Automated Quantification of Plasma Metabolites by NMR to Study Prostate Cancer Risk Biomarkers

Hanna E. Röhnisch
Faculty of Natural Resources and Agricultural Sciences
Department of Molecular Sciences
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2019

Acta Universitatis agriculturae Sueciae
2019:2

ISSN 1652-6880
ISBN (print version) 978-91-7760-322-1
ISBN (electronic version) 978-91-7760-323-8

Print: SLU Service/Repro, Uppsala 2019

Automated Quantification of Plasma Metabolites by NMR to Study Prostate Cancer Risk Biomarkers

Abstract

Targeted quantitative NMR-based metabolomics can be used to identify disease risk biomarkers. However, NMR-based metabolomics yields complex spectra with signals from many different metabolites. These signals are often located within close proximity and, therefore, signal interferences are observed. Such interferences must be accounted for to yield accurate metabolite concentrations. Quantifications can become very time-consuming, especially in large-scale studies.

In response to this, an Automated Quantification Algorithm (AQuA) was designed as the final step of the NMR-based metabolomics workflow. Implementation and evaluation was done for quantification of human plasma metabolites in samples collected using heparin as anticoagulant. AQuA enabled the quantification of 67 metabolites in 1342 samples within one second on a standard personal computer. AQuA performed with equal accuracy as a manual procedure for targeted profiling performed using a software package dedicated to metabolite quantification by NMR. In contrast to using heparin as anticoagulant, the use of EDTA introduced additional interferences. With some modifications, AQuA also quantified human plasma metabolites despite the presence of the high intensity signals from EDTA, some of which displayed inter-spectral deviations in signal positions and line widths.

To further demonstrate its usefulness, AQuA was utilised for risk biomarker discovery in a case-control study nested within the Northern Sweden Health and Disease Cohort. Plasma metabolites were quantified in samples from 1554 men, 777 whom were diagnosed with prostate cancer more than 5 years after sample collection (baseline), and 777 whom were matched controls. MS-based metabolomics was also employed to yield complementary information. Conditional logistic regression and correction for multiple testing were performed. Risk biomarkers for prostate cancer varied with baseline age and disease aggressiveness. For example, glycine and pyruvic acid were identified in younger subjects, while lipid species (e.g., lysophosphatidylcholines) associated with overall disease risk in older subjects and with risk of aggressive disease. A reverse cross-association could also be identified between risk of prostate cancer and type 2 diabetes at the metabolite level.

Keywords: Automated Quantification Algorithm (AQuA), mass spectrometry, Northern Sweden Health and Disease Cohort, nuclear magnetic resonance, prostate cancer, risk biomarkers, targeted metabolomics, type 2 diabetes

Author's address: Hanna E. Röhnisch, SLU, Department of Molecular Sciences, P.O. Box 7015, 750 07 Uppsala, Sweden

Dedication

To my family

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 Hanna E. Röhnisch*, Jan Eriksson, Elisabeth Müllner, Peter Agback, Corine Sandström, Ali A. Moazzami* (2018). AQuA: An Automated Quantification Algorithm for High-Throughput NMR-Based Metabolomics and Its Application in Human Plasma. *Analytical Chemistry*, 90(3), pp. 2095-2102.

II Hanna E. Röhnisch, Jan Eriksson, Lan Vi Tran, Elisabeth Müllner, Corine Sandström, Ali A. Moazzami. An Improved Automated Quantification Algorithm (AQuA) and Its Application to NMR-Based Metabolomics of EDTA-Containing Plasma (manuscript).

III Hanna E. Röhnisch, Cecilie Kyrø, Anja Olsen, Göran Hallmans, Ali A. Moazzami. Identification of Metabolites Associated with Prostate Cancer Risk. A Nested Case-Control Study in the Northern Sweden Health and Disease Cohort (manuscript).

* Corresponding author(s).

The contribution of Hanna E. Röhnisch to the papers included in this thesis was as follows:

I Performing the NMR-based metabolomics analyses
Design, implementation and evaluation of AQuA
Production of figures and writing the manuscript
Construction and compilation of the supporting information (SI)

II Design, implementation and evaluation of the modified AQuA
Production of figures and writing the manuscript
Construction and compilation of SI

III Performing the NMR-based metabolomics analyses
Performing the statistical analyses
Production of figures and writing the manuscript
Constructing and compilation of SI

List of tables

List of figures

Abbreviations

1D	One-dimensional
^{1}H	Proton
2D	Two-dimensional
ABTC	Alpha-Tocopherol, Beta-Carotene cancer prevention study
AICR	American Institute for Cancer Research
AQuA	Automated Quantification Algorithm
BMI	Body mass index
CI	Confidence interval
CPMG	Carr-Purcell-Meiboom-Gill
CSF	Cerebrospinal fluid
CV	Coefficient of variation
DFTJ	DongFeng-TongJi cohort
DSA	4,4-Dimethyl-4-silapentane-1-ammonium trifluoroacetate
DSS	Sodium 4,4-dimethyl-4-silapentane-1-sulfonate
EDTA	Ethylenediamine-tetra-acetic acid
EPIC	European Prospective Investigation into Cancer and Nutrition study
FDA	Food and Drug Administration
FDR	False discover rate
FFQ	Food-frequency questionnaire
FID	Free induction decay
FN	False negative
FP	False positive
F_q	(related to) Degree of interference
FSH	Framingham Heart Study
FWER	Family-wise error rate
FWHM	Full width at half maximum
GWAS	Genome-wide association studies
HMDB	Human metabolome database

IDF	International Diabetes Federation
IFG	Impaired fasting glucose
IGF-I	Insulin-like growth factor-I
IGT	Impaired glucose tolerance
J	Scalar coupling constant
JSNCD	JiangSu Non-Communicable Disease cohort
LC	Liquid chromatography
LOD	Limit of detection
LPC	Lysophosphatidylcholine
MDCS	Malmö Diet and Cancer Study
MRD	Mean relative deviation
MRM	Multiple reaction monitoring
MAD	Median absolute deviation
MS	Mass spectrometry
MWAS	Metabolome-wide association studies
m/z	Mass-to-charge ratio
NGT	Normal glucose tolerance
NIH	National Institutes of Health
NMR	Nuclear magnetic resonance
NSHDC	Northern Sweden Health and Disease Cohort
OGTT	Oral glucose tolerance test
OPLS	Orthogonal projections to latent structures
OR	Odds ratio
PCA	Principal component analysis
PC aa	Diacyl phosphatidylcholine
PC ae	Acyl-alkyl phosphatidylcholine
PLCO	Prostate, Lung, Colorectal, and Ovarian cancer screening trial
PREDIMED	PREvención con DIeta MEDiterránea trial
PLS	Partial least squares projection to latent structures
QC	Quality control
QQQ	Triple quadrupole
r^2	Goodness-of-fit
SM	Sphingomyelin
SM-OH	Hydroxysphingomyelin
S/N	Signal-to-noise
SI	Supporting information
T2D	Type 2 diabetes
TN	True negative
TP	True positive
TSP	Sodium 3-trimethylsilyl propionate

WCRF	World Cancer Research Fund
WHO	World Health Organization
δ	Chemical shift
Δ	Interference

1 Introduction

1.1 General introduction to metabolomics

1.1.1 Metabolomics

A traditional approach in molecular biology has been to select one or a few molecules for examination based on an existing hypothesis. However, recent advances in technology have resulted in evolvement towards global approaches, such as genomics, transcriptomics, proteomics and metabolomics. Global approaches take advantage of available analytical methods to facilitate more comprehensive and holistic examinations of the molecular content in biological samples (Elliott *et al.*, 2016; Reinhold, 2015; German *et al.*, 2005). Metabolomics specifically aim to examine molecules of low weight (<1500 Da). These molecules are called metabolites and together they build up the metabolome (Wishart *et al.*, 2007; Hollywood *et al.*, 2006). Metabolites operate on a level close to function, downstream of genes, transcripts and proteins. Hence, the metabolome can be viewed as a functional read-out of gene-environmental interactions (Figure 1) (Hollywood *et al.*, 2006; German *et al.*, 2005; Goodacre, 2005).

The Human Metabolome Database (HMDB) is an open source database that compiles information on human metabolites, including their quantities and disease-related properties. Thousands of endogenous metabolites have been identified and quantified in different biological samples such as blood (serum or plasma), urine, cerebrospinal fluid (CSF) and saliva (Wishart *et al.*, 2018; Wishart *et al.*, 2007).

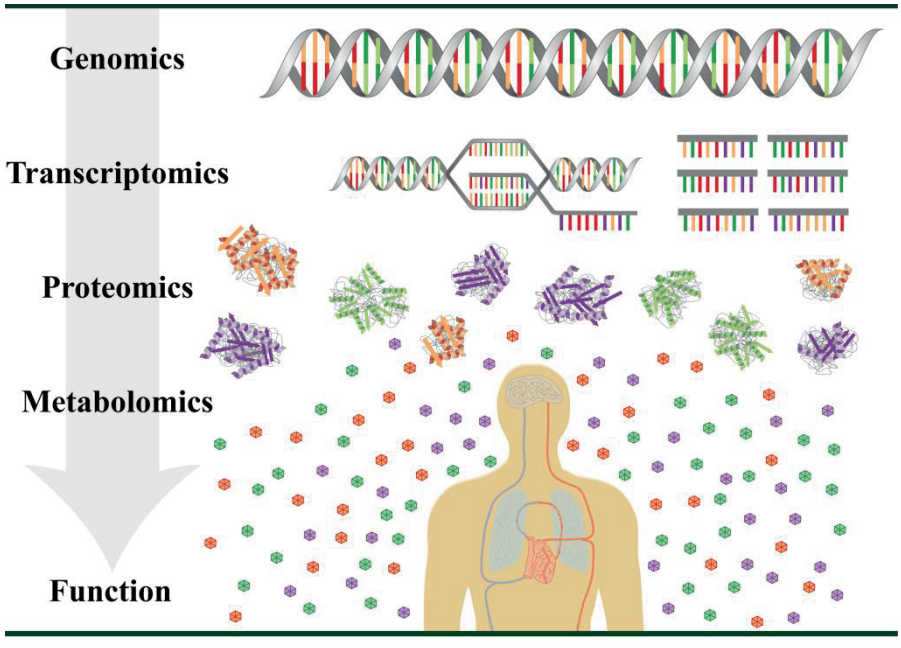

Figure 1. The different omics sciences and their relation to function.

1.1.2 Analytical methods

Metabolomics analysis can be employed using different analytical methods. Two commonly used methods in metabolomics are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) (Dunn & Ellis, 2015).

NMR spectroscopy utilises the quantum mechanical properties (spin) of certain atomic nuclei to study their behaviour in a magnetic field. In a static magnetic field, the spins of such nuclei will distribute to minimise the energy. This energy equilibrium can be disturbed by applying a radio frequency pulse. However, equilibrium is restored with time. The return to equilibrium can be recorded and mathematically transformed into an NMR spectrum. The spectrum generated via NMR analysis of a liquid solution with different molecules holds information related to their identity and quantity (Nagana Gowda *et al.*, 2017; Bharti & Roy, 2012; Claridge, 1999).

In MS, molecules are first converted into ions in an ion source. These ions, or fragments thereof, are then analysed with regard to their mass-to-charge ratio (m/z) in a mass analyser. Tandem MS can be used for controlled fragmentation of ions. The triple quadrupole (QQQ) is a mass analyser that is particularly useful in this context, since it can operate in different modes e.g., to detect (with high sensitivity) a set of target molecules. The aforementioned mode is called multiple reaction monitoring (MRM). Although a solution that

contains different molecules can be directly injected to the ion source, hyphenation with chromatographic systems, such as liquid chromatography (LC) can be used to separate the molecules prior to MS-analysis (Pitt, 2009).

NMR and MS suffer from different limitations, but they also display different advantages. For example, NMR is highly reproducible and quantitative in nature. However, NMR suffers from sensitivity issues. Therefore, a relatively large sample volume is typically required for the analysis. In contrast, MS display better sensitivity, but relatively poor reproducibility (Wishart, 2016).

The combined use of more than one analytical method increases the number of metabolites that can be analysed. This is since different analytical methods yield complementary information. Both NMR and MS have been used to study the metabolite content in human CSF, serum and urine. The use of isotopically labelled standards followed by direct flow injection tandem MS with MRM allowed for quantification of different lipid species (e.g., acylcarnitines, sphingomyelins and glycerophospholipids) in both urine and serum. Hyphenation with LC also allowed for detection of amino acids and biogenic amines. The use of LCMS was limited in detecting metabolites in CSF. The use of a single internal standard facilitated quantification by NMR of different amino acids, organic acids and sugars in CSF, serum and urine (Bouatra *et al.*, 2013; Psychogios *et al.*, 2011; Wishart *et al.*, 2008).

Although different biological fluids may be analysed with a given method, many metabolomics studies focus on one type of biofluid (Shi *et al.*, 2018; Ruiz-Canela *et al.*, 2018; Schmidt *et al.*, 2017; Huang *et al.*, 2016; Qui *et al.*, 2016; Kühn *et al.*, 2016; Drogan *et al.*, 2015; Mondul *et al.*, 2015; Mondul *et al.*, 2014; Wang *et al.*, 2011). The analysis of human blood metabolites is very common, which is reflected in statistics from the HMDB that reveal a superior number of detected and quantified metabolites in blood samples (serum and plasma) compared to e.g., CSF, urine and saliva (http://www.hmdb.ca/statistics).

1.1.3 Targeted and untargeted metabolomics

Metabolomics analysis can be performed in a targeted or untargeted fashion (Figure 2). Untargeted analysis aims to examine as many metabolites as possible, including previously unknown metabolites, while targeted analysis aim to quantify a set of pre-selected metabolites (Gorrochateugui *et al.*, 2016). The workflow of targeted analysis typically includes sample preparation, data collection, processing of data, metabolite identification and quantification. Hence, the data generated includes a list of metabolite concentrations in

different samples, which is used in the statistical analysis. The untargeted workflow typically includes sample preparation, data collection and data processing. Processing generates data other than absolute concentrations (e.g., signal areas or intensities) for different metabolite features in different samples. The workflow is often followed by metabolite identification steps. However, identification can be a key bottleneck in the untargeted workflow and may therefore be limited to specific features that displayed a promising statistical outcome. The untargeted workflow is explorative but not fully quantitative, while the targeted workflow is less explorative but quantitative (Matsuda, 2016; Schrimpe-Rutledge *et al.*, 2016; Alonso *et al.*, 2015).

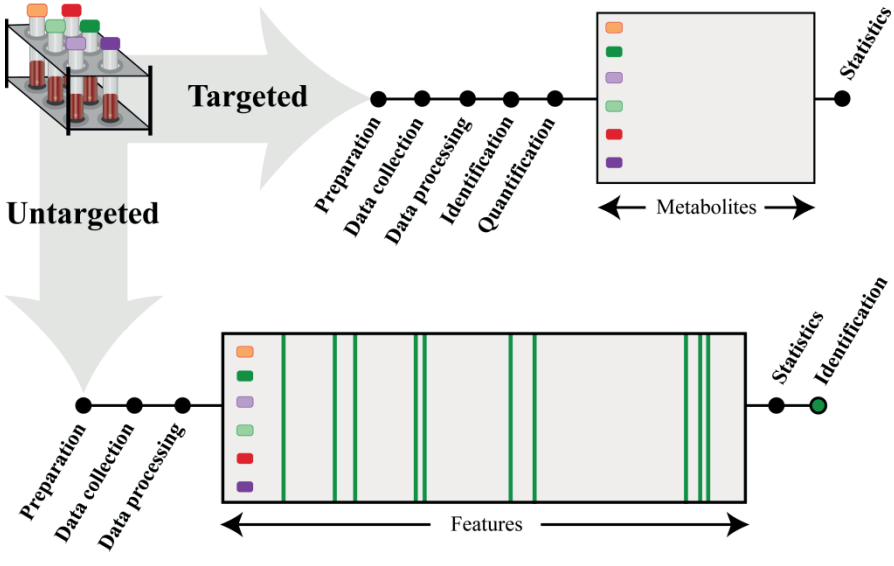

Figure 2. General workflows for targeted and untargeted metabolomics.

Although metabolomics yields much information, the generation of data is challenging. This often owes to the extensive and sometimes time-consuming processing steps required to convert raw data into data suitable for further statistical analysis (Bingol, 2018; Matsuda, 2016; Schrimpe-Rutledge *et al.*, 2016). NMR-based metabolomics of human plasma and serum can serve as a clarifying example. Close to seventy human blood metabolites have been identified by NMR. Unfortunately, NMR signals from different metabolites are sometimes difficult to distinguish from each other. This results in signal interferences, which hamper accurate quantification (Nagana Gowda *et al.*, 2015; Weljie *et al.*, 2006). Data for statistical analysis can be generated in a highly rapid manner when interferences are ignored. Each NMR spectrum can be divided into smaller units (buckets) and the intensities in different buckets

can be extracted and used in the statistical analysis. This untargeted workflow is called metabolic fingerprinting (Worley & Powers, 2014). In contrast, the processing time increases when interferences are accounted for to yield absolute concentrations (Weljie *et al.*, 2006). In order to combine the concept of high-throughput with accurate quantification, substantial effort has been done to increased level of automation of previously manual procedures that account for interferences. Still, further development is warranted to increase the throughput of metabolite quantification by NMR (Bingol, 2018).

1.1.4 Statistics in metabolomics

Statistical analysis, an essential step in metabolomics studies, can be done via univariate or multivariate methods (Alonso *et al.*, 2015). A key aspect in univariate statistics is hypothesis testing, where test statistics is performed and a pre-decided cut-off value for significance determines whether or not the null hypothesis can be rejected. Univariate statistics is easy to employ and the results are highly interpretable. The use of univariate statistic in metabolomics has some important implications. For example, several metabolites will be subjected to the same statistical test and this multiple testing will increase the rate of false positive results (Vinaixa *et al.*, 2012; Morshed *et al.*, 2009). This is problematic since the commonly used cut-off for statistical significance ($p < 0.05$) has been criticised for being prone to yield false positive outcomes to begin with (Benjamin *et al.*, 2018). Yet, multiple testing can be controlled for using different correction procedures (Alonso *et al.*, 2015).

Correction from multiple testing

It is complicated to assess the overall error rate in multiple testing scenarios since there is a risk for both false negatives and false positives in each individual test (Broadhurst & Kell, 2006; Storey, 2002). Bonferroni correction is the standard approach, which accounts for the family-wise error rate (FWER) – i.e., the probability to yield one or several false positive in a series of multiple hypotheses tested. In this approach, the α-level (0.05) is simply divided by the number of variables tested. This approach assumes all metabolites as independent variables (although many metabolites or features are inter-correlated). Bonferroni correction has been criticised for being too conservative and thereby increasing the rate of false negative results (Broadhurst & Kell, 2006; Storey, 2002). Rather than using the FWER to control for multiple testing, less stringent approaches may be applied. One example is the false discovery rate (FDR) approach (Storey & Tibshirani, 2003; Benjamini & Hochberg, 1995). The FDR is a quantity of the expected

proportion of false positives among rejected hypotheses. The FDR approach is a powerful yet more liberal approach compared to Bonferroni correction (Storey, 2002).

Multivariate statistics

Different metabolites (or metabolite features) can display high inter-correlations. In contrast to univariate statistics, where all variables are treated as independent variables, multivariate methods offer an alternative approach that considers inter-correlation between different variables (Alonso *et al.*, 2015). Multivariate statistical analysis can retain the global nature of metabolomics experiments to distinguish between different outcomes/groups on a metabolome level (Worley & Powers, 2013). Although there are many multivariate methods, they can broadly be divided into two categories, namely unsupervised or supervised methods.

Principle component analysis (PCA) is an unsupervised method. In PCA, the data from metabolomics analyses, X, is not related to any response variable(s), Y. The full dimensionality of X is simply reduced to a fewer number of principal components that capture important variation in the dataset (Greene *et al.*, 2014).

Partial least squares projection to latent structures (PLS) is a supervised method for multivariate regression analysis, where X is related to Y (Wold *et al.*, 2001). PLS can also be extended in various directions. Discriminant analysis takes into account a categorical response variable, while a multilevel direction takes into account paired data (Westerhuis *et al.*, 2010). Orthogonal projections to latent structures (OPLS) aims to separate the variation in X that is explained by Y, from the variation in X that cannot be explained by Y (Trygg & Wold, 2002). OPLS can also be extended in various directions (Jonsson *et al.*, 2015; Westerhuis *et al.*, 2010).

1.1.5 Application areas

Metabolomics has a wide range of application areas in molecular biology (Putri *et al.*, 2013). One example is the field of biomarker discovery where the use of metabolomics heavily relies upon its integration with observational study designs. This field of research has been termed molecular epidemiology and typically aim to establish association between exposure (metabolites) and outcome (disease) on a population-based level (García-Closas *et al.*, 2011; Hendriks *et al.*, 2011). Although the definition of a biomarker varies in existing literature, it has been broadly defined as a measurement of a normal or pathophysiological process, or a response to an intervention or exposure.

Biomarkers can be classified into different subtypes. Diagnostic biomarkers detect the presence of a disease, while risk biomarkers reflect an increased risk of disease development before its clinical manifestation (Califf, 2018; FDA-NIH Biomarker Working Group, 2016).

One example of molecular epidemiology applied for biomarker discovery are metabolome-wide association studies (MWAS), which collectively aim to investigate the association between the metabolic phenotype (metabotype) and disease risk (Nicholson *et al.*, 2008). These studies adopt the same concept as the earlier genome-wide association studies (GWAS) that aim to investigate association of genotype with disease risk. Weak association between genetic variation and disease risk were typically reported in early studies (Visscher *et al.*, 2012). Scientists have argued that the concept of MWAS is inherently more fruitful, since the development of many diseases depends on complex gene-environmental interactions, something that cannot be captured by GWAS (Nicholson *et al.*, 2011; Nicholson *et al.*, 2008).

Many clinical and epidemiological studies collect human blood samples and store them (as aliquots of plasma or serum) for future research purposes (Vaught *et al.*, 2009). Such sample repositories are highly suitable for biomarker discovery, especially since the level of different human blood metabolites are regulated by variation of environmental and genetic factors (Liu, 2014; Nicholson *et al.*, 2011).

The concept of MWAS has already been implemented to identify risk biomarkers for different diseases such as prostate cancer and type 2 diabetes (T2D) (Table 1).

1.1.6 Scope of thesis work

The scope of this thesis work can be divided into two parts. The first part regards methodological development to improve the throughput of metabolite quantification by NMR in human plasma samples. The second part regards integration of high-throughput workflows for targeted metabolomics into large-scale molecular epidemiology to identify disease risk biomarkers. Below follows a more detailed introduction to each respective part.

Table 1. *Example of studies employing the concept of MWAS* [a, b]

Reference	*Disease*	*Cohort*	*Samples*	*Method*	*Approach*
Shi *et al.*, 2018	T2D	NSHDC	Plasma	MS	Untargeted
Ruiz-Canela *et al.*, 2018	T2D	PREDIMED	Plasma	MS	Untargeted
Schmidt *et al.*, 2017	Prostate cancer	EPIC	Plasma	MS	Targeted
Huang *et al.*, 2016	Prostate cancer	PLCO	Serum	MS	Untargeted
Qui *et al.*, 2016	T2D	DFTJ & JSNCD	Plasma	MS	Targeted
Kühn *et al.*, 2016	Prostate cancer	EPIC Heidelberg	Plasma	MS	Targeted
Drogan *et al.*, 2015	T2D	EPIC Potsdam	Serum	MS	Untargeted
Mondul *et al.*, 2015	Prostate cancer	ABTC	Serum	MS	Untargeted
Mondul *et al.*, 2014	Prostate cancer	ABTC	Serum	MS	Untargeted
Wang *et al.*, 2011	T2D	FHS & MDCS	Plasma	MS	Targeted

[a] Cohorts: Alpha-Tocopherol, Beta-Carotene (ATBC) cancer prevention study; DongFeng-TongJi (DFTJ) cohort; European Prospective Investigation (EPIC) into cancer and nutrition study; Framingham Heart Study (FHS); JiangSu Non-Communicable Disease (JSNCD) cohort; Malmö Diet and Cancer Study (MDCS); Northern Sweden Health and Disease Cohort (NSHDC); PREvención con DIeta MEDiterránea (PREDIMED) trial; Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial.
[b] All studies employed the same epidemiological study design.

