetes in young men, *Diabetes Care*, 31, pp. 2032–2037.

Trifonova, O., Lokhov, P. and Archakov, A. (2013) Postgenomics diagnostics: metabolomics approaches to human blood profiling., *Omics : A Journal of Integrative Biology*, 17, pp. 550–9.

Uusitupa, M., Hermansen, K., Savolainen, M. J., Schwab, U., Kolehmainen, M., Brader, L., Mortensen, L. S., Cloetens, L., Johansson-Persson, A., Önning, G., Landin-Olsson, M., Herzig, K. H., Hukkanen, J., Rosqvist, F., Iggman, D., *et al.* (2013) Effects of an isocaloric healthy Nordic diet on insulin sensitivity, lipid profile and inflammation markers in metabolic syndrome - a randomized study (SYSDIET), *Journal of Internal Medicine*, 274, pp. 52–66.

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J. and Yanes, O. (2012) A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data, *Metabolites*, 2, pp. 775–795.

Van Dam, R. M. and Hunter, D. (2012) Biochemical Indicators of Dietary Intake, *Nutritional Epidemiology*, Oxforf University Press.

Warrack, B. M., Hnatyshyn, S., Ott, K. H., Reily, M. D., Sanders, M., Zhang, H. and Drexler, D. M. (2009) Normalization strategies for metabonomic analysis of urine samples, *Journal of Chromatography B*, 877, pp. 547–552.

WHO (2006) Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia, Report of A WHO/IDF Consultation.

Wishart, D. S. (2016) Emerging applications of metabolomics in drug discovery and precision medicine, *Nature Reviews Drug Discovery*, 15, pp. 473–484.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M. A., Forsythe, I., *et al.* (2007) HMDB: The human metabolome database, *Nucleic Acids Research*, 35, pp. 521–526.

World Health Organization (2016) Global Report on Diabetes.

Xi, B., Gu, H., Baniasadi, H. and Raftery, D. (2014) Statistical Analysis and Modeling of Mass Spectrometry-Based Metabolomics Data, in *Molecular Analysis and Genome Discovery*, pp. 333–353.

Yao, Q., Xu, Y., Yang, H., Shang, D., Zhang, C., Zhang, Y., Sun, Z., Shi, X., Feng, L., Han, J., Su, F., Li, C. and Li, X. (2015) Global Prioritization of Disease Candidate Metabolites Based on a Multi-omics Composite Network, *Scientific Reports*, 5,17201.

Yengo, L., Arredouani, A., Marre, M., Roussel, R., Vaxillaire, M., Falchi, M., Haoudi, A., Tichet, J., Balkau, B., Bonnefond, A. and Froguel, P. (2016) Impact of statistical models on the prediction of type 2 diabetes using non-targeted metabolomics profiling, *Molecular Metabolism*, 5, pp. 918–925.

Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S. and Liang, Y. (2016) Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, *Analytica Chimica Acta*, 914, pp. 17–34.

Yin, P. and Xu, G. (2014) Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications, *Journal of Chromatography A*, 1374, pp. 1–13.

Zhang, A.-H., Qiu, S., Xu, H.-Y., Sun, H. and Wang, X.-J. (2013) Metabolomics in diabetes., *Clinica Chimica Acta*, 429, pp. 106–110.

Zhang, A. H., Sun, H., Yan, G. L., Yuan, Y., Han, Y. and Wang, X. J. (2014) Metabolomics study of type 2 diabetes using ultra-performance LC-ESI/quadrupole-TOF high-definition MS coupled with pattern recognition methods, *Journal of Physiology and Biochemistry*, 70, pp. 117–128.

Zhao, J., Zhu, Y., Hyun, N., Zeng, D., Uppal, K., Tran, V. T., Yu, T., Jones, D., He, J., Lee, E. T. and Howard, B. V. (2015) Novel Metabolic Markers for the Risk of Diabetes Development in American Indians, *Diabetes Care*, 38, pp. 220–227.

Zheng, Y., Yu, B., Alexander, D., Steffen, L. M. and Boerwinkle, E. (2014) Human metabolome associates with dietary intake habits among African Americans in the atherosclerosis risk in communities study, *American Journal of Epidemiology*, 179, pp. 1424–1433.

Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., Bixby, H., Cowan, M. J., Ali, M. K., Taddei, C., Lo, W. C., Reis-Santos, B., Stevens, G. A., Riley, L. M., Miranda, J. J., *et al.* (2016) Worldwide trends in diabetes since 1980: A pooled analysis of 751 population-based studies with 4.4 million participants, *The Lancet*, 387, pp. 1513–1530.

Zhou, H., Li, Y., Liu, X., Xu, F., Li, L., Yang, K., Qian, X., Liu, R., Bie, R. and Wang, C. (2017) Development and evaluation of a risk score for type 2 diabetes mellitus among middle-aged Chinese rural population based on the RuralDiab Study, *Scientific Reports*, 7,42685.

Zhu, Z.-J., Schultz, A. W., Wang, J., Johnson, C. H., Yannone, S. M., Patti, G. J. and Siuzdak, G. (2013) Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database., *Nature Protocols*, 8, pp. 451–460.

Acknowledgements

I am deeply grateful to all the people that have helped me and supported me during the past four years along the wonderful journey.

