QUANTITATION IN UNTARGETED MASS SPECTROMETRY-BASED METABOLOMICS

PROEFSCHRIFT

Ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 21 mei 2014
klokke 13:45 uur

door
Frans Meindert van der Kloet
geboren te Joure in 1970

Promotiecommissie

Promotor:	Prof. dr. Th. Hankemeier
Co-Promotor:	dr. T.H. Reijmers
Overige leden:	Prof. dr. M. Danhof
	dr. W. Dunn
	Prof. dr. J. van der Greef
	Prof. dr. A.H.C. van Kampen
	Prof. dr. A.K. Smilde

ISBN/EAN: 978-90-74538-82-4

Cover by Anita van der Kloet

Printed by Ipskamp Drukkers B.V., Enschede, The Netherlands

CONTENTS

1

INTRODUCTION

In the Oxford Dictionary the metabolome is defined as: the **total number** of **metabolites** (the small molecules that are intermediates or products as a result of a metabolic reaction) present within an **organism**, **cell** or **tissue**. This definition covers the three key factors of metabolomics, the research field investigating the composition, role and function of the metabolome [140]. In the analysis of the definition of the metabolome we first identify the biological origin of the research field which can vary with regards to the type of biological question, and therewith connected, the type of samples that are analyzed. This can range from small individual cells to cells clusters to tissue slices to all sorts of biofluids like blood, urine or cerebrospinal fluid. Secondly, the term metabolite implies some form of identification of the chemical compounds being studied. Popular profiling and identification methods range from Nuclear Magnetic Resonance (NMR) to Mass Spectrometry (MS). Recently in particular fragmentation trees obtained with the MS^n [61, 106] approach are used to assign identities to the data features obtained with MS-based profiling techniques. The total number refers to the number of identified (and unidentified) chemical (metabolic) features and their concentration levels that are detected with the same analytical techniques mentioned before.

To obtain biological interpretable results these three important types of information (identity, quantity, and biological relevance) on a metabolite (Figure 1) are equally important and interact strongly. For a good understanding of the biological context the identity of the metabolites must be known. Conversely, identification of metabolites can be greatly improved by including biological information [58]. Furthermore proper (relative) quantification of the metabolites in question [139, 91] is necessary for a better understanding and modeling of the chemical processes of the biological system of interest, i.e. hypothesis generating metabolomics. Even though determination of the quantities of metabolites is not a necessity in all metabolomics experiments, it is very helpful for proper biological interpretation.

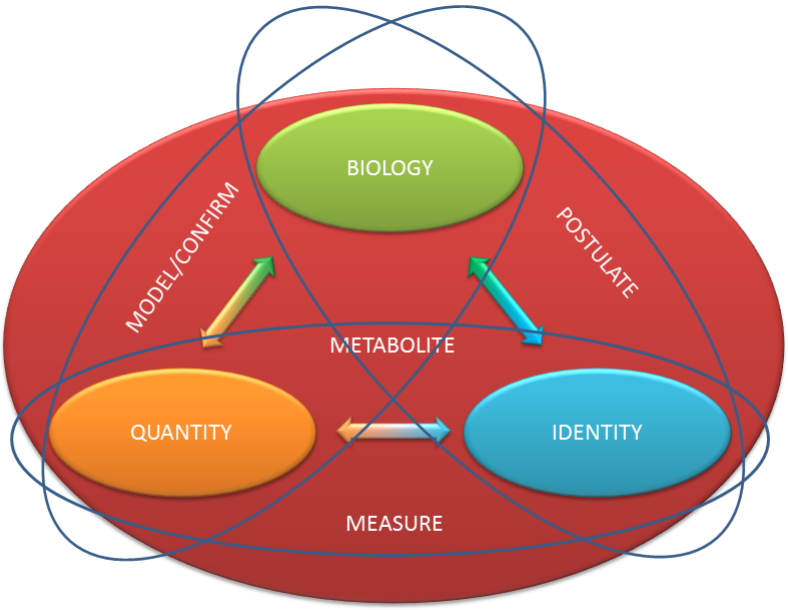


Figure 1: The three key factors of metabolomics: Biological relevance, Identity and Quantity and their interactions.

The word metabolome itself is a construct of the words metabolism and genome [47] and hints to the hierarchy within cell biology: the metabolome is the result of a whole range of chemical and regulation processes that are the result of the interaction of other biochemical organization levels such as the genome and their interaction with the environment. For example, changes in a cells physiological state as a result of gene deletion or overexpression are the complex result of processes at the transcriptome and the proteome and ultimately metabolome level[65, 126, 55]. For example, hard to detect multifactorial changes in the genome resulting in a disease may be easier detected by changes in the metabolite concentrations. This amplification of effects indicates the strength of metabolomics.

Contrary to proteomics and genomics the chemical structures and, therefore, physicochemical properties, observed within the metabolome are much more diverse. The proteome and genome consist of well-defined structural building blocks (i.e. amino acids and nucleotides respectively, although possible post-translational modification and epigenetics have to be taken into account). The diversity in the metabolome combined with the fact that metabolites are known to participate in many different biological pathways, reactions and processes challenges determination and biological interpretation in metabolomics. Without the proper biological knowledge there is no (bio-) logical explanation even if discriminating metabolites are found. The ubiquitous presence of fluids like for example blood at various places of possible biologically relevant processes, complicates the interpretation of metabolic activity in isolation even further[93] as they are not specific to any part of the body. To study a se-

lected part of cellular metabolic networks in a targeted manner, more recently tracer-based metabolomics has been developed as a new experimental data acquisition approach[77].

1.1 MASS SPECTROMETRY

Hyphenated mass-spectrometry (GC, CE or LC-MS) has become the predominant technology for determining metabolite abundances, mainly because of its sensitivity allowing the measurement of low abundant metabolites in small sample volumes. In Figure 2 the schematic of a time-of-flight (TOF) mass spectrometer (MS) detector is shown. The analytes are ionized in the ion source and separated by the applied electric field E (between grid A and B) in which the ions are differentially accelerated depending on their mass and charge. The time it takes to reach the detector (from B to C (length L)) is characteristic for the mass/charge ratio of an ion. Ions with a lower mass (having the same charge) are accelerated more and reach the detector earlier due to $E = \frac{1}{2}mv^2$ and consequently $m = \frac{2E}{v^2}$. v is the velocity i.e. the measured time (t) it takes for ions to travel the distance L.

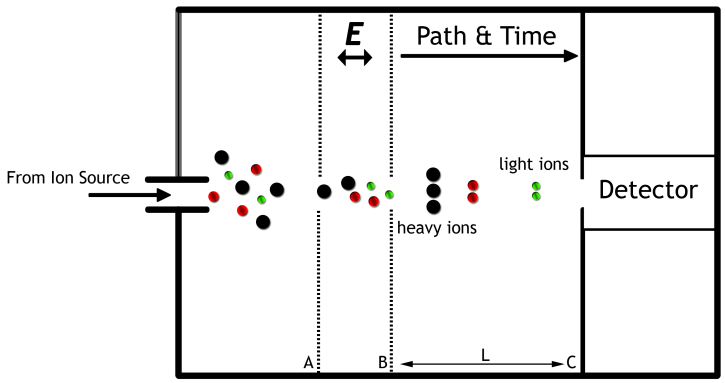


Figure 2: Schematic of a linear Time of Flight (TOF) mass spectrometer, heavier ions (with same charge) travel proportionally slower than lighter ions in an electric field.

When all ions have reached the detector a mass spectrum can be generated (Figure 3). The intensity on the y- axis corresponds to the number of ions that were detected with a specific mass (the x-axis).

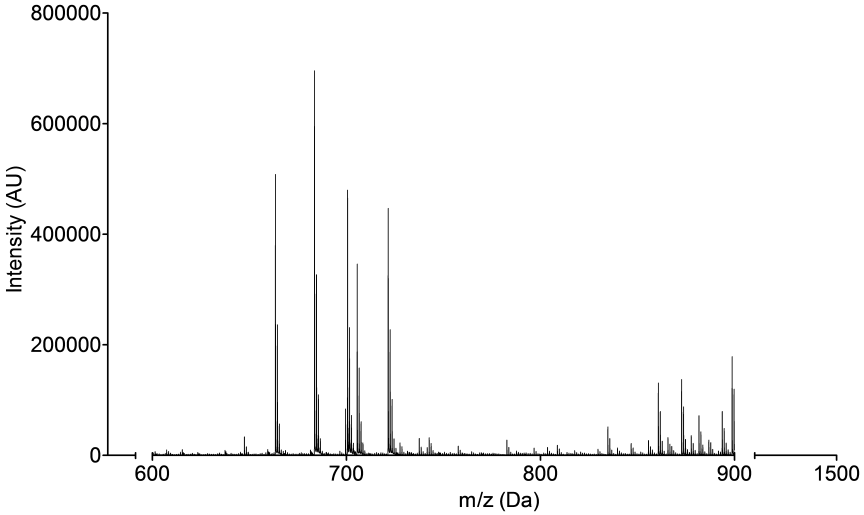


Figure 3: A typical (part of a) mass spectrum.

A drawback in MS is that each metabolite has its own response factor, i.e. the signal depends on the number of molecules but also on the type of molecule. For example two metabolites showing up in a mass spectrum with each of its (e.g.) protonated molecule having an intensity of 106 do not necessarily have the same concentration when they are introduced into the MS. This depends on factors like solubility, ionizability, fragmentation, etc. [4], which are different for the different metabolites. In addition, mass-dependent discrimination can occur due to the mass spectrometer. Furthermore, the response factor for a certain metabolite is matrix dependent, i.e. dependent on the composition of the solvent (in which various compounds can be present) when introduced into the MS, and consequently can vary over different samples creating differences in measured responses for identical metabolite concentrations[6, 79]. With other words, in two different human plasma samples the same metabolite with the same concentration can have different responses. The complex interactions between analyte and the matrix, in which it was measured, can have a significant effect on the response in the MS; this is often referred to as ion suppression/enhancement effects. To compensate for these variations, correction of the response using internal standards is needed. These internal standards should have the same chemical behavior as the analyte but should be detected separately from the analyte of interest. The best internal standard for a certain metabolite is the stable isotopically-labeled (D, ^{13}C or ^{15}N) metabolite itself[78]. Once added to the sample the response of the (isotopically labeled) internal standard can be used for correction of different kinds of chemical and instrumental variations like sample treatment differences, pipetting errors, storage effects, ion suppression etc.. The ratio between the peak intensities of the analyte and internal standard gives an indication of the relative (to the selected internal standard) concentration of the analyte. Absolute quantification of the actual concentration levels (e.g. $\frac{\mu mol}{\mu l}$, $\frac{g}{kg}$ etc.) in all samples of the study can only be calculated if a calibration line for the metabolite of interest was included during measurement. For increased separation the MS is often hyphenated to a separation technique, e.g. gas chromatography (GC),

liquid chromatography (LC) or capillary electrophoresis (CE). In addition to an improved separation of analytes of interest, possible matrix effects and consequently ion suppression effects may be significantly reduced this way.

1.2 QUANTIFICATION

To get reliable quantitation (preferably absolute) the observed differences between the different analyzed samples should not be hampered by analytical variation and should be attributed only to real biological differences of interest. Consequently any further (data) analysis then solely can focus on interpreting these differences. The quantifiable response for a metabolite is the product of its concentration and a metabolite specific response factor. The response factor however, is affected by matrix effects which necessarily need to be minimized. The common ways to characterize these matrix effects are either by post column infusion methods or post-extraction spiking methods[23, 87]. Because the first method only characterizes the matrix effects qualitatively, the quantitative assessment using the second method is more common. With both methods however, the characterisation is biased towards a set of known metabolites only. In metabolomics where typically hundreds to thousands of (also unidentified) metabolites are measured, it is very uncommon to measure (internal) standards for each of these metabolites. This would be very laborious and thus expensive. In addition, it is not known a priori, which metabolites are of interest for the study at hand. As a consequence often platforms are used that cover a wide range of metabolites whose identity is not known in advance (so called untargeted platforms). In these cases usually at least one internal standard per class of metabolites is included to enable relative quantitation (e.g. on lipid per lipid class in lipidomics[53]).

The choice of a proper internal standard influences the estimated (relative) concentration of the compounds in question. Figure 4a shows the peak areas of L-Leucine and two internal standards that were added in replicated (GC-MS) measurements[8, 48] of over 100 identical reference samples (technical replicates [32], i.e. the complete analytical process rather than repeat injections of the same sample). Because the measurements concern the same sample, the ratio between the analyte and the internal standard (IS) should remain constant. The ratios of L-Leucine with the internal standards are plotted in Figure 4b. It is clear that correction with Leucine-D3 generates an almost constant value. However, it is also obvious that correction with a less suitable internal standard (in this case Phenylalanine-D5) can have a dramatic effect on the estimated relative concentration of L-Leucine.

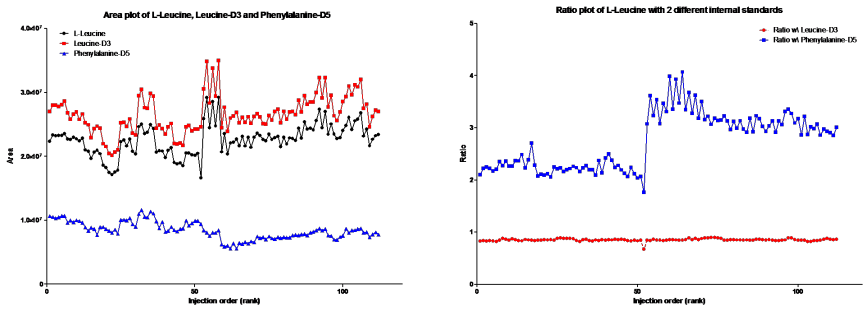


Figure 4: (a) The peak areas of L-Leucine and 2 internal standards for a series repeated measurements (112 technical replicates) of a QC sample. (b) The ratio plot of L-Leucine with each of the two internal standards.

If the ideal internal standard is not available or used, there are four levels of correction to consider to estimate the relative concentration: between analytes within one sample, between analytes over samples measured within one batch and, when many samples need to be measured that cannot be processed within one batch, between analytical batches of samples, and finally, when there is also a substantial time difference between measurements of sample sets, between studies correction. The common factor in all of these four levels is (acquisition) time and in specific all kinds of instrumental and environmental variations like matrix differences, sample degradation, different apparatus but also preprocessing/integration variation that have changed in this time. The challenge in metabolomics is to minimize these variations for as many as possible different metabolites. It is at this stage that metabolomics greatly benefits from statistics (e.g. experimental design [67], data analysis) but of course also from improved analytical sample preparation and analysis methods. With regards to analytical methods, one could think of using a different analytical setup (e.g. post-column infusion techniques [23]) that would quantify suppression effects for a whole range of metabolites but also other optimizations of experimental conditions like concentration levels of the added internal standard [104, 105, 10] can be considered. Statistically, a (mathematical) solution could be to construct virtual internal standards based on a (multivariate/linear) combination of internal standards to normalize the responses of unknown compounds. Finally, to improve comparison over analytical batches of samples and between studies appropriate reference samples could be used[54]. The choice which samples to use as a reference would be a clear result of the combined efforts in analytics and statistics.

1.3 INTEGRATION

Even if all analytical and instrumental settings are optimized, one issue in analyzing MS data that is often left untouched is the integration step itself. The principle to translate the area under the (unimodal and non-overlapping) curves to areas belonging to 2 different components as shown in Figure 5a is evident.

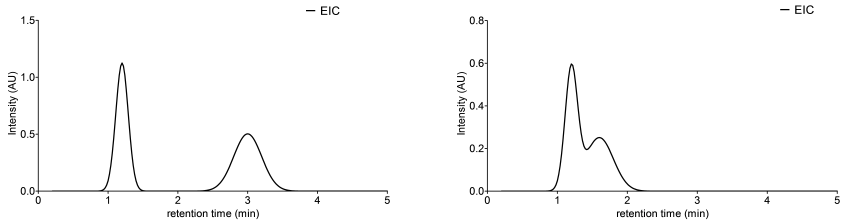


Figure 5: (a) Unimodal extracted ion chromatogram (EIC). (b) Bimodal EIC, (where) should the peaks be separated?

In case of bimodality or multi-modality curves (e.g. due to not fully separated isomers) things get more complicated and arbitrary decisions have to be made (Figure 5b); solutions are to calculate the sum of the total peak are under the curve, split them in the middle or try to fit the signal by (two or more) separate peaks (i.e. by deconvoluting them). Once a choice has been made the software has to be parameterized accordingly. Different (MS) vendors provide own software packages for integration and it is at this point where the different software packages show different outcomes for almost identical cases as depicted in Figure 6 (a and b).

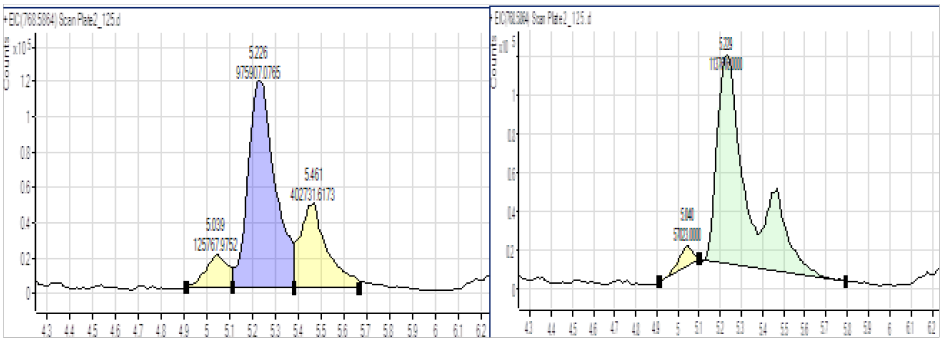


Figure 6: The unexpected behavior with automated integration software. (a) The peak is split in three separate peaks. (b) The two right peaks are combined.

In Figure 6 the same cases of multi-modal peaks are split in different ways using slightly different integration settings. Arguably in cases like this one option is to improve the chromatography but that is unfortunately not always possible, and is anyway time consuming. However, are such overlapping peaks really a problem? This depends on the type of research that is performed. If the aim is to extract known compounds/peaks only, the integration results can be validated by eye and manually adjusted if necessary, however, this is a time-intensive, and therefore expensive, process. If the approach is an untargeted profiling of analytes then there is no bias towards any specific analytes and consequently no (analyte) specific processing steps are there to configure. Visual optimization of integration parameters therefore is very difficult and

manual curation procedures as in targeted data processing is hardly feasible.

Despite the limited number of compounds reported and expensive manual data curation, targeted approaches are widely used. Obvious reasons are that the targeted metabolites/compounds are known which is very important for the data interpretation, and the possibility to quantify them (using internal standards and reference compounds) often with better precision and accuracy then in untargeted modes. To a large extent this is also due to the lack of appropriate software that would enable untargeted extraction and integration without introducing artifacts and errors. As a result, integration is often limited to a set of known metabolites (targets) only and in most cases vendor software is used for such targeted data processing.

1.4 STATISTICS AND DATA ANALYSIS

In metabolomics statistics are applied throughout the whole process of analyzing samples, from method development to data analysis. Experimental designs are applied to setup a study in such a way to minimize the number of experiments while retaining the maximum amount of information[67]. Repeated measurements of samples are used to statistically indicate whether or not the analytical platform functions within specification[37]. To this extent, often, for each specific metabolomics study, a pooled sample (a so-called Quality Control sample) from all samples of that study is created and repeatedly measured. As mentioned earlier, correction steps are necessary to compare metabolites between and over samples. Depending on the type of sample, the way it was measured etc., a whole range of statistically data pretreatment (normalization) methods are offered to improve ultimately the biological interpretation of the data[127]. Actually most, if not all, statistics (in metabolomics) are performed to remove/indicate analytical variation in metabolites (features) that are measured. Those features that do not meet the pre-defined criteria are usually removed from the dataset and further data analysis/interpretation is continued with a smaller set of reliable metabolites. This removal does not necessarily improve biological interpretation but the complexity of follow-up (data) analysis can be reduced considerably.