1.2 Targeted NMR-based metabolomics of human plasma

1.2.1 Basic theory of one dimensional proton NMR spectroscopy

The proton possesses a nuclear spin – i.e., its nuclei rotates around its own axis and it displays a magnetic moment. When applying a static magnetic field on a magnetic moment, it circulates about the applied field with a Larmor precession. The Larmor frequency is the rate of this precession and it describes the resonance frequency for each given nuclear species. Since two spin states exist for the proton nuclei, its magnetic moments may align parallel or anti-parallel to the applied magnetic field. A population of spins in a static magnetic field ($\overline{\mathbf{B}}_{\mathbf{0}}$) will distribute over the two spin states so there is an excess in the spin state of lowest energy (α). This results in a net magnetisation ($\overline{\mathbf{M}}_{\mathbf{0}}$) aligned in parallel to the magnetic field (Figure 3A). NMR occurs when the magnetic

moments alter spin states. This is achieved during NMR experiment by applying a radiofrequency pulse ($\bar{\mathbf{B}}_1$) (Figure 3B). To induce a change in spin states, the frequency of the applied pulse must match with the Larmor frequency of the spins. Then, the net magnetisation is perturbed ($\bar{\mathbf{M}}$) so the population difference is equalised and the spins possess phase coherence (Figure 3C). The effect of applying a radio frequency pulse disappears over time as the net magnetisation returns to equilibrium (Figure 3D). The disappearance of phase coherence over time can be measured as an exponentially decreasing signal, a free induction decay (FID) (Claridge, 1999). The FID is typically recorded by repeatedly applying the radiofrequency pulse after the net magnetisation has returned to equilibrium (i.e., multiple scans are recorded).

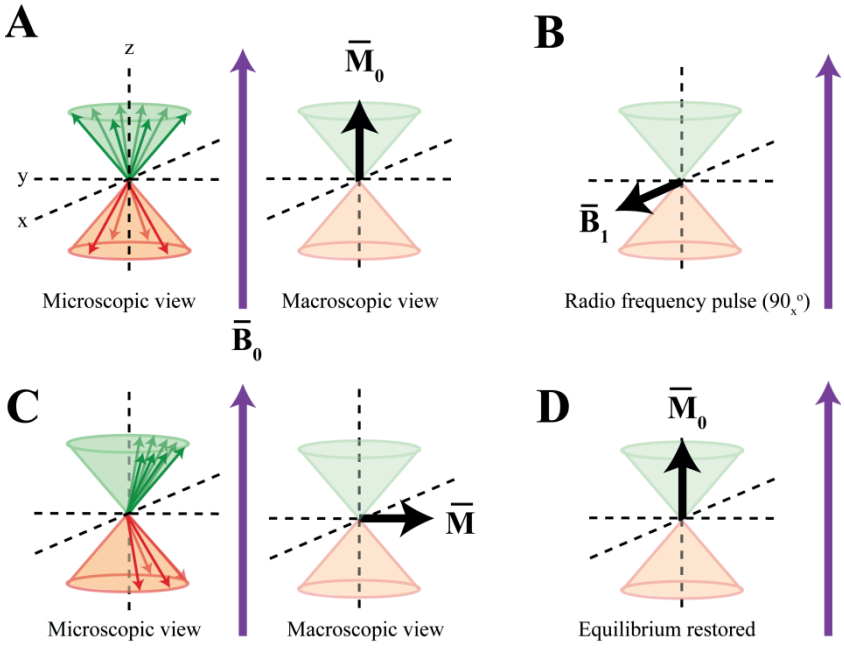

Figure 3. Illustration of events occurring during a one dimensional proton NMR experiment. (A) Distribution of spinning nuclei in a static magnetic field ($\bar{\boldsymbol{B}}_0$). Left: The microscopic view shows an excess (green arrows) in the spin state with lowest energy and a recess (orange arrows) in the spin state with highest energy. Right: The macroscopic view showing the bulk magnetisation vector ($\bar{\boldsymbol{M}}_0$). (B) Applying a radio frequency pulse (90_x^o) that match the Larmor frequency of the spins ($\bar{\boldsymbol{B}}_1$). (C) The effect of applying the radio frequency pulse (90_x^o). Left: The microscopic view shows an even distribution of spins that also display phase coherence. Right: The macroscopic view shows tilted bulk magnetisation vector $\bar{\boldsymbol{M}}$. (D) Net magnetisation returns to equilibrium.

In the most simple case, there is only one nuclear species – e.g., the protons in a solution of chloroform ($C\underline{H}Cl_3$). Then, all spins will have the same resonance frequency. However, a given molecule can also have several nuclear species, which results in several resonance frequencies. Each nuclear species has a unique chemical environment that affects the local magnetic field, which in turn affects the Larmor frequency. All nuclei of a species (e.g., different protons in a molecule) are excited simultaneously by the applied radio frequency pulse. As a result, several FIDs will be superimposed in an interferogram, where interferences between the different FIDs, are observed.

Fourier transformation is a mathematical operation that is applied in order to interpret the frequency domain of the interferogram (Friebolin, 1991a). The Fourier transform of an exponentially decreasing signal, such as a FID, has a Lorentzian line shape. The resonances (or signals) are located on a chemical shift scale, which is a relative frequency scale in parts per million (ppm) (Freeman, 1988a).

Figure 4 illustrates the one dimensional (1D) proton (^1H) NMR spectrum of ethanol and isopropanol (hydroxyl protons are not shown). The number of signals observed in the spectrum for a given molecule corresponds to the number of chemically equivalent protons. There are some general trends regarding the signal positions. For example, signals from methine protons generally display higher chemical shift values than protons in a methylene group and signals from methylene protons generally display a higher chemical shift value than signals from methyl protons. The relative integral area of signals corresponds to the number of protons that gives rise to each respective signal (Friebolin, 1991a; Friebolin, 1991b). Furthermore, signals display different spin-coupling patterns. This owes to connectivity (through covalent bonds) between different groups of chemically non-equivalent protons. For example, a signal with one neighbouring proton will split into a signal with two apices and the distance between the apices is called the scalar coupling constant (J) and it is measured in Hz. The signal from the methyl protons in isopropanol (doublet) only has one neighbouring (methine) proton. The signal from the methine proton in isopropanol is split into a seven apices (septet) due to the six neighbouring (methyl) protons. For ethanol, the two signals from the methylene and methyl protons are split into a quartet and a triplet, respectively (Figure 4). Signals display more complex splitting patterns as the number of protons located within close proximity increases (Carbajo & Neira, 2003).

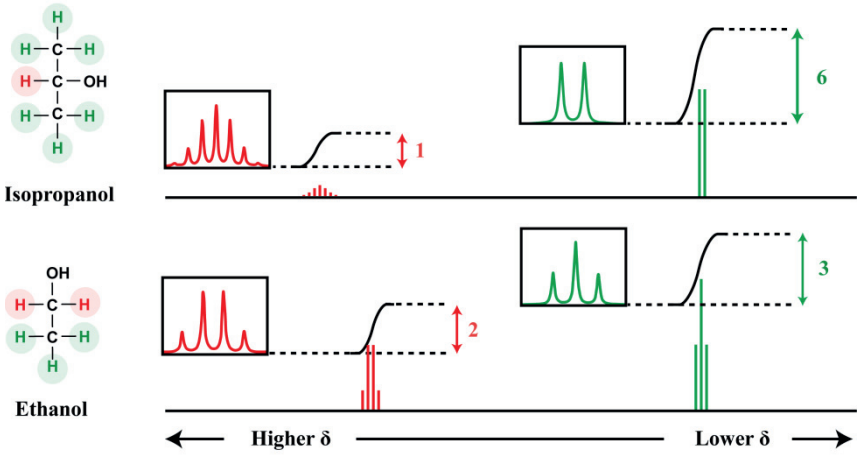

Figure 4. Interpretation of ^{1}H NMR signals from isopropanol and ethanol in relation to molecular structures. Signals from the hydroxyl protons are not shown.

1.2.2 Sample preparation

NMR is quantitative in nature (Ranjan & Sinha, 2018). However, macromolecules such as human plasma proteins generate broad and weak ^{1}H NMR signals that interfere with many metabolite signals. Therefore, macromolecules can hamper the quantitative analysis of metabolites (Wallmeier *et al.*, 2017; Tiziani *et al.*, 2008). Macromolecules can be removed from the sample prior to NMR analysis e.g., using ultrafiltration or precipitation with different solvents (Nagana Gowda & Raftery, 2014; Sheedy *et al.*, 2010).

Quantification by NMR requires calibration. Internal calibration is done by adding a standard in a known concentration to each sample. The internal standard can be used for chemical shift referencing and for quantification. The optimal internal standard should generate signal(s) in an otherwise signal free chemical shift region and it should preferably not interact with other analytes in the sample (Pauli *et al.*, 2012; Burton *et al.*, 2010; Ala-Korpela, 1995). Importantly, the use of one internal standard enables quantification of several metabolites and the standard does not need to be structurally related with the metabolites targeted for quantification (Pauli *et al.*, 2012). Different internal standards can be used in NMR-based metabolomics of human plasma such as 4,4-dimethyl-4-silapentane-1-ammonium trifluoroacetate (DSA), sodium 4,4-dimethyl-4-silapentane-1-sulfonate (DSS) or sodium 3-trimethylsilyl propionate (TSP). Substituting the hydrogens in the carbon chain with deuterium greatly reduce the intensity of the corresponding signals in the ^{1}H

NMR spectrum, while leaving one internal standard signal (singlet) with higher signal-to-noise (*S/N*) ratio at 0.0 ppm (Emwas *et al.*, 2018; Alum *et al.*, 2008; Nowick *et al.*, 2003). The internal standard is typically added after the removal of macromolecules. The commonly used internal standards can interact with macromolecules. This can cause non-concentration dependent variation in the area of the internal standard signal (Kriat *et al.*, 1992). Alternatives to using an internal standard exist (Bharti & Roy, 2012). For example, the reference signal can be synthesised electronically and calibrated against absolute quantities (Akoka, *et al.*, 1999).

The positions (and line widths) of some metabolite signals may vary between experimental spectra. This typically owes to differences in pH or ionic strength between samples. The inter-spectral deviation in signal positions makes quantification more difficult (Hao *et al.*, 2014). By adding a buffer solution to each sample, the variation in pH between samples may be reduced. Sample dilution may also reduce the ionic strength differences between samples (Bharti & Roy, 2012).

1.2.3 Acquisition of NMR data

Signal-to-noise ratio

A sufficient *S/N* ratio is required to yield precise quantification result. The *S/N* ratio can be improved by increasing the magnetic field strength ($\overline{\mathbf{B}}_{\mathbf{0}}$) or by increasing the number of scans (Bharti & Roy, 2012). The frequency for observation of ^{1}H NMR was only 40 MHz in the 1950s (Becker, 1993). Nowadays, spectrometers operating at more than tenfold higher observation frequency are used in metabolomics studies (Louis *et al.*, 2017).

Suppression of unwanted resonances

The water content in plasma is about 90%, which corresponds to a water proton concentration of about 100 M. The water signal dominates the chemical shift region around 4.7 ppm. The signal area of water will be many orders of magnitude larger than the area of metabolite signals. The large variation in dynamic range between the water signal area and the metabolite signal areas can introduce errors in metabolite quantification (Zheng *et al.*, 2011; Ala-Korpela, 1995). Therefore, the resonance from water is typically supressed during the experiment. Suppression of the water resonance can be done with different methods (e.g., excitation with pulsed field gradients and 1D NOESY pre-saturation) (Zheng & Price, 2010).

Although sample preparation of human plasma typically includes the physical removal of macromolecules, it should be noted that quantification of metabolites can be done even in the presence of proteins. Resonances from proteins can be suppressed using the Carr-Purcell-Meiboom-Gill (CPMG) echo train acquisition. However, difficulties are encountered due to the interaction between internal standard and proteins (Wallmeier *et al.*, 2017).

Processing

Processing is typically required to enhance the spectral quality. Firstly, the interferogram can be multiplied with a window function prior to Fourier transformation. To facilitate quantification, the interferogram is multiplied with an exponential window function, where the line broadening parameter is set to determine the rate of the function decay. This improves *S/N* ratio. Typically a value between 0.3 to 1.0 Hz is recommended for metabolomics studies.

Zero filling can be done to improve the resolution (Bharti & Roy, 2012). Briefly, since the algorithm for Fourier transformation requires 2^n data points, the number of measured data points in the interferogram is increased to the closest 2^n value by adding zeros at the end of the interferogram. This results in a narrower distance between the data points in the spectrum (Freeman, 1988b).

Furthermore, phase correction is required to obtain the desired (Lorentzian) appearance of signals. Phase errors or flawed attempts to correct for such can introduce error in quantification. Although more time consuming, manual phase correction often yield better results than auto correction.

A proper baseline that is completely flat and horizontal facilitates accurate integration of signals. Correction can be done to remove baseline distortions. Several computer-assisted programs for baseline correction exist. They are typically semi-automated. However, fully automated baseline correction has also been implemented for use in metabolomics studies; but the successful use of such depends on the spectral complexity (Emwas *et al.*, 2018; Cobas *et al.*, 2006; Chen *et al.*, 2002).

Post-processing steps, such as spectral alignment, data reduction, scaling and normalisation, can also be integral parts in metabolomics studies (Emwas *et al.*, 2018).

1.2.4 Metabolite identification

Metabolite identification and robust signal assignment are required to achieve accurate metabolite quantification (Tredwell *et al.*, 2011). Plenty of research has been conducted to assign experimental signals observed by ^{1}H NMR and metabolite libraries are available in open access and commercially available databases (Ranjan & Sinha, 2018). Utilising the vast amount of information available in such databases aid the signal assignment and identification processes (Nagana Gowda & Raftery, 2017). Importantly, visualisation of metabolite identification and signal assignment can be done. Rather than comparing database values with experimental values, it is possible to do the comparison visually via signal pattern recognition (Weljie *et al.*, 2006). The concept of visual signal pattern recognition is illustrated in Figure 5.

Some metabolites are easily identified by 1D ^{1}H NMR, while other metabolites present a larger challenge (e.g., multiple hits may be found in database searches). Additional experiments – e.g., two-dimensional (2D) NMR and spike-in experiments – are typically used to confirm the identity of metabolites (Nagana Gowda & Raftery, 2017). A relatively recent study identified 67 human blood metabolites using a combination of 1D and 2D NMR experiments (Nagana Gowda *et al.*, 2015).

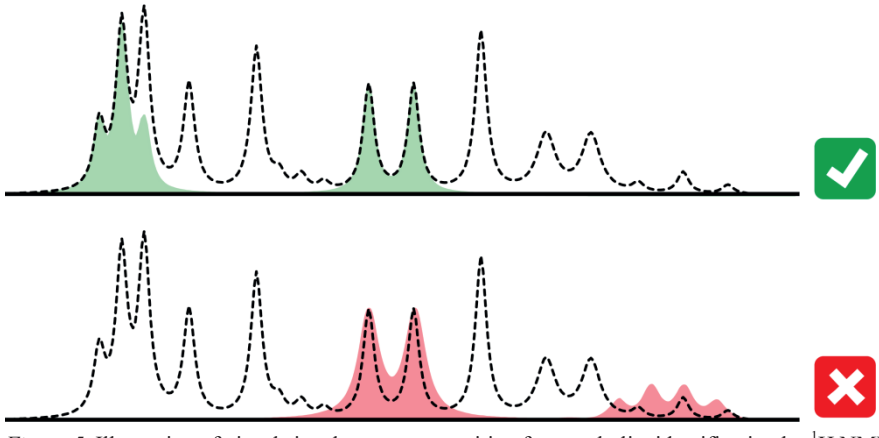

Figure 5. Illustration of visual signal pattern recognition for metabolite identification by ^{1}H NMR. Top: Identification and assignment of putative metabolite (green) is supported. Bottom: Identification and assignment of putative metabolite (red) is not supported.

1.2.5 Metabolite quantification

Many human blood metabolites display signals with similar chemical shift values (Nagana Gowda *et al.*, 2015). The overlap in positions results in signal interferences. Unless accounted for properly, metabolite quantification will be

hampered (Weljie *et al.*, 2006). The observed signal pattern in an experimental ^{1}H NMR spectrum may be viewed as the total sum of intensity contributions from the individual metabolite signals, or, the signal pattern in the experimental spectra can be viewed as a linear combination of individual metabolite spectra (Ravanbakhsh *et al.*, 2015; Zheng *et al.*, 2011).

Deconvolution of the signal pattern into the contribution from individual metabolite signals may be done utilising spectral information from a metabolite library. Quantification can be done manually in the ChenomX NMR Suite software package (ChenomX Inc., Edmonton, Canada). In this approach for targeted profiling, a signal pattern that matches the experimental ^{1}H-NMR spectrum is built by a step-wise addition and of spectra from a metabolite library. For each added library spectrum, its positions and intensities may be adjusted to yield the optimal match (Figure 6). The line widths of the library signals are matched to the experimental spectra using the internal standard for calibration (Weljie *et al.*, 2006). Manual targeted profiling, done with this software, is frequently employed in metabolomics studies and has even been referred to as a near golden standard approach (Hao *et al.*, 2014). Manual targeted profiling is very time-consuming, especially when employed on many experimental spectra. Hence, manual targeted profiling can be a major bottleneck in the workflow in large-scale metabolomics studies.

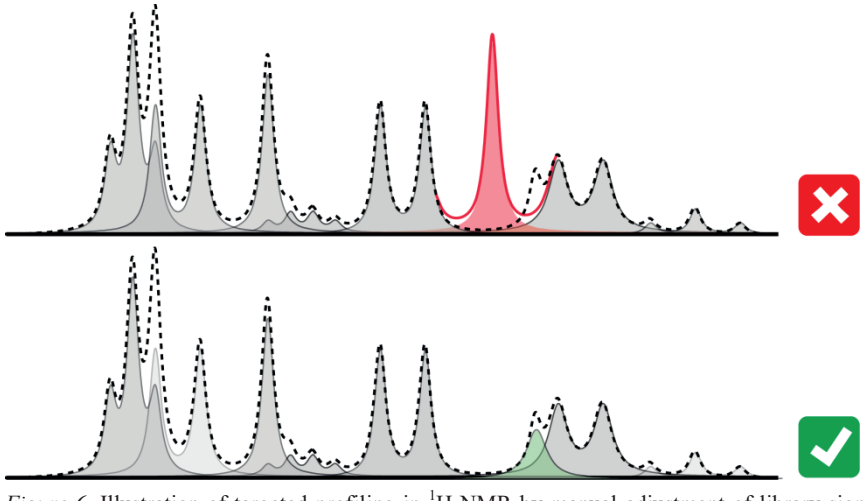

Figure 6. Illustration of targeted profiling in ^{1}H NMR by manual adjustment of library signals. Top: Non adjusted library signal (red). Bottom: Adjusted library signal (green).

Attempts to automate the deconvolution process in order to increase the efficiency have been done (Bingol, 2018). BATMAN, BAYESIL and *BQuant* are examples of automated algorithms that are based on Bayesian modelling.

Briefly, Bayesian modelling utilise a 'best-guess' approach with regards to signal characteristics in the library and iterations are performed until convergence with the experimental signal pattern (Ravanbakhsh *et al.*, 2015; Hao *et al.*, 2014; Zheng *et al.*, 2011). Still, processing of larger datasets is a main obstacle for such algorithms, especially if the dataset includes spectra with highly complex signal pattern. This is since the increased complexity results in increased computational burden. Varying signal positions between experimental spectra (e.g., caused by pH and ionic strength differences between samples) specifically increase the computational burden (Hao *et al.*, 2014). Both manual and automated deconvolution procedures are more suitable to apply on spectra of less complexity. In this context, experimental spectra from urine samples can be more problematic than spectra from plasma and serum (Emwas *et al.*, 2018). Firstly, more metabolites can be identified in urine by NMR (Bouatra *et al.*, 2013; Psychogios *et al.*, 2011). Secondly, signal positions of urine metabolites vary more between spectra due to larger pH and ionic strength differences between samples (Rist *et al.*, 2013).

1.2.6 Impact of different anticoagulants

Blood serum is obtained when the coagulation process has occurred, while blood plasma is obtained when it has been prevented by the addition of an anticoagulant – e.g., ethylenediamine-tetra-acetic acid (EDTA) or heparin (Tuck *et al.*, 2009). Allowing or hindering the coagulation process has an impact on the results from metabolomics analysis. For example, a previous study revealed that serum and plasma metabolite levels are highly correlated. Higher reproducibility was observed in plasma in repeated measurements, while better sensitivity was observed in serum when comparing groups of individuals with different phenotypes (Yu *et al.*, 2011). Furthermore, the use of different anticoagulants also has an impact on the result from metabolomics analysis (Gonzalez-Covarrubias *et al.*, 2013; Yin *et al.*, 2013; Barton *et al.*, 2010). The use of heparin as anticoagulant is often favoured in NMR-based metabolomics. The appearance of (the broad and weak) NMR signals from a macromolecule such as heparin may be avoided e.g., using ultrafiltration (Casu *et al.*, 2015; Daykin *et al.*, 2002). The physical removal of the anticoagulant prior to NMR analysis avoids additional interferences between heparin and metabolite signals. EDTA generates several characteristic signals due to its binding with different cations (e.g., H^+, Ca^{2+} and Mg^{2+}). These signals are located within close proximity to signals from many human plasma metabolites. Hence, interferences between EDTA signals and metabolite signals will be observed (Barton *et al.*, 2010). Metabolite quantification by

NMR in plasma that contains EDTA may still be desired, especially since such samples are frequently collected and stored for future research purposes (Tuck *et al.*, 2009; Vaught *et al.*, 2009).

1.3 Molecular epidemiology

1.3.1 Observational study designs

Observational study designs typically assess associations between exposure and outcome. In the context of molecular epidemiology, the exposure of interest is of biological origin (e.g., metabolites) and the outcome of interest is often a disease (García-Closas *et al.*, 2011; Hendricks *et al.*, 2011). There are three main types of observational study designs, namely the cross-sectional design, the case-control design and the cohort design (Belbasis & Bellou, 2018). In the cross-sectional study design, exposure and outcome are measured at the same point in time (Setia, 2016). In the case-control study design, the participants are selected based on an outcome. Exposure status is then assessed in both groups (either at the present or at a previous point in time) (Schulz & Grimes, 2002). In a cohort study design, individuals are selected based on exposure status and followed in time with regards to outcome. Using this prospective design, the temporal relation of cause and outcome can often be established (Grimes & Schulz, 2002). There is a large potential for prospective cohort studies in molecular epidemiology due to worldwide efforts in collecting and storing biological specimens (Vaught *et al.*, 2009).