Rikard Landberg--- Thank you very much for the fantastic opportunities, great supervision, all the support and help that you have given me. Thank you also for the always constructive and speedy feedback which guide my progression on lab work, statistics and scientific writing. I really appreciate your patience, encouragements and understanding during the past 4 years. I am amazed by your enthusiasm and attitude to the research. You embody the role model of a researcher I always aspire to.

Carl Brunius---I sincerely express my appreciation to you for the awesome Lessons of Statistics and Life that you have taught me. I cannot imagine how shall I deal with such amount of data without your supervision and encouragements. I really enjoy the time when we have discussions and the way you express ideas: ‘paper and pen’! Will never forget that we wrote down the smart idea regarding metabolites annotation on the napkin paper in the airport.

Kati Hanhineva---Thank you very much for sharing your valuable knowledge of metabolites identification with me. It is a such complex task and I really have learnt a lot from you. I had great time in Kuopio and I would not forget your beautiful garden and lovely sheep! I am very amazed by your high work efficiency and quality!

Ali Moazzami--- I would like to show my profound gratitude for the all the knowledge about NMR metabolomics that I've gained from you. This was my first experience in metabolomics field. Your guidance and advice are very helpful and make the study progress smoothly. We may have different ideas but I really enjoy all the discussions between us.

I sincerely thank all my collaborators and co-authors, Johan A Westerhuis, Ingvar A. Bergdahl, Olov Rolandsson, Ingegerd Johansson, Bernt Lindahl,Johan Rosén, Marko Lehtonen, Seppo Auriola, for huge effort and important contribution to the work involved in the thesis.

Big thanks to my wonderful officemates in Uppsala: Izabela Biskup---Thank you so much for being my 'Iza mom'. You always encourage me, and make me keep calm and carry on when I am depressed and crying. I love our 'Diabetes prevention program' (Time for Polish Vodka, lemon flavor please)! I thank you for all that you have done for me. 'Happiness is letting go of the negative stuff'. José Luis Vázquez ---Many thanks to for being my office mate and most importantly, Spanish Sangria! You always call me 'little princess' and you are definitely my super dancing King!! I also thank Alex Dragovic for the paintings and my wonderful portrait. Roksana Wierzbicka---Thank for your encouragement and the wonderful time we have in the office and in the swimming pool. Souad Abou Shameh ---Thank you for all nice chats, laughter and inspiration. Per Åman--- I have huge respect to your brilliant work and your attitude to the research. Your words mean a lot to me.

Bingjie Li, Jing Lu, and Fengping Yang---Thank you all for being so nice to me. You laugh with me when we share happiness and always inspire me during a difficult time when I need words of encouragement. What will always be unforgettable are the dinner time, awesome trips we enjoyed together, and our laughter.

I want to thank all the people at the former Department of Food Science, SLU, Uppsala for the welcoming and supportive working environment. The particular thanks goes to Lena Dimberg, Janicka Nilsson, Roger Andersson, Galia Zamaratskaia, Jana Pickova for all the help you have provided to me along the way. Many thanks to Margaretha Wijkström, Stina Stiernström, Karin Vikström, Jaana Evander, and Camilla Löfgren for all the administrative support. Huaxing Wu, Anne Kirstine Eriksen, Kia Noehr Iversen, Adila Omar, Isabella Lee, Ken Cheng, Xue Zhao, Daniel Johansson, Susanne Djurle, Maria Karlsson, I really enjoy working with you in the big Food Science family.

I would like to show my sincere gratitude to all the people at the Food and Nutrition Science, Chalmers University of Technology, Gothenburg, for the nice working environment, Fika, and after work activities. Particular thanks to Alastair Ross, Otto Savolainen and Lillie Cavonius for the kind help and support in research and to Linnea Qvirist, Niklas Engström, Nuria

Majaliwa and Serafina Vilanculos for the warm and lovely office environment.

Warm thanks to Sangeet lamichhaney, Durga Chapagain Lamichhaney, Aahana Shrestha, Merina Shrestha, and to Chinese friends I met in Sweden for hot-pot, BBQ, dumplings: Ci Ren, Ying Ye, Xing Liu, Xu Sun, Lingzhao Fang, Shizhi Wang, Tian Xia, Hanqian Zhang, Bing Zhu, Shengjie Li, Ming Zhao, Xinmei Feng, Qing Liang, Li Sun, Siming Huang, Tong Liu, Xin Zeng, Pengcheng Wang, Wenbo Dong, Jun Yi, Hua Jiang, Mingliang Fei, Qinsong Liu, Yanjun Zan, Yan Ji, Na Xiu, Fei Huang, Chengjuan Liu, Mingyi Li……

I am also grateful to my **'Metabolomics'** friends I met in China, Sweden, UK, Ireland and Australia. We encourage each other, study and make progress together. You make my life rich and colorful.

The China Scholarship Council, Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and the Swedish Research Council—Medicine, Lantmännen research foundation, Swedish Research Council for Health, Working Life and Welfare are very much acknowledged for financial support of projects.

I want to thank my beloved family for all the love, support and care. To my dear mom and grandma, my heart just keeps thanking you for everything you have done for me. Showing my sincere gratitude to my darling, Tao Ni--- Your love and support makes me brave. You have wiped my tears, fought with me over stupid things that we now laugh about, grew with me, and have taught me how to love. The distance, 6,834 km between Xi' an and Sweden dose not dilute our love, but makes the heart grow fonder. Without you, I would not be where I am today.

Lastly, I embrace my flaws and weaknesses and thank my attitudes and emotions.

----不要让未来的我，讨厌现在的自己.