In univariate statistics the focus is on one variable at a time and the results are relatively easy to interpret from a statistical point of view (e.g. the effect of the variable is significant or not using T-tests[89]). As a consequence univariate statistics methods -are widely accepted, especially in clinical settings[147, 100]. Because changes in biological samples are often multifactorial[147, 100], metabolomics data should be analyzed using multivariate statistics as well. In contrast to univariate statistics multivariate statistics focusses on simultaneously analyzing a set of variables. The (relative) importance of the individual variables in answering the biological question is not always that straightforward and easy to determine but the multivariate profiles however, often do reveal important variables that would have not appeared relevant based on univariate statistics only. After proper quantification, principal component analysis (PCA) is often used to do pattern recognition and visualize observed group differences[86] and methods like partial least squares − discriminant

analysis (PLS-DA)[142] are commonly used to relate these differences to specific metabolites. Using statistical modeling of properly quantified metabolites, (multivariate) metabolic networks can even be inferred[44]. Because of the limited number of samples in comparison to the huge amount of variables (e.g. metabolites) that are measured, multivariate models easily lead to overfitted results (i.e. perfect fits are found, but the predictive power of the model is limited). The results are hugely aided by variable selection methods to select the important from the less important variables and cross-validation and permutation[142] procedures to prevent this overfitting when building predictive multivariate models.

1.5 SCOPE AND OUTLINE OF THIS THESIS

In the previous paragraphs some typical challenges were discussed that researchers are faced with when handling data from metabolomics studies using untargeted mass spectrometry based data. The aim of this thesis was to develop concepts and methods to extract qualitative and quantitative information about metabolites from untargeted mass spectrometric data. For this, different methods were developed to obtain quantitative metabolite data in large studies using GC-MS, LC-high resolution MS (HR-MS) and direct infusion high resolution mass spectrometry. The different methods address different parts in the metabolomics workflow, i.e. data -acquisition, data pre-processing up to data-analysis.

As the performance of analytical systems can vary, different methods of normalization to improve quantification for known and unknown compounds were developed. In **Chapter 2** it is demonstrated that for (relative) quantification of metabolites in GC-MS metabolomics studies, in the absence of matched stable isotopes, per metabolite normalization based on a single internal standard is not enough to correct for analytical batch-to-batch differences. This is especially troublesome in large scale metabolomics studies where many samples need to be measured and consequently many analytical batches are needed. Furthermore, even within a single analytical batch a clear trend in the response for specific metabolites was observed. A statistical procedure based on repetitive measurements of identical samples (i.e. technical replicates) is suggested that corrects for these batch-to-batch differences even for metabolites without a proper internal standard.

In the search for biomarkers for Diabetic Kidney Disease (DKD) in **Chapter 3** LC-MS data of urine samples of an epidemiological study were analyzed. Data acquisition was for that data set unfortunately suboptimal, and various variations in the data were present making (relative) quantification of this untargeted data set difficult. Still, after extensive data preprocessing, a clean data set was obtained suitable for data analysis. It was shown that multivariate statistical modeling was advantageous over univariate modeling for the discovery of biomarkers for this data set. Penalized logistic regression models were used to create a predictive model. Double-cross validation was used to reveal potential new biomarkers.

In **Chapter 4** a method has been developed and demonstrated for the processing of another type of very complex metabolomics data, i.e. metabolomics data obtained by direct infusion mass spectrometry. It was demonstrated that with the preprocessing method that was developed, biological relevant results, i.e. the characterization of different development stages of zebrafish embryos, could be extracted from these very complex metabolomics data. Feature identification was solely based on accurate mass and therefore the samples were recorded with a very high mass resolution. The method developed was based on the binning tools developed for LC-MS (Chapter 5) by aligning the masses over samples which enabled further automated data analysis. Internal standard correction for the unknown features was based on the same strategy as described in Chapter 2. In the absence of quality control samples however, the relative standard deviation (RSD) was calculated using replicated measurements.

The integration problems that were observed during pre-processing of untargeted LC-MS data from earlier experiments (including those reported in Chapter 3), led to the awareness of the lack of good software to integrate peaks in such data sets. The freely available software options required much expertise to configure and were not robust enough to quantify metabolites present at low intensities with good precision and accuracy. In **Chapter 5** therefore a new approach was introduced to integrate samples acquired using LC-time-of-flight-MS. The samples were automatically processed one-by-one to facilitate (future) parallel processing. With only a few parameters that need to be set the user interaction is kept to a minimum, but at the same time obtaining reliable quantitative data on peak areas of known and unknown metabolites.

2

ANALYTICAL ERROR REDUCTION USING SINGLE POINT CALIBRATION FOR ACCURATE AND PRECISE METABOLOMIC PHENOTYPING

ABSTRACT

Analytical errors caused by suboptimal performance of the chosen platform for a number of metabolites and instrumental drift are a major issue in large scale metabolomics studies. Especially for MS-based methods, which are gaining common ground within metabolomics, it is difficult to control the analytical data quality without the availability of suitable labeled internal standards and calibration standards even within one laboratory. In this paper we suggest a workflow for significant reduction of the analytical error using pooled calibration samples and multiple internal standard strategy. Between and within batch calibration techniques are applied and the analytical error is reduced significantly (increase of 25% of peaks with *RSD* lower than 20%) and does not hamper or interfere with statistical analysis of the final data.

Kloet, F.M. Van Der, Bobeldijk, I, Verheij, E.R and R.H. Jellema. Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. Journal of Proteome Research. 2009 Nov;8(11):5132-41

2.1 INTRODUCTION

Recently there has been an explosion of analytical methods developed and applied in different metabolomics related research areas, such as nutrition research[90, 13, 135, 35] , drug discovery[63], optimization of fermentation processes [132, 131] and for breeding [24] of plants. In all these applications it is important to be able to understand and control factors that contribute to errors in the data and result in poor data quality. The total variation in a dataset is a function of different sources of variation[128]. The biological variation is present by design of the study and selection criteria of the subjects. In some cases, additional 'biological' variation can be introduced by differences in sample collection and sample storage [111, 25] Samples drawn from biological systems such as a microbiological fermentation or from body fluids like blood or urine are highly susceptible to changes due to biological reactions that take place, especially when the environment of the sample changes. It is therefore essential that changes in metabolites are minimized during sampling and sample preparation [131] in order to obtain a snap-shot representation of the biological system at the time of sampling. From an analytical point of view the data is of a much higher quality than years ago thanks to the efforts of instrument vendors to obtain more reproducible data, but it is still not enough. The analytical errors should be controlled as much as possible and reduced to a minimum and should not be confused with biological differences within the studied system.

2.1.1 *Sources of analytical variation*

A large part of the analytical variation is caused by suboptimal performance of the chosen platform for (sub-)sets of metabolites and instrumental drift. The ability of a method to detect a specific metabolite (i.e. its performance) is a complex interplay of its physical and chemical properties and is also partially dependent on the sample composition (matrix, e.g. ion suppression in MS based systems [11]) and in many cases on its concentration. Analytical variation for an individual analyte caused by differences in sample composition (matrix effect during extraction, derivatisation and analysis) can be removed only by using stable isotope labeled internal standards. The isotope labeled equivalent of the analyte performs the same as the original analyte and therefore differences in measured peak intensities for the isotope can directly be related to instrumental errors or sample preparation errors. Internal standards that are added before any sample preparation has been performed allow for correction of instrumental drift and sample preparation errors. Instrumental drift is especially important when a large number of samples are concerned, albeit within one batch or when they have to be divided over a number of batches. If instruments need to be cleaned within a series of measurements on samples of the same study, the data suffers from systematic differences between the batches. The severity of the systematic differences of course depends on the platforms being used and to a great extent on the chemical properties of the individual metabolites. Instrumental drift or offset between batches can be corrected for by internal standards or by calibration standards as is often done in bio-analysis or in targeted metabolite profiling [14]. In metabolomics

where hundreds of (also unidentified) metabolites are measured it is very uncommon to measure calibration standards for each of these metabolites. This would be very laborious and thus expensive, but equally important is the fact that beforehand it is not always clear which metabolite is of interest for the study at hand. In order to assess the data quality of all of these metabolites the use of pooled study (QC) samples has been described recently in the literature [28, 37]. In this approach pooled samples are analyzed regularly in between the individual study samples, several times within each batch. As a pooled study sample reflects the average metabolite concentrations within a study, this sample contains the same compounds (e.g. metabolites) as all the other samples. The performance of the analytical platform for all the compounds can be assessed by calculating the relative standard deviation (population standard deviation divided by the population mean) in these pooled samples [129, 28]. Various approaches to detect artifacts are described in Burton et al.[18]. Most of them however depend on visual inspection. Descriptions of data quality improvements are quite scarce. This paper describes a workflow for significant reduction of the analytical error using these pooled QC samples. Based on a multiple internal standard strategy and moreover between and within batch calibration techniques the analytical error is reduced significantly and will thus not hamper statistical analysis of the final data. Although this paper focuses on GC/LC-MS measurements and peak areas, it should be noted that the solution presented here is generic.

2.2 WORKFLOW AND METHODS

For an effective removal of different sources of analytical variation the preprocessing steps should follow a specific sequence. The first step is the data normalization using an internal standard. This step reduces the differences in sample extraction (which can be caused by slight differences in the composition of the samples) and also differences in the volumes injected. Especially the latter issue is of importance when injecting such low volumes as 1-2 μl. The second step is the removal of between-batch and within-batches batch offsets and drifts. This step can only be omitted if each metabolite has a structural analogue IS that corrects for all offsets and drifts, which is not the case in metabolomics analysis. The final steps consist of the combination of data from replicate sample analysis and removal of noise [13] and biomass correction [141]. The biomass correction neutralizes differences in response due to sample weight or volume (e.g. weighed liver tissue or dry matter within a suspension). As the biomass correction is a per-sample multiplicative correction it does not matter at which stage this is performed. The current paper focuses on the normalization using a single internal standard from a set of internal standards and the removal of between and even within batch differences by means of single point calibration using pooled quality control (QC) samples. These calibration techniques effectively render these QC samples, now used for calibration, useless for an independent assessment of the systems performance. To give an independent (unbiased) performance index it is suggested that in addition to these calibration samples additional QC samples should be measured as well which are used for validation of the results. The systematic approach of data pre-processing allowed the workflow to be implemented in a

fully automated environment.

2.2.1 *Used symbols and terminology*

Terminology

Internal standard: There are several definitions of internal standards and surrogates in the literature describing analytical methods. In this publication internal standard is a compound added to the sample before a critical step in the analysis. Depending on the method, the internal standard can be added before or during the extraction of the sample, derivatisation steps, etc.. An internal standard is not necessarily an isotope labeled version of an analyte but can also be structurally related to one or more analytes but not naturally occurring in the samples of interest. If a method covers analytes from different compound classes, multiple internal standards preferably covering all classes should be used.

Batch: a group of samples that has been extracted, derivatised (if applicable) and analysed together at the same time and using the same chemicals, same storage conditions.

QC sample: sample prepared by pooling aliquots of individual study samples, either all or a subset representative for the study. The QC sample has (should have) an identical or a very similar (bio) chemical diversity as the study samples. If insufficient sample volumes are available (e.g. rodent studies), samples collected outside the study but from a similar origin can be used. The QC samples are evenly distributed over all the batches and are extracted, derivatised (if applicable) and analysed at the same time as the individual study samples as part of the total sequence order.

QC calibration sample: sample chemically identical to the QC sample (from the same pool), prepared in the same way as QC samples. QC calibration samples are used for external calibration.

QC validation sample: sample chemically identical to the QC sample (from the same pool), prepared in the same way as QC samples. QC validation samples are solely used to monitor the result of all the data pre-processing steps and the quality of the full method. They are not used for external calibration.

Peak: For the purpose of this publication we use a broader definition of peak. A peak can be a single feature (intensity of a mass/ion or a different signal at a retention time or shift) or can be a sum of features (summed intensity of several ions at the same retention time). In our examples one peak represents one compound or metabolite detected in the data.

Analytical performance: The ability of an instrument to accurately detect a specific chemical component.

Assumptions

The method described in the procedure below focusses on peaks. The study samples vary in concentration for several peaks. For the purpose of this publication and all the procedures described here, we assume that response factors and the analytical performance of the internal standards and individual analytes are not influenced by the sample differences. In other words, the analytical performance of the method for all the individual analytes observed in the pooled QC sample will be the same in all other individual study samples.

Used symbols

i	index number for samples
p	Chromatographic peak (chemical component)
$C_{p,i}$	Concentration of peak p for sample i
F_p	Response factor for peak p
$F_p(t)$	Response factor for peak p at time point t
$F_{p,i}$	Response factor for peak p for sample i
$G_{p,i}$	Transformed form of $F_{p,i}$ after internal standard correction
$X_{p,i}$	Measured response of peak p for sample i
$X_{is,i}$	Measured response of internal standard peak is for sample i
$X'_{p,i}$	Relative response after internal standard calibration of peak p
$X'_{qc,p,b}$	Relative response after internal standard calibration of peak p for QC calibration samples in batch b
$X''_{p,i}$	Relative response after internal standard calibration and batch calibration of peak p for sample i
$A_{p,b}$	Average amplification relative response factor for peak p in batch b
$cf_{p,b}$	Calibration factor for peak p in batch b
$\alpha_{p,b,qc}$	Slope for linear estimate of QC calibration values for peak p in batch b
$\beta_{p,b,qc}$	Intercept for linear estimate of QC calibration values for peak p in batch b
$G_{p,b,i}$	Linear estimate of QC calibration values for peak p for sample i in batch b
Z	Smoothed estimate of QC calibration values for a single peak in a single batch

2.2.2 *Internal Standard normalization*

When focussing on MS analysis, it is generally difficult to model the extraction (derivatisation), MS ionisation and fragmentation variability of a compound by the behaviour of an internal standard with very different physical-chemical properties. This is especially the case when compound and reference belong to chemically different classes (e.g. glucose-d7 is a good representative for glucose but chances are high that it is not suitable for valine or a lipid). Theoretical and practical experiences indicate a positive effect from the use of a cocktail of stable isotope labelled internal standards with the same chemical diversity as the metabolites detected in the samples and exploit the close chem-

ical similarity (e.g. glucose-d7 is also expected to be a good internal standard for fructose and other hexoses). One could argue which internal standards, if more are included, should be used to adequately correct the errors in the measured responses of the individual metabolites. Sysi-Aho et al.[122] suggest selecting the best internal standard based on similarity between the distributions of the available internal standards and the compounds that are measured. Measurements that are performed at a later stage are then corrected using the preferred internal standard. This way however, real-time analytical variation is not included in the internal standard selection process which may result in suboptimal error correction. We suggest using analyte responses in quality control (QC) samples, regularly analyzed in between the study samples, as means to find the best internal standard. Using the relative standard deviations (*RSDs*) of the analyte response in the QC samples to quantify the amount of analytical variation, the best internal standard is the one that gives a minimum relative standard deviation.

In general, the response of a detector for a peak p can be defined as a product of its concentration C_p and a response factor F_p specific to this compound. For sample i the measured response $X_{p,i}$ is defined as shown in Equation 1.

$$X_{p,i} = C_{p,i} \cdot F_p \tag{1}$$

In an ideal situation F_p is constant and therefore measurements with a constant $C_{p,i}$ have identical responses. The *RSD* of QC samples for each peak would then be zero.
Internal standards are routinely used to correct systematic errors in the measured response by transforming the measured response $X_{p,i}$ into a relative response $X'_{p,i}$ using the measured response X_{is} of the internal standard is as denoted in Equation 2.

$$X'_{p,i} = \frac{X_{p,i}}{X_{is,i}} \tag{2}$$

This transformation and its error correcting effect is based on the assumption that for a perfect internal standard the sensitivity of the instrument for compound p is directly related to the sensitivity of the instrument for internal standard IS. In case of a non ideal standard, the corrective effect is not predictable. It may range from almost as good as the ideal internal standard to an actual increase of the error. In typical metabolomics methods the corrective effect of internal standards is highly variable because the number and the chemical diversity of the analytes exceed that of the internal standards (only a few metabolites form a perfect pair with a certain internal standard in a typical dataset).

The *RSD* for the QC samples is calculated using Equation 3, in which the standard deviation ($\sigma_{X'_{p,qc}}$) (after internal standard correction) is divided by the average (<>) relative response after internal standard correction($< X'_{p,qc} >$)

$$RSD_{p,qc} = \frac{\sigma_{X'_{p,qc}}}{< X'_{p,qc} >} \tag{3}$$

The best internal standard is the one that results in a minimal *RSD*. This *RSD* is calculated per peak. When measurements are divided over multiple batches the relative standard deviation is calculated over all QC samples.

2.3 BATCH CALIBRATION

Adjustments of the analytical instrument (e.g. maintenance, cleaning, tuning etc.) between batches of samples can be the cause of analytical errors that cannot be corrected for using solely internal standard calibration. This behaviour exerts itself in different response factors between and even within batches. We suggest that QC samples describe this type of analytical variation adequately and as a result, QC samples can be used as a means to correct for it. This type of correction is referred to as batch calibration.

2.3.1 *Mean and median correction (between-batch)*

Assuming that the measurement errors in a single batch are randomly distributed, then different batches can be compared and corrected using the average or median value of the QC samples in a batch. The average amplification relative response factor per peak per batch ($A_{p,b}$) can be written as the average (¡ ¿) of the responses after internal standard correction of the QC samples per batch (Equation 4).

$$A_{p,b} =< X'_{p,qc,b} > \tag{4}$$

The between batch calibration concerns the adjustment of the amplification factor $cf_{p,b}$ per peak with respect to a reference batch (in Equation 5 batch 1 is taken as reference).

$$cf_{p,b} = \frac{A_{p,1}}{A_{p,b}} \tag{5}$$

The error between measurements in a single batch is assumed to be a homoscedastic and random effect and therefore the same offset correction factors obtained from the QC samples can be transferred to the samples that are measured in between the different QC samples per batch.

$$X''_{p,b,i} = cf_{p,b} \cdot X'_{p,b,i} \tag{6}$$

As an alternative, the median can be used for determining the batch correction factor (in Equation 4) instead of the average. This has advantages over the mean in being a more robust measure. However, most parametric (statistical) tests, that for example facilitate outlier detection, are focussed on averages which makes the use of the average advantageous.