The nested case-control study design offers a cost and time efficient alternative to full cohort studies. Cases are identified within a cohort and a limited number of controls from the same cohort are matched to each case. Exposure status is assessed at study baseline considering both groups. This design results in reductions with respect to data collection and analyses efforts, with only minor compromise on the statistical efficiency (Califf, 2018; Ernster, 1994). The nested case-control study design is particularly useful in identifying biological precursors of disease (Califf, 2018; Ernster, 1994).

1.3.2 Bias in observational studies

Observational studies typically include a randomly selected sample from a larger target population. Based on the outcome from this sample, general conclusions are drawn for the target population (Morshed *et al.*, 2009). However, observational studies can both be affected by bias that reveals

associations that are incorrect (internal validity) and bias that affects the generalisability of results (external validity). Although there are many different sources of bias in observational studies they can broadly be categorised into three subgroups, namely selection bias, information bias and confounding (Delgado-Rodríguez & Llorca, 2004).

Selection bias can occur if the sample does not represent the target population (Delgado-Rodríguez & Llorca, 2004). This is especially a problem in case-control studies, where recruitment can be problematic, resulting in the inclusion of controls that does not represent the population from which the cases origin. Such selection bias is a smaller problem in nested case-control studies, since controls are selected from the same cohort as the cases (Ernster, 1994). However, exposure and/or outcome related selection into the cohort might still cause some bias (Munafó *et al.*, 2018). Cohort studies are often prone to selection bias due to loss to follow-up (Delgado-Rodríguez & Llorca, 2004; Grimes & Schulz, 2002). Loss to follow-up is, however, a minor issue in Nordic countries due to record linkages by unique personal numbers (Maret-Ouda *et al.*, 2017).

Information bias refers to errors that occur during data collection. For example, if there are difficulties in distinguishing individuals based on outcome and/or exposure level, then systematic misclassification may occur (e.g., non-exposed are wrongfully classified as exposed and vice versa). (Delgado-Rodríguez & Llorca, 2004)

Confounding may occur if a variable affects both exposure and outcome, without being related to the causal relation between exposure and outcome (Delgado-Rodríguez & Llorca, 2004). Efforts to reduce or account for bias can be built into the design. For example, confounding can be adjusted for in statistical analysis or in the sample selection step by matching. However, an unmeasured confounder cannot be accounted for (Jepsen *et al.*, 2004; Geenland, 1996).

1.3.3 Statistics in observational studies

Univariate statistical analysis of observational data generally consists of two parts, namely estimation and hypothesis testing. The first part involves the calculation of a point estimate for the association of interest and its precision (i.e., a confidence interval; CI). Hypothesis testing is done to determine the probability to observe an association of interest even if it does not exist (null hypothesis). As mentioned in the general introduction to metabolomics, test statistics is performed and the decision of a cut-off for significance determines if the null hypothesis can be rejected or not (Morshed *et al.*, 2009).

The odds ratio (OR) is one example of a point estimate of the association between exposure and outcome, which is commonly used in case-control studies. There are four possible groupings of individuals, namely exposed cases (true positive: TP), unexposed cases (false positive: FP), exposed controls (false negative: FN) and unexposed controls (true negative: TN). The OR is the ratio between (TP × TN) and (FP × FN). Hence, an OR >1 indicates a positive association, while an OR <0 indicates an inverse association (Broadhurst & Kell, 2006; Breslow & Day, 1980). Logistic regression can be used to calculate the OR for a binary outcome to investigate its association with metabolite level (Morshed *et al.*, 2009; Bewick *et al.*, 2005). If such an association exist, then the probability of a given outcome should vary with metabolite level. For example, the probability of a given outcome (p) may increase with an elevation in metabolite level (x) or vice versa. However, this relationship typically follows an S-shaped curve. The logit function, logit(p) can transform such an S-shaped curve to approximate a linear form, $k \times x + m$, where the OR is Eulers number to the power of k. The parameters (k and m) cannot be assessed using linear regression since the underlying distribution for a binary outcome assumedly follows a binomial distribution. The parameters, are usually derived by maximum likelihood estimation. Iterative techniques performed using computer-assisted packages can be required (Bewick *et al.*, 2005).

1.3.4 Risk biomarkers for type 2 diabetes

In 2017, about 425 million people suffered from diabetes worldwide and incidences are expected to increase even more in the future (IDF, 2017a). T2D is characterised by insulin resistance, a metabolic condition where the production of insulin first increases in order to the lower blood glucose level; but, in time the production becomes insufficient, which results in an elevated blood glucose level. Elevated blood glucose levels are currently used to diagnose T2D and pre-diabetic states such as impaired fasting glucose (IFG) and impaired glucose tolerance (IGT) (IDF, 2017b). This is done using an oral glucose tolerance test (OGTT) where glucose is administered orally and the blood levels measured pre- and post-ingestion. The World Health Organization (WHO) provides with recommended cut-off values to differentiate between normal glucose tolerance (NGT), glucose intolerance (IFG, IGT) and T2D (WHO, 2017; WHO, 1999).

Several nested case-control studies have been conducted to identify plasma or serum metabolites associated with risk of T2D (Shi *et al.*, 2018; Ruiz-Canela *et al.*, 2018; Qui *et al.*, 2016; Drogan *et al.*, 2015; Wang *et al.*, 2011).

For example, higher levels of some branched chain amino acids and aromatic amino acids associated with elevated risk of T2D, while lower level of some lysophosphatidylcholines associated with elevated risk of T2D.

1.3.5 Risk biomarkers for prostate cancer

Prostate cancer is the second most common cancer among men worldwide. In 2018, Sweden was among the ten countries in the world with the highest rate of prostate cancer (Bray *et al.*, 2018). Prostate cancer typically develops over a long period of time and is most frequently diagnosed in men older than 60 years of age. The disease severity varies, ranging from indolent tumours, which may be of small clinical significance, to more aggressive stage tumours. Risk factors for prostate cancer have been frequently studied within epidemiology. One of the associations, with strongest evidence in the scientific literature, is between obesity and risk of advanced stage prostate cancer. An association between high consumption of dairy products and elevated risk of prostate cancer has been suggested; but, with limited evidence (WCRF/AICR, 2018). Meta-analysis of observational studies suggests that there is an association between T2D and reduced risk of prostate cancer (Bansal *et al.*, 2013). Furthermore, insulin-like growth factor-I (IGF-I) is considered a modifiable risk factor for prostate cancer (WCRF/AICR, 2018) (Travis *et al.*, 2016). Interestingly, the association between higher circulatory levels of IGF-I and elevated risk of prostate cancer vary with age, the association being stronger in relatively young subjects (<59 years) (Stattin *et al.*, 2004).

Studies have revealed that several serum and plasma levels of some amino acids, carnitines and glycerophospholipids may differ between prostate cancer patients and healthy individuals (Kelly *et al.*, 2016). This includes different lysophosphatidylcholines with a saturated fatty acid chain. Yet, contradictory findings have been reported for lysophosphatidylcholines (i.e., prostate cancer has been linked to both higher and lower levels) (Zhou *et al.*, 2012; Lokhov *et al.*, 2010; Osl *et al.*, 2008).

The studies mentioned above should be clearly distinguished from MWAS. Rather than comparing individuals diagnosed with prostate cancer and healthy individuals, MWAS aims to identify metabolites associated with risk of developing the disease in the future. Several nested case-control studies have been conducted to identify plasma or serum metabolites associated with risk of prostate cancer (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014). Lower levels of 1-stearoylglycerol (an intermediate in lipid metabolism) associated with elevated risk of prostate cancer in a study nested within the ATBC (Mondul *et al.*, 2014). However, this

finding was not replicated in a later study within the same cohort (Mondul *et al.*, 2015). Furthermore, lower levels of several glycerophospholipids associated with elevated risk of advanced stage prostate cancer in a study nested within EPIC (multicenter) (Schmidt *et al.*, 2017).

Sample size, fasting status and follow-up time – i.e., the time between sample collection (baseline) and diagnosis with prostate cancer – are some important factors to consider in MWAS. Firstly, the sample size should be large enough so the statistical power (i.e., the probability that the statistical test yields significant results) is adequate (Houle *et al.*, 2005). Secondly, it has been shown that non-fasting samples display lesser temporal stability than fasting samples. Stability over time is essential when using a single measurement to assess association between metabolites and disease risk. Hence, the use of fasting samples is favored in MWAS (Caraloy *et al.*, 2015). Thirdly, the nested case-control study design can be used to ensure that the exposure (here, metabolite data) is collected prior to the outcome of disease rather than at the presence of disease. Therefore, it is a useful study design in MWAS (Nicholson *et al.*, 2008; Ernster, 1994). The association between metabolites and risk of prostate cancer may differ depending on the follow-up time. For example, a previous study reported different statistical outcome when stratifying the statistical analyses based on follow-up time (<5 years, >5 years) (Schmidt *et al.*, 2017).

Further MWAS on prostate cancer (employing a nested case-control study design) are warranted since none of the previous studies have included both (1) large sample size, (2) entirely fasting samples and (3) a long follow-up (>5 years). Furthermore, none of the previous studies have reported whether association between metabolites and prostate cancer risk vary with baseline age (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014). Still, this may be highly relevant since the association between IGF-I (a modifiable risk factor for prostate cancer) and risk of prostate cancer vary with baseline age (Stattin *et al.*, 2004).

2 Objectives

Methods for quantifying plasma metabolites by NMR should (1) be based on simplistic and easy to understand principles on how to account for signal interferences, (2) only require minor computational efforts and (3) account for inter-spectral deviations in signal positions and line widths when such exist. In response to this, an Automated Quantification Algorithm (AQuA) was developed for quantification of human plasma metabolites in samples collected using heparin and EDTA as anticoagulant, respectively (papers I and II). Targeted metabolomics was also employed in case-control study nested within the NSHDC with the aim of identifying risk biomarkers of prostate cancer (paper III).

- **Paper I**: Design of an AQuA focused on (1) to (2) and its application on NMR spectra from plasma samples collected using heparin as anticoagulant.

- **Paper II:** Modification of the AQuA with focus on (3) and its application on NMR spectra from plasma samples collected using EDTA as anticoagulant.

- **Paper III**: Identification of risk biomarkers for prostate cancer in the NSHDC via the use of targeted MS and NMR-based metabolomics employed on plasma samples collected using heparin as anticoagulant.

3 AQuA

In paper I, an Automated Quantification Algorithm (AQuA) was designed, implemented and evaluated as the final step in the targeted NMR-based workflow. A principle for targeted profiling was introduced for manual quantification of selected compounds by ^{1}H NMR. The general principle of AQuA was designed using the (manual) principle for targeted profiling as a starting point. AQuA was implemented for quantification of human plasma samples collected using heparin as anticoagulant. AQuA was evaluated by comparison with targeted profiling and by computation of different quality indicators.

3.1 Methods

3.1.1 Design of targeted profiling

The principle for targeted profiling, designed for manual quantification of target compounds by ^{1}H NMR, accounts for signal interferences in experimental spectra via deconvolution using a compound library. The principle is illustrated below for five putative compounds.

One signal in the library (i.e., reporter signal) is selected for the quantification of each compound. In this example, the signal at target position δ_1 is selected as reporter for compound *1*, δ_2 for compound *2*, δ_3 for compound *3*, δ_4 for compound *4* and δ_5 for compound *5* (Figure 7A). An integration order is also determined. In the current example, the integration order *5-4-3-2-1* is set (Figure 7A). Targeted profiling is then applied on each experimental spectrum by adjusting the height of each reporter signals in the pre-determined integration order so the sum of library signals (i.e., the reporter signals and signals from interfering compounds) match with the corresponding experimental signals (i.e., target signals) (Figure 7B). Importantly, the

integration order is set so that interferences that affect each given reporter are accounted for. For example, the reporter signal for compound *5* is integrated prior to the reporter signal for compound *3*. This is since a (non-reporter) signal from compound *5* interferes with the reporter signal for compound *3*. Adjustment of library signal positions can be done if necessary.

This manual principle for targeted profiling can be applied on different experimental spectra to yield concentrations of the target compounds in different samples.

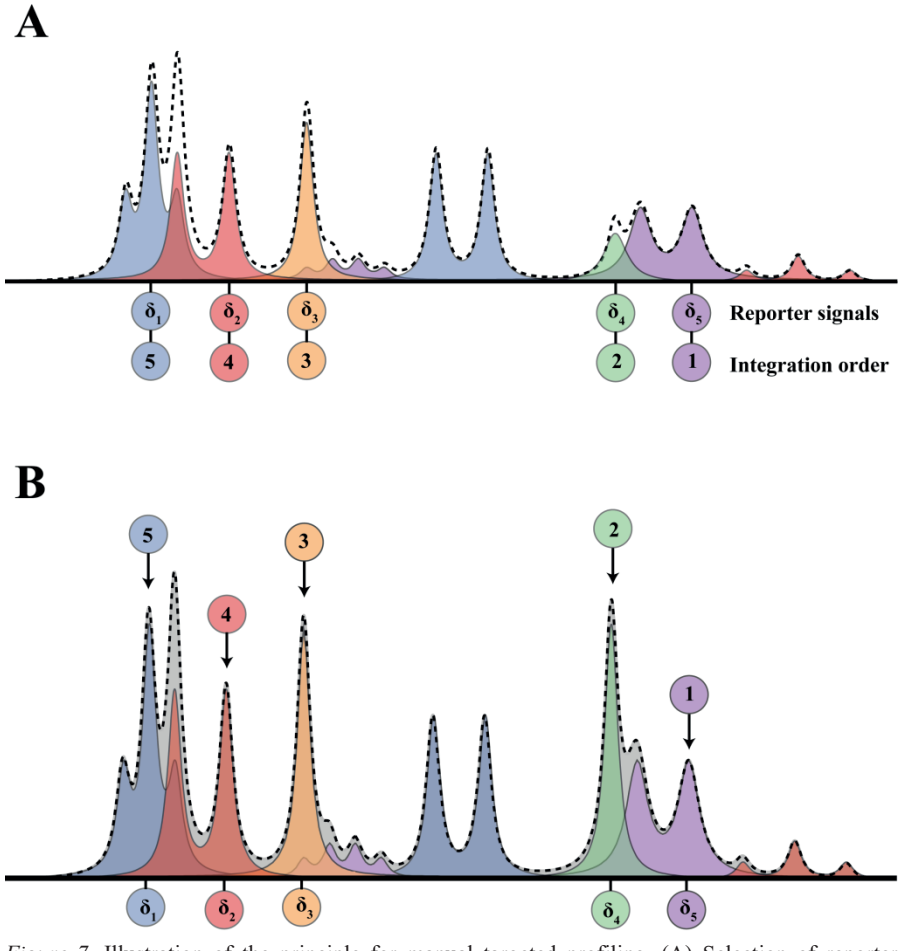

Figure 7. Illustration of the principle for manual targeted profiling. (A) Selection of reporter signals and setting the integration order. (B) The outcome of manual targeted profiling when applied on one experimental spectrum.

3.1.2 Design of AQuA

The general principle of AQuA is illustrated below, again for five putative compounds. The first step in the design of AQuA is to select one reporter signal per compound, which is to be used for its quantification. This step is identical to the first step in the principle for targeted profiling (Figure 7A). The library spectrum for each compound *i* is normalised by dividing all intensities with the height found at target position δ_i. Data reduction is done by extracting the intensities at all target positions δ_1, δ_2, δ_3, δ_4 and δ_5 (Figure 8A). The extracted values for each compound are arranged as columns in a (5 × 5) $\bar{\bar{\mathbf{m}}}$ matrix (Figure 8B).

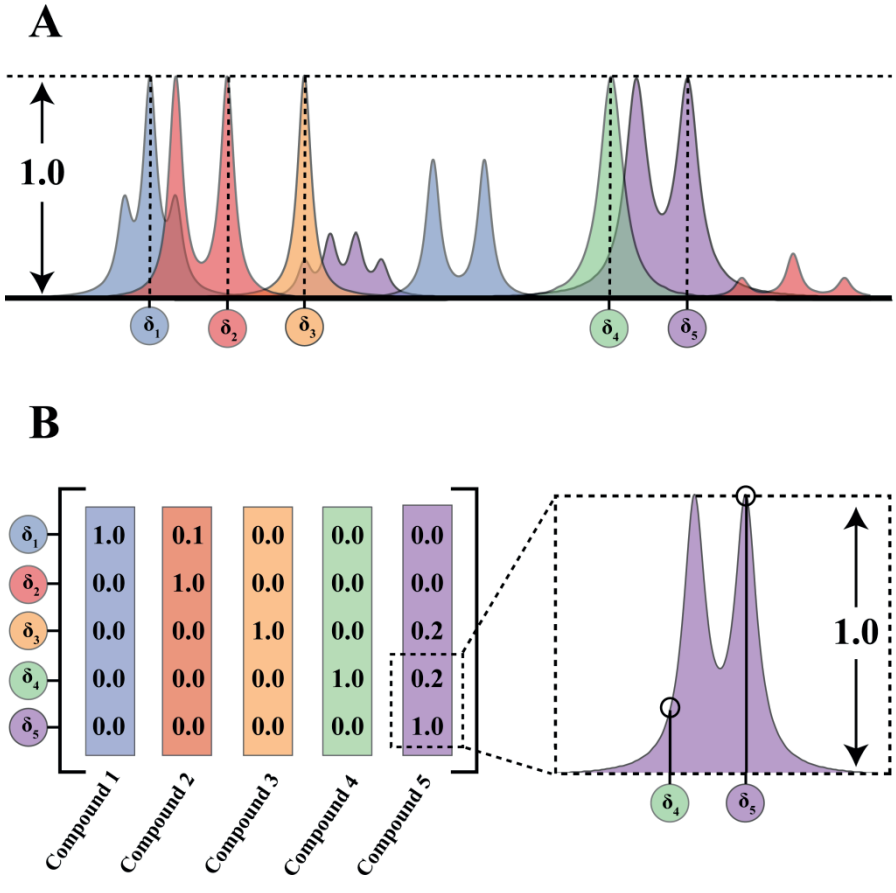

Figure 8. Illustration of generating the $\bar{\bar{\mathbf{m}}}$ matrix. (A) Normalisation and data reduction of the compound library. (B) Extracted values (normalised intensities at all target positions) for each compound arranged as columns in the matrix.

Data reduction is employed on the experimental spectrum (Figure 9A). In this step, the height at the corresponding target positions δ_1, δ_2, δ_3, δ_4 and δ_5 (i.e., the height of all target signals) are extracted from the experimental spectrum and organised into a (5 × 1) $\bar{\mathbf{y}}$ vector. The $\bar{\bar{\mathbf{m}}}$ matrix and the $\bar{\mathbf{y}}$ vector are used as input in the AQuA computation (Eq. 1A) to generate a (5 × 1) $\bar{\mathbf{x}}$ vector. This vector corresponds to the same reporter signals as those obtained via the manual targeted profiling (Figure 7B).

$$\bar{\mathbf{y}} = \bar{\bar{\mathbf{m}}} \times \bar{\mathbf{x}} \qquad \text{(Eq. 1A)}$$

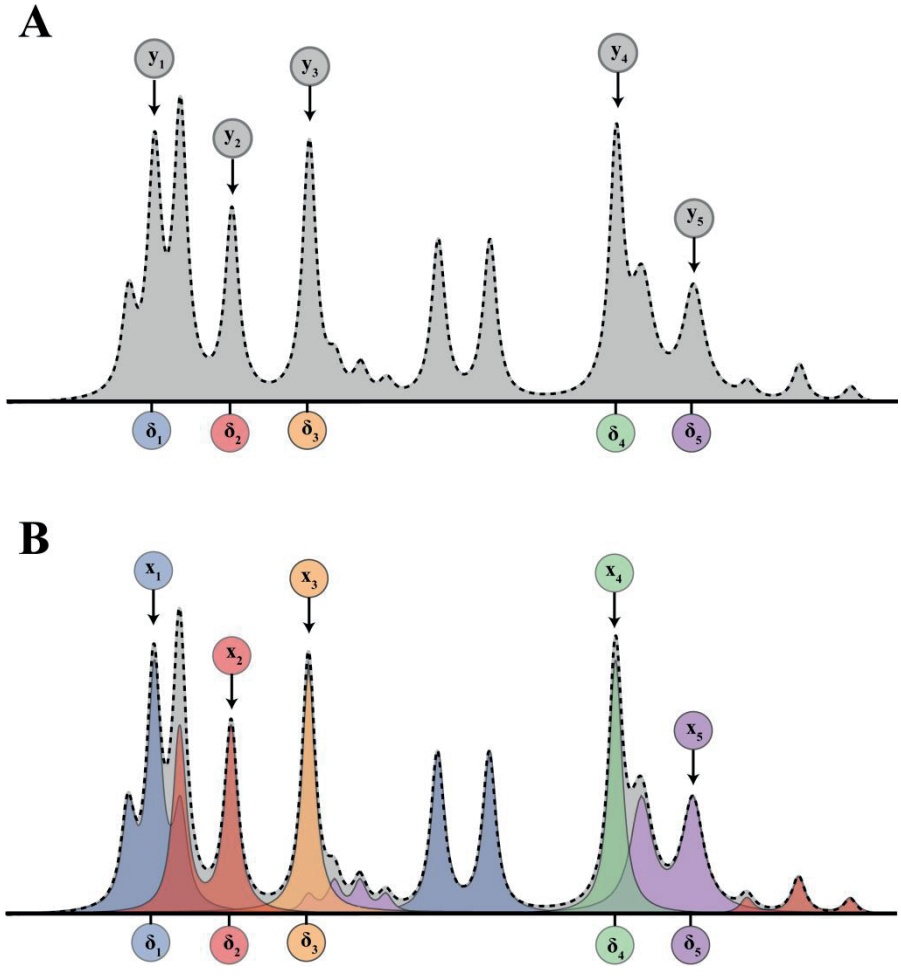

Figure 9. Illustration of the AQuA computation. (A) Data reduction of the experimental spectrum. (B) The outcome from the AQuA computation.

The AQuA computation (Eq. 1A) can also be expressed as in Eq. 1B. The height of each target signal (y_i at δ_i) is modelled as the sum of contributions from: (1) the reporter signal (x_i at δ_i) and (2) interfering compounds (≥0). Each individual contribution is expressed as a product of a matrix element and a reporter signal ($m \times x$) (Eq. 1). Although only the reporter signals are generated in computation, the individual contribution from each interfering compound (≥0) may also be derived. For example, the individual contribution from compound *5* to the target signal height used for quantification of compound *4* is ($m_{4,5} \times x_5$). Note that $m_{4,5}$ refers to the normalised interference from compound *5* at position δ_4 (see Figure 8).

$$
\begin{aligned}
y_1 &= m_{1,1} \cdot x_1 + m_{1,2} \cdot x_2 + m_{1,3} \cdot x_3 + m_{1,4} \cdot x_4 + m_{1,5} \cdot x_1 \\
y_2 &= m_{2,1} \cdot x_1 + m_{2,2} \cdot x_2 + m_{2,3} \cdot x_3 + m_{2,4} \cdot x_4 + m_{2,5} \cdot x_2 \\
y_3 &= m_{3,1} \cdot x_1 + m_{3,2} \cdot x_2 + m_{3,3} \cdot x_3 + m_{3,4} \cdot x_4 + m_{3,5} \cdot x_3 \\
y_4 &= m_{4,1} \cdot x_1 + m_{4,2} \cdot x_2 + m_{4,3} \cdot x_3 + m_{4,4} \cdot x_4 + m_{4,5} \cdot x_4 \\
y_5 &= m_{5,1} \cdot x_1 + m_{5,2} \cdot x_2 + m_{5,3} \cdot x_3 + m_{5,4} \cdot x_4 + m_{5,5} \cdot x_5
\end{aligned}
\qquad \text{Eq. 1B}
$$

The total contribution from interfering compounds to the target signal for a given compound *i* is the difference between its target and reporter signal height (y_i - x_i). The interference (Δ_i) for a given compound *i* is defined as the relative sum of contribution from other compounds (Eq. 2).

$$\Delta_i = \frac{y_i - x_i}{y_i} \qquad \text{Eq. 2}$$

AQuA can be designed for quantification of any number of compounds. The addition of one compound results in a one unit increase in the dimension of the linear equation system (Eq. 1). The computation may also be applied for different $\bar{\mathbf{y}}$ vectors, to generate corresponding $\bar{\mathbf{x}}$ vectors – i.e., quantifying compounds in different samples. For each experimental spectrum *n* (with vector $\bar{\mathbf{y}}_\mathbf{n}$), the AQuA computation yields the corresponding set of reporter signals (vector $\bar{\mathbf{x}}_\mathbf{n}$), which represents the quantitative mixture of compounds in sample *n* (Eq. 3).

$$\bar{\mathbf{y}}_\mathbf{n} = \bar{\bar{\mathbf{m}}} \times \bar{\mathbf{x}}_\mathbf{n} \quad \text{(Eq. 3)}$$

Note that the AQuA (as well as manual targeted profiling) is a targeted approach – i.e., pre-selected compounds are targeted for quantification.

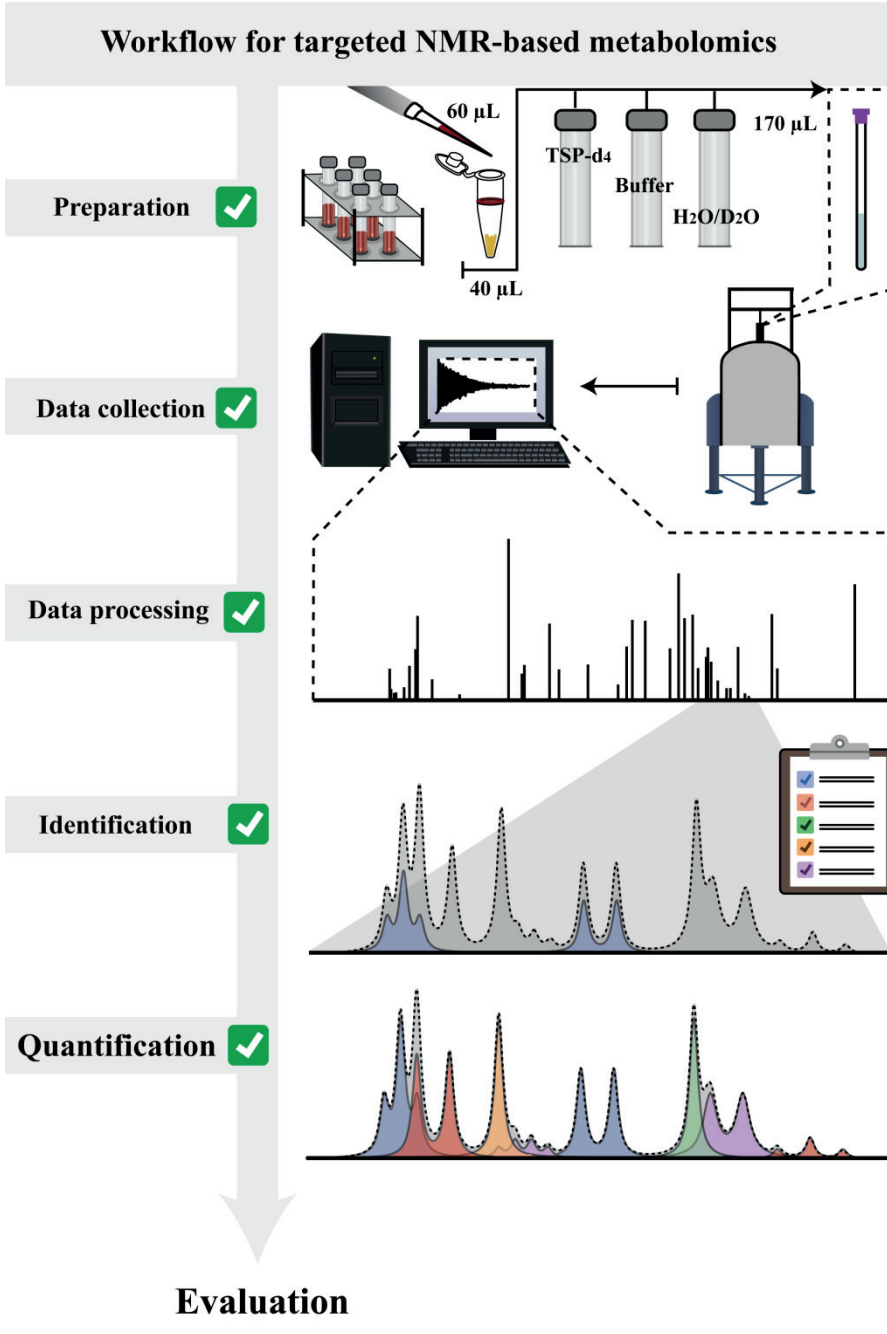

Figure 10. Workflow for targeted NMR-based metabolomics applied on human plasma samples.

3.1.3 Implementation of AQuA

Targeted NMR-based metabolomics was employed on 1342 human plasma samples collected using heparin as anticoagulant. The samples were analysed according to the workflow presented in Figure 10.

Macromolecules (such as plasma proteins and heparin) were removed using ultrafiltration (Tiziani *et al.*, 2008). A solution with deuterated TSP (internal standard), phosphate buffer and H_2O/D_2O was mixed with each plasma filtrate in the ratio of 4.25:1. The internal standard was added to each sample to allow for quantification, while the buffer was added to reduce pH differences between samples (Bharti & Roy, 2012). An experimental spectrum was obtained for each NMR sample on a spectrometer operating at 600 MHz using a 1D ^{1}H NMR experiment (512 scans) that suppressed the water resonance (Hwang & Shaka, 1995). Processing – e.g., phase correction and application of an exponential window function – was done manually to allow quantification by signal heights (Hays & Thompson, 2009). This workflow generated a dataset consisting of 1342 experimental spectra. This dataset is hereafter referred to as Dataset Heparin.

Metabolite identification and assignment of experimental NMR signals in Dataset Heparin was done using signal pattern recognition. This resulted in a set of compounds – i.e., human plasma metabolites – targeted for quantification with manual targeted profiling and with AQuA. The procedure for manual targeted profiling was implemented in ChenomX NMR Suite (version 7.5, ChenomX Inc.). AQuA (Eq. 3) was implemented in MATLAB (version R2012b, MathWorks Inc.). Table 2 compiles an explanation of the values, which were used and generated in AQuA computations. A majority of these values have already been introduced in the method section above, but the table provides with an overview. After employing both procedures, two concentration estimates (μM) were generated for each metabolite *i* in each sample *n*. Final plasma concentrations were derived by accounting for the dilution during sample preparation.

Table 2. *Explanation of important values in AQuA* [a, b, c]

Name	*Explanation*
Vector elements determined in each experimental spectrum *n*	
Target signal (y_i)	The signal used for quantifying compound i; sum of intensity contributions from reporter i and interfering compounds $\neq i$
Target position (δ_i)	The corresponding position of target signal i
Vector elements generated in each AQuA computation *n*	
Reporter signal (x_i)	The interference-free signal height of compound i at target position i
Interference (Δ_i)	The relative contribution from all interfering compounds ($\neq i$) at target position i (Eq. 2; ≥ 0). $(x_i / y_i) + \Delta_i = 1$ (100%).
Concentration (C_i)	The concentration of compound i in the NMR sample
Matrix ($\bar{\bar{\mathbf{m}}}$) used in AQuA computations as derived from a compound library	
Row ($m_{i,}$)	Normalised intensities of all compounds at target position i
Column ($m_{,j}$)	Normalised intensities of compound j at all target positions
Element ($m_{i,j}$)	Normalised intensity of compound j at target position i
Diagonal	Normalised intensity of the reporter at its target position ($i = j$, $m = 1$)
Non-diagonal	Normalised intensity of an interfering compound j at target position i ($i \neq j$, $m \geq 0$)

[a] In paper I, the compound library consisted of one calibration spectrum per human plasma metabolite targeted for quantification with AQuA (with characteristic ^{1}H NMR signals positions and line widths adjusted *in silico* to match with the experimental spectra).
[b] Experimental signals were modelled via the library signals. Each experimental signal was therefore predicted to have the same position and line width as the corresponding library signal.
[c] Reporter signals generated in AQuA computation (Eq. 3) were converted to metabolite concentrations using a set calibration factors (for each metabolite, the ratio between its library concentration and its reporter height in the library before normalisation).

3.1.4 Evaluation of AQuA

Mean relative deviation and goodness-of-fit

The accuracy of AQuA was evaluated against manual targeted profiling in a subset of 30 randomly selected spectra from Dataset Heparin. After employing both procedures, two concentration estimates (μM) were generated for each metabolite *i* in each sample *n* ($C_{i,n}$ where the superscript indicates the quantitative procedure used, namely *auto* for AQuA and *manual* for targeted profiling). The mean relative deviation (*MRD*) was calculated for each respective metabolite using the concentrations generated by both procedures (Eq. 4; N =30).

$$MRD_i = \frac{1}{\mathrm{A}} \times \sum_{\mathrm{i=1}}^{\mathrm{N}} \left(\frac{C_{i,n}^{auto} - C_{i,n}^{manual}}{C_{i,n}^{manual}} \right) \qquad \text{(Eq. 4)}$$

Furthermore, the values used for computing the *MRD* for each metabolite were also used in linear regression analysis to compute the goodness-of-fit (r^2) for each metabolite.

Quality indicators

The evaluation also included the computation of three quality indicators for each respective metabolite – i.e., *occurrence*, *positional deviation* and F_q values (related to degree of interference). Each quality indicator was computed based on values inherently generated by AQuA (Table 3).

The *occurrence* was defined as the percentage (%) of reporter signals found to be above the limit of detection (LOD) in a given dataset. The *positional deviation* was defined as the distance (± bins, where 1 bin =0.0002 ppm) from the median target position that accounted for 95% of target signals in a given dataset. The F_q value for a given metabolite was defined as the fraction of spectra where the interference (Eq. 2) exceeded the pre-determined value q. F_q was computed for q =0.05 (5%) and q =0.50 (50%).

Table 3. *Explanation of quality indicators* [a]

Quality indicator	*Explanation*
Occurrence	
Unit	Percent (%)
Based on	Reporter signal of metabolite *i* (x_i)
Distribution	$x_i > (3 \times \text{LOD})$
Cut-off	% of values found to be above the detection limit
Results	Metabolite *i* has 70% occurrence
Interpretation	The reporter signals of metabolite *i* are above the detection limit in 70% of the experimental spectra
Positional deviation	
Unit	± bins (1 bin =0.0002 ppm)
Based on	The target position for metabolite *i* (ppm)
Distribution	Median-centered
Cut-off	95% of experimental spectra
Results	The positional deviation for metabolite *i* is (±) 100 bins
Interpretation	For metabolite *i*, the deviation from the median target position is within 100 bins in 95% of experimental spectra
F_q values	
Unit	Fraction
Based on	The interference (Δ; Eq. 2)
Distribution	$\Delta > q$ (for q =0.05 and q =0.50)
Cut-off	Fraction of spectra where Δ exceeds a pre-determined value q
Results	Metabolite *A* displays $F_{0.50}$ =1
Interpretation	For metabolite *A*, the contribution from interfering metabolites signals exceeds 50% in all experimental spectra

[a] The LOD was determined in each experimental spectrum from the baseline noise in a signal free region.

3.2 Results and discussion

3.2.1 Accuracy

Comparison with targeted profiling

AQuA and a manual procedure for targeted profiling was employed on a subset of experimental spectra in Dataset Heparin (30 out of 1342). The *MRD* and r^2 derived for each respective metabolite was used as a measure of the quantitative accuracy of AQuA relative to manual targeted profiling (Figure 11). A majority of metabolites displayed an r^2 >0.99 and *MRD* within ±0.05. Hence, the two quantitative procedures typically generated results that were in excellent agreement. The accuracy of AQuA was thus tested and confirmed in comparison with manual targeted profiling. The latter procedure was performed using a software package dedicated to metabolite identification and quantification, which has been generally approved by the metabolomics community (Hao *et al.*, 2014; Weljie *et al.*, 2006).

Quality indicators

Although the results from AQuA showed excellent agreement with results from a manual procedure for targeted profiling, evaluation by such a comparison was somewhat limited since the manual procedure was too time-consuming to employ on the entire dataset. Manual targeted profiling was therefore only employed on a subset of experimental spectra in Dataset Heparin. Instead, an alternative approach for evaluating AQuA was presented. This evaluation included the computation of three quality indicators – i.e., *occurrence*, *positional deviation* and F_q values (Figure 11).

Most metabolites displayed an *occurrence* ≥90% in the subset. The few metabolites with low *occurrence* typically displayed poor r^2 and *MRD*, while metabolites with high *occurrence* typically displayed r^2 ≥0.99 and *MRD* within ±0.05. Hence, evaluation of AQuA by comparison with a manual procedure for targeted profiling yielded similar information as investigating the distribution of *occurrences* in the subset. Furthermore, the distribution of *occurrences* in the subset was representative of the *occurrence* distribution in the entire dataset. This shows that metabolites with low *occurrence* may be difficult to quantify accurately with AQuA. However, this issue is not specific to AQuA since signals of low intensity are difficult to quantify accurately regardless of the procedure used.

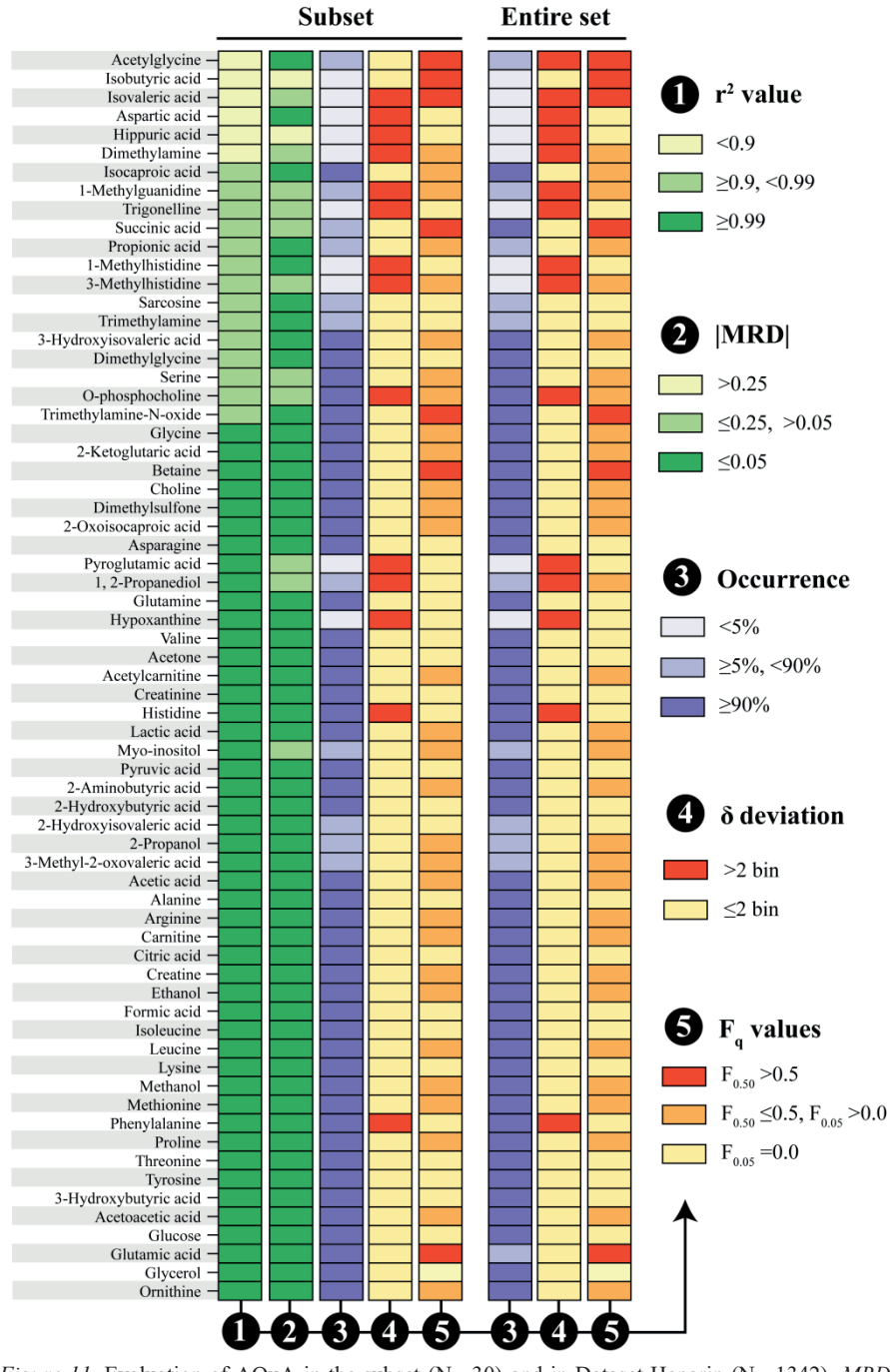

Figure 11. Evaluation of AQuA in the subset (N =30) and in Dataset Heparin (N =1342). *MRD* and r^2 values: derived by comparison of AQuA and targeted profiling. Quality indicators (*occurrence*, *positional deviation* and F_q values): computed from values generated by AQuA.

Most metabolites displayed limited *positional deviation* (within ±2 bins or ±0.0004 ppm) in the subset. Note that metabolites with low *occurrence* displayed some *positional deviation* (i.e., noise detection resulted in *positional deviation*). Few metabolites with high *occurrence* displayed *positional deviation* (e.g., histidine and phenylalanine). Inter-spectral positional deviations of ^{1}H NMR signals often owes to variation in pH and/or ionic strength between samples. But, such variations are limited in spectra from plasma compared to more complex spectra (e.g., from urine) (Emwas *et al.*, 2018; Bouatra *et al.*, 2013; Rist *et al.*, 2013).

Many metabolites displayed a low ($F_{0.05}$ =0) or intermediate ($F_{0.05}$ >0 and $F_{0.50}$ ≤0.5) degree of interference in the subset. Only a limited number of metabolites (e.g. betaine, glutamic acid, succinic acid and trimethylamine-N-oxide) displayed a high degree of interference ($F_{0.50}$ >0.5). Most of these metabolites displayed relatively low r^2 values considering the entire distribution. Hence, metabolites with high degree of interference may be somewhat more difficult to quantify with AQuA compared to metabolites with low or intermediate degree of interference. The sources of interference are listed in Table 4 for metabolites with $F_{0.05}$ >0 and >5% *occurrence*.

AQuA (Eq. 3) utilised one matrix for all computations. Due to this, the experimental signal positions (and line widths) were assumed to have, and modelled as if they displayed, limited inter-spectral deviations. This assumption may lead to quantitative errors e.g., if target signals are located in regions where both interference and inter-spectral deviations in signal positions are observed. In contrast, the manual procedure for targeted profiling allowed for adjustment of library positions. Metabolites with diminished quantitative accuracy, due to the assumption in AQuA regarding experimental signal positions, should thus be distinguished in the evaluation via: (1) a higher *MRD* and/or a lower r^2 as well as (2) intermediate (or high) degree of interference and (3) *positional deviation*. Yet, such metabolites were rarely identified. It was possible to identify metabolites with r^2 >0.99 and *MRD* within ±0.05 that displayed (1) *positional deviation*, but low degree of interference, (2) intermediate degree of interference, but limited *positional deviation* and (3) limited *positional deviation* and low degree of interference. This demonstrates that AQuA could accurately quantify metabolites that belong to one of the groups (1) to (3). The distributions of the quality indicators in the subset were highly similar to the distributions in the entire dataset (Dataset Heparin) (Figure 11). This implies that the level of accuracy is highly similar between the entire dataset and the subset (i.e., comparable with manual targeted profiling). Hence, via a suitable selection of target/reporter signals, error in AQuA computations may be avoided. The evaluation in paper I show that this

is possible in experimental spectra from plasma collected using heparin as anticoagulant.

Table 4. *Interfering ^{1}H NMR signals from different human plasma metabolites*

Target signal (δ ppm): other metabolite signals located within close proximity
Acetylglycine (δ 2.05): 2-hydroxyisovaleric acid, glutamic acid, isovaleric acid, proline, pyroglutamic acid
Isocaproic acid (δ 0.88): 2-hydroxybutyric acid, 3-methyl-2-oxovaleric acid
1-Methylguanidine (δ 2.83): aspartic acid, asparagine
Succinic acid (δ 2.41): 3-hydroxybutyric acid, glutamine, pyroglutamic acid
Propionic acid (δ 1.06): isobutyric acid, valine
3-Hydroxyisovalerate (δ 1.27): isoleucine, threonine
Serine (δ 3.97): 1-methylhistidine, 3-methylhistidine, hippuric acid
O-phosphocholine (δ 3.22): 3-methylhistidine, arginine, carnitine, choline, glucose, tyrosine
Trimethylamine-N-oxide (δ 3.27): arginine, betaine, glucose, histidine, myo-inositol, phenylalanine
Glycine (δ 3.57): 1, 2-propanediol, glucose, glycerol, myo-inositol, threonine
2-Ketoglutaric acid (δ 3.01): lysine
Betaine (δ 3.27): arginine, glucose, histidine, myo-inositol, phenylalanine, trimethylamine-N-oxide
Choline (δ 3.21): 1-methylhistidine, 3-methylhistidine, acetylcarnitine, arginine, o-phosphocholine, tyrosine
Dimethylsulfone (δ 3.16): 1-methylhistidine, histidine, phenylalanine
2-Oxoisocaproic acid (δ 0.95): isoleucine, leucine
Acetylcarnitine (δ 3.20): 1-methylhistidine, arginine, choline, o-phosphocholine, tyrosine
Lactic acid (δ 1.33): threonine
Myo-inositol (δ 4.07): choline, creatinine
2-Aminobutyric acid (δ 0.98): 2-hydroxyisovaleric acid, leucine, valine
2-Propanol (δ 1.18): 3-hydroxybutyric acid, ethanol
3-Methyl-2-oxovaleric acid (δ 0.90): 2-hydroxybutyric acid, isocaproic acid, isovaleric acid
Acetic acid (δ 1.92): 2-aminobutyric acid, arginine, lysine, ornithine
Arginine (δ 1.67): 2-hydroxybutyric acid, leucine
Carnitine (δ 3.23): 3-methylhistidine, arginine, glucose, histidine, o-phosphocholine, tyrosine
Creatine (δ 3.04): creatinine, lysine, ornithine, tyrosine
Ethanol (δ 1.19): 2-propanol, 3-hydroxybutyric acid
Leucine (δ 0.98): 2-aminobutyric acid, 2-hydroxyisovaleric acid, valine
Methanol (δ 3.37): proline
Methionine (δ 2.65): aspartic acid, citric acid
Proline (δ 4.15): 3-hydroxybutyric acid
Acetoacetic acid (δ 2.29): valine
Glutamic acid (δ 2.36): proline
Ornithine (δ 3.06): 1-methylhistidine, creatinine, lysine, tyrosine

3.2.2 Efficiency

Quantification with AQuA was completed in <1 second on a standard personal computer. This included the quantification of 67 human plasma metabolites in 1342 samples. Hence, AQuA computations were extremely rapid. The high efficiency of AQuA was thus demonstrated.