2.3.2 *Linear regression (within-batch)*

In many cases the response of the QC samples is not randomly distributed within a sequence of measurements and a notable drift can exist. In such cases the mean or median correction method will quite adequately correct for differences between batches but poorly for samples within a batch. If it is assumed that the behaviour between two consecutive QC samples is linear it can be modelled using first order regression. Mathematically, Equation 1 still holds but now F_p is not a constant factor but dependent on the time point

at which the sample was measured within a sequence. Equation 1 has to be rewritten to Equation 7.

$$X_{p,i}(t) = C_{p,i}(t) \cdot F_p(t) \tag{7}$$

Assuming that the analysis time of each sample is the same, time is equivalent to injection order and Equation 7 reduces to Equation 8.

$$X_{p,i} = C_{p,i} \cdot F_{p,i} \tag{8}$$

After internal standard normalization has been performed the definition of QC calibrated data follows Equation 9, in which $G_{p,i}$ is the transformed form of $F_{p,i}$ after internal standard correction. The data are not calibrated for between batch differences. This is done as the final step.

$$X''_{p,i} = cf_{p,i} \cdot X'_{p,i} = X'_{p,i} \cdot \frac{1}{G_{p,i}} \tag{9}$$

Because for each batch the correction factor is different, Equation 9 translates into Equation 10.

$$X''_{p,b,i} = cf_{p,b,i} \cdot X'_{p,b,i} = X'_{p,b,i} \cdot \frac{1}{G_{p,b,i}} \tag{10}$$

The estimated trend of the (relative) response for the QC samples per peak within a batch can be written as a function of injection order that is adjusted for slope β, and an intercept α (Equation 11).

$$X'_{p,b,qc} = \beta_{p,b,qc} \cdot i_{b,qc} + \alpha_{p,b,qc} \tag{11}$$

In case of a first order regression the factors α and β can be calculated using regular linear regression. Although higher order regression methods can be applied they heavily depend on the number of QC samples that are measured and are more sensitive to outliers. Using the regression coefficients from Equation 11 an estimate of the QC response can be calculated at each injection point i within a batch. In order to make a good estimation the QC samples should be distributed evenly within the measurements in a batch to ensure a good representation of the total drift during a batch and measured at each start and end of a batch to prevent extrapolation (errors).

$$G_{p,b,i} = \beta_{p,b,qc} \cdot i_b + \alpha_{p,b,qc} \tag{12}$$

Using this estimated trend, the relative response after internal standard calibration, per batch, is divided by this trend (Equation 13)

$$X''_{p,b,i} = cf_{p,b,i} \cdot X'_{p,b,i} = \frac{X'_{p,b,i}}{G_{p,b,i}} \equiv X''_{p,i} \tag{13}$$

2.3.3 *Linear smoother*

The assumption that the data between consecutive QC samples, within a batch, behave in the exact linear manner has a drawback if only a few QC samples measurement points are available or the QC samples exhibit too much

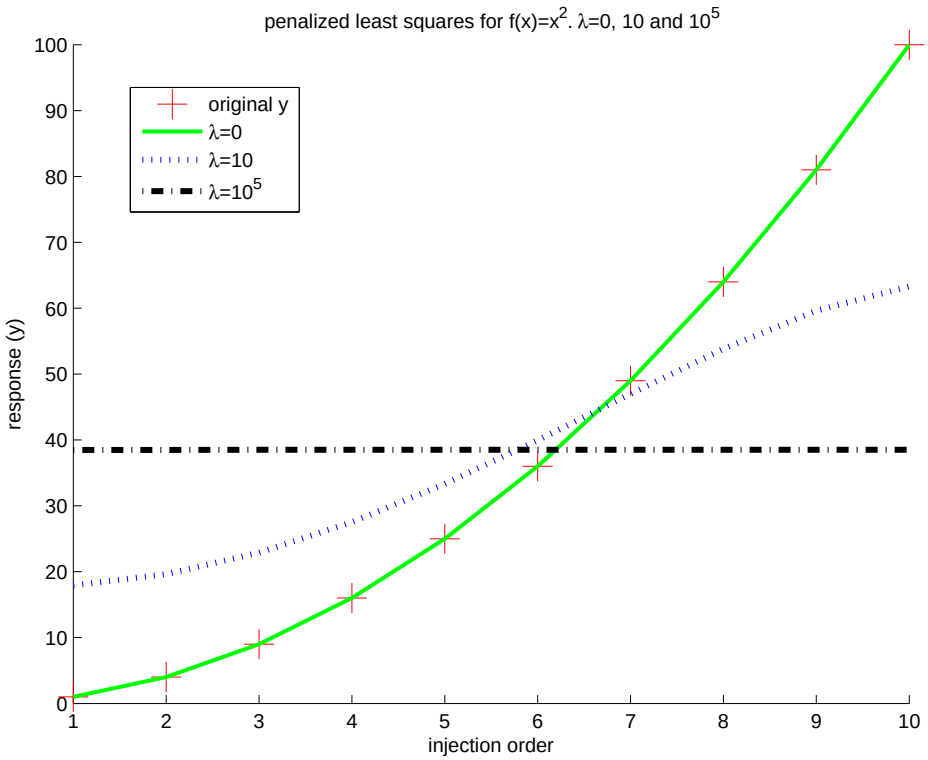


Figure 1: The effect of different values of λ on the smoothed estimate Z of an arbitrary trend exhibited by QC samples (e.g. $f(x) = x^2$). The black stippled line, where $\lambda = 10^5$ is used, is a horizontal line through the mean value of the QC samples. The blue dotted line, $\lambda = 10$, is a smoothed line that follows the general trend of the QC samples. The green line, $\lambda = 0$, follows the exact pattern of the QC samples.

variation (noisy data) for a significant linear trend. Linear correction would still improve the overall quality but could also introduce new analytical variation. In such cases the drift would best be described by a smoothed trend. Eilers[29] has shown that discrete penalized least squares can be used to estimate a smoothed trend. λ is the smoothing parameter where a larger λ results in a smoother estimate of the regression line Z. For really large values of λ, Z will result in a horizontal line (Figure 1, the black stippled line). This is a favourable characteristic because in cases of these large λs it is the only assumption that can be made (i.e. there's no overall linear relation between the QC samples). For small values of λ however, Z follows the trend exhibited by the QC samples (Figure 1, the blue dotted line). When no penalty is imposed the smoothed estimate Z follows the exact pattern of the QC samples (Figure 1 the green line).

To find an appropriate value for the penalty, λ_p has been made proportional to the residual error of the linear estimate ($G_{p,b}$) and the actual QC sample values ($X''_{p,b,qc}$) (Equations 11 and 12). Anything else than a perfect linear fit

results in a smoothed estimate of the trend between consecutive QC samples. The final QC trend is removed via

$$X''_{p,b,i} = \frac{X'_{p,b,i}}{Z'_{p,b,i}} \tag{14}$$

Finally the calibrated response is calibrated for between batch differences using the math as described in Equations 4 through 6. In this case however, $X'_{p,i}$ is substituted by $X''_{p,i}$.

2.4 EXPERIMENTAL

2.4.1 *Data sets*

To demonstrate the use of QC samples for the determination of the best internal standard and batch (between and within) calibration techniques two different datasets were used.
Data processing was performed using the MSD ChemStation E02.00.493 (Agilent technologies, Santa Clara, CA, USA). Based on many previous studies a target table is pre-defined containing matrix and study specific (metabolites with known and unknown identities). Each peak is characterized by its retention time and selected specific m/z value. For each study an update of the retention times (and in limited cases m/z) values is prepared. A limited number of selected chromatograms are compared to a chromatogram of a reference sample (sample that is analysed in each study). Peaks that have not been observed previously are added to the target table. Artefacts of the method are removed from the data as well as (multiple) entries for a single (identified) metabolite caused by different derivatization products of which the performance is known to be irregular. For plasma, typically 120-200 metabolites are reported. Even though this procedure is quite time-consuming, we believe it gives more reliable data than peak picking procedures or most deconvolution procedures with less missing values and less peaks. During acquisition of both datasets the following compounds were used as internal standards: Alanine d4 (ALA-D4), Cholic acid d4 (CA-D4), Leucine d3 (LEU-D3), Phenylalanine d5 (PHE-D5), Glutamic acid d3 (GLU-D3), Dicyclohexylphtalate (DCHP), Difluorobiphenyl (DFB), Trifluoroacetylantracene (TFAA). All compounds were purchased from Sigma (Zwijndrecht, the Netherlands).

Example metabolomics study 1

A nutritional intervention study that involved 36 volunteers. Volunteers were divided into 4 groups and received 4 different treatments: A, B, C and D, including placebo. A cross-over design was used in this study, with each group receiving each of the treatments, in a randomized order [7]. At the end of each treatment period each subject received an oral lipid challenge test, after which several blood samples were collected. Plasma samples collected within this study were analysed using different metabolomic platforms including the GC-MS method as described by Koek et al.[72] From the challenge test, only samples from treatment groups A and B were analysed by GC-MS

and this data is used for the application demonstration of the developed workflow and methods. Plasma samples (100 μl) were extracted with methanol and after evaporation the metabolites were derivatized (oximation and silylation). 8 different internal standards were added to the samples before the different sample preparation steps. The number of individual study samples analysed by GC-MS was 504. The samples were analysed in 18 batches, each batch contained 28 study samples (all timepoints from two subjects, randomized per subject) and 3 pooled QC samples. Each study sample was injected once; each QC sample was injected twice per batch. The QC injections were distributed evenly in the batch: at the start of the batch, after approximately every 6 samples and at the end of each batch. Besides these QC samples, additional QC validation samples were included. Each batch contained 1 QC validation sample. At the end of the analysis, a 19th batch was included, which contained a real replicate analysis of the complete time profile of two selected subjects. For this purpose a separate aliquot of the samples was extracted, derivatised and analysed. Data processing was performed as described above. 145 peaks (excluding internal standards) were reported.

Example metabolomics study 2

An inflammation modulation study with placebo and diclofenac was performed in parallel [144]. Each group had 10 volunteers. 19 volunteers completed the treatment. Blood samples were taken after an overnight fast on days 0, 2, 4, 7 and 9. Subjects underwent an oral glucose tolerance test (OGTT) on day 0 and day 9 of the study. Blood samples were taken just before (0 minutes) and 15, 30, 45, 60, 90, 120 and 180 minutes after the administration of the glucose solution (75 grams). The samples taken at day 9 were analysed using the same analytical methods described for example study 1. The number of individual study samples analysed by GC-MS was 361 (19 volunteers, 19 timepoints per volunteer), each sample was analysed twice, resulting in 722 sample injections. The samples were analysed in 26 batches, each batch contained 35 injections including QC samples. Batch 1 started with all samples of one of the subjects, timepoints randomized, followed by the randomized replicate measurements of the same subject until the maximum of 29 (sample) injections was reached. The next batch started with the remaining (replicate) samples of the previous subject followed by the, timepoint randomized, full set of samples of the next subject etc. In this way for each subject at least one replicate of the full time profile was analysed within one batch. The QC injections were distributed evenly in the batch: at the start of the batch, after approximately every 6 sample injections and at the end of each batch. No additional QC validation samples were measured. To assess the effect of the different preprocessing steps the replicated measurements were used. Data processing was performed as described above. 137 peaks (excluding internal standards) were reported.

Data extraction

Both example studies were processed using a target approach. The target table was adjusted 3 times for retention time shifts caused by shortening the column by several centimeters after each batch.

Data processing

Prototyping of the correction methods was executed in Matlab version 2007b[50]. The final implementation of the software was done in SAS version 9.1.3[52] as stored procedures complementary to a data warehouse (SAS) in which the study data were captured.

2.4.2 *Results and discussion*

Internal Standard normalization

The number of internal standards is dependent on the analytical method with a minimum of 1 and no maximum. To mimic the behaviour of the analytes structure analogues and stable isotope labelled compounds were used added as internal standards. In our example study we used 8 internal standards and applied our selection method to select the most suitable one. Table 1 shows the results of the selection method on the *RSD* values of the QC calibration samples. That the *RSD* indeed seems to be functioning as a good criterion for the selection of the best internal standard is shown in Table 2 in which examples are shown of the internal standard that was selected as best for a number of identified metabolites in metabolomics study 1. In all cases where an analyte had an own deuterated internal standard, this standard was selected, as it also gives the lowest *RSD* in the QC validation samples. For compounds with no deuterated analogue a structurally related IS was selected, e.g. LEU-D3 was selected for the corrections of both leucine and isoleucine, for alanine, ALA-D4 is selected. For some aminoacids within this study, the the deuterated analogue gives a slightly higher *RSD* than a different deuterated IS (8% vs 5%). Within this study this is the case for glutamate, for this compound PHE-D5 is selected as the IS giving the lowest *RSD* in the QC validation samples. Within this study, not all the reported metabolites have a suitable IS structurally. For example the detected fatty acids are less volatile and elute later in the chromatogram. For these matabolites the described procedure chooses apolar and/or late eluting internal standards such as DFB or DCHP as the most suitable. It should be noted that each metabolite might have several suitable IS within the approach giving similar *RSD* values. Therefore the chosen IS can differ from study to study, depending on the dataset.

RSD	Corrected for 1 IS (DCHP)	corrected for best IS
0%-10%	26%	58%
10%-20%	32%	24%
20%-30%	27%	10%
>30%	16%	8%

Table 1: Frequency distribution of *RSD* values of QC calibration samples from the first example study. The effect of the 'best' internal standard clearly translates into more peaks with lower *RSD* values.

Each of the peaks was assigned a best internal standard (1 out of 8) using the criterion as described in Equation 3. After the normalization step with the appropriate internal standard, the PCA score plot (Figure 2) shows cluster-

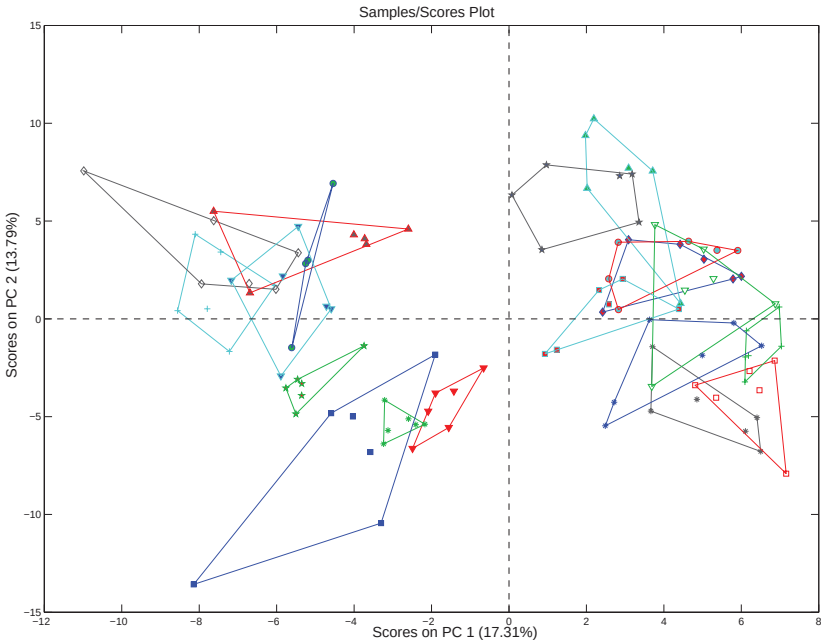


Figure 2: PCA score plot of QC calibration samples from the first example study after internal standard normalization. Different colours refer to the different batches. The data are autoscaled. Two clusters are visible; one before cleaning the MS source (left) and one after the cleaning procedure (right).

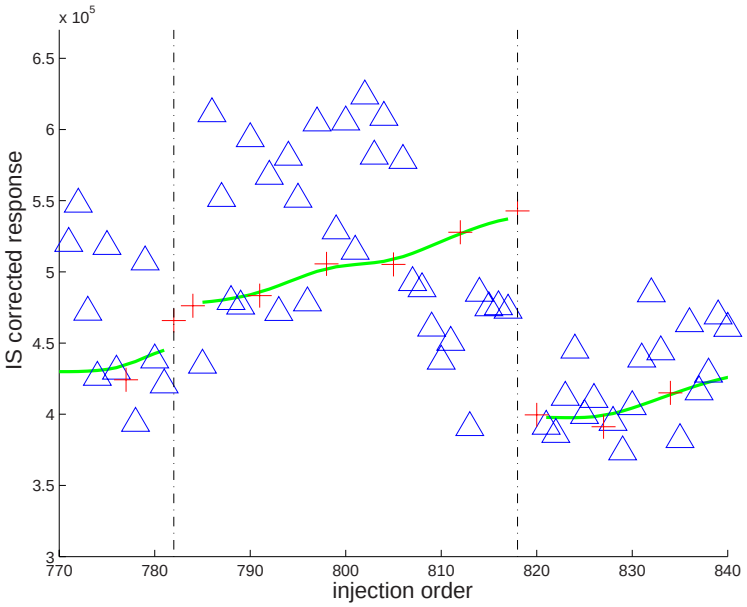
ing of the different QC samples per batch. Furthermore, the plot shows that there's a significant difference between two clusters of batches. Investigation reveals that the group on the left hand side are batches 1-9 whilst the remaining batches 10-19 are plotted on the right hand side; it coincides with the fact that the MS source was cleaned after the 9th batch. This behaviour emphasizes that QC samples (indeed) characterize the systems state (or can be used to do so). It also leads to the conclusion that remaining variation due to between-batch and/or within-batch differences is insufficiently corrected for using the normalization procedure with internal standards.

Metabolite	IS selected as best
Alanine	ALA-D4
Leucine	LEU-D3
Isoleucine	LEU-D3
Glutamic acid	PHE-D5
C16:1 Fatty acid	DFB
C16:0 Fatty acid	DCHP
C17:0 Fatty acid	DCHP

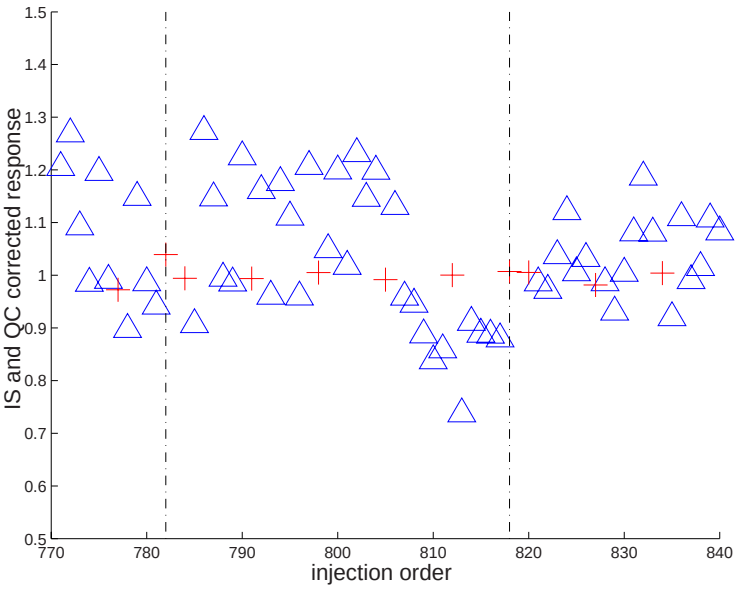
Table 2: Examples of the selected best internal standard for a selection of metabolites found in the first example study.

Batch calibration

Figure 3a shows a part of a time order profile plot of a specific metabolite from the second example study. The real samples are represented by the blue triangles and the QC samples by the red crosses, the green lines represent the smoothed estimate for the trend exhibited by the QC samples. Each vertical dashed line represents the start of a new batch. The figure clearly shows the effect of analytical errors that are introduced during measurements. Even though the data was corrected for the best internal standard, for some metabolites such as this example, large between-batch and within-batch differences still exist for the QC samples. The study was not set up for quantification purposes but it is apparent that for any further (statistical) analysis this type of error should be removed. In order to do so, a smoothed trend, per batch, was fitted through the QC samples. Figure 3b shows the results of the same metabolite after the batch calibration step. The *RSD* value of this metabolite for the QC samples dropped from 13.7% after internal standard correction to 2.1% after the additional batch correction step. The QC samples follow an almost horizontal line indicating that the within batch calibration was applied successfully. Furthermore, it also shows that the offset differences between the batches have been removed. The resulting variation is mainly due to the actual compositional differences between the samples (different subjects and different timepoints).



a



b

Figure 3: A part of the time order plot of a specific metabolite from the second example study after internal standard normalization (a) and batch calibration (b). The study samples are represented by triangles and the QC samples by crosses. The vertical dashed lines represent the start/end of a batch. The green lines represent the smoothed estimate of the QC samples. Fig. a shows the data after best internal standard normalization. Offset differences and within batch trends are clearly visible. Fig. b shows the same data but now after additional batch calibration. The QC samples clearly follow a horizontal line and no offset differences between batches are visible.