Although many metabolites display several characteristic ^{1}H NMR signals, it is sufficient to use one signal for its quantification (Bharti & Roy, 2012). AQuA is based on this principle. This principle does not only ensure rapid computations via data reduction, but also avoids processing of spectral regions e.g., where both interference and *positional deviation* occurs. This is utilised by AQuA to focus the quantitative process towards target regions where the selected signals used for quantification display favourable characteristics (e.g., high *S/N* ratios and small inter-spectral deviation in positions). In contrast, available alternatives for manual or automated quantification often consider all experimental signals (Ravanbakhsh *et al.*, 2015; Hao *et al.*, 2014; Weljie *et al.*, 2006). This dramatically increases the demands on computational efforts in automated procedures and makes manual procedures time-consuming. For example, quantification of 50 serum metabolites required about 5 minutes with BAYESIL, while BATMAN required about 13 minutes to quantify 24 metabolites (Ravanbakhsh *et al.*, 2015; Hao *et al.*, 2014). About 2 minutes was required to quantify 14 metabolites using the ChenomX software (Weljie *et al.*, 2006). However, a manual procedure becomes very time-consuming, especially when employed on large datasets (Schleif *et al.*, 2011). The superior efficiency of AQuA and the low computational burden required for the computations facilitates large-scale studies.

3.3 Conclusions

An Automated Quantification Algorithm (AQuA) was designed, implemented and evaluated for quantification of human plasma metabolites by NMR in samples collected using heparin as anticoagulant. The accuracy of AQuA was tested and confirmed by comparison with a manual procedure for targeted profiling. The results were in excellent agreement, but AQuA performed the quantifications at a superior rate. Due to its high efficiency, AQuA can facilitate large-scale studies.

Beyond the rapid generation of quantitative data on human plasma metabolites, AQuA generated information on NMR signal characteristics (e.g., detection limits, intensities, positions and interferences). This information was utilised to generate different quality indicators (*occurrence*, *positional deviation* and F_q values). The evaluation of the quality indicators yielded similar information as the comparison of AQuA with a manual procedure for targeted profiling. Hence, the quality indicators provide a way to evaluate the results from AQuA without the need for comparison to an independent quantification procedure. Accurate quantification with AQuA can be expected for all metabolites, except those that display low *occurrence*, high degree of interference ($F_{0.50}$ >0.5) or a combination of *positional deviation* and interference ($F_{0.05}$ >0).

4 Improved AQuA

In paper II, the automated quantification algorithm (AQuA) was modified to also handle inter-spectral deviations in signal positions and line widths. The modified AQuA was implemented for quantification of metabolites in human plasma samples collected using EDTA as anticoagulant. The results from the modified AQuA were evaluated using quality indicators (Table 3).

4.1 Methods

4.1.1 Design

The modification of AQuA can be employed to specific compounds that display inter-spectral deviations in signal positions and/or line widths. Compounds targeted for quantification may therefore be divided into two subgroups: (1) compounds that do not display inter-spectral deviations in signal positions or line widths and (2) compounds that display the aforementioned deviations.

For each compound in the second subgroup, its corresponding spectrum in the compound library is recreated as, and replaced by, a set of Lorentzian functions that match with the original library spectrum. The condition for each compound (i.e., its actual signal positions and line widths) is monitored in each experimental spectrum *n*. The values generated for each experimental spectrum *n* are used as input in the corresponding Lorentzian function to generate a new library spectrum *n*. The constant part of the library (originating from compounds in the first subgroup), alongside each generated library spectrum *n* (originating from compounds in the second subgroup), is then subjected to normalisation and data reduction as described in paper I. This generates a new matrix $\bar{\bar{\mathbf{m}}}_{\mathbf{n}}$ for each spectrum *n*. The extraction of target signals to each $\bar{\mathbf{y}}_{\mathbf{n}}$ vector is done in an identical manner as presented in paper I.

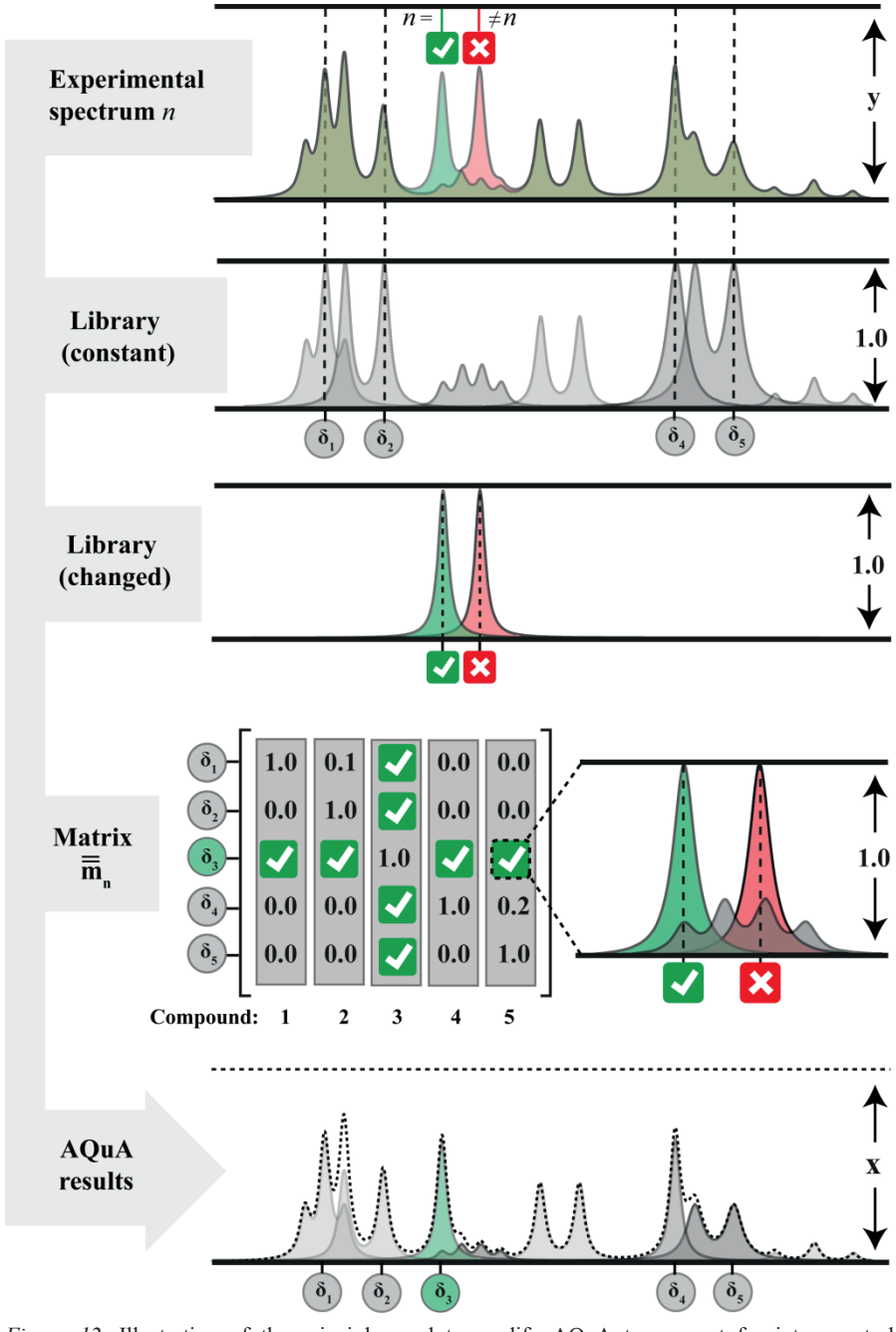

Figure 12. Illustration of the principle used to modify AQuA to account for inter-spectral deviations in signal positions and line widths. For compound *3*, the conditions in experimental spectrum *n* is monitored and used to create a corresponding library spectrum *n*. Normalisation and data reduction of the compound library (including library spectrum *n*) generates a matrix $\bar{\bar{\mathbf{m}}}_\mathbf{n}$ that is used in the AQuA computation (Eq. 5).

The AQuA computation is done by solving Eq. 5. This principle for modifying AQuA will account for inter-spectral deviations in signal positions and/or line widths for selected compounds.

$$\bar{\mathbf{y}}_{\mathbf{n}} = \bar{\bar{\mathbf{m}}}_{\mathbf{n}} \times \bar{\mathbf{x}}_{\mathbf{n}} \quad \text{(Eq. 5)}$$

The principle to modify AQuA is illustrated in Figure 12, using the same five putative compounds as in paper I. Here, the principle is illustrated for compound *3*, which only has one characteristic NMR signal. Its position (and linewidth) is determined in experimental spectrum *n*. This empirically derived position (and line width) is used as input in a Lorentzian function to generate a new library spectrum *n*. Upon data reduction and normalisation, this generates an optimised matrix $\bar{\bar{\mathbf{m}}}_{\mathbf{n}}$ that accurately models the (normalised) interferences in experimental spectrum *n*.

4.1.2 Implementation

In paper II, AQuA was implemented for quantification of human plasma metabolites in samples collected using EDTA as anticoagulant. Targeted NMR-based metabolomics was employed on 772 samples. Each sample was analysed with the workflow presented in Figure 10 (i.e., a similar workflow as employed to generate the experimental spectra in Dataset Heparin). This generated a dataset with 772 experimental spectra, which hereafter is referred to as Dataset EDTA. The original design of AQuA (Eq. 3) was implemented and all compounds included were human plasma metabolites. This AQuA was extended to also include non-metabolites, namely free EDTA (H-EDTA^{3-}) and two EDTA-complexes (Ca-EDTA^{2-} and Mg-EDTA^{2-}) (Barton *et al.*, 2010). The strategy to account for inter-spectral deviations in signal positions and line widths was employed on the (two) signals from free EDTA (Eq. 5). This modification was necessary since free EDTA signals displayed inter-spectral deviations in signal positions and line widths (see the Results and Discussion section further below).

4.1.3 Evaluation

Quality indicators (Table 3) were computed for each respective metabolite based on values inherently generated by the modified AQuA. The evaluation of quality indicators was not done for metabolites with <5% *occurrence* in Dataset EDTA. For the remaining metabolites, values below LOD were excluded (except when computing *occurrences*). Importantly, F_q values were

computed after separation into: (1) interference from other metabolite signals and (2) interference from EDTA signals, respectively.

4.2 Results and discussion

Metabolite quantification by NMR may be desired in plasma that contains EDTA since such samples are frequently collected and stored in clinical and epidemiological studies (Tuck *et al.*, 2009; Vaught *et al.*, 2009). However, EDTA generates several characteristic NMR signals due to its binding with different cations (e.g., H^+, Ca^{2+} and Mg^{2+}). These signals interfere with many different metabolite signals (e.g., acetylcarnitine, arginine, carnitine, choline, creatine, creatinine, glucose, glycerol, histidine, ornithine, phenylalanine, tyrosine and valine) (Barton *et al.*, 2010).

In paper II, targeted NMR-based metabolomics was employed on 772 plasma samples collected using EDTA as anticoagulant. This generated an experimental dataset (Dataset EDTA) used for quantification with AQuA. AQuA can facilitate quantification of metabolites in the presence of EDTA signals by inclusion of free EDTA (H-$EDTA^{3-}$) and EDTA metal ion complexes (Ca-$EDTA^{2-}$ and Mg-$EDTA^{2-}$) in the AQuA computations in a similar manner as the other set of compounds selected for quantification (here, human plasma metabolites). For each added compound, there will be a one unit increase in the dimension of the linear equation system used in AQuA computations (Eq. 3). If such implementation is done, then additional interferences between EDTA and metabolite signals will also be considered in each AQuA computation (beyond consideration of signal interferences between metabolites).

As discussed in paper I, the use of a constant matrix (Eq. 3) can be suboptimal for compounds that display interference and inter-spectral deviation in positions and/or line widths (however, only if these signals are located in a target region that is utilised in the AQuA computations). The EDTA signals were automatically monitored in Dataset EDTA to identify whether such issues were observed (Figure 13). It was found that the two high intensity signals from free EDTA displayed inter-spectral deviations in both signal positions and line widths. These deviations can be pH-dependent (Mónico *et al.*, 2017; Bharti & Roy, 2012; Barton *et al.*, 2010). Thus, despite the use of a buffer solution to reduce the variation in pH between samples, some deviations were still observed. Therefore, Dataset EDTA was an excellent test system to demonstrate the strategy for modifying AQuA to handle inter-spectral deviations in signal positions and line widths.

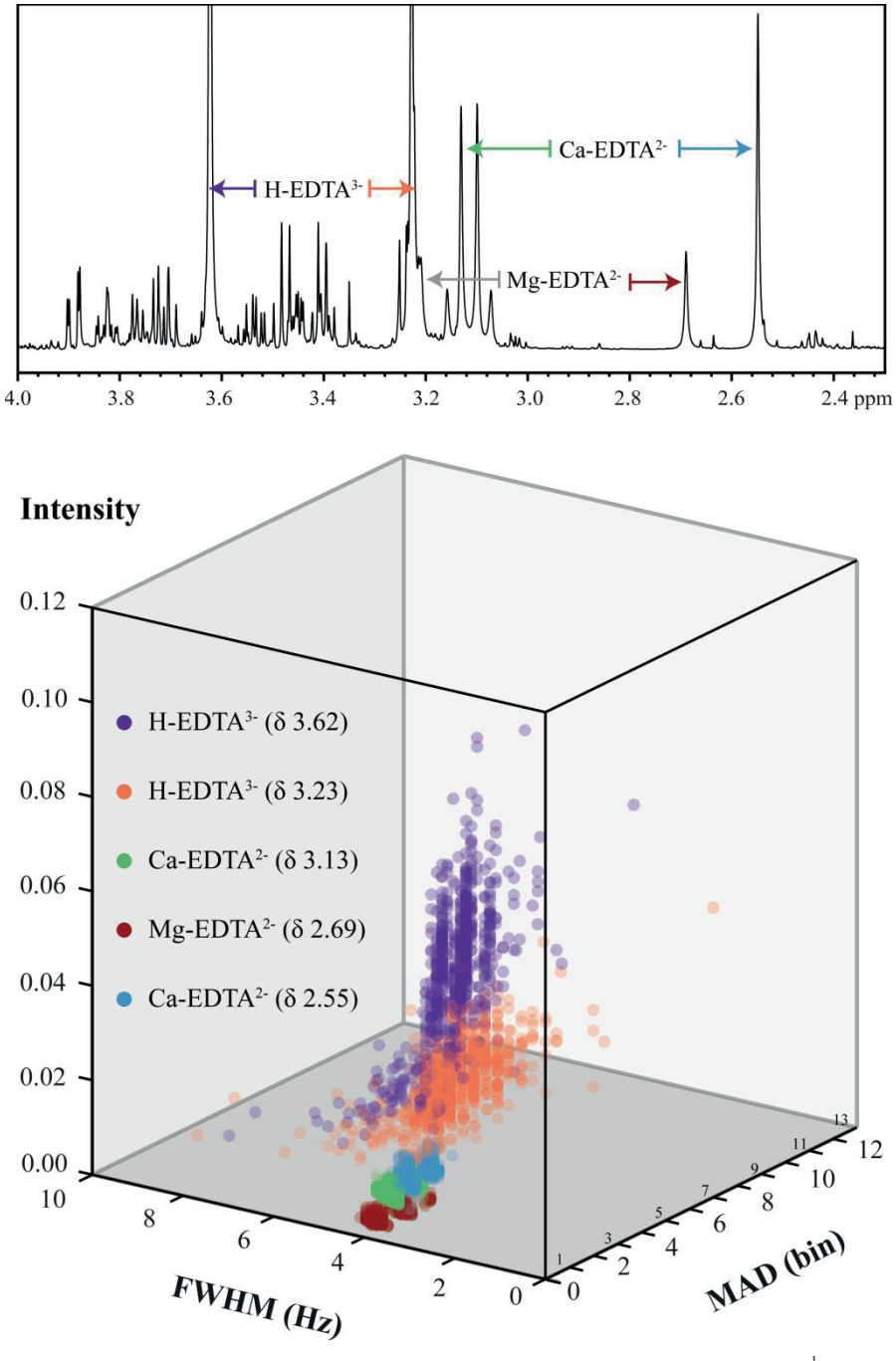

Figure 13. Anticoagulant signals identified in Dataset EDTA (N = 772). Top: Average ^{1}H-NMR spectrum. Bottom: heights, intensities and positions. MAD: Median absolute deviation. FWHM: Full width at half maximum. One signal from Mg-EDTA^{2-} is not presented in the bottom figure.

4.2.1 Accuracy

Quality indicators

Evaluation of the modified AQuA was done using the quality indicators: *occurrence*, *positional deviation* and F_q values (Figure 14). In paper II, F_q values were computed separately based on (1) interferences from other metabolites and (2) interferences from EDTA. Furthermore, interpretation of quality indicators in relation to quantitative accuracy was guided by the conclusions in paper I. In paper I, it was concluded that metabolites with low *occurrence*, a high degree interference ($F_{0.50}$ >0.50) or a combination of *positional deviation* and interference ($F_{0.05}$ >0 and $F_{0.50}$ ≤0.5), may be prone to quantitative errors when using AQuA.

It should be noted that the respective distributions of *occurrence*, *positional deviation* and the degree of interference (based on F_q values from metabolites) were highly similar in Dataset Heparin and Dataset EDTA (see Figure 11 and Figure 14). For example, most metabolites displayed high *occurrence* and *positional deviation* was limited. This indicates that the level of quantitative accuracy is comparable between the two datasets when only considering these three quality indicators. However, additional interferences caused by EDTA were identified via larger F_q values (from EDTA). This may lead to difficulties in quantifying those metabolites, but only if the degree of interference is high.

It was possible to identify a few metabolites with a high degree of interference (e.g., carnitine and dimethylsulfone), which may be difficult to quantify with the modified AQuA in plasma that contains EDTA. However, most metabolites, which were affected by additional interference from EDTA, still displayed an intermediate degree of interference. This implies that the modified AQuA can accurately quantify most metabolites affected by interference from EDTA signals.

4.2.2 Comparison between different AQuA implementations

In paper II, two AQuAs were implemented. A non-modified AQuA was implemented as described in paper I – i.e., only inter-metabolite signal interferences were accounted for. This AQuA was extended to account for the additional interferences caused by EDTA. This implementation also included handling the inter-spectral deviations in signal positions and line widths of free EDTA signals (modified AQuA). Both these implementations were employed on Dataset EDTA for comparative purposes.

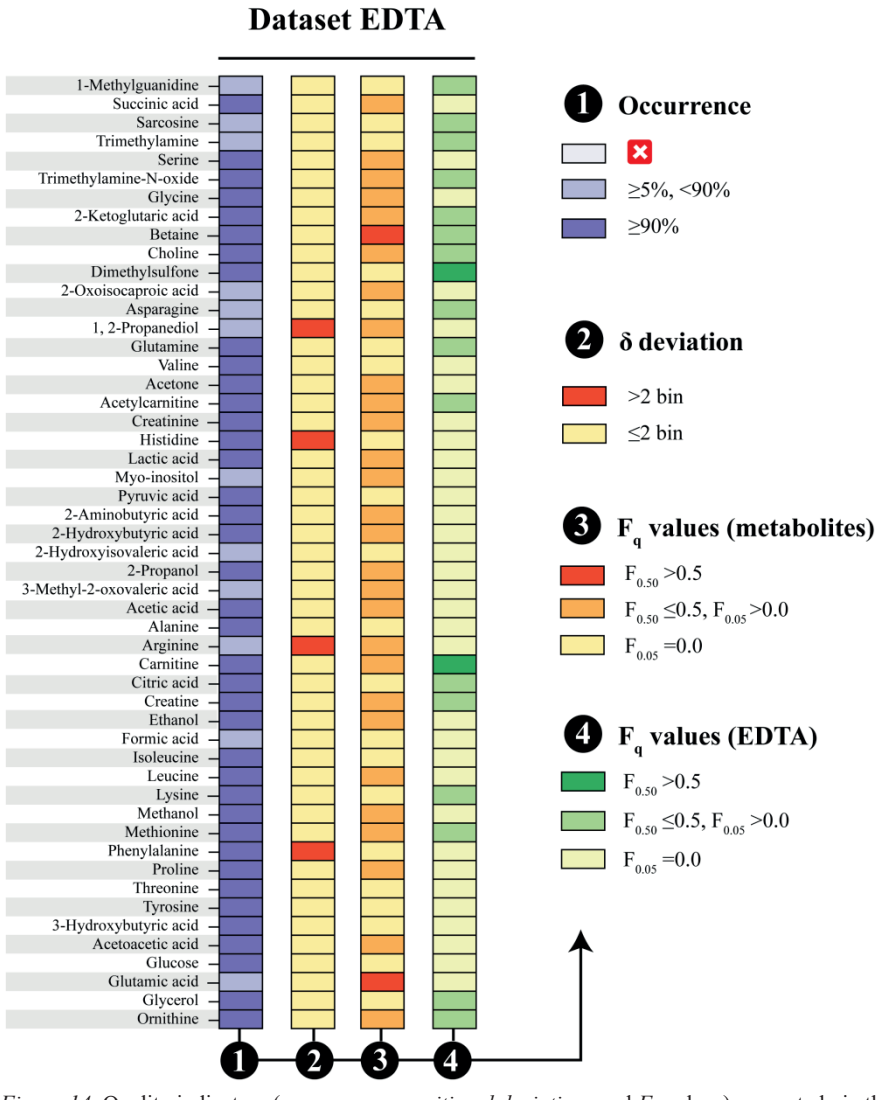

Figure 14. Quality indicators (*occurrences*, *positional deviations* and F_q values) generated via the modified AQuA when employed on Dataset EDTA. F_q values computed separately based on the interferences from other metabolites and from EDTA, respectively.

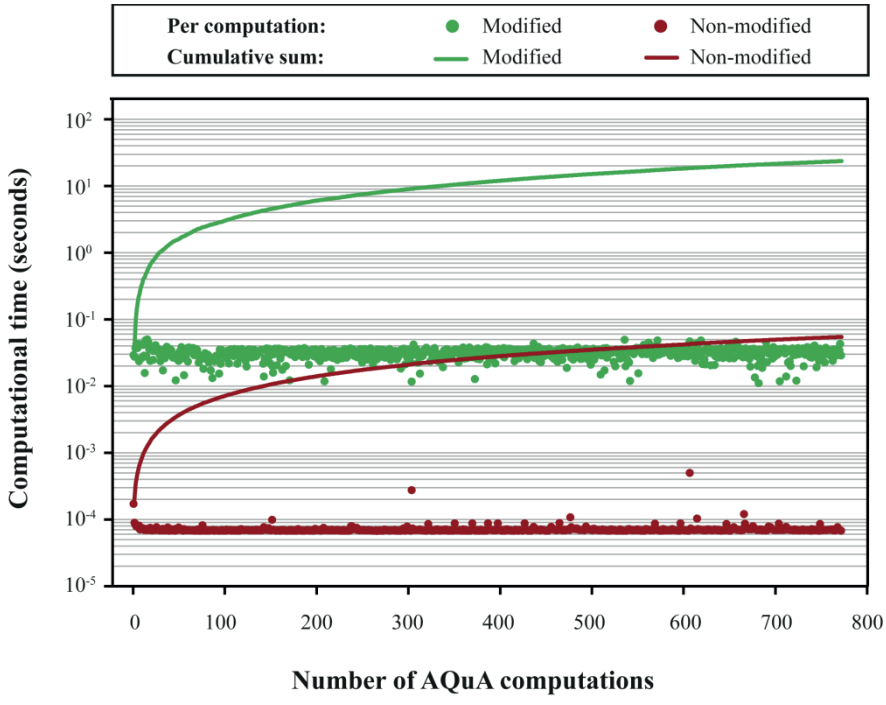

Figure 15. Time required for AQuA computations in dataset EDTA before and after its modification.

Efficiency

The efficiencies of the two implementations were compared (Figure 15). Quantification of human plasma metabolites with the modified AQuA (Eq. 5) required less than 50 ms per experimental spectrum. Decreased efficiency was observed compared to the non-modified AQuA. This can be explained by the generation of a new $\bar{\bar{\mathbf{m}}}_n$ matrix for each experimental spectrum *n*, compared to the use of one $\bar{\bar{\mathbf{m}}}$ matrix for all experimental spectra – i.e., it is a direct consequence of accounting for inter-spectral deviations in positions and line widths (here, for signals from free EDTA). Still, the modified AQuA performed extremely rapid, since all computations were done in <25 seconds on a standard personal computer. As mentioned in paper I, automated alternatives require at least five minutes to process one experimental spectrum (Ravanbakhsh *et al.*, 2015; Hao *et al.*, 2014).

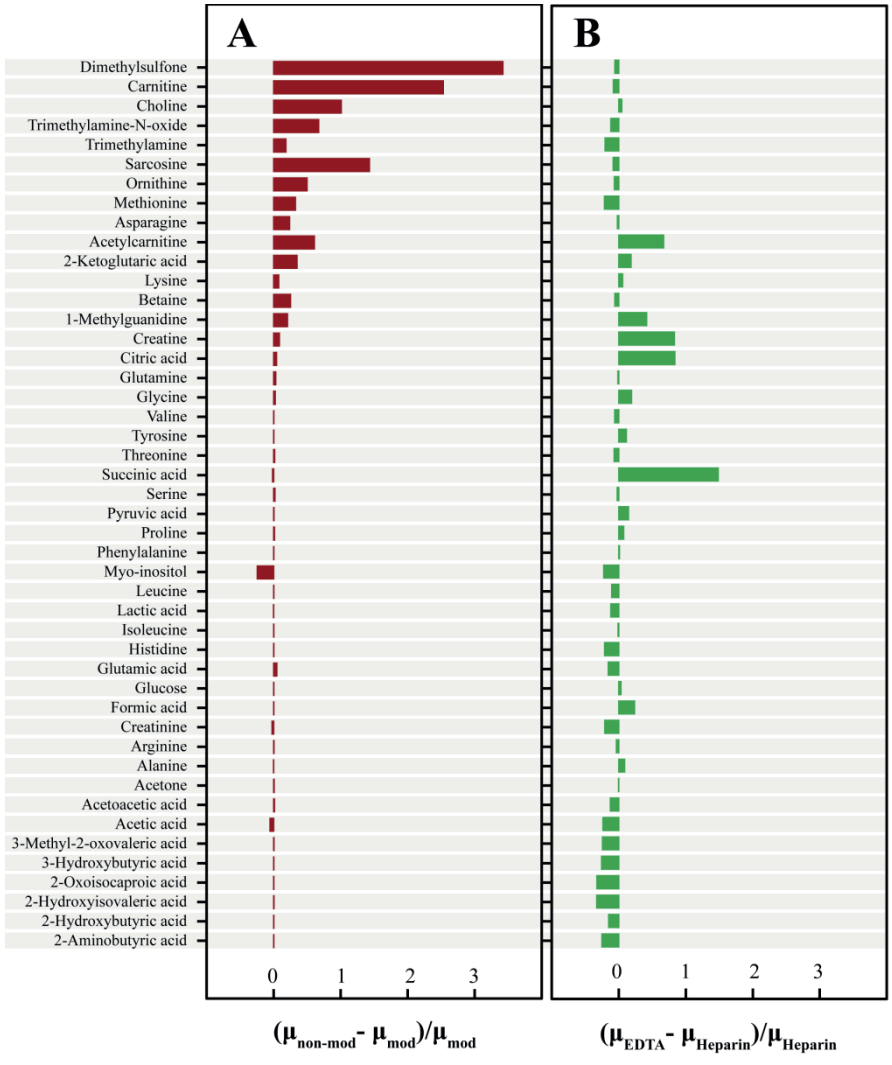

Figure 16. Comparison of mean sample concentrations (μ) generated for each respective metabolite via different AQuA implementations. (A) Calculated values of ($\mu_{non\text{-}mod}$ - μ_{mod})/μ_{mod}, where μ_{mod} is the results from the non-modified AQuA and μ_{mod} is the results from the modified AQuA, when employed on Dataset EDTA. (B) Calculated values of (μ_{EDTA} - $\mu_{Heparin}$)/$\mu_{Heparin}$, where μ_{EDTA} the identical to μ_{mod} in (A) and $\mu_{Heparin}$ is the results from the (non-modified) AQuA when employed on Dataset Heparin.

Mean sample concentrations

For each metabolite, the mean sample concentration (μ) derived from each respective implementation was used to calculate ($\mu_{non\text{-}mod}$ - μ_{mod})/μ_{mod}, where $\mu_{non\text{-}mod}$ is the results from the non-modified AQuA and μ_{mod} is the results from the modified AQuA when employed on Dataset EDTA. This comparison showed that the non-modified AQuA overestimated the concentrations of metabolites that were affected by interference from EDTA (Figure 16A).

The results from the modified AQuA when employed on Dataset EDTA (μ_{mod} in Figure 16A or μ_{EDTA} in Figure 16B) were also compared with the results from paper I ($\mu_{Heparin}$), where the non-modified AQuA was employed on Dataset Heparin. This comparison was done by calculating (μ_{EDTA} - $\mu_{Heparin}$)/$\mu_{Heparin}$ (Figure 16B). As can be seen in Figure 16, the size of the overestimations, due to EDTA signals being ignored, often exceeded the difference between the datasets – i.e., differences between two target populations. Hence, using the modified AQuA is clearly required to yield more accurate concentration estimates. Alcohols are not shown in Figure 16 due to sample preservation issues, which can results in unreliable alcohol concentrations (Psychogios *et al.*, 2011).

4.3 Conclusions

Paper II introduced a strategy for modifying AQuA to handle inter-spectral deviations in NMR signal positions and line widths. The modified AQuA was implemented for quantification of human plasma metabolites in samples collected using EDTA as anticoagulant. EDTA-containing plasma was a good test system for this modification since NMR signals from EDTA displayed inter-spectral deviations in both position and line widths. Evaluation of quality indicators showed that the modified AQuA generated accurate concentrations of most human plasma metabolites. The principle for modification of AQuA may be implemented on other datasets where inter-spectral deviation in signal positions and line widths are observed. Due to its rapid performance, the modified AQuA is suitable for large-scale studies.

5 Identification of disease risk biomarkers

In paper III, a nested case-control study was employed within the NSHDC to identify risk biomarkers for prostate cancer among plasma metabolites measured with targeted NMR and MS-based metabolomics. Importantly, AQuA (paper I) facilitated the quantitative process of human plasma metabolites measured with NMR.

5.1 Methods

5.1.1 Study design

Details on the NSHDC have been presented elsewhere (Norberg *et al.*, 2010; Johansson *et al.*, 2002). Participants were enrolled to the NSHDC at 40, 50 or 60 years of age. Enrolled subjects underwent a baseline examination that included: (1) performing of an OGTT, (2) answering validated food-frequency questionnaire (FFQ) and (3) donating a blood sample after overnight fasting. The samples were stored at -80^{o}C in plasma aliquots.

A total number of 777 case-controls were selected for the nested case-control study in paper III. Inclusion criteria for cases were: (1) no T2D at baseline, (2) at least 5 years between baseline and the diagnosis with prostate cancer, (3) no previous incidence of cancer. Each case was matched with one healthy control based on age, BMI (body mass index) and sample storage time. Heparin was used as anticoagulant in the plasma samples. This study, including metabolomics analysis of the plasma samples, was approved by the Regional Ethical Review Board (2013/124) (Uppsala, Sweden).

5.1.2 Targeted metabolomics

The plasma samples were analysed with targeted metabolomics. Quality control (QC) samples were included in the metabolomics workflow to assess the analytical coefficient of variation (CV). For targeted NMR-based metabolomics, the method presented in paper I was employed. 1342 plasma samples were included in both paper I and paper III. For targeted MS-based metabolomics, the AbsoluteIDQ p180 assay (BIOCRATES, Innsbruck, Austria) was used. A total number of 188 human plasma metabolites were quantified. This included 40 acylcarnitines (Cx:y), 21 amino acids, 20 biogenic amines, 38 diacyl phosphatidylcholines (PC aa Cx:y), 39 acyl-alkyl phosphatidylcholines (PC ae Cx:y), 14 lysophosphatidylcholines (LPC Cx:y), 5 hydroxysphingomyelins (SM-OH Cx:y), 10 sphingomyelins (SM Cx:y) and one hexose (sum of several isomers including glucose). In Cx:y, x corresponds to the number of carbons, and y the number of double bonds in the fatty acid chain(s). The methodology has been presented in detail elsewhere and it has been used for similar study designs (Schmidt *et al.*, 2017; Kühn *et al.*, 2016; Bogmuil *et al.*, 2008).

5.1.3 Statistical analyses

Quality control was employed on all metabolites measured to identify plasma metabolites that displayed relatively low occurrence (<50%) and/or relatively large coefficient of variation in the QC samples (CV_{QC} >15%). These metabolites were excluded from statistical analyses. The association between the baseline plasma level of each individual metabolite and risk of prostate cancer was investigated using conditional logistic regression (log2 transformed metabolite data) conditioned on matching factors. Statistical analysis generated the OR (95% CI) and the corresponding p-value for each association. Correction for multiple testing was done at two levels of stringency. The highly conservative Bonferroni correction was used as well as a less stringent FDR approach. In the FDR approach, the level of significance was set at 20% based on q-values generated from the p-values (Storey & Tibshirani, 2003). Statistical analyses were repeated after stratification by baseline age (40–50 years, 60 years) and disease aggressiveness (non-aggressive, aggressive).

Aggressive prostate cancer was defined as: (1) poorly differentiated tumour (WHO grade 3 or Gleason score 8-10), (2) non-localised tumour T3-4, (3) lymph node metastasis (N1), (4) bone metastasis (M1) or (5) serum prostate specific antigen (PSA) concentration >50 ng/mL. Cases with fatal prostate cancer (until March 2007) were also included in the group of aggressive cases.

Cases that did not qualify as having an aggressive disease were included in the non-aggressive group.

5.2 Results

5.2.1 Metabolites measured by NMR

Table 5 lists potential risk biomarkers for prostate cancer among metabolites measured with NMR. Although several metabolites were significantly associated with risk of prostate cancer with a nominal p-value below 0.05 ($p < 0.05$), only two associations remained significant after correction for multiple testing (FDR, 20%). Higher concentrations of glycine associated with elevated risk of overall prostate cancer in younger subjects, while lower concentrations of pyruvic acid associated with elevated risk of overall prostate cancer in younger subjects. The evaluation of quality indicators in paper I showed that these metabolites could be accurately quantified with AQuA in samples collected using heparin as anticoagulant (Figure 11).

Table 5. *Potential risk biomarkers for prostate cancer measured by NMR* [a, b]

Disease group	*Age group*	*Potential risk biomarkers*
Overall	40 – 60 years	Pyruvic acid (↓)
	40 – 50 years	**Glycine*** (↑), **pyruvic acid*** (↓)
	60 years	Glutamine (↓), histidine (↓)
Non-aggressive	40 – 60 years	Pyruvic acid (↓)
	40 – 50 years	Glycine (↑), pyruvic acid (↓)
	60 years	Choline (↑)
Aggressive	40 – 60 years	Glutamine (↓), histidine (↓), ornithine (↓)
	40 – 50 years	-
	60 years	Carnitine (↓), glutamine (↓), histidine (↓)

[a] Metabolites significant after correction for multiple testing (FRD, 20%) are indicated in bold and with (*).
[b] The direction of each association is indicated with an arrow. OR >1: ↑, OR <1: ↓.

5.2.2 Metabolites measured by MS

Table 6 lists potential risk biomarkers for prostate cancer among the metabolites measured with targeted MS-based metabolomics. Higher plasma concentrations of LPC C17:0 and LPC C18:0 associated with elevated risk of overall prostate cancer. These findings were observed in the subgroup that included all subjects as well as the subgroup with older subjects. Furthermore, higher plasma concentrations of PC ae C38:3, PC ae C38:4, PC ae 40:2, LPC

C17:0, LPC C20:3 and LPC C20:4 associated with elevated risk of aggressive prostate cancer. Similar findings were identified for LPC C17:0 in older cases with aggressive disease and their matched controls. These aforementioned findings were significant after correction for multiple testing (FDR, 20%). Importantly, the association between LPC C17:0 and elevated risk of overall prostate cancer was also significant after Bonferroni correction in the subgroup of older subjects.

Table 6. *Potential risk biomarkers for prostate cancer measured by MS* [a, b, c]

Disease group	*Age group*	*Identified risk biomarkers*
Overall	40 – 60 years	C3, LPC C(16:0, **17:0***, **18:0***, 18:1, 20:4), PC aa C38:5, PC ae C(36:5, 38:4, 40:1, 40:5) (↑)
	40 – 50 years	PC ae C40:5 (↑) & arginine (↓)
	60 years	C3, LPC C(16:0, **17:0****, **18:0***, 18:1, 20:4), PC ae C(36:2, 40:1), taurine (↑) & ornithine (↓)
Non-aggressive	40 – 60 years	LPC C(17:0, 18:0), PC ae C(40:1, 40:3) (↑) & arginine (↓)
	40 – 50 years	Arginine (↓)
	60 years	C3, LPC C17:0, LPC C18:0, PC ae C40:0, PC ae C40:1 (↑)
Aggressive	40 – 60 years	LPC C(16:0, **17:0***, 18:0, **20:3***, **20:4***, PC aa C40:4, PC ae C(36:1, 36:2, **38:3***, **38:4***, **40:2***) (↑) & **C18:2** (↓)
	40 –50 years	C3, C4, C5, PC ae (40:6, 42:2) (↑)
	60 years	LPC C(**17:0***, 20:3, 20:4), PC ae C(36:1, 38:3, 38:4, 40:2) (↑) & C18:2 (↓)

[a] Metabolites significant after correction for multiple testing (FRD, 20%) are indicated in **bold** and with (*).
[b] Metabolites significant after correction for multiple testing (Bonferroni) are indicated in **bold** and with (**).
[c] The direction of each association is indicated with an arrow. OR >1: ↑, OR <1: ↓.

5.2.3 Additional findings for lysophosphatidylcholines

A majority of potential risk biomarkers identified in paper III were LPCs (Table 6). In paper III, it was shown that the LPCs could be divided into two clusters, where LPCs that belonged to the same cluster displayed high inter-sample correlations and where LPCs from different clusters displayed low inter-sample correlation. The first cluster includes LPCs with ≤20 carbons in the fatty acid chain and the second cluster included LPCs with >20 carbons in the fatty acid chain. Note that LPC C17:0, LPC C18:0, LPC C20:3 and LPC C20:4 associated with risk of prostate cancer and these metabolites belonged to the first cluster. It was also shown in Paper III that adjusting for the total sum plasma level of LPCs with ≤20 carbons weakened the associations between individual LPCs and the risk of prostate cancer. However, associations between LPC C17:0 and risk of prostate cancer were still observed after this adjustment. These results distinguished LPC C17:0 from the other risk biomarker LPCs identified in paper III. Potential determinants for LPC C17:0 were sought. A low to moderate correlation was observed in paper III between plasma LPC C17:0 and corresponding dietary fatty acid ($r = 0.19$, $p < 0.0001$). This suggests that plasma concentration of LPC 17:0 can, to some extent, be influenced by consumption of food items that are high in fatty acid C17:0.

Paper III also showed that LPCs with ≤20 carbons displayed negative correlations with glucose values measured at baseline during an OGTT. The negative correlations observed between LPCs and glucose values from the OGTT were stronger for post-load glucose values (2h) compared to fasting glucose values (0h). LPC C17:0 displayed the strongest correlation with post-load glucose values ($r = -0.20$, $p < 0.0005$).

The results in paper III also revealed that IGT associated with lower risk of prostate cancer. In paper III, the statistical analysis was repeated for the identified risk biomarkers for prostate cancer after limiting to case-controls pairs with NGT at baseline. The results showed that the associations often appeared stronger after this stratification (for lipid species and with overall prostate cancer risk). This may imply that baseline IGT obscures the metabotype associated with risk of prostate cancer.

5.3 Discussion

5.3.1 Strengths and limitations

The study in paper III has several advantages such as the large sample size, the entirely fasting samples and the long follow-up between baseline and diagnosis with prostate cancer (>5 years; this only applies to the cases). The advantage of fulfilling these criteria was mentioned in the introduction (see section 1.3.5 Risk biomarkers for prostate cancer).

Furthermore, paper III assessed whether association between metabolites and prostate cancer risk varied with baseline age, which was not done in previous studies (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014).

Additionally, two analytical methods (NMR and MS) were employed in paper III to yield complementary information on metabolites. Previous studies have only used MS (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014). Importantly, the use of AQuA facilitated accurate and rapid quantification of plasma metabolites measured by NMR in paper III. Existing alternative, such as manual targeted profiling, would have been extremely time-consuming to employ on such a large set of samples (Weljie *et al.*, 2006).

Participant enrolled to the NSHDC also underwent an OGTT and answered an FFQ at baseline, which enabled the unique possibility for further investigations – e.g., correlation analyses between metabolites and OGTT and FFQ-results (Norberg *et al.*, 2010).

The study in paper III also has some limitations. For example, due to several stratifications, the sample size was somewhat limited in specific subgroups (e.g., older cases with aggressive disease and their matched controls), which results in uncertain risk estimates. Furthermore, the results may be difficult to generalise since the study population only included individuals from one geographical region (northern Sweden).

5.3.2 Comparison with previous studies

Previous studies of similar design did not report pyruvic acid and glycine as potential risk biomarkers for prostate cancer (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014). However, these previous studies did not employ targeted NMR-based metabolomics, which was used to identify these risk biomarker candidates in paper III.

The finding for acylcarnitine C18:2 was in agreement with a previous study, where lower concentrations of acylcarnitine C18:2 associated with elevated risk of advanced stage prostate cancer. Similar design and MS methodology were used in this previous study; however, the study population included individuals from several European countries (Schmidt *et al.*, 2017). This previous study also reported that lower concentrations of some PCs associated with elevated risk of advanced stage disease. This may seem contradictory to some of the findings for PCs in paper III. However, the associations in the previous study were typically observed for different diacyl PCs, which were not identified in paper III (Schmidt *et al.*, 2017).

The findings for LPC C17:0 have not been reported in previous studies (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014). However, lower concentrations of LPC C18:0 associated with elevated risk of overall and advanced stage prostate cancer in previous studies (Schmidt *et al.*, 2017; Kühn *et al.*, 2016). This is somewhat contradictory to the findings in paper III. In contrast to paper III, these studies included some cases with relatively short follow-up until prostate cancer diagnosis (<5 years). Short follow-up may result in inclusion of cases with subclinical disease at baseline. It can be speculated that the metabotype for diagnosed prostate cancer may obscure the metabotype for prostate cancer risk. This may hamper the statistical outcome for specific metabolites that are known to differ between prostate cancer patients compared to healthy controls (such as LPC C18:0) (Zhou *et al.*, 2012; Lokhov *et al.*, 2010; Osl *et al.*, 2008). Stratified analysis was done with regards to follow-up time in one of these previous studies, and LPC C18:0 was not significantly associated with risk of overall prostate cancer in the subgroup of cases with >5 years until diagnosis and their matched controls (Schmidt *et al.*, 2017).

5.3.3 Dairy consumption and risk biomarkers for prostate cancer

The main findings in paper III regarded LPC C17:0. The key sources of odd-chain fatty acids are endogenous production, gut microbiota and consumption of specific food items (e.g., dairy products) (Crown *et al*., 2015; Jenkins *et al.*, 2015). LPC C17:0 has been suggested as a biomarker for dairy consumption in humans (Nestel *et al.*, 2014). The results in paper III can support LPC C17:0 as a potential dietary marker. In line with this observation, high consumption of dairy products has been suggested as a risk factor for prostate cancer (with limited evidence) (WCRF/AICR, 2018).

Interestingly, there is a relationship between higher consumption of dairy products and elevated circulatory levels of IGF-I, a modifiable risk factor for

prostate cancer (Travis *et al.*, 2016; Qin *et al.*, 2009). Notably, a previous study in the NSHDC identified an association between higher circulatory levels of IGF-I and elevated risk of prostate cancer (Stattin *et al.*, 2004). Furthermore, a previous study has shown that plasma lipids, which contain saturated odd-chain FAs, display positive correlation with IGF-I levels in men (Knacke *et al.*, 2016). Hence, risk biomarkers for prostate cancer identified in the same cohort (i.e., IGF-I and LPC C17:0) may be linked via consumption of dairy.

Previous nested case-control studies on prostate cancer have not reported LPC C17:0 as a potential risk biomarker for prostate cancer (Schmidt *et al.*, 2017; Huang *et al.*, 2016; Kühn *et al.*, 2016; Mondul *et al.*, 2015; Mondul *et al.*, 2014). Under the assumption that dairy consumption influence the metabotype associated with elevated risk of prostate cancer, such risk biomarkers are likely to be revealed in a population where the consumption of dairy is high. In agreement with this assumption, a previous study reported that individuals from northern Sweden consume more dairy products compared to individuals from many other European countries (Hjartåker *et al.*, 2002). However, more research is required to draw conclusions.

5.3.4 Glucose intolerance and risk biomarkers for prostate cancer

The use of saturated fatty acid chain with seventeen carbons as a marker for dairy consumption has been questioned. Instead, it has been shown that biosynthesis may be the key source of fatty acid C17:0, while other odd-chain fatty acids (e.g., C15:0) may reflect dairy consumption more accurately. Fatty acid C17:0 has been distinguished from C15:0 via its link to glucose intolerance (Jenkins *et al.*, 2017). Also, lower level of LPC C17:0 associated with elevated risk of impaired glucose tolerance in a previous study as well as elevated risk of T2D in the NSHDC (Shi *et al.*, 2018; Wang-Sattler *et al.*, 2012). Interestingly, lower circulatory levels of IGF-I also associate with elevated risk of IGT (Sandhu *et al.*, 2002). Furthermore, T2D (as well as insulin resistance in the NSHDC) associated with lower risk of prostate cancer (Bansal *et al.*, 2013; Stocks *et al.*, 2007). Hence, risk biomarkers for prostate cancer identified in the same cohort (i.e., IGF-I and LPC C17:0) seems to be linked, even without the consideration of the aforementioned link to dairy consumption, via impaired glucose metabolism.

5.3.5 PI3K/AKT signalling

The downstream signalling pathway for IGF-I (PI3K/AKT) has implications in both prostate cancer and T2D pathophysiology. For example, enhanced PI3K/AKT signalling promotes prostate cancer, while inactivation can promote T2D (Huang *et al.*, 2018; Yue *et al.*, 2014). Activation of the downstream signalling of the PI3K/AKT pathway can affect the level of plasma lipids via enhanced cellular uptake of lipoproteins (Yue *et al.*, 2014). It can be speculated that enhanced PI3K/AKT signalling may result in increased plasma levels of LPCs.