Validating the results

The availability of QC calibration, QC validation samples and replicated sample measurements allow for a triple validation check. The performance statistics used are as follows:

1. Improvement of *RSD* in the QC calibration samples
2. Improvement of *RSD* in the (independent) QC validation samples
3. Improvement of differences between samples with different composition (representative replicates)

The effect induced by the different calibration steps on the *RSD* value of the QC validation samples from the first example study is shown in Table 3. The table shows the distribution of the number of metabolites when the *RSD* range is divided into 4 classes. For replicated measurements the results are shown in Table 4. The results in Table 3 and Table 4 are comparable, the removal of between and within batch differences using the real-time variation information embedded within pooled QC samples shows a significant impromevent in observed *RSD* values.

RSD	Raw data	IS calibrated	IS + batch calibrated
0%-10%	12%	58%	81%
10%-20%	59%	24%	16%
20%-30%	16%	10%	3%
>30%	13%	8%	0%

Table 3: Frequency distribution of *RSD* values of the QC validation samples from the first example study. The IS normalization and batch calibration steps clearly have a favourable effect on the *RSD* frequency distribution of these samples.

RSD	Raw data	IS calibrated	IS + batch calibrated
0%-10%	49%	53%	72%
10%-20%	36%	32%	21%
20%-30%	8%	8%	7%
>30%	8%	7%	1%

Table 4: Frequency distribution of *RSD* values of duplicated measurements of real samples from the second example study. The calibration steps show the same positive effect on the frequency distribution of '*RSD*'s of these replicate measurements as the QC validation samples.

Visualizing information potential

To get a global idea about the information potential within the study data, a scatter plot of the analytical performance versus its reproducibility for all metabolites in a given data set is a remarkably elegant and simple method for

depicting the improved data quality obtained with the QC calibration method. An example for GC-MS data from the second example study is shown in an Information Density plot (ID plot) (Figure 4 a and b). The x-axis shows the *RSD*, and the y-axis shows the correlation between replicates. The correlation is both a measure of data quality and the range of observations. For example, a metabolite with a relatively large *RSD* (e.g. 50%) and a large concentration range (e.g. one order of magnitude between lowest and highest) typically has a good replicate correlation. On the other hand a metabolite with an excellent *RSD* (e.g. <10%) may have a very poor replicate correlation if its concentration in all samples is identical.

These plots visualize the proportion of good quality data (left of the vertical line / limit) and the proportion of metabolites with a wide concentration range (above the horizontal line / limit). The upper-left corner contains the most information rich metabolite data and it is obvious that the QC correction indeed shifts metabolites towards the lower (good) *RSD* region and results in more metabolites in the upper left hand corner region. Similar results are obtained if the concentration range is used instead of the replicate correlation coefficient. Depending on the objective of a metabolomics study these plots may be used in various ways for variable selection prior to statistical analysis of the data.

It is important to understand that the variability of the QC samples should represent the variability of the study samples. Therefore, the QC samples should be handled as if they were study samples which means for example that reuse of 'old', already extracted and derivatised QC samples is not a very good idea because it will induce an extra source of variation that is not present in the study samples. For large, long duration studies we suggest preparing sufficient number of QC aliquots, such that an identical sample is used throughout the whole study. Hereby we assume that the influence of the storage of the QC sample over longer periods of time has a negligible effect on the composition of the sample

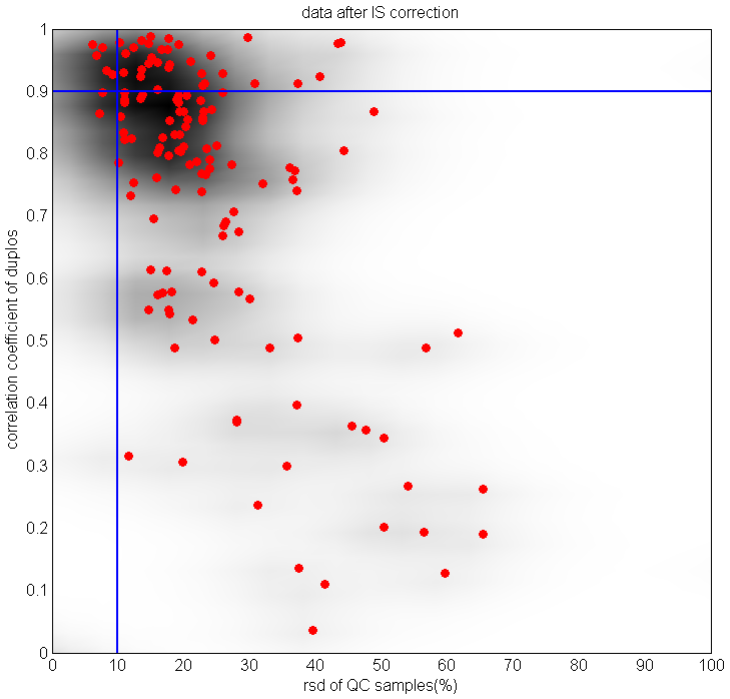
2.4.3 *Recommendations*

Internal Standard normalization

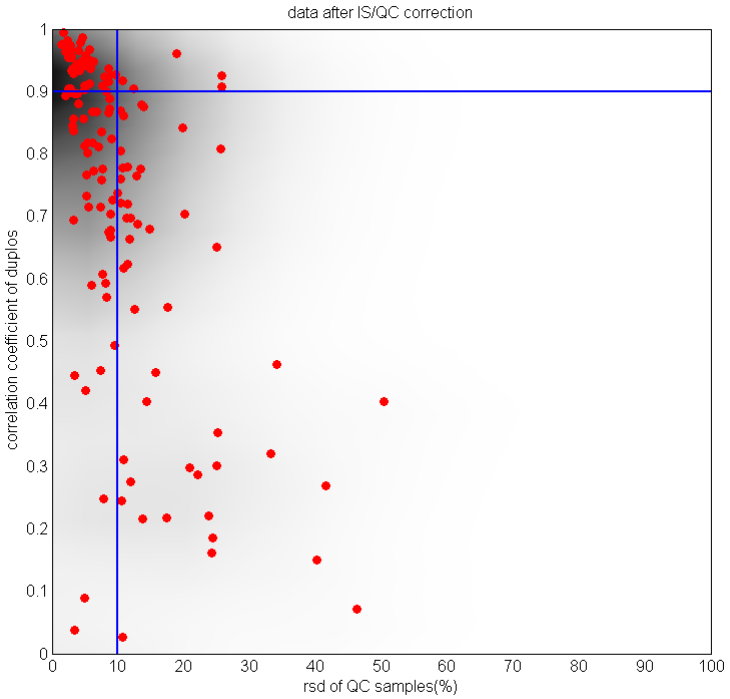
The most important descriptors of data quality are accuracy and precision. Standards and calibration curves are typically not used in unbiased metabolomics. This makes it impossible to assess the accuracy of the method for all the metabolites measured, identified and unidentified, and as a consequence it is equally impossible to improve the accuracy for one or more metabolites. The described procedure focuses on improving data precision, which is equivalent to minimizing the *RSD*. From the point of view of optimising data quality it would be beneficial to use a cocktail/mix of deuterated internal standards that have structures that are analogue to the ones that have to be analyzed.

Batch calibration

QC samples that are generally used to assess the performance of the system are now used for calibration purposes. To obtain a calibration model, at least 2 QC calibration samples should be measured per batch, one at the beginning and one at the end of each batch. To increase the robustness and reliability of



a



b

Figure 4: Information Density plot (ID plot), a scatter plot of the analytical performance versus its reproducibility for all metabolites for data from the second example study. Fig a. shows the results after internal standard normalization, Fig b. shows the results after internal standard normalization and batch calibration.

such a model more QC calibration samples should be measured per batch. The actual number of QC samples depends on the individual analytical method, its robustness and performance characteristics and the availability of QC material. There should be a balance between the number of QC samples and the number of study samples in order to keep the analysis efficient and low cost. The number of replicate injections per QC sample should be equal or similar to the number of injections per individual study sample. It is also clear that outliers in QC samples can adversely affect the outcome.

2.5 CONCLUSION

Results from metabolomics studies can be improved using a single point calibration based upon results obtained from pooled QC samples that are repeatedly measured in between study samples. This one point calibration is indispensable when large scale metabolomics studies are performed and both within and between batch differences become a problem due to instrumental and environmental changes during measurements. We have shown that two types of QC samples are required whereby the first type, calibration QC samples, is used to perform a one point calibration, and the second type, validation QC samples, is used to assess how well the calibration procedure improved the data quality. We have shown that it is feasible to increase the number of metabolites with a relative standard deviation for replicated measurements below 10% from 49% of the peaks to 72%. As a result, the induced or biological variation in the study samples becomes more apparent and more meaningful statistical models can be build from the corrected data.
We have also shown that the *RSD* in the QC samples before and after internal standard correction is a good measure to find the best match between a given metabolite and the set of internal standards that were spiked in the sample. This is a practical alternative to using a separate (isotope labeled) internal standard for each metabolite, which is not feasible due to costs and availability. For large data sets, it is a difficult task to obtain an idea about the information content of the data and to make a comparison between a before and after correction situation. For this purpose, we suggest and demonstrate a new method for presentation of the total dataset focusing on the analytical variability (*RSD*) and the concentration or intensity range of the metabolites. The example shown in this paper clearly shows an improved information content after correction of the raw data with internal standards and QC samples.

The methodology presented here was applied to GC-MS data but is applicable also to other datasets obtained with other analytical techniques. A future perspective for further development of the methodology lies in the fusion of datasets obtained from different studies or different instruments. A copy of the MATLAB[50] prototyping code is available on request from the corresponding author.

ACKNOWLEDGEMENT

The authors would like to acknowledge Bas Muilwijk and Marc Tienstra for the GCMS data. The help of Gertruud Bakker and Carina Rubingh with the data analysis and randomization schemes is acknowledged.

3

DISCOVERY OF EARLY-STAGE BIOMARKERS FOR DIABETIC KIDNEY DISEASE USING MS-BASED METABOLOMICS (FINNDIANE STUDY)

ABSTRACT

Diabetic kidney disease (DKD) is a devastating complication that aects an estimated third of patients with type 1 diabetes mellitus (DM). There is no cure once the disease is diagnosed, but early treatment at a sub-clinical stage can prevent or at least halt the progression. DKD is clinically diagnosed as abnormally high urinary albumin excretion rate (AER). We hypothesize that subtle changes in the urine metabolome precede the clinically signicant rise in AER. To test this, 52 type 1 diabetic patients were recruited by the FinnDiane study that had normal AER (normoalbuminuric). After an average of 5.5 years of follow-up half of the subjects (26) progressed from normal AER to microalbuminuria or DKD (macroalbuminuria), the other half remained normoalbuminuric. The objective of this study is to discover urinary biomarkers that differentiate the progressive form of albuminuria from non-progressive form of albuminuria in humans. Metabolite proles of baseline 24 h urine samples were obtained by Gas Chromatography-Mass Spectrometry (GC-MS) and Liquid Chromatography-Mass Spectrometry (LC-MS) to detect potential early indicators of pathological changes. Multivariate logistic regression modeling of the metabolomics data resulted in a profile of metabolites that separated those patients that progressed from normoalbuminuric AER to microalbuminuric AER from those patients that maintained normoalbuminuric AER with an accuracy of 75% and a precision of 73%. As this data and samples are from an actual patient population and as such, gathered within a less controlled environment it is striking to see that within this profile a number of metabolites (identified as early indicators) have been associated with DKD already in literature, but also that new candidate biomarkers were found. The discriminating metabolites included acyl-carnitines, acyl-glycines and metabolites related to tryptophan metabolism. We found candidate biomarkers that were univariately significant different. This study demonstrates the potential of multivariate data analysis and metabolomics in the field of diabetic complications, and suggests several metabolic pathways relevant for further biological studies.

Kloet, F.M. van der, F.W.A. Tempels, N. Ismail, R. van der Heijden, P.T. Kasper, M. Rojas Chertó, R. van Doorn, G. Spijksma, M. Koek, J. van der Greef , V.P. Mäkinen, C. Forsblom, H. Holthöfer, P.H. Groop, T.H. Reijmers and T. Hankemeier, 2011. Discovery of earlystage biomarkers for diabetic kidney dis-

ease using ms-based metabolomics (FinnDiane study). Metabolomics, 2012 Feb;8(1):109-119.

3.1 INTRODUCTION

As the number of patients with diabetes increases, diabetic kidney disease (DKD) is a growing public health problem. An estimated third of type 1 diabetic patients will develop DKD over the course of several decades after diabetes onset [33, 40]. These diabetes patients have a 10-fold risk of premature death due to cardiovascular and other circulatory diseases, and the risk increases even more for those who develop renal failure [83, 39, 118]. Typical clinical manifestations of DKD include increased urinary albumin excretion rate (AER) and rising blood pressure, histological manifestations include the Kimmelstiel-Wilson nodules in the glomerulus [68]. DKD cannot be cured at present, but improved glycemic control and aggressive treatment of high blood pressure can halt the progression of the disease, especially when administered at an early stage of DKD [5, 124]. AER is the primary diagnostic biomarker for DKD in clinical practice. Elevated levels of AER measured at follow-up times with respect to the AER levels measured at baseline indicate if a patient is suffering a progressive form of DKD. For healthy individuals it is not uncommon to find elevated levels of AER which are clinically at the edge of microalbuminuria, but have no damage of the kidney function, therefore AER is not an early predictive marker or a quantitative measure of the kidney function at an early stage of kidney disease [20]. It may be possible that subtle alterations in metabolic pathways precede the changes that manifest as macroalbuminuria. These changing levels of the related low-molecular weight metabolite proles may therefore be useful as early markers of a progressive form of DKD. Mass Spectrometry-based metabolomics has been extensively applied in disease diagnostics [123, 56]. Nevertheless, there are only a handful of similar studies of human DKD [85, 84, 146] most of which were based on serum samples and differentiated between patients suffering from DKD and a healthy control group. In this study, urine samples from 52 Type 1 diabetic patients from the FinnDiane Study that were clinically defined as having a normal AER ($<$30 mg/24h, [85]) were profiled. Half of this group (26 patients) suffered from the progressive form of albuminuria; the other half did not show a progression in albumin excretion. Both Gas ChromatographyMass Spectrometry (GC-MS) and Liquid ChromatographyMass Spectrometry (LC-MS) were used to analyze a wide range of metabolites in these urine samples. The data was from an actual patient population measured within a less controlled environment. As changes in biological samples are often multifactorial [147, 100], both univariate and multivariate data analyses were used. The multivariate metabolite profiles to differentiate between the two groups were found using logistic regression (LR) with variable selection. Based on MSn fragmentation experiments (LC-MS only), manual interpretation combined with database searches we could identify several of the discriminating compounds that may be relevant for further biological studies.

3.2 EXPERIMENTAL

3.2.1 *Samples*

At baseline, Type 1 diabetic patients were recruited by the Finnish Diabetic Nephropathy Study Group (FinnDiane). The initial data collection was cross-sectional (serum and urine samples), but with longitudinal records of albuminuria and clinical history. These study patients suering from Type 1 diabetes mellitus had an age of onset below 35 years and a transition to insulin treatment that occurred within a year after onset. The classication of renal status was made centrally according to urinary albumin excretion rate (AER) in at least two out of three consecutive overnight or 24 h-urine samples. Absence of diabetic kidney disease was dened as AER within the normal range (AER <20 g/min or <30 mg/24h). Prospective clinical data were available for a subset of patients, all being male. There were 26 subjects that progressed from normal AER to microalbuminuria (PR) and had urine samples available at the time having normal AER. 26 clinically group-matched (age, diabetes duration, baseline albuminuria status, sex) non-progressive AER (NP) subjects were selected as study reference. Subjects for this group (NP) that had a long follow-up were preferentially included. Table 1 shows some clinical characteristics of these subjects for the non-progressive AER and progressive AER groups. A more detailed table is included in the supplement (Table S31).

	Clinical parameters (Normoalbuminuric)	
	Non-progressive	Progressive
Number of samples (Male)	26	26
Age (years)	36±9	35±10
Blood pressure (mm Hg)	80/132±8/10	83/132±11/15
BMI (kg/m^2)	25±2	27±3
Serum creatinine (μmol/l)	88±14	86±13
AER (mg/24 h)	12±6	14±7
HbA_{1c}(%)	8±1	9±1
Diabetes duration (years)	27±6	18±11
Follow-up time (years)	6±1	5±2

Table 1: Clinical characteristics of the subjects at baseline.

GC-MS

All urine samples were processed and analyzed once using a randomized sample sequence over multiple batches. After every 6^{th} study sample a Quality Control (QC) sample was injected. The QC sample was obtained by taking an aliquot of the same volume of all urine samples from this study. They were prepared once and measured in duplicate. All urine samples were analyzed with GC-MS according to the method described below.

LC-MS

The urine samples of the normal AER subjects were split into two aliquots, which were independently processed by the sample-pre-treatment. Each extract was subsequently analyzed by LC-MS once so that for each sample from the normal AER group in total two LC-MS analyses were obtained. A pooled QC sample (the same as for GC-MS) was analyzed after every 6^{th} sample.

3.2.2 *Materials*

GC-MS

For the GC-MS analysis, pyridine and N-methyl-N-trimethylsilyl triuoroacetamide were obtained from Mallinckrodt Baker BV (Deventer, The Netherlands) and Alltech (Breda, The Netherlands), respectively. Standards were purchased from Sigma-Aldrich (Zwijndrecht, The Netherlands).

LC-MS

For the LC-MS method, LC-MS grade acetonitrile (AcN) and MS-grade water were obtained from Biosolve (Valkenswaard, The Netherlands). All standards were purchased from Sigma-Aldrich, except for phenylalanine-d5, which was from C/D/N Isotopes Inc. (PointeClaire, Quebec, Canada). Acetic acid, formic acid and sodium hydroxide were obtained from Biosolve (Valkenswaard, The Netherlands), Acros Organics (Geel, Belgium) and Merck (Darmstadt, Germany), respectively.

3.2.3 *Sample preparation*

Initial sample preparation

Urine samples (stored at -80 °C) from FinnDiane were thawed at room temperature, homogenized using a vortex and centrifuged at 7500 g for 20 minutes. Specic volumes from the supernatant (typically 100 or 50 L) were taken and 10 vol% aliquots of a 1M acetic acid solution (adjusted to pH 6.0 with solid sodium hydroxide) were added and stored in vials. Samples were then stored at -80 °C prior to method specic sample preparation.

GC-MS specic sample preparation

Sample preparation for GC-MS analysis was done similar to the method described elsewhere [73]. In short, samples and standard urine solutions were thawed at room temperature and homogenized using a vortex. All 80 μL samples were mixed with 10 μL solutions containing the internal quality standards leucine-d3, glutamic acid-d3, phenylalanine-d5 and cholic acid-d4 (each present at a concentration of about 250 μg/mL in methanol/water (1:4 v/v)) and subsequently lyophilized at 37 °C in autosampler vials. The internal quality standards alanine-d4 and glucose-d7 in pyridine (each about 250 μg/mL) were added to the dry extracts prior to oximation. Oximation (90 min at 40 °C) was performed after adding 20 μL of a 56 mg/mL ethoxyamine hydrochloride solution in pyridine and 20 L of pyridine to the extracts. Next, a 10 μL mixture

of triuoracetylanthraceen, dicyclohexylphthalate (DCHP) and diuorobiphenyl (each at a concentration of 250 μg/mL in pyridine) was mixed with the extracts and the mixtures were silylated for 50 min at 40 °C with 200 μL of N-methyl-N-trimethylsilyl triuoroacetamide. Then the samples were centrifuged (500 g for 20 min) and the supernatant was taken for analysis by GC-MS. The nal GC-MS prepared samples were containing standards at a concentration of 10 ng/mL each. More details are given in the supplement.