Several metabolites that were identified as potential risk biomarkers for prostate cancer in paper III (e.g., LPC C17:0, LPC C18:0 and glycine) have also been linked with IGT and T2D risk, but with associations of opposite directions (Wang-Sattler *et al.*, 2012). The results in paper III appear to reveal the reverse cross-association between the risk of these two diseases (prostate cancer and T2D) at the metabolite level.

5.4 Conclusions and future studies

Several risk biomarker candidates for prostate cancer could be identified in a case-control study nested within the NSHDC. A reverse cross-association between risk of prostate cancer and T2D was revealed on the metabolite level. The findings should be validated in an independent cohort. Although a potential role for dairy consumption was distinguished, further research is needed to draw final conclusions. One step can be to target specific lipids that contain different odd-chain fatty acids (e.g., C15:0 and C17:0) for additional metabolomics analysis.

6 Concluding remarks and future perspectives

Targeted NMR-based metabolomics is most useful when being both rapid and accurate. A key bottleneck in the workflow is the quantification step. An automated Quantification Algorithm (AQuA) was introduced in response to this. The successful use of AQuA for rapid and accurate metabolite quantification was demonstrated in NMR spectra from human plasma samples collected using different anticoagulants (paper I and II). The use of AQuA was further demonstrated in the context of biomarker discovery (paper III).

AQuA has a great potential for use beyond the application to plasma presented in the current thesis work. A natural step would be to implement AQuA for other biofluid samples (e.g., urine), which generate more complex NMR spectra where inter-spectral deviations in signal positions and line widths occur more frequently.

Further development of AQuA is required, such as a more universal application of its principles in order to move away from conscious decision-making and move towards automation. The use of AQuA principles is not restricted to NMR signals, but can also be applied on quantitative signals generated by other analytical methods.

Furthermore, a broader use of AQuA would be facilitated by increased user-friendliness. I envision the development of an AQuA software package with a graphical user interface, which is dedicated to automated quantification of compounds.

References

Ala-Korpela, M. (1995). 1H NMR spectroscopy of human blood plasma. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 27(5-6), pp. 475-554.

Akoka, S., Barantin, L. & Trierweiler, M. (1999). Concentration measurement by proton NMR using the ERETIC method. *Analytical Chemistry*, 71(13), pp. 2554-2557.

Alum, M.F., Shaw, P.A., Sweatman, B.C., Ubhi, B.K., Haselden, J.N. & Connor, S.C. (2008). 4,4-Dimethyl-4-silapentane-1-ammonium trifluoroacetate (DSA), a promising universal internal standard for NMR-based metabolic profiling studies of biofluids, including blood plasma and serum. *Metabolomics*, 4(2), pp. 122-127.

Alonso, A., Marsal, S., Julià, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3, 23.

Bharti, S.K. & Roy, R. (2012). Quantitative 1H NMR spectroscopy. *TrAC Trends in Analytical Chemistry*, 35, pp. 5-26.

Bansal, D., Bahnsali, A., Kapil, G., Undela, K. & Tiwari P. (2013) Type 2 diabetes and risk of prostate cancer: a meta-analysis of observational studies. *Prostate Cancer and Prostatic Diseases*, 16(2), pp. 151-158.

Barton, R.H., Waterman, D., Bonner, F.W., Holmes, E., Clarke, R., Procardis Consortium, Nicholson, J.K. & Lindon, J.C. (2010). The influence of EDTA and citrate anticoagulant addition to human plasma on information recovery from NMR-based metabolic profiling studies. *Molecular bioSystems*, 6(1), pp. 215-224.

Becker, E.D. (1993). A brief history of nuclear magnetic resonance. *Analytical Chemistry*, 65(6), pp. 295-302.

Belbasis, L. & Bellou, V. (2018). 'Introduction to epidemiological studies', in Evangelou, E. *Genetic Epidemiology*. NY: Humana Press, pp. 1-6.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafő, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J. & Johnson, V.E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, pp. 6-10.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), pp. 289-300.

Bewick, V., Cheek, L. & Ball, J. (2005). Statistical review 14: logistic regression. *Critical Care*, 9(1), pp. 112-118.

Bingol, K. (2018). Recent advances in targeted and untargeted metabolomics by NMR and MS/NMR methods. *High-throughput*, 7(2), 9.

Bogmuil, R., Koal, T., Weinberger, K.M. & Dammeier, S. (2008). Targeted metabolomics: fast, standardized mass spectrometric analysis of blood plasma with a kit. *Laborwelt*, 2, pp. 17-23.

Bouatra, S., Aziat, F., Mandal, R., Guo, A.C., Wilson, M.R., Knox, C., Bjorndahl, T.C., Krishnamurthy, R., Saleem, F., Liu, P., Dame, Z.T., Poelzer, J., Huynh, J., Yallou, F.S., Psychogios, N., Dong, E., Bogumil, R., Roehring, C. & Wishart, D.S. (2013). The human urine metabolome. *PLoS One*, 8(9), e73076.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, in press. Available at: http://gco.iarc.fr/

Breslow, N.E. & Day, N.E. (1980) 'General considerations for the analysis of case-control studies', in *Statistical Methods in Cancer Research*, Lyon: IARC, pp. 94-99.

Broadhurst, D.I. & Kell, D.B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4), pp. 171-196.

Califf, R.M. (2018). Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3), pp. 213-221.

Carayol, M., Licaj, I., Achaintre, D., Sacerdote, C., Vineis, P., Key, T.J., Onland Moret, N.C., Scalbert, A., Rinaldi, S. & 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(8), e0135437.

Carbajo, R.J. & Neira, J.L. (2003). 'Spectroscopic parameters in nuclear magnetic resonance', in *NMR for Chemists and Biologists*, Dordrecht: Springer, pp. 39-42.

Casu, B., Naggi, A. & Torri, G. (2015). Re-visiting the structure of heparin. *Carbohydrate Research*, 403, pp. 60-68.

Chen, L., Weng, Z., Goh, L. & Garland, M. (2002). An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1-2), pp. 164-168.

Claridge, T.D.W. (1999). 'Introducing high-resolution NMR', in *High-Resolution NMR Techniques in Organic Chemistry*, Amsterdam: Elsevier, pp. 13-24.

Cobas, J.C., Bernstein, M.A., Martín-Pastor, M. & Tahoces, P.G. (2006). A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of Magnetic Resonance*, 183(1), pp. 145-151.

Crown, S.B., Marze, N. & Antoniewicz, M.R. (2015). Catabolism of branched chain amino acids contributes significantly to synthesis of odd-chain and even-chain fatty acids in 3T3-L1 adipocytes. *PLoS One*, 10(12), e0145850.

Daykin, C.A., Foxall, P.J., Connor, S.C., Lindon, J.C. & Nicholson J.K. (2002). The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Analytical Biochemistry*, 304(2), pp. 220-230.

Delgado-Rodríguez, M. & Llorca, J. (2004). Bias. *Journal of Epidemiology and Community Health*, 58(8), pp. 635-641.

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., Wareham, N.J. & Boeing, H. (2015). Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case control study. *Clinical chemistry*. 61(3), pp. 487-497.

Dunn, W.B. & Ellis, D.I. (2005) Metabolomics: current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24(4), pp. 285-294.

Elliott, K.C., Cheruvelil, K.S., Montgomery, G.M. & Soranno, P.A. (2016). Conceptions of good science in our data-rich world. *Bioscience*, 66(10), pp. 880-889.

Emwas, A.H., Saccenti, E., Gao, X., McKay, R.T., Dos Santos, V., Roy, R. & Wishart, D.S. (2018). Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, 14(3), 31.

Ernster, V.L. (1994). Nested case-control studies. *Preventive Medicine*, 23(5), pp. 587-590.

FDA-NIH Biomarker Working Group (2016). *BEST (Biomarkers, EndpointS, and other Tools*). S.I: Food and Drug Administration (US), pp. 4-16.

Freeman, R. (1988a). 'Free induction decay', in *A handbook of nuclear magnetic resonance*, Harlow: Longman, pp. 87-88.

Freeman, R. (1988b). 'Zero filling', in *A handbook of nuclear magnetic resonance*, Harlow: Longman, pp. 302-305.

Friebolin, H. (1991a). 'The physical basis of NMR spectroscopy', in *Basic One- and Two-Dimensional NMR Spectroscopy*, Weinheim: VHC, pp. 1-35.

Friebolin, H. (1991b). 'The chemical shift', in *Basic One- and Two-Dimensional NMR Spectroscopy*, Weinheim: VHC, pp. 37-40.

García-Closas, M., Vermeulen, R., Cox, D., Lan, Q., Caporaso, N.E. & Rothman, N. (2011). Population-based study designs in molecular epidemiology. *IARC Scientific Publications*, (163), pp. 241-259.

German, J.B., Hammock, B.D. & Watkins, S.M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, 1(1), pp. 3-9.

Gonzalez-Covarrubias, V., Dane, A., Hankemeier, T. & Vreeken, R.J. (2013). The influence of citrate, EDTA, and heparin anticoagulants to human plasma LC-MS lipidomic profiling. *Metabolomics*, 9(2), pp. 337-348.

Goodacre, R. (2005). Metabolomics – the way forward. *Metabolomics*, 1(1), pp. 1-2.

Gorrochategui, E., Jaumot, J., Lacorte, S. & 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.

Greene, C.S., Tan, J., Ung, M., Moore, J.H & Cheng C. (2014). Big data bioinformatics. *Journal of Cellular Physiology*, 229(12), pp. 1896-1900.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25(6), pp. 1107-1116.

Grimes, D.A & Schulz, K.F. (2002). Cohort studies: matching towards outcomes. *Lancet*, 359(9303), pp. 341-345.

Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy J.G. & Ebbels, T.M. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols*, 9(6), pp. 1416-1427.

Hays, P.A. & Thompson, R.A. (2009). A processing method enabling the using of peak height for accurate and precise proton NMR quantitation. *Magnetic Resonance in Chemistry*, 47(19), pp. 819-824.

Hendriks, M.M.W.B., van Eeuwijk, F.A., Jellema, R.H., Westerhuis, J.A., Reijmers, T.H., Hoefsloot, H.C.J. & Smilde, A.K. (2011). Data-processing strategies for metabolomics studies. *TrAC Trends in Analytical Chemistry*, 30(10), pp. 1685-1698.

Hjartåker, A., Lagiou, A., Slimani, N., Lund, E., Chirlaque, M.D., Vasilopoulou, E., Zavitsanos, X., Berrino, F., Sacerdote, C., Ocké, M.C., Peeters, P.H., Engeset, D., Skeie, G., Aller, A., Amiano, P., Berglund, G., Nilsson, S., McTaggart, A., Spencer, E.A., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Linseisen, J., Schulz, M., Hemon, B. & Riboli, E. (2002). Consumption of dairy products in the European prospective investigation into cancer and nutrition (EPIC) cohort: data from 35 955 24-hour dietary recalls in 10 European countries. *Public Health Nutrition*, 5(6), pp. 1259-1271.

Hollywood, K., Brison, D.R. & Goodacre R. (2006). Metabolomics: current technologies and future trends. *Proteomics*, 6(17), pp. 4716-4723.

Houle, T.T., Penzien, D.B. & Houle, C.K. (2005). Statistical power and sample size estimation for headache research: an overview and power calculation tool. *Headache*, 45(5), pp. 414-418.

Huang, X., Liu, G., Guo, J. & Su, Z. (2018). The PI3K/AKT pathway in obesity and type 2 diabetes. *International Journal of Biological Sciences*, 14(11), pp. 1483-1496.

Huang, J., Mondul, A.M., Weinstein, S.J., Koutros, S., Derkach, A., Karoly, E., Sampson, J.N., Moore, S.C., Berndt, S.I. & Albanes, D. (2016). Serum metabolomic profiling of prostate cancer risk in the prostate, lung, colorectal, and ovarian cancer screening trial. *British Journal of Cancer*, 115(9), pp. 1087-1095.

Hwang, T.L. & Shaka, A.J. (1995). Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *Journal of Magnetic Resonance, Series A*, 112(2), pp. 275-279.

IDF (2017a). 'The global picture', in (8 ed.) *IDF Diabetes Atlas*. Brussels: International Diabetes Federation, pp. 40-63.

IDF (2017b). 'What is diabetes?', in (8 ed.) *IDF Diabetes Atlas*. Brussels: International Diabetes Federation, pp. 14-24.

Jenkins, B.J., Seyssel, K., Chiu, S., Pan, P.H., Lin, S.Y., Stanley, E., Ament, Z., West, J.A., Summerhill, K., Griffin, J.L., Vetter, W., Autio, K.J., Hiltunen, K., Hazebrouck, S., Stepankova, R., Chen, C.J., Alligier, M., Laville, M., Moore, M., Kraft, G., Cherrington, A., King, S., Krauss, R.M., de Schryver, E., Van Veldhoven, P.P., Ronis, M. & Koulman, A. (2017). Odd chain fatty acids; new insights of the relationship between the gut microbiota, dietary intake, biosynthesis and glucose intolerance. *Scientific Reports*, 7, 44845.

Jenkins, B., West, J.A. & Koulman, A. (2015). A review of odd-chain fatty acid metabolism and the role of pentadecanoic acid (c15:0) and heptadecanoic acid (c17:0) in health and disease. *Molecules*, 20(2), pp. 2425-2444.

Jepsen, P., Johnsen, S.P., Gillman, M.W. & Sørensen, H.T. (2004). Interpretation of observational studies. *Heart*, 90(8), pp. 956-60.

Johansson, I., Hallmans, G., Wikman, A., Biessy, C., Riboli, E. & Kaaks, R. (2002). Validation and calibration of food-frequency questionnaire in the northern Sweden health and disease cohort. *Public Health Nutrition*, 5(3), pp. 487-496.

Jonsson, P., Wuolikainen, A., Thysell, E., Chorell, E., Stattin, P., Wikström, P. & Antti, H. (2015). Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples. *Metabolomics*, 11(6), pp. 1667-1678.

Kelly, R.S., Vander Heiden, M.G., Giovannucci, E. & Mucci, L.A. (2016). Metabolomic biomarkers of prostate cancer: prediction, diagnosis, progression, prognosis, and recurrence. *Cancer Epidemiology, Biomarkers & Prevention*, 25(6), pp. 887-906.

Knacke, H., Pietzner, M., Do, K.T., Römisch-Margl, W., Kastenmüller, G., Völker, U., Völzke, H., Krumsiek, J., Artati, A., Wallaschofski, H., Nauck, M., Suhre K., Adamski, J. & Friedrich, N. (2016). Metabolic fingerprints of circulating IGF-1 and the IGF-1/IGFBP-3 ratio: a multifluid metabolomics study. *The Journal of Clinical Endocrinology and Metabolism*, 101(12), pp. 4730-4742.

Kriat, M., Confort-Gouny, S., Vion-Dury, J., Sciaky, M., Viout, P. & Cozzone, P.J. (1992). Quantitation of metabolites in human blood serum by proton magnetic resonance spectroscopy. A comparative study of the use of formate and TSP as concentration standards. *NMR in Biomedicine*, 5(4), pp. 179-184.

Kühn, T., Floegel, A., Sookthai, D., Johnson, T., Rolle-Kampczyk, U., Otto, W., von Bergen, M., Boeing, H. & Kaaks, R. (2016). Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Medicine*, 14, 13.

Liu, A. (2014). Biobank as an important tool for biomarker discovery and validation. *JSM Biomarkers*, 1(1), p. 1001.

Lokhov, P.G., Dashtiev, M.I., Moshkovskii, S.A. & Archakov, A.I. (2010). Metabolite profiling of blood plasma of patients with prostate cancer. *Metabolomics*, 6(1), pp. 156-163.

Louis, E., Cantrelle, F.X., Mesotten, L., Reekmans, G., Bervoets, L., Vanhove, K., Thomeer, M., Lippens, G. & Adriaensens, P. (2017). Metabolic phenotyping of human plasma by 1H-NMR at high and medium magnetic field strengths: a case study for lung cancer. *Magnetic Resonance in Chemistry*, 55(8), pp. 706-713.

Munafó, M.R., Tilling, K., Taylor, A.E., Evans, D.M. & Davey Smith, G. (2018). Collider scope: when selection bias can be substantially influence observed associations. *International Journal of Epidemiology*, 47(1), pp. 226-235.

Matsuda, F. (2016). Technical challenges in mass spectrometry-based metabolomics. *Mass Spectrometry*, 5(2), S0052.

Maret-Ouda, J., Tao, W., Wahlin K. & Lagergren J. (2017). Nordic registry-based cohort studies: possibilities and pitfalls when combining Nordic registry data. *Scandinavian Journal of Public Health*, 45(17), pp 14-19.

Mondul, A.M., Moore, S.C., Weinstein, S.J., Karoly, E.D., Sampson, J.N. & Albanes, D. (2015). Metabolomic analysis of prostate cancer risk in a prospective cohort: the alpha-tocolpherol, beta-carotene cancer prevention (ATBC) study. *International Journal of Cancer*, 137(9), pp. 2124-2132.

Mondul, A.M., Moore, S.C., Weinstein, S.J., Männistö, S., Sampson, J.N. & Albanes, D. (2014). 1-Stearoylglycerol is associated with risk of prostate cancer: results from serum metabolomic profiling. *Metabolomics*, 10(5), pp. 1036-1041.

Mónico, A., Martínez-Senra, E., Cañada, F.J., Zorrilla, S. & Pérez-Sala, D. (2017). Drawbacks of dialysis procedures for removal of EDTA. *PLoS One*, 12(1), e0169843.

Morshed, S., Tornetta, P., 3rd. & Bhandari M. (2009). Analysis of observational studies: a guide to understand statistical methods. *The Journal of Bone and Joint Surgery*, 91(3), pp. 50-60.

Nagana Gowda, G.A. & Raftery, D. (2014). Quantifying metabolites in protein precipitated serum using NMR spectroscopy. *Analytical Chemistry*, 86(11), pp. 5433-5440.

Nagana Gowda, G.A. & Raftery, D. (2017). Recent advances in NMR-based metabolomics. *Analytical Chemistry*, 89(1), pp. 490-510.

Nagana Gowda, G.A., Gowda, Y.N. & Raftery, D. (2015). Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. *Analytical Chemistry*, 87(1), pp. 706-715.

Nestel, P.J., Straznicky, N., Mellett, N.A., Wong, G., De Souza, D.P., Tull, D.L., Barlow, C.K., Grima, M.T. & Meikle, P.J. (2014). Specific plasma lipid classes and phospholipid fatty acids indicative of dairy food consumption associate with insulin sensitivity. *American Journal of Clinical Nutrition*, 99(1), pp. 46-53.

Nicholson, J.K., Holmes, E. & Elliott P. (2008). The metabolome-wide association study: a new look at human disease risk factors. *Journal of Proteome Research*, 7(9), pp. 3637-3638.

Nicholson, G., Rantalainen, M., Maher, A.D., Li, J.V., Malmodin, D., Ahmadi, K.R., Faber, J.H., Hallgrímsdóttir, I.B., Barrett, A., Toft, H., Krestyaninova, M., Viksna, J., Neogi, S.G., Dumas, M.E., Sarkans, U., The MolPAGE Consortium, Silverman, B.W., Donnelly, P., Nicholson, J.K., Allen, M., Zondervan, K.T., Lindon, J.C., Spector, T.D., McCarthy, M.I., Holmes, E., Baunsgaard, D. & Holmes, C.C. (2011). Human metabolic profiles are stably controlled by genetic and environmental variation. *Molecular Systems Biology*, 7, 525.

Norberg, M., Wall, S., Boman, K. & Weinehall, L. (2010). The Västerbotten intervention programme: background, design and implications. *Global Health Action*, *3*.

Nowick, J.S., Khakshoor, O., Hashemzadeh, M. & Brower, J.O. (2003). DSA: a new internal standard for NMR studies in aqueous solution. *Organic Letters*, 5(19), pp. 3511-3513.

Osl, M., Dreiseitl, S., Pfeifer, B., Weinberger, K., Klocker, H., Bartsch, G., Schäfer, G., Tilg, B., Graber, A. & Baumgartner, C. (2008). A new rule-based algorithm for identifying metabolic markers in prostate cancer using tandem mass spectrometry. *Bioinformatics*, 24(24), pp. 2908-2914.

Pauli, G.F., Gödecke, T., Jaki, B.U. & Lankin, D.C. (2012). Quantitative 1H NMR: development and potential of an analytical method - an update. *Journal of Natural Products*, 75(4), pp. 834-851.

Pitt, J.J. (2009). Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical Biochemist, Reviews*, 30(1), pp 19-34.

Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E., Huang, P., Hollander, Z., Pedersen, T.L., Smith, S.R., Bamforth, F., Greiner, R., McManus, B., Newman, J.W., Goodfriend, T. & Wishart, D.S. (2011). The human serum metabolome. *PLoS One*, 6(2), e16957.

Putri, S.P., Nakayama, Y., Matsuda, F., Uchikata, T., Kobayashi, S., Matsubara A. & Fukusaki E. (2013). Current metabolomics: practical applications. *Journal of Bioscience and Bioengineering*, 115(6), pp. 579-589.

Qin, L.Q., He. K. & Xu, J.Y. (2009). Milk consumption and circulating insulin-like growth factor-I level: a systematic literature review. *International Journal of Food Sciences and Nutrition*, 60(7), pp. 330-340.

Qin, 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., He, M., Yang, M., Hu, Z., Tang, H., Shen, H., Hu, F.B., Pan, A. & Wu, T. (2016). Plasma metabolomics identified novel metabolites associated with risk of type 2 diabetes in two prospective cohorts of Chinese adults. *International Journal of Epidemiology*. 45(5), pp. 1507-1516.

Ranjan, R. & Sinha, N. (2018). Nuclear magnetic resonance (NMR)-based metabolomics for cancer research. *NMR in Biomedicine*, e3916.

Ravanbakhsh, S., Liu, P., Bjorndahl, T.C., Mandal, R., Grant, J.R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R. & Wishart, D.S. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One*, 10(5), e0124219.

Reinhold, W.C. (2015). Current dichotomy between traditional molecular biological and omic research in cancer biology and pharmacology. *World Journal of Clinical Oncology*, 6(6), pp. 184-188.

Rist, M.J., Muhle-Goll, C., Görling, B., Bub, A., Heissler, S., Watzl, B. & Luy, B. (2013). Influence of freezing and storage procedure on human urine samples in NMR-Based metabolomics. *Metabolites*, 3(2), pp. 243-258.

Ruiz-Canela, M., Guasch-Ferré, M., Toledo, E., Clish C.B., Razquin, C., Liang, L., Wang, D.D., Corella, D., Estruch, R., Hernáez, Á., Yu, E., Gómez-Gracia, Zheng, Y., Arós, F., Romaguera, D., Dennis, C., Ros, E., Lapetra, J., Serra-Majem, L., Papandreou, C., Portoles, O., Fitó, M., Salas-Salvadó, J., Hu, F.B. & Matrínez-González, M.A. (2018). Plasma branched chain/aromatic amino acids, enriched Mediterranean diet and risk of type 2 diabetes: case-cohort study within the PREDIMED trial. *Diabetologia*, 61(7), pp. 1560-1571.

Sandhu, M.S., Heald, A.H., Gibson, J.M., Cruickshank, J.K., Dunger, D.B. & Wareham, N.J. (2002). Circulating concentrations of insulin-like growth factor-I and development of glucose intolerance: a prospective observational study. *Lancet*, 359(9319), pp. 1740-1745.

Schleif, F.M., Reimer, T., Börner, U., Schnapka-Hille, L. & Cross, M (2011). Genetic algorithm for shift-uncertainty correction in 1-D NMR-based metabolite identifications and quantifications. *Bioinformatics*, 27(4), pp. 524-533.