LC-MS specic sample preparation

The FinnDiane urine samples were thawed at room temperature and subsequently mixed using a vortex. Next, the urine was centrifuged (7500 g for 10 minutes) at room temperature. To obtain the LC-MS samples, 35 μL of the supernatant was transferred into an autosampler vial and mixed with 10 μL of the internal standards mix (valine-d8, phenylalanine-d5, tryptophan-d5, thymine-d4 and reserpine in water at a concentration of 21, 21, 21, 35 and 7 μg/mL, respectively) and 25 μL water. For the validation of the LC-MS method (see supplement), two urine samples of one normoalbuminuric diabetic and one healthy volunteer were mixed in a 1:1 ratio. To 500 μL urine, 100 μL of a solution of 1 M acetic acid (adjusted to pH 6.0 with sodium hydroxide), 50 μL of the internal standards mix (valine-d8, phenylalanine-d5, tryptophan-d5, thymine-d4 and reserpine in water at a concentration of 60, 60, 60, 100, 20 μg/mL, respectively) and a varying volume of the calibration mix (phenylalanine, tryptophan and salicylamide in water at a concentration of 100 μg/mL) were added. Subsequently, water was added so that nally 1000 μL of validation sample was obtained for each calibration concentration. The samples were centrifuged (7500 g for 10 min) and the supernatant was analyzed by the LC-MS method. More details are given in the supplement.

3.2.4 *Identification of metabolites*

High resolution mass spectra were acquired using the 1200 Agilent gradient LC system coupled to a linear ion trapFourier transform (LTQ-FT) hybrid mass spectrometer and a LTQ-orbitrap mass spectrometer (both from Thermo Fisher Waltham, MA) Both systems were equipped with an ionmax ESI source. Spectra were recorded only in positive ESI centroid ion mode, with a source temperature of 275 °C, source voltage of 4 kV and a sheath gas of 40 AU.

FT

LTQ-FT was setup for a MS3 scanning method. Resolution was set to 12500 for all events to decrease scan time. Scan event one was a full scan with a scan range from 120 to 1000 m/z. Scan event two was set to fragment one of the targeted masses with a CID energy of 35% and isolation width of 1.5 m/z. Scan event three was set to Data Dependent Scan where it was set to fragment the most intense ion from scan event two with CID of 35% and isolation width of 1.5 m/z. All spectra were recorded in FT-mode with a typical mass accuracy of $<$1.5 ppm for both full scan and MS/MS spectra.

Orbitrap

The LTQ-orbitrap was setup for (MS^3) HCD fragmentation measurements. Resolution was set to 7500 for all events to decrease scan time. Scan event one was setup for a full scan with a scan range from 120 to 600 m/z. Scan event two was set to fragment one of the targeted masses in HCD fragmentation mode with a HCD energy of 30% and isolation width of 1.5 m/z. Scan event three was set to fragment one of the targeted masses in HCD fragmentation mode with a HCD energy of 50% and isolation width of 1.5 m/z. All spectra were recorded in FT-mode with lock masses 279.15909 and 391.28429 m/z in full scan. Mass accuracy in full scan spectra was <2 ppm and in HCD MS-MS spectra <4ppm.

Interpretation of the spectra

The spectra were collected and molecular formulae were then calculated by an in-house developed software tool using MS^3 fragmentation data. Spectra together with molecular formulae were further interpreted manually. The compounds that were identified were confirmed by comparison to spectra found in databases such as HMDB and by authentic standards where possible.

3.2.5 *Data Processing and analysis*

GC-MS

After sample analysis with GC-MS a target table of all relevant peaks (with known and unknown identity) was constructed. For this, a standard target table containing over 300 entries of endogenous plasma metabolites and urine specific peaks was used ultimately leading to a target list of 144 compounds. These compounds were integrated using a reconstructed ion chromatogram of a characteristic ion of each compound. The internal standards were quantified in the same manner. The GC-MS data have been corrected for internal standard response using DCHP followed by QC correction as described earlier [1].

LC-MS

After analysis of all samples using LC-MS a small subset of samples that contained 2-3 samples of each albuminuric class was used for screening compounds using software provided by Bruker-Daltonics (DataAnalysis). The m/z resolution was set to 0.01 Da, the minimum S/N ratio was set to 3 and to prevent integration problems the retention time window was set to 20 seconds. This resulted in a target list of about 600 features, where each feature was characterized by a retention time and a mass, and which could represent a metabolite. One metabolite can have multiple features. This target list was used as input for the software package Quant-Analysis by Bruker-Daltonics to create Extracted Ion Chromatograms (EICs) for all peaks for all LC-MS samples. In this way a peak table was constructed that was further narrowed down by removing features that had lots of missing values or no response at all in the QC samples [13]. As the dead time of the LC-MS method was 1 minute, features detected with retention times lower than 2 minutes were discarded.

Features with large differences in retention time (RSD higher than 5%) were removed as well. This resulted in a final peak table of 106 features. For the LC-MS data the response has been corrected using the most optimal internal standard as described earlier [1].

GC-MS and LC-MS sample normalization

For further statistical analysis the feature representing glucose (indicative for diabetes) was removed from the data. For healthy subjects the response of urine samples is often normalized by its creatinine level. In the case of varying degrees of kidney failure (i.e. microalbuminuric or macroalbuminuric), the creatinine levels cannot be used for normalization because they show irregular behavior due to diabetes and/or medication. Furthermore, a recent report shows that the creatinine concentration may vary over a large scale and that variances in urinary metabolite concentrations are not due to urine dilution effects; rather, they reflect actual metabolite variances [56, 109]. To circumvent this, as a means of normalization per sample, row scaling (i.e. subject normalization) was applied by taking the sum of the peak areas of all the components measured and dividing the response of each metabolite in a sample by this sum [66]. For LC-MS measurements duplicates were averaged per metabolite after visual inspection. All data were auto-scaled prior to further multivariate statistical analysis.
All multivariate data analyses were performed using Matlab 2008a [51]. The PCA charts were created using the PLS-toolbox 5.5.2 from Eigenvector Research [49].

3.3 RESULTS AND DISCUSSION

3.3.1 *Data processing and quality*

The LC-MS method was set up and successfully validated for urine (for details, see supplement), while the GC-MS method validation results were published earlier [73]. To monitor the stability of the analytical system, quality control (QC) samples were measured during both GC-MS and LC-MS analyses. In these QC samples the responses of each compound should be constant over time. The stability of the response per compound is expressed in Relative Standard Deviation (RSD) values, in which case for each metabolite the standard deviation of the response in all QC samples is divided by the average of the response in all QC samples. Large RSD values of response indicate poor repeatability, which can be assigned to instability of the analytical system or due to other variations such as instability of the analyte, etc. For GC-MS a total of 144 compounds were measured of which 106 had an RSD of less than 10% in the QC samples. The remainder was predominantly in the RSD range of 10-20% (for 9 compounds the RSD value was larger than 30%). For LC-MS, 106 features were found of which 65 had an RSD value of 10-20% and of which the remainder was predominantly in the 20-30% RSD range (for 8 compounds the RSD value was larger). 130 compounds (GC-MS) and 89 features (LC-MS) were selected for further data analyses based on RSD values less than or equal to 25%.

3.3.2 *GC-MS results*

Univariate tests for significant difference between non-progressive and progressive AER subjects (t-tests and Wilcoxon tests [89] were executed for each compound at a 95% significance level. The probability of a Type I error (i.e. the error of rejecting a null hypothesis when it is actually true) was further reduced using the Bonferroni approach by setting the significance level to α=0.05/130≈0.00039. None of the tested compounds showed a significant difference.

Principal Component Analysis (PCA), an unsupervised multivariate data analysis method, was used to investigate whether there was an apparent metabolomic separation between the non-progressive normo AER subjects vs. the progressive AER subjects. Using the GC-MS data of all metabolites, PC1 vs. PC2 did not show any clear separation between the two groups, whereas PC1 vs. PC3 (Figure 1) showed some clustering of the two classes.

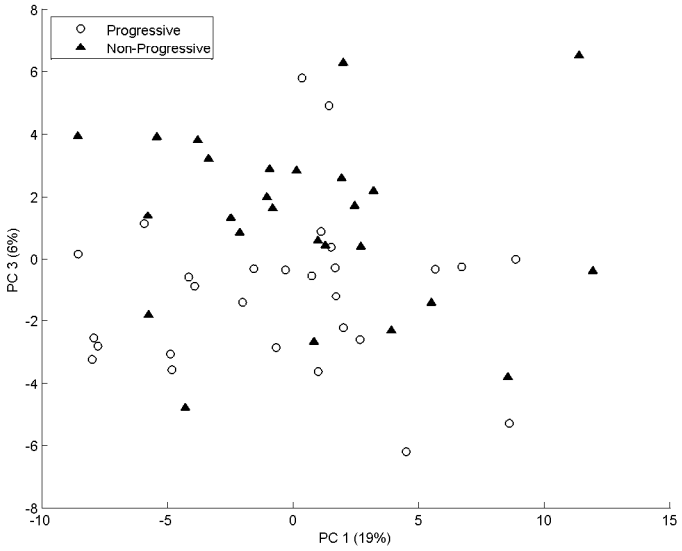


Figure 1: PCA score plot of the GC-MS data of urine from normal AER subjects using 130 compounds.

Because no clear separation was visible in the PCA score plot, supervised multivariate data analysis was used. As the classification problem had a dichotomous outcome (progressive or non-progressive), we used multivariate Logistic Regression (LR) in which the data was modeled in such a manner that the predicted outcome was always bounded between zero and one (corresponding with the 2 groups) [142, 41]. The choice of Logistic Regression was made because an ordinary linear regression method assumes that in the population a normal distribution of error values around the dependent variable is associated with each independent variable, and that the dispersion of the error values for each of these independent values is the same. However, the distribution of errors for any independent value cannot be normal when the

distribution has only two values [95]. To prevent over-fitting we used cross-validation followed by permutation tests. The choice of Logistic Regression in combination with variable selection allowed us to deal with the heterogeneity of the data and obtain stable cross validated models (for details see supplement).

The comparison between normal AER PR and normal AER NP subjects rendered a cross-validated Logistic Regression model with an accuracy of 65% and a precision of 64%. Ultimately 65 out of the 130 available metabolites were left in the final model. Accuracy can be viewed as the overall effectiveness of a classifier and precision as class agreement of the prediction with a specific class (e.g. progressive, non-progressive) [119]. Details regarding the cross-validation, the exact definition of accuracy and precision can be found in the supplement (Table S33).

To obtain a list of candidate biomarkers that form the predictive metabolic profile using this multivariate model, the significant contribution of each of these biomarkers to the model and the model itself was evaluated as described below. Permutation tests by means of randomizing the class membership vector were performed to evaluate the significance of the Logistic Regression model differentiating the non-progressive from the progressive normoalbuminuric subjects (see Supplement). The unpermuted model showed a tendency towards being significant (16 out of 100 permuted models gave equal or better classification results). Furthermore, using the regression vectors obtained from the permutation tests the significance of the contribution to the Logistic Regression model for each of the 65 compounds was determined. In total 34 compounds were found to be significant with a p-value lower than 0.05 (at most 5 out of 100 permuted models had a larger regression coefficient than the unpermuted model, see Supplement). Table 2 lists those compounds (21 in total) that were identified from this list of 34 compounds, ranked by their significance, together with their up-regulation (i.e. the relative concentration increased for the PR samples compared to the NP samples), their multivariate p-value and t-test p-value.

Compound	Up-regulation[a]	Significance[b]	t-test *P*-value
4-Oxoproline[c]	-1	0	0.03
Pseudouridine[d]	1	0	0.41
3,4,5-Trihydroxypentanoic acid[e]	-1	0	0.09
Deoxyfructose[c]	1	0	0.80
3-Hydroxy-3-(3-hydroxyphenyl) propanoic acid[e]	1	0	0.58
L-Valine[d]	1	0.01	0.25
2,3-Dihydroxy-3-methylbutanoate[c]	-1	0.01	0.27
5-Hydroxymethyl-2-furancarboxylic acid[e]	-1	0.02	0.09
Galactonic acid[e]	-1	0.02	0.01
2-Hydroxyvaleric acid[c]	1	0.02	0.58
N-formylproline or N-ethylproline[c]	-1	0.02	0.00
2-Hydroxyglutaric acid[d]	1	0.02	0.51
N-(3-hydroxybenzoyl)glycine[e]	1	0.02	0.47
L (+) Arabinose[d]	-1	0.03	0.13
Benzoic acid[d]	-1	0.03	0.04
3-Hydroxyphenylacetic acid[d]	-1	0.03	0.01
Glucuronide compound[c]	1	0.03	0.32
D-Glutamicacid[d]	-1	0.05	0.04
Gluconicacid[d]	1	0.05	0.04
Glycolicacid[d]	-1	0.05	0.05
L-Cystine[d]	1	0.05	0.37

[a] Increased concentration for progressive subjects, i.e. positive for progression: 1, decreased concentration for progressive subjects: -1

[b] The number of times that the metabolite in a permuted model had a larger regression coefficient than the unpermuted model divided by the total number of permutations executed

[c] Compounds were characterized, and only the class of metabolite could be suggested

[d] Compounds were identified, and the identity confirmed by an authentic standard

[e] Compounds were annotated based on elemental composition and by comparison to reference libraries

Table 2: Metabolites discriminating progressive and non-progressive normal AER subjects using the logistic regression of GCMS data

3.3.3 *LC-MS results*

With a Bonferroni corrected of 0.00056 (0.05/89), 3 out of the 89 features showed significantly different group means with either a t-test or a Wilcoxon test (Figure 2). As most of the statistically relevant features so far had not been identified in our lab for the LC-MS method, they were subjected to identication using high resolution MS and multi-stage MS/MS (see Experimental). Table 3 lists these 3 compounds, their respective p-values and their up-regulation (i.e. the relative concentration increased for the PR samples compared to the NP samples).

	P-value		
Compound	t-test	Wilcoxon	Up-regulation[a]
Substituted carnitine[b]	0.00000592	0.00000368	1
Hippuric acid[c]	0.00003066	0.00004267	-1
S-(3-oxododecanoyl) cysteamine[c]	0.00004540	0.00003714	1

[a] 1 means increased concentration for progressive subjects, i.e. positive for progression, -1 means decreased concentration for progressive subjects
[b] Compounds were characterized, and only the class of metabolite can suggested
[c] Compounds were annotated based on elemental composition and MS/MS (e.g. by comparison to reference libraries)

Table 3: The 3 metabolites that show a statically relevant difference between the group means of the progressive group and non-progressive normoalbuminuric group

The PCA score plot of LC-MS results of the normal AER subjects (Figure 3) showed some clustering along the diagonal of the rst and second principle component of the progressive and non-progressive subjects. Analogue to the GC-MS data analysis, LR with variable selection was used. The resulting model contained 42 features. The accuracy of the cross-validated Logistic Regression model for the binary classication of NP vs. PR was 75% with a precision of 73%. To evaluate the significance of metabolites contributing to the LC-MS based LR model differentiating between non-progressive and progressive normoalbuminuric subjects, and the model itself, permutations test were performed. The model was found to be significant; only 4 out of 100 permuted models gave equal or better classification results (see Supplement). Using the regression vectors from the permutation tests, the significance of each of the 42 features was determined. 14 features were significant (at most 5 out of 100 permuted models had a larger regression coefficient than the unpermuted mode, see supplement). High resolution MS and multi-stage MS/MS (see Experimental) were used to identify these features. Table 4 lists 8 of these 14 features that were identified ranked by significance, together with their up-regulation (i.e. the relative concentration increased for the PR samples compared to the NP samples) and the univariate t-test p-value. Literature study revealed that several compounds that were identified with a multivariate significance higher than a p-value of 0.05 could be linked to DKD. As these compounds also con-

tributed to the model, these compounds were added to Table 4. Note that the compounds that showed a univariate significance were also included in the LR model (Table 4).

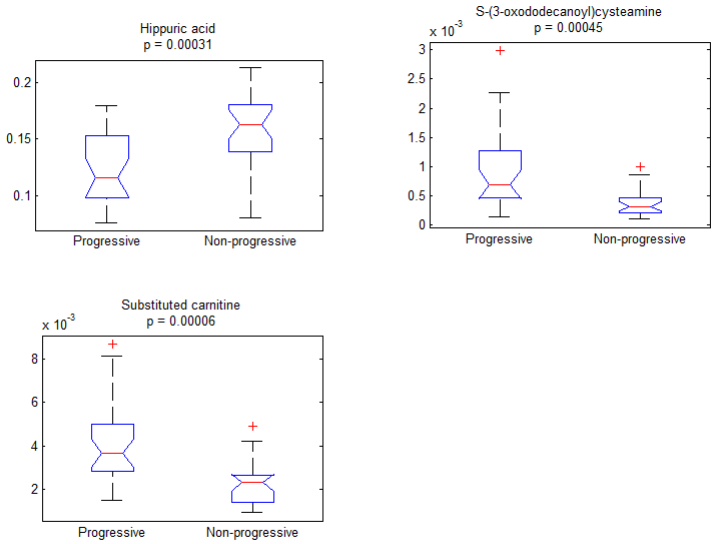


Figure 2: Boxplots of the 3 compounds that showed significant group means.

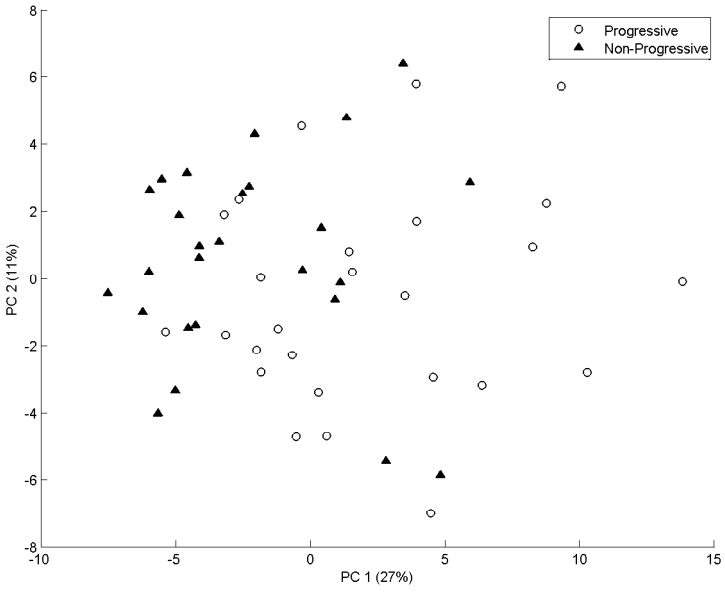


Figure 3: PCA score plots of the LC-MS data of the urine samples from Normal AER subjects using 89 features.

Compound	Up-regulation[a]	Significance[b]	t-test *P*-value
Tryptophan[c]	1	0	0.00
Salicyluric acid[c]	1	0.01	0.09
Substituted carnitine[d]	1	0.01	0.00
S-(3-oxododecanoyl) cysteamine[e]	1	0.01	0.00
Hippuric acid[e]	-1	0.02	0.00
Substituted carnitine[d]	1	0.02	0.00
N-methyl guanosine[e]	1	0.04	0.00
Substituted carnitine[d]	1	0.05	0.00
Kynurenic acid[c]	1	0.08	0.18
2-(2-phenylacetoxy)propionylglycine[e]	1	0.09	0.45
substituted carnitine[d]	-1	0.09	0.36
Indoleacetic acid[c]	-1	0.1	0.15
3-methylcrotonylglycine[e]	1	0.1	0.03
Heptanoylcarnitine[d]	1	0.1	0.00

[a] Increased concentration for progressive subjects, i.e. positive for progression: 1, decreased concentration for progressive subjects: -1

[b] The number of times that the metabolite in a permuted model had a larger regression coefficient than the unpermuted model divided by the total number of permutations executed

[c] Compounds were characterized, and only the class of metabolite could be suggested

[d] Compounds were identified, and the identity confirmed by an authentic standard

[e] Compounds were annotated based on elemental composition and by comparison to reference libraries

Table 4: Identified compounds from the metabolites discriminating most between progressive and non-progressive normoalbuminuric subjects using logistic regression model of LCMS data

3.3.4 *Strengths and weaknesses*

Both GC-MS and LC-MS data suffer from heterogeneity in the (samples of the) study population. The origin of this heterogeneity is the result of multiple factors, the main ones are: (i) the diculty to obtain an exact kidney phenotype, as discussed in the introduction; AER is the primary diagnostic biomarker in clinical practice, but its usefulness as an early marker is limited due to high natural variance [20] ; and (ii) the uncontrolled environment in which the urine samples were taken (e.g. urine metabolite concentrations depend on diet, variation due to slight differences in sampling 24 hour urine, etc.). This heterogeneity and the small number of prospective samples severely complicated proper statistical analyses. In cross-validating the binary classication models, the leave two out strategy was used (1 sample was left out for each class). In

order to select those features/compounds in the model that were specific for the whole dataset and not just for a few samples, a variable selection method (see above) was included. It turned out that when a subset of variables was used more accurate predictive models were obtained that were less susceptible to dierent training/test-set schemes (data not shown). Although the number of prospective samples is small the nature of the data (i.e. urine samples at an early stage of diabetic kidney disease) and the fact that some of the found biomarkers were already related to DKD the obtained results certainly give rise to future research.

3.3.5 *Biological context of the new candidate biomarkers*

Reviewing the metabolites in Table 2 it is interesting to note that many of the GC-MS compounds are carboxylic compounds and/or acidic metabolites that are prevalently detected in urine [76] (5-hydroxyme-thyl-2-furancarboxylic acid , benzoic acid and hippuric acid). Others are endogeneous amino acids (valine, serine). From the identified compounds, we have not found any documented direct relation to DKD. Deoxyfructose could be related to a phenomenon called diabetic stress in which deoxyglucosone is converted to the less reactive deoxyfructose [71]. Galactonic acid has been associated with diabetic retinopathy [59]. Rainey et al. already suggested an interaction of 5-hydroxymethyl-2-furancarboxylic acid with galactonic acid [92].
The metabolites in LC-MS (Table 4) are either: (1) acylcarnitines, (2) acylglycines, (i.e. salicyluric acid, hippuric acid, (2-phenylacetoxy- propionyl) glycine and 3-methylcrotonylglycine) and (3) compounds related to the tryptophan metabolism (i.e tryptophan, indoleacetic acid and kynurenic acid).
It is perhaps not that straightforward to link the metabolites to a specific pathway as a recent study already demonstrated that many of the acylglycines and tryptophan metabolites in mammalian blood have shown a relation to gut microbiome [143] which could explain a presence in urine by regular excretion. In general, acyl-carnitines are formed in the fatty acid metabolism pathway to transport the long-chain acyl groups of fatty acids into mitochondria, where these groups are broken down through β-oxidation to acetate to obtain energy via the citric acid cycle. Under normal homeostasis conditions, carnitine is eliminated by excretion in urine, in both free and esteried forms, mainly as acetylcarnitine [22]. A higher acylcarnitine to carnitine ratio in urine in relation to plasma is suggested to be the result of a less ecient reabsorption of acylcarnitines or of a renal acylation of carnitine followed by leakage of the locally formed acylcarnitine product into urine [133, 134, 103] .
Glycine conjugation is an eective detoxication system for preventing accumulation of acyl-CoA esters in several inherited metabolic disorders. Acylglycines in urine have been reported as the direct expression of accumulation of the correspondent acyl-CoA esters in the mitochondrion [15].
Tryptophan metabolism changes with DKD have been reported before [15], where tryptophan plasma concentration in animals with experimental renal failure decreased while a simultaneous increase of metabolites related to the kynurenine pathway (e.g. kynurenic acid) in plasma were observed. It was demonstrated in animals that the content of kynurenic acid in kidneys is the highest among all tissues [81]. Furthermore, it has been known that kynurenic

acid is the main metabolite excreted from organism (rats) by means of tubular secretion. In renal failure this mechanism is considerably impaired, which in consequence leads to excessive accumulation of this substance in the organism [98].
Of interest are the elevated levels of 2-(2-phenylacetoxy)propionyl-glycine in the progressive patients. In the absence of medium chain acyl-CoA dehydroxygenase (MCAD) phenylpropionacid is converted to phenylpropionglycine instead of benzoic acid. Phenylpropionglycine is detected only in the urine of MCAD-deficient patients and has been used as a biomarker for the diagnosis of this condition [143]. However, in summary, it should be mentioned that further research is required to investigate the biochemical context of all the candidate biomarkers in more detail.

3.4 CONCLUSIONS

It was demonstrated that based on LC-MS measurements of urine samples a statistically significant multivariate model could be constructed to distinguish between progressive and non-progressive subjects within the normal AER group with an accuracy of 75%. Many of the compounds contributing to the model could be grouped in one of three classes, i.e. acyl-carnitines, acyl-glycines and compounds related to the tryptophan metabolism. All of the compounds that were measured that show a univariate significant difference between the two groups were included in the metabolic profile defined by the multivariate model. The metabolic profile also included metabolites that did not show a univariate significant difference and emphasizes the additional benefit of multivariate statistics over univariate statistics alone in preventing overlooking candidate biomarkers. Future research will focus on the discovery of additional biomarkers using complementary metabolomics platforms and the validation of the explorative biomarker proles with a validation set. In addition, more effort will be directed to the biological interpretation: it will be investigated which pathways were involved in the biochemical changes associated with the onset, development and progression of DKD, and whether these changes are the same during onset and progression, or if different changes of biochemistry occur at the different stages of DKD, e.g. due to the disease pathology or due to the use of medication after onset of DKD. In summary, the results obtained demonstrate the potential of metabolomics in the study of diabetic complications, as subtle changes in the urine metabolome precede the clinically signicant rise in AER.

ACKNOWLEDGEMENTS

This project was supported by the EU DiaNa project (EU-FP6:LSHBCT-2006-037681). The identification of metabolites was supported by the Netherlands Metabolomics Centre (NMC), which is a part of The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. We thank SARA Computing and Networking Services (www.sara.nl) for their support in using the Life Science Grid.

S3

SUPPLEMENT TO DISCOVERY OF EARLY-STAGE BIOMARKERS FOR DIABETIC KIDNEY DISEASE USING MS-BASED METABOLOMICS (FINNDIANE STUDY)

S3.1 EXPERIMENTAL

S3.1.1 *Material*

Table S31 shows the detailed clinical characteristics of the subjects for the non-progressive AER and progressive AER groups.

S3.1.2 *Data acquisition*

GC-MS

GC-MS analysis is performed as described in [73]. The derivatized samples were analyzed with an Agilent 6890 gas chromatograph (Agilent Technologies, Waldbronn, Germany) coupled to an Agilent 5973 mass selective detector. Sample volumes of 1 uL were injected in a DB5-MS capillary column (30 m x 250 μm i.d., 0.25- μm film thickness; J&W Scientific, Folson, CA, USA) using PVT injection (Gerstel CIS4 injector) in the splitless mode. The temperature of the PTV was 70 °C during injection, and 0.6 min after injection, the temperature was raised to 300 °C at a rate of 2 °C/s and held at 300 °C for 20 minutes. Initially, the GC oven temperature was 70 °C, 5 minutes after injection the GC oven temperature was increased at a rate of 5 °C/min to 320 °C and then held constant for 5 minutes. Helium was used as carrier gas and pressure programmed such that the helium flow was kept constant at a flow rate of 1.7 mL/min. Detection was achieved using MS detection in the electron impact mode and full scan monitoring mode (m/z 15-800). During the elution of urea, the detector was shortly switched off. The temperature of the ion source was set at 250 °C and that of the quadrupole at 200°C.

LC-MS

The pretreated urine samples were analyzed using a 1200 Agilent gradient LC system, consisting of a degasser, a binary pump and an autosampler. A T3 column (C18, 2.1 x 100 mm, 3 μm particles; Waters, Milford, MI, USA) was employed at room temperature for analysis. Mobile phase A consisted of a mixture of water/acetonitrile/formic acid in a ratio of 99:1:0.1 (v/v/v) and mobile phase B was a mixture of water/acetonitrile/formic acid in a ratio of

	Normoalbuminuric	
	non-progressive	progressive
Number of samples	26	26
Age (years)	36 ± 9	35 ± 10
Diastolic blood pressure (mm Hg)	80 ± 8	82 ± 11
Systolic blood pressure (mm Hg)	132 ± 10	132 ± 15
BMI (kg/m^2)	25 ± 2	27 ± 3
Serum creatinine (mol/l)	88 ± 14	86 ± 13
Albumin Excretion Rate (mg/24 hr)	12 ± 6	14 ± 7
HbA_{1c} (%)	8 ± 1	9 ± 1
Diabetes duration (years)	27 ± 6	18 ± 11
Follow-up time (years)	6 ± 1	5 ± 2
Urine volume in 24 hr (ml)	2236 ± 709	2153 ± 612
Collection time of urine in 24 hr (minutes)	1396 ± 78	1399 ± 83
Albumin concentration (mg/l)	6 ± 4	7 ± 3
Antihypertensive medication	6	2
ACE Inhibitor	4	2
Angiotensin 2 receptor blocker	0	0
Betablocker	2	0
Diuretic medication	1	1
Other antihypertensive	0	0
Lipid lowering medication	4	4
Calcium blocker	1	2
Waste Hip Ratio	1 ± 0	1 ± 0
Age onset (years)	9 ± 7	17 ± 8
Insuline dose per kg body weight	1 ± 0	1 ± 0
esitmated Glucose Disposal Rate (mg/(kg*min)	6 ± 3	6 ± 2
Trigliceride	1 ± 0	2 ± 2
Total cholesterol (mmol/l)	5 ± 1	5 ± 1
HDL cholesterol (mmol/l)	1 ± 0	1 ± 0
Smoker	7	6

Table S31: Detailed clinical characteristics of all subjects.

1:99:0.1 (v/v/v). A gradient program with a total flow rate of 250 μL/min was used (Table S32). Urine samples were in the autosampler tray at 4 oC and for each run 5μL sample was injected. A short piece of PEEK tubing (0.005 i.d.) connected the column with the divert valve of the BrukerMicrOToF (BrukerDaltonics, Bremen, Germany). The divert valve directed a calibration solvent (isopropanol/water/formic acid/1M sodiumhydroxide) in a 50:50:0.2:0.5 (v/v/v/v) ratio, delivered at a flow rate of 50 μL/min by an LKB pump (Pharmacia, Uppsala, Sweden)) to the LC sprayer on the mass spectrometer during the first and last minutes of the run, while the column effluent went to waste. During the remainder of the time, the valve directed the column effluent to the LC sprayer, while the calibration solvent went to waste. Spray settings of the ToF detector were: endplate offset -500 V, capillary voltage -4900 V, nebulizer gas pressure 1.6 bar, dry gas flow 8.0 L/min and dry temperature 150 °C.

Time(min)	A(%)	B(%)
0	100	0
2	100	0
10	85	15
15	40	60
20	40	60
22	2	98
27	2	98
28	100	0
30	100	0

Table S32: LC-MS gradient program.

S3.1.3 *Multivariate Logistic Regression and variable selection*

Logistic regression in combination with regularization was implemented using the Matlab code written by M. Schmidt [112]. All calculations were performed in Matlab 2008a ([51]).

Ordinary linear regression method assumes a normal distribution of error values around the predicted variable associated with each independent variable and that the dispersion of the errors is the same. When the distribution of the predicted variable only has two possible outcomes, the distribution of errors for any independent value cannot be normal. As a consequence linear regression faces a problem when dealing with a dependent variable with a ceiling and a floor: the same change in an independent variable has a different effect on the dependent variable (Y) depending on how close a predicted value is to the real value of Y (i.e. is non-linear) [95]. Using Logistic regression (LR) this non-linearity problem has been overcome using an appropriate transformation function (Logit).

To explain the LR, it is useful to assume that a sample (i) has a probability (P_i) of belonging to a class, where P_ii=1 means i belongs to class 1 and P_i=0, i does not belong to the class (i.e. the dichotomous dependent variable that is

modeled). The odds (O) of a sample (i) belonging to a class is defined as the ratio of the probability to one minus the probability:

$$O_i = [\frac{P_i}{1 - P_i}] \tag{15}$$

Consequently, the probability is expressed as:

$$P_i = [\frac{O_i}{1 + O_i}] \tag{16}$$

Based on these formulas, the probability can never equal or exceed one: no matter how large the odds become in the numerator, they will always be smaller by one than the denominator. Of course, as the odds become large, the gap between the odds and the odds plus 1 will become relatively small and the probability will approach (but not reach) one. Conversely, the probability can never fall below zero. As long as the odds equal or exceed 0, the probability must equal or exceed zero. This way, the probability can never equal or exceed one and conversely, the probability can never fall below zero. The smaller the odds in the numerator become, the larger the relative size of the 1 in the denominator. The probability comes closer and closer to zero as the odds come closer and closer to zero.

In order to use a regress the dependent variable the floors and ceilings have to be removed. Taking the natural log of the odds (L_i, the Logit) eliminates the floor of 0 and transforming probabilities into odds eliminates the ceiling of 1.

$$L_i = ln[\frac{P_i}{1 - P_i}] \tag{17}$$

Consequently, the probability is expressed as:

$$P_i = \frac{e^{L_i}}{1 + e^{L_i}} \tag{18}$$

Without this floor and ceiling, a linear relationship between the independent variables (x_i) and Logit can be computed. In our case, we use multivariate linear regression, in which the Logit is modeled as

$$L_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k} \tag{19}$$

The probability of a sample belonging to a class is calculated as

$$P_{i,\beta} = \frac{e^{L_i}}{1 + e^{L_i}} = \frac{e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k}}}{1 + e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_k x_{i,k}}} \tag{20}$$

The modeling process minimizes the sum between the probability Pi and the real class assignment (yi) class for all samples (i) by selecting a proper β (the regression vector) via:

$$min = \sum_{i=1}^{n} |y_i - P_{i,\beta}|^2 \tag{21}$$

In order to reduce the number of variables and to create a more robust model (i.e. select those features/compounds that were specific for the whole dataset

to model and not just for a few samples), a penalty (λ) is added [41], Equation 21 now changes to

$$min = \sum_{i=1}^{n} |y_i - P_{i,\beta}|^2 + \lambda|\beta|^2 \quad (22)$$

As the penalty λ becomes higher, the search for a minimum will force coefficients to be smaller and ultimately set to 0 (i.e. certain variables will not contribute to the model). Variables that have a small or no effect at all on the classification results are the first that will be removed from the regression model by setting the coefficient to zero. Dependent on the value of λ more or less variables are included in the model.

S3.1.4 *Cross-validation*

Proper values for penalty λ were obtained by using cross-validation together with variable selection. This can be viewed as a two-step process:

Step 1 In case of the non-progressive vs. progressive data 26 non-progressive and 26 progressive samples were used. The leave-two-out method was used (1 sample from each class). The data was split into 26 subsets. 25 out of the 26 subsets were used to create a logistic regression model. With this model a prediction of the left-out-set was made. This process was repeated 26 times (i.e. for each subset) so ultimately for each sample a prediction was available. The regression vector of each model was stored.

Step 2 Because the Logistic Regression was penalized to reduce the number of variables, the modeling from step 1 was repeated several times. Each time with a different penalty λ ranging from exp(-2) to exp(2) in steps of 0.1 (in total 41 times). The optimal value for λ was determined by plotting the classification error (for the left-out-sets) against λ and locating the minimum. In case of the non-progressive vs. progressive data, 26 regression vectors were obtained. The average over all these regression vectors was taken as the final classification result. From this average regression vector the importance or ranked contribution of each variable (i.e. metabolite) to the classification model was calculated.

S3.1.5 *Accuracy and precision*

To explain these terms it is useful to setup a confusion matrix (Table S33) for two possible outcomes [119].

The ability of the classification model to correctly predict the class a subject belongs to is described by its accuracy. The accuracy is defined as the proportion of true results (both true positive and true negative) in the whole population.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (23)$$

An accuracy of 100% means that the predicted values are exactly the same as the observed values. For a classification model the number of false positives (i.e. subjects that do not progress but are predicted as suffering from the progressive form of albumin excretion) has to be minimal and is indicated by its precision. The precision is defined as the proportion of the true positives (i.e.

subjects that suffer from progression that are correctly predicted) against all the positive predicted results (both true positives and false positives).

$$precision = \frac{TP}{TP + FP} \tag{24}$$

S3.2 RESULTS

S3.2.1 *Results of method validation*

The suitability of the LC-TOFMS system for quantitative and semi-quantitative measurements was evaluated by constructing a calibration curve from a mixture of tryptophan, phenylalanine and salicylamide that was spiked in urine samples, as described in the manuscript. These samples, spiked in a concentration range of 0 10 ng/μL, were analyzed in the presence of deuterium-labeled analogues (d5-tryptophan and d5-phenylalanine) as internal standards or an internal standard analogue eluting close to the calibrant (d5-tryptophan in the case of salicylamide) in duplicate, while each level was also prepared in duplicate. For all calibrants, two calibration curves were realized by plotting the peak area against the injected concentration and the area ratio (analyte vs. complementary internal standard) against the injected concentration. Good linearity (peak area of the calibrant versus concentration) was observed for all calibrants over the measured concentration range. Limits of detection (LODs) and limits of quantification (LOQs) were determined using the slopes of the calibration curves and the peak area standard deviation of the 0.3 ng/μL spiked urine sample (for salicylamide), or, because of significant endogenous levels of calibrants that are present, the peak area standard deviation of a urine sample spiked with 0.3 ng/μL deuterium-labeled calibrant (for tryptophan and phenylalanine). LODs were defined as 3.3 times the ratio of standard deviation and calibration curve slope, while LOQs were 10 times that ratio. For salicylamide, tryptophan and phenylalanine, LODs (in μg/ml) were 0.04, 0.05 and 0.05, respectively. LOQs were determined to be 0.11, 0.15 and 0.15 for salicylamide, tryptophan and phenylalanine, respectively. In order to assess possible losses

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Table S33: Confusion matrix for two possible outcomes p (positive) and n (negative)

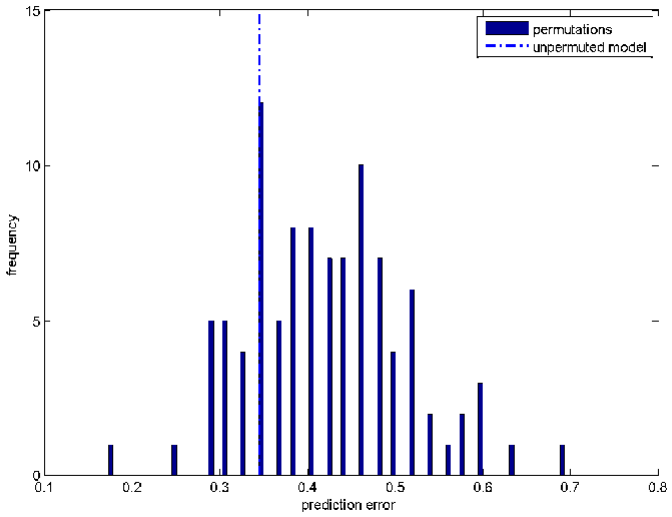


Figure S31: Prediction results obtained by permutation test of GC-MS model for non-progressive vs. progressive AER.

during the sample preparation steps (i.e. mixing, centrifugation and dilution), urine samples were spiked with calibrants before and after these steps (having a concentration of 3 ng/μL of all calibrants). On each of three days, the samples were prepared three times and all measured in duplicate. Comparison of the samples spiked before and after the preparation steps revealed that peak areas differed less than 10 percent for the three calibrants. These results indicated an acceptable deviation of the measured concentrations due to the performed sample preparation steps. Repeatability was evaluated by analyzing three preparations of the urine sample with calibrant concentrations of 3 ng/μL in duplicate on three consecutive days. Within each day, repeatability of the calibrant peak areas was well within 5 percent, while the repeatability between the days was 10 to 15 percent.