Schmidt, J.A., Fensom, G.K., Rinaldi, S., Scalbert, A., Appleby, P.N., Achaintre, D., Gicquiau, A., Gunter, M.J., Ferrari, P., Kaaks, R., Kühn, T., Floegel, A., Boeing, H., Trichopoulou, A., Lagiou, P., Anifantis, E., Agnoli, C., Palli, D., Trevisan, M., Tumino, R., Bueno-de-Mesquita, H.B., Agudo, A., Larrañaga, N., Redondo-Sánchez, D., Barricarte, A., Huerta, J. M., Quirós, J.R., Wareham, N., Khaw, K. T., Perez-Cornago, A., Johansson, M., Cross, A. J., Tsilidis, K. K., Riboli, E., Key, T.J. & Travis, R.C. (2017). Pre-diagnostic metabolite concentrations and prostate cancer risk in 1077 cases and 1077 matched controls in the European prospective investigation into cancer and nutrition. *BMC Medicine*, 15(1), 122.

Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. & McLean, J.A. (2016). Untargeted metabolomics strategies - challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12), pp. 1897-1905.

Schulz, K.F. & Grimes, D.A. (2002). Case-control studies: research in reverse. *Lancet*, 359(9304), pp. 431-434.

Setia, M.S. (2016). Methodology series module 3: cross-sectional studies. *Indian Journal of Dermatology*, 61(3), pp. 261-264.

Sheedy, J.R., Ebeling, P.R., Gooley, P.R. & McConville, M.J. (2010). A sample preparation protocol for 1H nuclear magnetic resonance studies of water-soluble metabolites in blood and urine. *Analytical Biochemistry*, 398(2), pp. 263-265.

Shi, L., Brunius, C., Lehtonen, M., Auriola, S., Bergdahl I.A., Rolandsson O., Hanhineva, K. & Landberg, R. (2018). Plasma metabolites associated with type 2 diabetes in a Swedish population: a case-control study nested in a prospective cohort. *Diabetologia*, 61(4), pp. 849-861.

Stattin, P., Rinaldi, S., Biessy, C. Stenman, U.H, Hallmans, G. & Kaaks, R. (2004). High levels of circulating insulin-like growth factor-I increase prostate cancer risk: a prospective study in a population-based nonscreened cohort. *Journal of Clinical Oncology*, 22(15), pp. 3104-3112.

Stocks, T., Lukanova, A., Rinaldi, S., Biessy, C., Dossus, L., Lindahl, B., Hallmans, G., Kaaks, R. & Stattin, P. (2007). Insulin resistance is inversely related to prostate cancer: a prospective study in northern Sweden. *International Journal of Cancer*, 120(12), pp. 2678-2686.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64(3), pp. 479-498.

Storey, J.D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), pp. 9440-9445.

Tiziani, S., Emwas, A.H., Lodi, A., Ludwig, C., Bunce, C.M., Viant, M.R. & Günther, U.L. (2008). Optimized metabolite extraction from blood serum for 1H nuclear magnetic resonance spectroscopy. *Analytical Biochemistry*, 377(1), pp. 16-23.

Travis, R.C., Appleby, P.N., Martin, R.M., Holly, J.M.P., Albanes, D., Black, A., Bueno-de-Mesquita, H.B.A., Chan, J.M., Chen, C., Chirlaque, M.D., Cook, M.B., Deschasaux, M., Donovan, J.L., Ferrucci, L., Galan, P., Giles, G.G., Giovannucci, E.L., Gunter, M.J., Habel, L. A., Hamdy, F.C., Helzlsouer, K.J., Hercberg, S., Hoover, R.N., Janssen, J.A.M.J.L., Kaaks, R., Kubo, T., Le Marchand, L., Metter, E.J., Mikami, K., Morris, J.K., Neal, D.E., Neuhouser, M.L., Ozasa, K., Palli, D., Platz, E.A., Pollak, M., Price, A.J., Roobol, M.J., Schaefer, C., Schenk, J.M., Severi, G., Stampfer, M.J., Stattin, P., Tamakoshi, A., Tangen, C.M., Touvier, M., Wald, N.J., Weiss, N.S., Ziegler, R.G., Key, T.J. & Allen, N.E. (2016). A meta-analysis of individual participant data reveals an association between circulating levels of IGF-I and prostate cancer risk. *Cancer Research*, 76(8), pp. 2288-2300.

Tredwell, G.D., Behrends, V., Geier, F.M., Liebeke, M. & Bundy, J.G. (2011). Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. *Analytical Chemistry*, 83(22), pp. 8683-8687.

Trygg, J. & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3), pp. 119-128.

Tuck, M.K., Chan, D.W., Chia, D., Godwin, A.K., Grizzle, W.E., Krueger, K.E., Rom, W., Sanda, M., Sorbara, L., Stass, S., Wang, W. & Brenner, D. E. (2009). Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *Journal of Proteome Research*, 8(1), pp. 113-117.

Vaught, J., Kelly, A. & Hewitt, R. (2009). A review of international biobanks and networks: success factors and key benchmarks. *Biopreservation and Biobanking*, 7(3), pp. 143-150.

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J.J. & Yanes, O. (2012). A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, 2(4), pp. 775-795.

Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1), pp. 7-24.

Wallmeier, J., Samol, C., Ellmann, L., Zacharias, H.U., Vogl, F.C., Garcia, M., Dettmer, K., Oefner, P.J., Gronwald, W. & GCKD Study Investigators. (2017). Quantification of metabolites by NMR spectroscopy in the presence of protein. *Journal of Proteome Research*, 16(4), pp. 1784-1796.

Wang, T.J., Larson, M.G., Vasan, R.S., Cheng, S., Rhee, E.P., McCabe, E., Lewis, G.D., Fox, C. S., Jacques, P.F., Fernandez, C., O'Donnell, C.J., Carr, S.A., Mootha, V.K., Florez, J.C., Souza, A., Melander, O., Clish, C.B. & Gerszten, R.E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, 17(4), pp. 448-453.

WCRF/AICR (2018). Diet, nutrition, physical activity and prostate cancer: a global perspective. *Continuous Update Project Expert Report*. Available at: http://dietandcancerreport.org/

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

Weljie, A.M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C.M. (2006). Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry*, 78(13), pp. 4430-4442.

Westerhuis, J.A., van Velzen, E.J., Hoefsloot, H.C. & Smilde, A.K. (2010). Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics*, 6(1), pp. 119-128.

Wishart, D.S (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 15(7), pp. 473-848.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C. & Scalbert, A. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(1), pp. 608-617.

Wishart, D.S., Lewis, M.J., Morrissey J.A, Flegel M.D., Jeroncic, K., Xiong, Y., Cheng, D., Eisner R., Gautam., B., Tzur, D., Sawhney, S., Bamforth, F., Greiner, R. & Li., L. (2008). The human cerebrospinal fluid metabolome. *Journal of Chromatography, B.* 871(2), pp. 164-173.

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., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., Macinnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J. & Querengesser L. (2007). HMDB: the human metabolome database. *Nucleic Acid Research*, 35, pp. 521-526.

WHO (2017). 'Annexes', in *Global report on diabetes*, Geneva: WHO press, pp. 82-84.

WHO (1999). 'Clinical staging of diabetes mellitus and other categories of glucose tolerance', in *Definition, diagnosis, and classification of diabetes mellitus and its complications: report of a WHO consultation*, Geneva: WHO press, pp. 14-16.

Wold, S., Sjöström, M. & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp. 109-130.

Worley, B. & Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1), pp. 92-107.

Worley, B. & Powers, R. (2014). MVAPACK: a complete data handling package for NMR metabolomics. *ACS Chemical Biology*, 9(5), pp. 1138-1144.

Yin, P., Peter, A., Franken, H., Zhao, X., Neukamm, S.S., Rosenbaum, L., Lucio, M., Zell, A., Häring, H.U., Xu, G. & Lehmann, R. (2013). Preanalytical aspects and sample quality assessment in metabolomics studies of human blood. *Clinical Chemistry*, 59(5), pp. 833-845.

Yu, Z., Kastenmüller, G., He, Y., Belcredi, P., Möller, G., Prehn, C., Mendes, J., Wahl, S., Roemisch-Margl, W., Ceglarek, U., Polonikov, A., Dahmen, N., Prokisch, H., Xie, L., Li, Y., Wichmann, H.E., Peters, A., Kronenberg, F., Suhre, K., Adamski, J., Illig, T. & Wang-Sattler, R. (2011). Differences between human plasma and serum metabolite profiles. *PLoS One*, 6(7), e21230.

Yue, S., Li, J., Lee, S.Y., Lee, H.J., Shao, T., Song, B., Cheng, L., Masterson, T.A., Liu, X., Ratliff, T.L. & Cheng, J.X. (2014). Cholesteryl ester accumulation induced by PTEN loss and PI3K/AKT activation underlies human prostate cancer aggressiveness. *Cell Metabolism*, 19(3), pp. 393-406.

Zheng, G. & Price, W.S. (2010). Solvent signal suppression in NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 56(3), pp. 267-288.

Zheng, C., Zhang, S., Ragg, S., Raftery, D. & Vitek, O. (2011). Identification and quantification of metabolites in 1H NMR spectra by bayesian model selection. *Bioinformatics*, 27(12), pp. 1637-1644.

Zhou, X., Mao, J., Ai, J., Deng, Y., Roth, M.R., Pound, C., Henegar, J., Welti, R. & Bigler, S.A. (2012). Identification of plasma lipid biomarkers for prostate cancer by lipidomics and bioinformatics. *PLoS One*, 7(11), e48889.

Popular science summary

A traditional line of action in research within molecular biology has been to select a few molecules for analysis based on an existing hypothesis. After analysis and statistics, attempts are done to draw conclusions in relation to the hypothesis. But, advances in technology have allowed for more global approaches where many molecules are analysed simultaneously. Such approaches are used more frequently nowadays, which partly owes to their potential to return large amounts of information.

Metabolomics is a global approach used to measure small biomolecules such as amino acids, sugars and different lipids. These small biomolecules are called metabolites and they have important implications for human health. For example, abnormal elevation of blood glucose, measured after oral intake of a sugar-containing solution, can be used to diagnose type 2 diabetes. Smaller, but still abnormal, elevations in blood glucose can also be observed in undiagnosed individuals, which are at larger risk of being diagnosed in the future. Metabolites other than glucose have also been linked to increased risk of disease. Such metabolites are so called disease risk biomarkers. Metabolomics has been used frequently to identify risk biomarkers for different diseases.

Nuclear magnetic resonance (NMR) is a spectroscopic method often used in metabolomics. Each biofluid sample analysed by NMR generates a spectrum with hundreds of signals, which originate from different metabolites in the sample. Signals from a specific metabolite can be converted to its actual level in the sample. However, the quantitative process is problematic since signals from different metabolite are sometimes positioned closely. The resulting overlap can be ignored, at the expense of overestimating metabolite levels. Overlap can be accounted for to generate accurate metabolite levels. Although automated procedures have been developed, they require large computational power, especially when many samples have been analysed or if the position and/or line widths of signals vary between spectra. Hence, more efficient procedures are required.

The work in this thesis can be divided into two parts. The first part regards methodological development to enable rapid and accurate quantification of metabolites by NMR. The second part regards the use of metabolomics – including the developed method for quantification – to identify risk biomarkers for prostate cancer amongst older men.

In the methodological part, an automated algorithm was designed for quantification of metabolites by NMR. The algorithm was based on a principle where one specific signal was selected for the quantification of each metabolite and where the overlap was accounted for by a simple computation. The approach was tested on human blood samples. Evaluation was done to demonstrate its high speed and accuracy. Additionally, a strategy was presented to allow the algorithm to handle more complex NMR spectra where overlap between signals also was affected by changes in signal positions and/or line widths between spectra.

Next, metabolomics was used to identify risk biomarkers for prostate cancer amongst older men in a Swedish cohort. The age at the time of blood collection was an important factor to consider, since different risk biomarkers could be identified in younger men (40 – 50 years) compared to older men (60 years). For example, glycine was identified as a potential marker for prostate cancer in younger men, while several lipids where identified amongst older men. Higher levels of these lipids related to elevated risk of prostate cancer. Similar studies rarely compare different subgroups divided with regards to age. This is since population-based cohorts typically collect samples from individuals of various ages. In this study, samples have instead been collected from individuals at a specific age (40, 50 or 60 years). Beyond this, it was also found that aggressive disease was linked to more risk biomarkers, while risk biomarkers for non-aggressive disease were difficult to identify. Similar studies have often identified different risk biomarkers depending on disease aggressiveness.

Previous studies have shown that individuals with type 2 diabetes are at lower risk of developing prostate cancer. Similar metabolites to those identified in the present study (e.g., lipids) have been identified as potential risk biomarkers for type 2 diabetes. However, lower rather than higher levels related with type 2 diabetes risk in previous studies. In this study, negative correlations were found between some of the potential risk biomarkers for prostate cancer and blood glucose after oral intake of a sugar-containing solution. Additionally, impaired glucose tolerance related to a lower risk of prostate cancer. Together, these findings imply that a reverse relationship between the two diseases (type 2 diabetes prostate cancer) may be reflected at the metabolite level.

Populärvetenskaplig sammanfattning

Det traditionella tillvägagångssättet inom molekylärbiologisk forskning bygger på ett hypotes-baserat val av ett fåtal molekyler för analys. Därefter följer genomförandet av analysen, statistik och försök att dra slutsatser utifrån den ursprungliga hypotesen. Teknologiska framsteg har möjliggjort för mer globala tillvägagångssätt där massvis av molekyler analyseras samtidigt. Globala tillvägagångssätt används mer och mer inom forskningen, vilket till stor del beror på deras potential till att ge mycket information.

Metabolomik är ett globalt tillvägagångssätt för att mäta små biomolekyler, såsom aminosyror, socker och olika lipider. Dessa molekyler kallas för metaboliter och dessa har viktig betydelse för hälsan. Exempelvis kan en onormal förhöjning av glukoshalten i blodet, efter intag av en sockerlösning, användas för diagnos av typ 2 diabetes. Det går att även att uppmäta mindre, men ändå onormalt förhöjda, halter av glukos hos individer utan diabetes som löper större risk att drabbas av sjukdomen senare i livet. Även andra metaboliter har kopplats samman med ökad risk för sjukdom. Metaboliter som sammankopplas med ökad risk för sjukdom brukar kallas för risk-biomarkörer. Metabolomik har ofta används för att identifiera dessa.

Kärnmagnetisk resonans (NMR) är en spektroskopisk metod som ofta används för metabolomik. Varje prov som analyseras med NMR genererar ett spektrum som består av hundratals signaler från de olika metaboliterna som finns i provet. Signalerna kan omvandlas till metaboliternas halter i provet, men en sådan process är problematisk eftersom signaler från olika metaboliter är svåra att skilja åt då de överlappar. Överlappet kan ignoreras, men det leder till falskt förhöjda halter. Det går också att ta hänsyn till signalernas överlapp för att få fram riktiga provhalter, men detta tar lång tid att göra manuellt. Automatiska metoder har utvecklats för att öka på snabbheten och därmed möjliggörs haltbestämning av metaboliter i fler prover. Automatisk haltbestämning kräver mycket datorkraft, speciellt när många prover ska haltbestämmas och speciellt om signalernas position och/eller linjebredd

varierar mellan olika spektra. Därför finns behov av mer effektiva metoder. Detta avhandlingsarbete kan delas in i två delar. Den första delen avser metodutveckling för att möjliggöra snabb och riktig haltbestämning av olika metaboliter via NMR. Den andra delen i avhandlingsarbetet avser att använda metabolomik – inklusive den utvecklade metoden för haltbestämning – för att identifiera risk-biomarkörer för prostata cancer bland äldre män.

I metodutvecklingen designades en automatisk algoritm (AQuA) för haltbestämning av metaboliter via NMR som baserades på en princip där en signal specifikt valts ut för haltbestämning av varje metabolit och där överlapp hanteras genom en enkel beräkning. Algoritmen testades på olika blodprover och evaluerades för att visa på dess snabbhet och riktighet. Dessutom presenterades ett tillvägagångssätt som möjliggjorde att algoritmen också kunde hantera mer komplexa spektra där överlappet mellan signaler också påverkas av att deras position och/eller linjebredd varierade mellan spektra.

Därefter användes metabolomik för att identifiera potentiella risk-biomarkörer för prostata cancer bland äldre män i en svensk kohort. Ålder vid provtagning var en viktig faktor, eftersom olika risk-biomarkörer identifierades för yngre män (40 – 50 år) jämfört med äldre män (60 år). Exempelvis identifierades glycin som en potentiell risk-biomarkör för prostata cancer bland yngre män, medan vissa lipider identifierades bland äldre män. Högre nivåer av dessa lipider relaterade till ökad risk för sjukdom. Liknande studier brukar inte jämföra olika subgrupper uppdelade med avseende på ålder. Detta beror på att provsamlingar i många kohorter kommer från en population med brett åldersintervall. Just i denna kohort har man däremot valt att samla in prover från individer vid specifika åldrar (40, 50 samt 60 år). Utöver detta upptäcktes det även att en mer aggressiv sjukdom var kopplat till flertalet risk-biomarkörer (lipider) medan risk-biomarkörer för icke-aggressiv sjukdom var svåra att identifiera. Liknande studier brukar också identifiera olika risk-biomarkörer beroende på sjukdomens aggressivitet.

Tidigare studier har konstaterat ett omvänt förhållande mellan typ 2 diabetes och risk för prostata cancer. Dessutom har tidigare studier påvisat att liknande metaboliter (lipider) är potentiella risk-biomarkörer för typ 2 diabetes. I fallet typ 2 diabetes är dock lipidernas nivåer lägre hos individer som löper större risk för att utveckla sjukdomen. I denna studie hittades negativa korrelationer mellan vissa av risk-biomarkörerna för prostata cancer (lipider) och glukosnivåer i blodet efter intag av sockerlösning. Dessutom hittades en minskad risk för prostata cancer hos individer med nedsatt tolerans för glukos vid intag av sockerlösning. Dessa resultat indikerar att det omvända förhållandet mellan de två sjukdomarna (T2D och prostata cancer) kan synliggöras via metabolomik.

Acknowledgements

I am deeply grateful to all the people that have supported and helped me during the years since I started my doctoral studies. I would like to express my sincere gratitude to:

Ali Moazzami, my main supervisor. Thank you for selecting me as a PhD student in your group and for the possibility to work with such interesting research projects. I would also like to thank you for giving me the opportunity to pursue some of my own ideas and incorporating such into the thesis work, and also for you many good advices during these years.

Jan Eriksson, my co-supervisor. Thank you for the countless hours you have spent helping me with various aspects of the thesis work! I don't know how to express my gratitude in words. Without your help, this would not have been possible.

Corine Sandström, my co-supervisor. I am so grateful for all the encouragement and support you have given during this process. I would also like to thank you for taking your time in helping me out with writing manuscripts and this thesis!

Anja Olsen, my co-supervisor, thank you for allowing me to visit the Danish Cancer Society Research Center during the beginning of my doctoral studies to learn about statistics in epidemiology. In addition, thank you so much for all the help that you have provided since, with regards to reading manuscripts and the thesis.

I would also like to thank the external collaborators, **Cecilie Kyrø**, **Elin Thysell** and **Göran Hallmans**, for your important contributions to the work included in this thesis! Similarly, I would like to sincerely thank (both present and former) colleagues at the department, **Elisabeth Müllner**, **Peter Agback** and **Lan Vi Tran**, for your respective contributions.

Jag vill också passa på att tacka alla övriga kollegor som jag lärt känna under de senaste fem åren här på SLU och på institutionen. Jag ber om ursäkt redan nu, då jag säkert kommer att glömma någon!

Först vill jag nämna de (tidigare eller nuvarande) PhD studenter som jag lärt känna. **Elizabeth**, som började sina PhD studier något tidigare än mig, tack för allt ditt stöd som betytt väldigt mycket för mig. Likaså, tack till **Martin**, jag har alltid uppskattat våra diskussioner och ditt sätta att tänka. Till **Klara, Yan och Elin**, stort tack och lycka till i era doktorandstudier. **Josefin**, även om vi bara jobbat ihop litegrann, så vill jag ändå tacka dig för att du varit så trevlig och positiv! **Lin**, thank you for the cute presents, which randomly appeared on my desk! **Fredric**, tack för alla pratstunder och dina (icke-fungerade!) försök till att få med mig till gymmet. Tack **Anja**, för att man både kunnat prata allvarigt och skrattat tills man gråter i ditt sällskap. **Frida**, vi har läst kurser, undervisat, fikat och lunchat tillsammans. Tack för din vänskap! **Gustav**, min tidigare kontors-granne, tack för att du varit en bra källa till inspiration, samt att du stått ut med mitt ständiga knappande på skrivbords-tangenterna! **Christina**, tack för din härliga personlighet och alla konstiga samtalsämnen du kommer på! **Johnny**, du har verkligen varit en klippa sen du började här! Tack för din positiva attityd och allt stöd du visat. **Lena**, jag vill verkligen tacka dig för allt du gjort för mig sedan vi blev kontors-grannar. Du är en fantastisk person och väldigt bra förebild!

Thank you very much **Aahana**, for your well-needed help provided in the beginning of my studies and, also, best of good luck with your own doctoral studies!

Anders S, stort tack för att jag fått möjligheten att undervisa! Bitvis har det krävt en del tid, men det har det varit väl värt, då det varit både roligt och lärorikt!

Sist men inte minst vill jag tacka alla övriga kollegor som haft kontor på samma våning som mig. **Suresh**, **Anders B**, **Ingmar**, **Vadim**, **Gulaim**, **Daniel**, **Gunnar**, **Eva** och **Ievgen**, tack för att ni bidragit till den fina arbetsmiljön.

Det finns så många personer bland vänner och familj som jag måste tacka! **Jenny**, utan dig hade jag inte klarat studierna då vi båda började på universitetet för tio år sedan. Tack för den roliga tiden som kombos i Eriksberg och för att vi fortfarande är vänner idag! **Randi** och **Sandra**, jag är så tacksam att vi stått sida vid sida under universitetstiden, både innan och under våra doktorandstudier. Jag hoppas att jag fungerat som ett liknande stöd som ni varit för mig under denna tid. Det stödet ni gett har varit ovärderligt för mig! **Mamma** och **pappa**, ni ska ha det största tacket. Som ingenjör och lärare har

ni alltid framhållit vikten av akademiska studier, vilket till stor del påverkat mina val här i livet. Detta är jag väldigt tacksam över. Men något ännu viktigare är de värderingar ni lärt mig kring att aldrig ge upp och vikten av att vara omtänksam samt att stötta och hjälpa varandra. Min lillasyster (och i mångt och mycket också min förebild) **Kajsa**, tack för att du valde att flytta till Uppsala för åtta år sedan. Det beslutet har inneburit att jag haft en del av familjen nära också under stora delar av mitt vuxna liv. Min kära lillasyster **Filippa**, tack för att du också valde att flytta till Uppsala, samt att du vill umgås så ofta med mig. Du ska ha en extra eloge för att du korrekturläst denna avhandling. Mina fina lillebröder **Hampus**, **Kasper** och **Linus**, tack för att just ni är mina bröder och för att vi alltid har så roligt tillsammans! Jag vill också tacka min älskade farmor samt far- och morbror och kusiner för att ni är en del av mitt liv.

Sist men inte minst vill jag nämna min sambo **Karl**, tack för det enorma stöd och all den hjälp som du gett under mina doktorandstudier, och tack för att du förgyller mitt liv förövrigt!