S3.2.2 *Permutation tests*

In order to test the predictive ability of the Logistic Regression models for non-progressive vs. progressive Normal AER, permutation tests were performed. This was done for both GC-MS as LC-MS data sets. The permutation test consisted of a randomization of the class vector followed by the creation of a Logistic Regression model with variable selection. This was repeated 100 times. Figure S31 shows the distribution of the classification errors for all 100 permutations for the GC-MS data.

In case of the GC-MS based model, 16 out of 100 permuted models gave lower classification results than the unpermuted model. The significance of the unpermuted model hence shows a tendency towards significance (p-value = 0.16 = 16/100).

Figure S32 shows the distribution of the classification errors for all 100 permutations but now for the LC-MS data. In this case only 4 out of 100 permuted

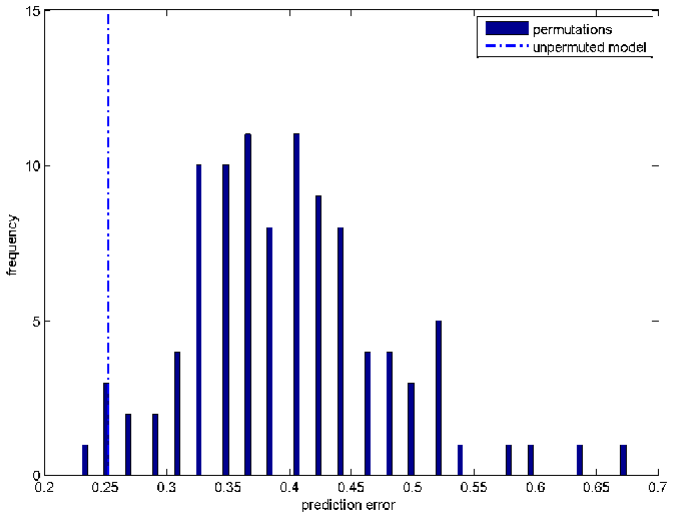
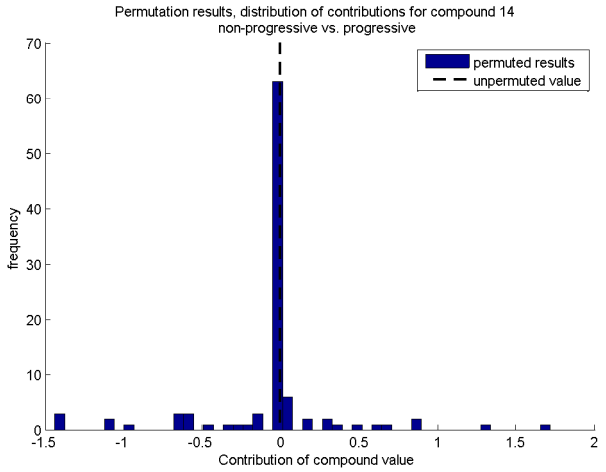


Figure S32: Prediction results obtained by permutation test of LC-MS model for non-progressive vs. progressive AER.

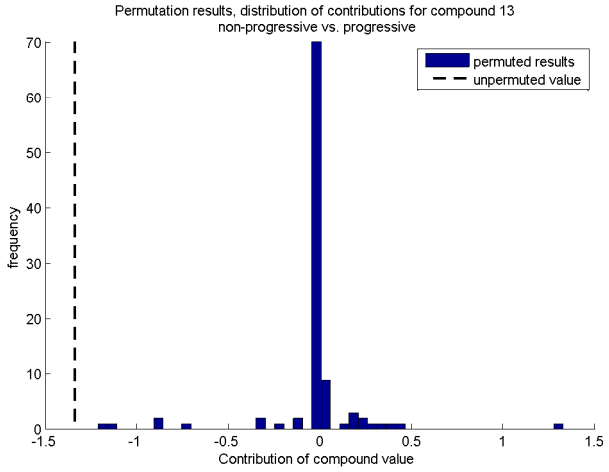
models gave better classification results than the unpermuted model which makes this model significant (p-value = 0.04 = 4/100).

S3.2.3 *Significance testing of the compounds/features*

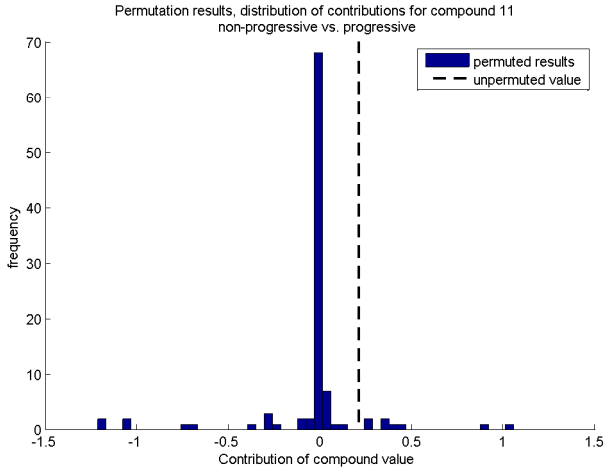
In order to test the significance of a coefficient in a regression vector its contribution was compared to the contributions to regression vectors generated by permutated models. The hypothesis is that if coefficients found in the unpermuted model do not differ from those generated by permuted models there is no significant contribution. Figure S3 shows 3 individual cases in which a) there is no significant contribution (p-value $\gg$ 0.05 and b) there is significant contribution (p-value $\leq$ 0.05) and c) there is almost a significant contribution (p-value $>$0.05 and $\leq$0.10). The Logistic Regression model for non-progressive vs. progressive AER based on GC-MS data showed 34 significant compounds. The Logistic Regression model based on LC-MS data showed 14 significant compounds.



(a) no significance



(b) significant



(c) almost significant

Figure S33: significance testing of variables using permutation results. a) no significance, b) significant and c) almost significant

4

RAPID METABOLIC SCREENING OF EARLY ZEBRAFISH EMBRYOGENESIS BASED ON DIRECT INFUSION-NANOESI-FTMS

4.1 ABSTRACT

Single zebrafish eggs were rapidly profiled using High Resolution-Direct Infusion-nanoelectrospray-Mass Spectrometry with limited sample preparation and without separation. The analysis time per sample is around 1 minute. Using this approach the different developmental stages of zebrafish eggs can be characterized by their active metabolites. Five different development stages with distinct metabolic fingerprints could clearly be observed when untargeted analysis is performed and the data are plotted using Principal Component Analysis (PCA). Using this approach early embryogenesis is followed with a time resolution of 1 hour and 102 features proved relevant. Of these, significant number of putatively identified compounds has not been reported earlier to have any association with early zebrafish embryogenesis yet. The onset of gene expression and the increase in energy requirement is reflected by the measured metabolome complementing earlier reported transcriptomics studies from a systems biology point of view. By deyolking and dechorionation eggs at two early developmental stages, we were able to observe distinct changes in localized metabolism.

Robert-Jan Raterink*, Frans Meindert van der Kloet*, Jiajie Li, Niels Abraham Wattel, Marcel Johannes Maria Schaaf, Herman Peter Spaink, Ruud Berger, Robert Jan Vreeken, Thomas Hankemeier. Rapid metabolic screening of early zebrafish embryogenesis based on direct infusion-nanoESI-FTMS, Metabolomics, 2013, 9(4):864-873

* Equally contributing authors

4.2 INTRODUCTION

In the past decade the Zebrafish (*Danio Rerio*) has become a popular vertebrate model system for studying human development and disease [2, 88, 38]. This is because the development of the zebrafish is very similar to the embryogenesis in higher vertebrates, including humans. But unlike mammals, zebrafish develop from a fertilized egg to an adult outside the female in a transparent egg. This makes it possible to observe developing embryos from the single cell to the entire organism level [110]. The development time of the embryo is fast (after about 2 days most common vertebrate specific body features can be seen including brain, eyes, ears and all internal organs) and the number of offspring is large (100-200 eggs per mating). Moreover, because they are small, available in large numbers and maintained at low cost, zebrafish embryos are ideal as model systems in high-throughput whole organism preclinical drug screening and toxicology studies as less drug is required, a larger number of animals can be used and less ethical issues are associated [97, 121]. Therefore, the zebrafish embryo can bridge the gap between cell assays and rodent assays. The external and rapid development as well as transparency of zebrafish embryos is ideal for observable phenotype-based screening at cellular and organ-tissue level using live imaging, for example using a Complex Object Parametric Analyzer (COPAS)[21]. Although these phenotypic changes are adequate for specific drug-induced biological responses, they dont reveal system-wide responses and are not sufficient in elucidating the mode of action or potential toxicity of drugs[121]. With the ever increasing interest in studying biological systems in a holistic manner (systems biology) the necessity for delivering qualitative and quantitative data of complete biological systems for which zebrafish offer many advantages becomes clear [3]. Metabolomics is a powerful tool within systems biology and investigates the complex interactions of the metabolism and metabolic networks [31, 130, 19]. One of the greatest strengths of metabolomics is the ability to capture a molecular snapshot of metabolites as reactants, intermediates or products of (enzyme-mediated) biochemical reactions. Metabolomics complements genomics, transcriptomics and proteomics for metabolites are in a unique position as they are building blocks for all other biochemical structures including proteins (amino acids), genes and transcripts (nucleotides) and cell walls (lipids) [107, 27].

So far, the number of metabolomics studies on zebrafish embryos is limited [94, 42, 96, 116, 43]. Most of these studies included chromatography which is time consuming and therefore less suitable for fast & high-throughput metabolic screening. Recent studies outlined the power of High-Resolution-DI-MS (HR-DI-MS) metabolomics on other complex samples [34, 82, 12, 26]. Most of these studies were performed using flow injection ESI-MS and therefore did not explore the advantages of nanoESI over normal ESI with respect to ionization efficiency and ion suppression effects [30]. In this paper we describe a rapid metabolic profiling method of zebrafish eggs, based on lysis of single zebrafish eggs and subsequent HR-DI-nanoESI-MS analysis. A clear metabolic distinction between developmental stages in early embryogenesis is described and up-and down regulation of some important primary metabolites is shown.

By deyolking and dechorionating the embryos, we were able to highlight localized and time resolved metabolism.

4.3 MATERIALS AND METHODS

4.3.1 *Chemicals and materials*

Methanol was from Biosolve (Valkenswaard, The Netherlands). Water was obtained from a Millipore high purity water dispenser (Billerica, MA, USA). All solvents were HPLC grade. The labeled amino acids ("Cell Free" amino acid mix (20 AA) (U-13C, 98%+; U-15N, 98%)) were bought from Cambridge Isotope Laboratories (Andover, MA, USA) and added to the samples at a concentration of 1μg/mL. Reserpine (also used as internal standard) was supplied by Fluka (Buchs, Switzerland) and added to the samples at a concentration of 500ng/mL.

4.3.2 *Whole embryo experiments*

Two wild type (strain A/B) parent groups were maintained under standard zebrafish aquarium conditions [75]. Since many cell differentiation and phenotopic processes occur within 48 hours after fertilization, the following 5 early developmental stages were chosen: the 4-cell (1hpf), 64-cell (2hpf), 1k (3hpf), 50%Epiboly (5hpf) and the 18-somites stage (18hpf) For each of the five developmental stages 8 eggs were analyzed separately in triplicate.

Lysation protocol for the whole embryo

1. The egg was optically selected under a microscope in the relevant developmental stage and pipetted to a 1.5 mL eppendorf tube. As much as possible of the egg water was removed.
2. The egg was washed 3 times with 1 mL demineralized H_2O.
3. 100μL (9:1, v/v methanol:water) including the internal standards (1μg/mL Labeled amino acids+ 500ng/mL Reserpine) was added. Immediately the sample was snap-frozen in liquid nitrogen for 2 minutes. This step should quench metabolism and precipitate the proteins.
4. After snap-freezing the sample was sonicated for 2 minutes to lyse and to homogenize the sample and visually inspected to confirm homogeneity. In case of non-homogeneity the sample was snap-frozen and sonicated a second time.
5. Precipitated proteins were spun down by centrifuging the lysate at 16.1rcf at 0 °C. The supernatant (80μL) was used for DI-MS analysis.

4.3.3 *Deyolking and dechorionating experiments*

Two developmental stages 1kcell (3hpf) and somite (18hpf)) of the same wild type were measured (pool of n=10) in triplicate.

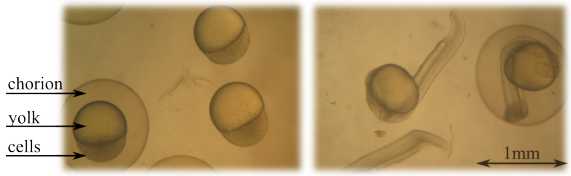


Figure 1: Photomicrograph of zebrafish embryos: (left) 1k-cell stage and (right) somite stage. In both stages some of the embryos were dechorionated in order to perform the deyolking

Dechorionating

1. The chorion of the egg was removed mechanically by a tweezer and was put in a 1.5mL eppendorf tube. Using this methodology only the membrane of the chorion could be analysed. The same lysation protocol as described above was used however, instead of 100 μL , now 1mL of solvent+IS (since it was a pool of 10 egg instead of a single egg) was added and no washing was done (since it would remove parts of the sample).

Deyolking

1. After dechorionating the egg only consists of the yolk with its cells (see Figure 1). The yolk was removed by adding a Ringer solution to the egg and pipetting it up-and down a few times.
2. Subsequently, the sample was centrifuged for 5 minutes at 0.8rcf (to segregate the cells) and the supernatant (including the yolk) was pipetted away. This way only the cells of the embryo could be analyzed.[110].
3. 1 mL of IS was added to the pallet and the same lysation protocol was used as described above.

Deyolking

1. First the non-fertilized egg (only 3hpf stage) was dechorionated resulting in only the yolk with its fluid. This way also the fluid in the yolk could be analyzed.
2. 1 mL of IS was added and the same lysation protocol was used as described above.

In the deyolking experiments again also whole-eggs were included as a reference.

4.3.4 *MS Analysis*

The analyses were performed by DI-nanoESI-MS in the positive ion mode using the automated Advion NanoMate Triversa system (type A chip) coupled to a LTQ-FT Ultra (Thermo Fisher Scientific). Eppendorf 96 well plates were used on which all the samples were randomly distributed. Of each sample 5μL was infused using a pressure of 0.2PSI and an electrospray voltage of 1.48kV in the positive ion mode.

Mass spectra were recorded using three scan ranges containing 20 scans: 50-250; 250-500; 500-1000 m/z (in this order) at a resolution of 100.000. Separate scan ranges instead of one full scan range was chosen in order to enhance sensitivity. The MS was tuned with inlet capillary temperature of 120°C, capillary voltage of 35V and the tube lens voltage of 50V. For tandem MS as well as for the deyolking experiments an LTQ-Orbitrap XL (Thermo Fisher Scientific) was used with inlet capillary temperature of 120°C, capillary voltage of 25V and the tube lens voltage of 80V.

4.3.5 *Data processing*

The first 10 scans (approximately 10 seconds) of every sample were averaged using XCalibur software (version 2.0.7; Thermo Fischer). These average scans were stored in separate files. In some samples no spray or no stable spray was obtained, i.e. no average scan could be made. These samples were discarded from further processing. Using a resolution of 100.000 (at 400 m/z) each sample typically resulted in 5 to 10 thousand unique features in the average spectrum. In contrast to hyphenated MS data, the drawback of DI-MS data is that all masses co-elute. Aligning features over samples now is solely based on accurate mass. To align the masses across the different samples we created mass-bins of very small sizes (0.0003 Da) and assigned the different masses to the nearest bin. For a mass range of 50 to 250 m/z this generated approximately 650.000 mass bins.
All data analysis was done using Matlab (version 2011a 64bit; MathWorks). In order for the data to be analyzed in Matlab the Xcalibur averaged spectra files were converted to mzXML format (ReAdW version 4.3.1).

4.4 RESULTS AND DISCUSSION

Our primary goal was to explore the possibility of discerning developmental stages in zebrafish embryogenesis by applying HR-DI-nanoESI-MS analysis on lysed eggs. Therefore, reproducible and high quality MS data should be generated and preferably an automated data processing tool is required.
The implementation of these (automated) tools, viz., acquisition and processing strategy, forms an integral part of the proposed method. After binning, empty and almost empty features were removed. This reduced the number of features from 650.000 to 30.000. To reduce analytical variation as much as possible, the data was normalized by selecting the optimal internal standard for each compound from the mix of internal standards. This selection was based on the minimization of the RSD of the response of replicate measurements [1]. Beckmann et al. [12] normalized the data by using the total ion current, however with our data this would reduce group differences which is not desirable. To enable statistical interpretation, features that showed consistency per group were selected [46]. Features were considered consistent if they were either present or absent in all samples for a group. However, as we were interested in changes between developmental stages, we allowed for, at most, one missing feature in the replicates of a particular developmental stage. This step further reduced the number of possible features to 5.000. As we anticipate differences

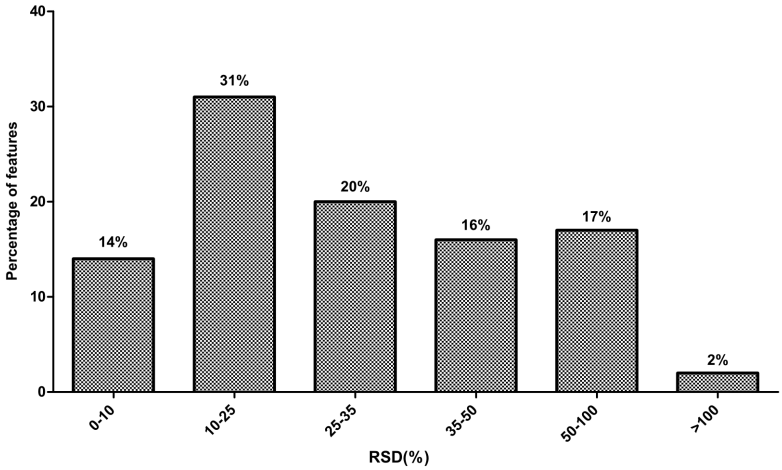


Figure 2: Histograms showing the technical reproducibility of the measurements. 45% of the 102 features has an RSD<25%.

between the samples to be apparent in the most abundant metabolites for that sample, we focused on the top 100 features per sample reducing the number of features of interest to 200. Out of these 200 features, 102 showed significant differences in concentration between the different developmental stages. As observed in other complex samples in other research [82], the lowest mass range (50-250 m/z) showed a clear clustering of the different developmental stages (see figure 3). This would indicate that at least small molecules relate to metabolic variations between the ages of the embryo.

4.4.1 *Robustness/validation of the analytical method*

Metabolomics often deals with differential analysis of fingerprints([82]) to highlight biomarkers. Especially, when a large number of analytes is taken into consideration absolute quantitation is only reported in some cases. In order to be able to give some indication of the analytical rigor of the method, the RSD of repeated measurements, for example using QC samples is often reported. In the absence of QC samples , the RSD was calculated for each feature using a RSD approach based on multiple sets of replicated measurements ([86]) and classified in one of the six possible classes (0-0.1,0.1-0.25,0.25-0.35,0.35-0.5,0.5-1 and >1). The result of this inter-assay reproducibility for the 102 features is displayed in Figure 2. As can be seen almost half of these features showed an RSD lower than 25%.

One could argue that by focusing on the top N features too much information is being discarded. For example, tryptophan, hypoxanthine, carnosine, methionine, aspartic acid, propionylcarnitine, dimethyllysine, methyllysine, acetyllysine and serine were discarded (see supplementary Figure S41). However, including lesser abundant features invariably led to higher RSD values for these new features. This could be explained by ion suppression effects: as we measured crude, complex samples containing various compound classes and compound sizes as a result of lysis of the whole embryo. As a comparison: yeast has an estimated 1100 metabolites which is expected to be significantly

feature	M+H	M+Na	M+K	$M+NH_4$	new
acetylcarnitine	0.33	0.29	n.d.	n.d.	
alanine	n.d.	n.d.	0.31	n.d.	
arginine	0.11	n.d.	n.d.	n.d.	✓
asparagine	n.d.	n.d.	0.07	n.d.	
aspartylphosphate /acetylaspartic acid	n.d.	n.d.	0.16	n.d.	✓
carnitine	0.28	0.27	n.d.	n.d.	✓
creatine	0.26	0.19	0.09	n.d.	
dimethylarginine	0.11	n.d.	n.d.	n.d.	✓
dopamine /vanillyamine	n.d.	n.d.	0.16	n.d.	✓
FAPy-adenine	n.d.	n.d.	0.2	n.d.	✓
glutamine	n.d.	n.d.	0.06	n.d.	
histidine	0.19	n.d.	0.22	n.d.	
homoserine /threonine	n.d.	n.d.	0.34	n.d.	

feature	M+H	M+Na	M+K	$M+NH_4$	new
indoleacetic acid	0.05	n.d.	n.d.	n.d.	✓
inositol cyclic phosphate	0.16	n.d.	n.d.	n.d.	
iso/nor leucine	0.14	0.12	0.09	n.d.	
iso valeraldehyde	n.d.	n.d.	n.d.	0.2	✓
lysine	0.12	n.d.	n.d.	n.d.	
phenylalanine	n.d.	0.13	0.09	n.d.	
phosphoethanolamine	n.d.	n.d.	0.2	n.d.	✓
proline	n.d.	n.d.	0.12	n.d.	
quinone	n.d.	0.37	n.d.	n.d.	✓
safrole	n.d.	0.33	n.d.	n.d.	✓
spermine dialdehyde	0.25	n.d.	n.d.	n.d.	✓
tyrosine	n.d.	n.d.	0.07	n.d.	
valine	n.d.	0.2	0.13	n.d.	
vinylacetylglycine	0.08	n.d.	n.d.	n.d.	✓

Table 1: RSD values of the 27 (putatively) identified features. The features that were not found in the cited references were ticked as new in the last column.

less than the metabolome of a whole zebrafish embryo([27]). This together with using nanoESI lead to quite some sample-loss due to, e.g., clogging of the nanoESI emitter, which resulted in a loss of replicates. Nevertheless, by discarding the missing variables from the dataset, still 102 features were extracted which revealed the developmental stage differences (Figure 3). (FDA suggest that for LC-MS profiling an RSD of 20% is acceptable [27]).

4.4.2 *Identification of metabolites*

Out of the significant 102 features 27 separate metabolites (36 features; including all adducts) were putatively identified (see table 1). Identification was based on 1) search of HR-MS data against HMDB (Human Metabolome project Data Base) and 2) searching MS/MS data against HMDB comparison to standards. After the first HMDB search 171 possible structures were found of which 21 were unique and 150 isomers. From those 150 remaining isomers, 6 were identified using tandem MS and subsequent database search or comparison to standards.

Table 1 shows all the putatively identified features, with their associated RSD values. Remarkably, the potassium adducts nearly always shows the best RSD, followed by the sodium adducts and finally the protonated molecule. As our samples are not acidified, alkali adducts are not suppressed. The reason for the pre-dominant potassium adducts is most likely due to relative high potassium concentration inside the cell as opposed to the relative high concentration of sodium outside living cells

From several of the 27 (putatively) identified features the boxplots with their up-or down regulation are shown in Figure4. Still 63% of the 102 significant features remained unidentified. This confirms that further research is required to expand the zebrafish metabolome database in order to increase identification using DI-MS methods([96]).

4.4.3 *Biological relevance*

In order to evaluate the possibility of using this technology for early fingerprinting of zebrafish embryos, different stages after fertilization were analyzed using this approach. 1,2,3,5 and 18 hours post fertilization (hpf) were tested.

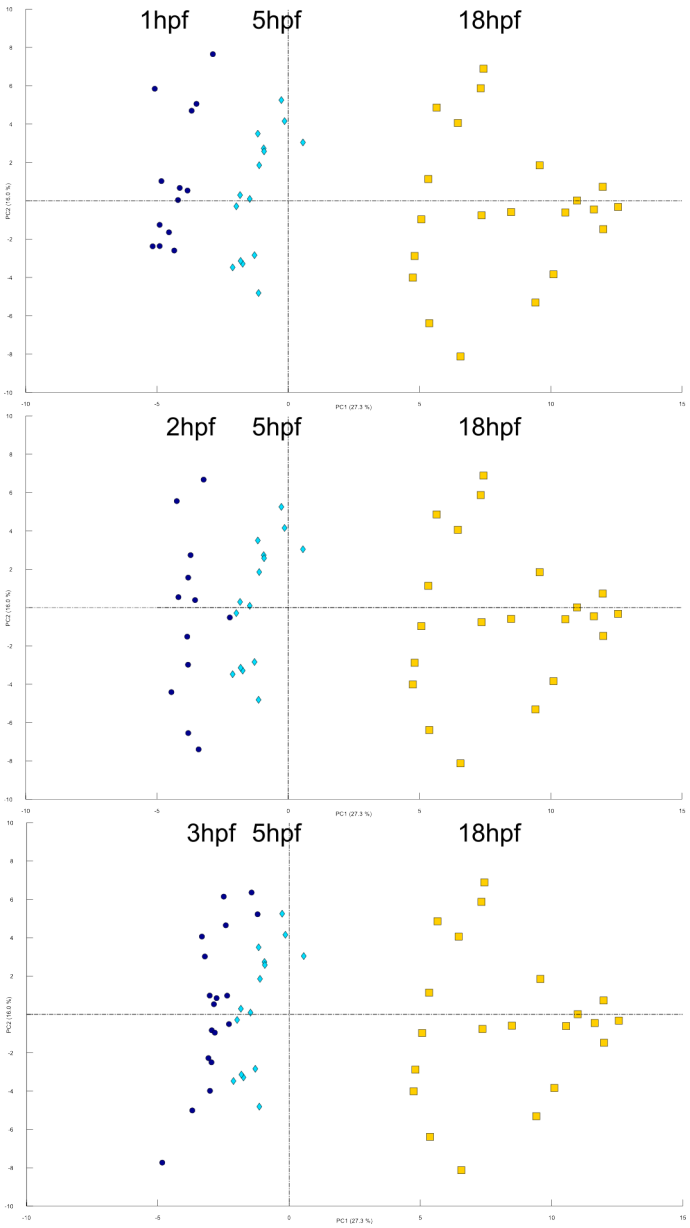


Figure 3: 3 PCA score plots indicating a visible difference between 5 different developmental stages. For these plots only one PCA model was created using all samples and only those features with an RSD <30%. The blocks represent the 18hpf stage, the diamonds the 5hpf stage. The circles represent the 1, 2 and 3 hpf respectively

The PCA in figure 3 shows the somite stage (18hpf) as the most distinctive group as well as having the most within-group variation. The fact that this stage is most distinctive from the other groups can be explained by the fact that different tissue types and organs (i.e. nervous system, skin, blood, and heart) begin to form at this stage. Since small time differences result already in differences of organ formation this will lead to biological variations in the different samples, explaining why this time point has the most variation in the metabolome measurements. Thus, as Chen et al. [96] remarked, the differences in the metabolome between the early embryonic stages may reflect the embryological properties of the cells. These results show that the zebrafish embryonic metabolome reflects differentiation. With our method we were able to observe a post fertilization time-trend with a time resolution of 1hpf. The metabolic shift between 1, 2, 3 and 5hpf can be clearly observed indicating that in early embryogenesis the metabolome changes quickly and significantly. As can be seen, the metabolome of the first three stages (1, 2 and 3hpf) closely resemble each other. This suggests that zygotic gene transcription only begins at the onset of the midblastula transition (3hpf)[60].

The boxplots in Figure 4 allows us to go more into depth regarding the biology during early embryogenesis. Interestingly, the increase in concentration of dimethylarginine can be a result of the enhanced methylation and acetylation metabolism of lysine and arginine which play an important role in histone activity and gene expression [113, 136] When we searched for other methylated and acetylated forms of lysine and arginine by discarding the described missings data cleaning step (resulting in a larger data subset) we also found increasing trends for dimethyllysine, methyllysine and acetyllysine (see supplemental Figure S41). Although we are aware these analytes did not make it to the best 102 features, these findings correlate to the onset of gene expression at 3 hpf. The increased concentration of spermine dialdehyde could reflect the turnover of the polyamine spermine which plays a role in normal and neoplastic growth as well as the (uni)directional transport of molecules by GAP junctions which is an important process in early embryogenesis [99, 137].

It can also be observed that the concentrations of several amino acids and biogenic amines are increasing during early embryogenesis. Apparently the embryos are able to release amino acids from storage proteins to provide the cells with building blocks and energy already in an early stage. Some of these reveal more than a 10-fold change going from 1hpf to 18hpf. In the 18hpf stage the rise of most of the amino acids is the largest. This could be explained by the increase in energy requirement for evoked muscle contraction starting at around 18hpf [108]. Isovaleraldehyde, vinylacetylglycine could reflect the degradation of branched chain amino acids like valine and (iso)leucine, both associated with the increase in energy requirement. The concentration of hypoxanthine (see supplemental Figure S41) is decreasing which can be explained by the enhanced DNA/RNA synthesis via xanthine, as also indicated by the decrease of guanine.

Some metabolites, like acetylcarnitine and creatine, show up- as well as down regulation within the five developmental stages. The observed trends of almost all of our identified features is supported by the observations of previous publications [42, 116]. Moreover, features that are indicated as new in table 1(arginine, acetylasparctic acid, carnitine, dimethylarginine, dopamine, FAPy-

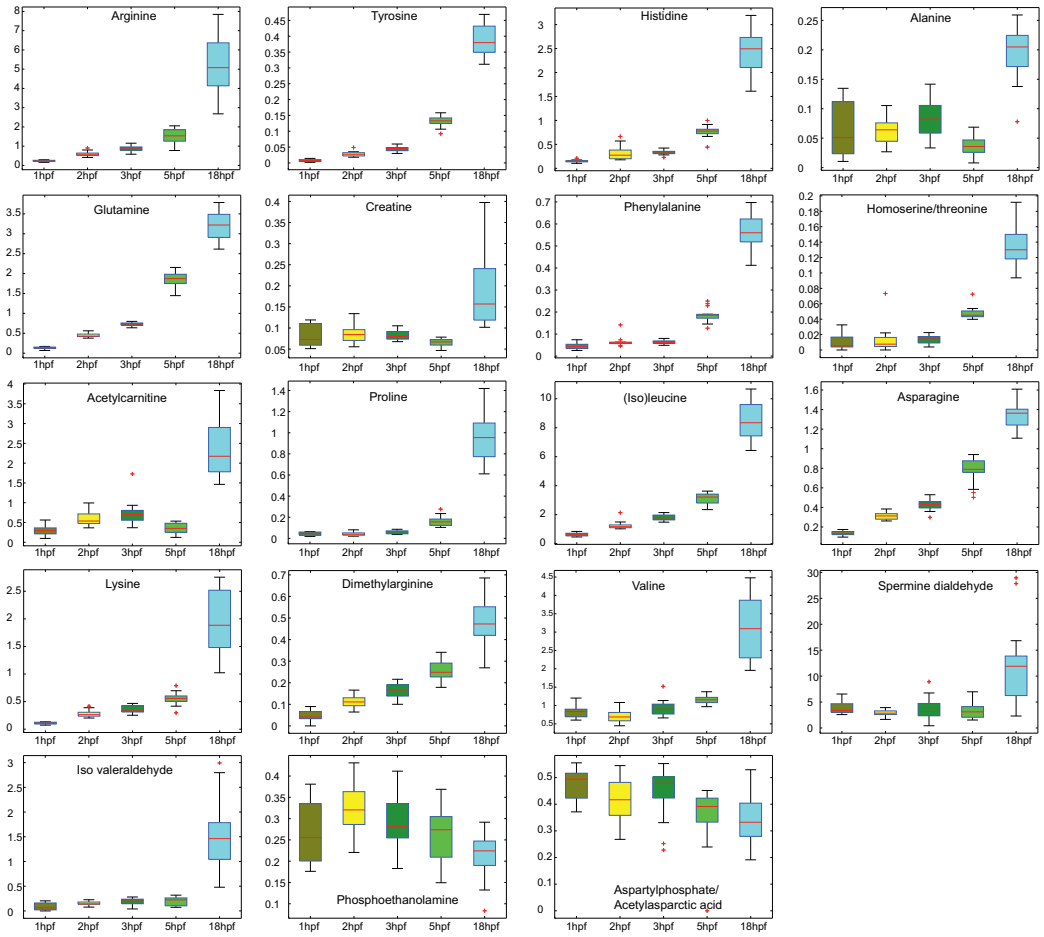


Figure 4: Boxplots of the up-and down regulation of some of the (putatively) identified metabolites through the five different developmental stages (for arg, his, actylcarnitine, lys, dimethylarg and spermine dialdehyde the $(M+H)^+$ plot is shown, for tyr, glu, creatine,phe, pro, leu, val, ala, homoserine, asparagine, phophoethanolamine, acetylasparctic acid the $(M+K)^+$ plot is shown, and for isovaleraldehyde the $(M+NH_4)^+$ plot is shown. The y-axis represents the ratio of the analyte/ optimal internal standard

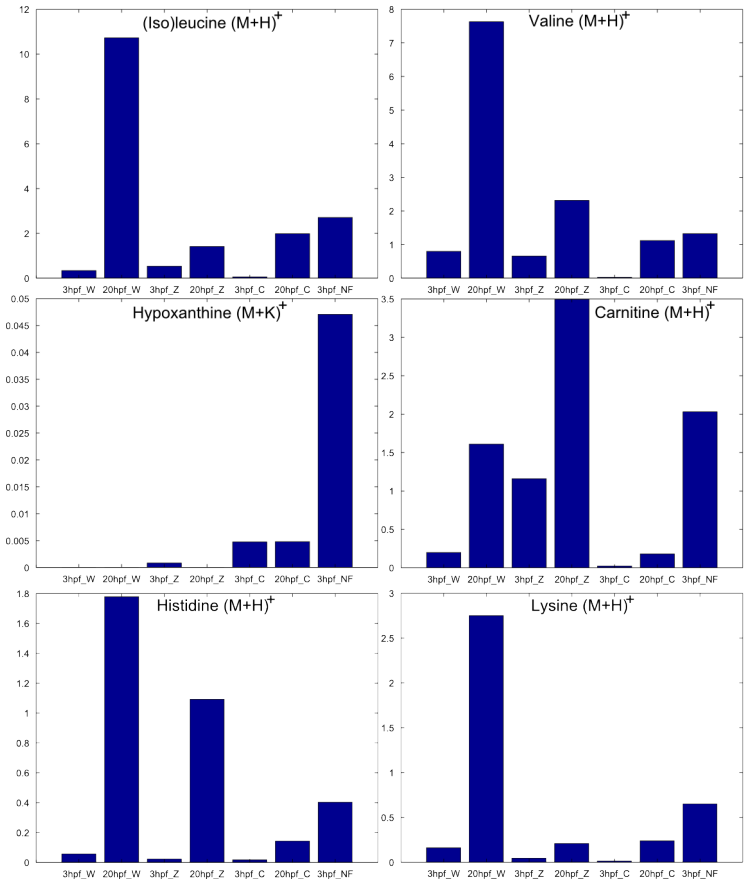


Figure 5: Bar-plots showing the mean of the different embryonic location at the age of 3hpf and 20hpf. W=whole embryo; C=chorion membrane; Z=zygote part; NF=dechorionated non-fertilized embryo (=yolk). The y-axis represents the ratio of the analyte/ optimal internal standard.

adenine, indoleacetic acid, isovaleraldehyde, phosphoethanolamine, quinone, safrole, spermine dialdehyde, vinylacetylglycine) were exclusively found with our method and not in the aformentioned references. This indicates the potential of HR-DI-MS for metabolic profiling purposes.

4.4.4 *Deyolking*

To obtain more insight in the biology of early embryogenesis, a series of deyolking experiments were performed in order to zoom in on the localization of metabolism.

Figure 5 shows several bar-plots of the group means of metabolites that were also (putatively) identified in the previous section. Hypoxanthine and carnitine (metabolites which were discarded using our data cleanup steps (see supplemental Figure **??**)) showed a down-regulation trend during embryogenesis). The charts confirm the same developmental trends as observed earlier. Because of the limited number of replicates that was measured (3 times) the statistical power is limited but ANOVA calculations showed that no significant difference between the zygote part and the whole egg could be detected. This

could indicate that most of the metabolites and most of the metabolic conversions take place in the zygote part.

4.5 CONCLUDING REMARKS

Using our rapid metabolic fingerprinting method we were able to distinguish metabolic profiles of early developmental stages of zebrafish embryos. Interestingly, the onset of gene expression and the increase in energy requirement is reflected by the measured metabolome confirming that from a systems biology point of view metabolomics complements transcriptomics. After data cleanup, only those features were selected that showed consistent behavior within each developmental stage resulting in 102 features. PCA revealed that periods of 1 hour time shifts post fertilization could be differentiated from each other. In total 27 out of the 102 features were (putatively) identified. Although unambiguous identification is beyond the scope of our approach, identification on 6 of these 27 extracted features was pursued using standards and tandem MS. Several trends of the putatively identified metabolites are included and almost all of these findings are supported by previous publications. Moreover our method exclusively found several new features. By deyolking and dechorionating we showed the potential of this method to enable more in-depth studies on localization of metabolism. We conclude that HR-DI-MS is suitable for rapid metabolic profiling on zebrafish embryos. However, to improve robustness and obtain more high-quality features fast sample preparation methods are required.

ACKNOWLEDGEMENTS

This work is (co)financed by the Netherlands Metabolomics Centre (NMC) which is a part of The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research and partly by the European community projects ZF-Health (FP7-Health-2009-242048).

S4

SUPPLEMENT TO RAPID METABOLIC SCREENING OF EARLY ZEBRAFISH EMBRYOGENESIS BASED ON DIRECT INFUSION-NANOESI-FTMS

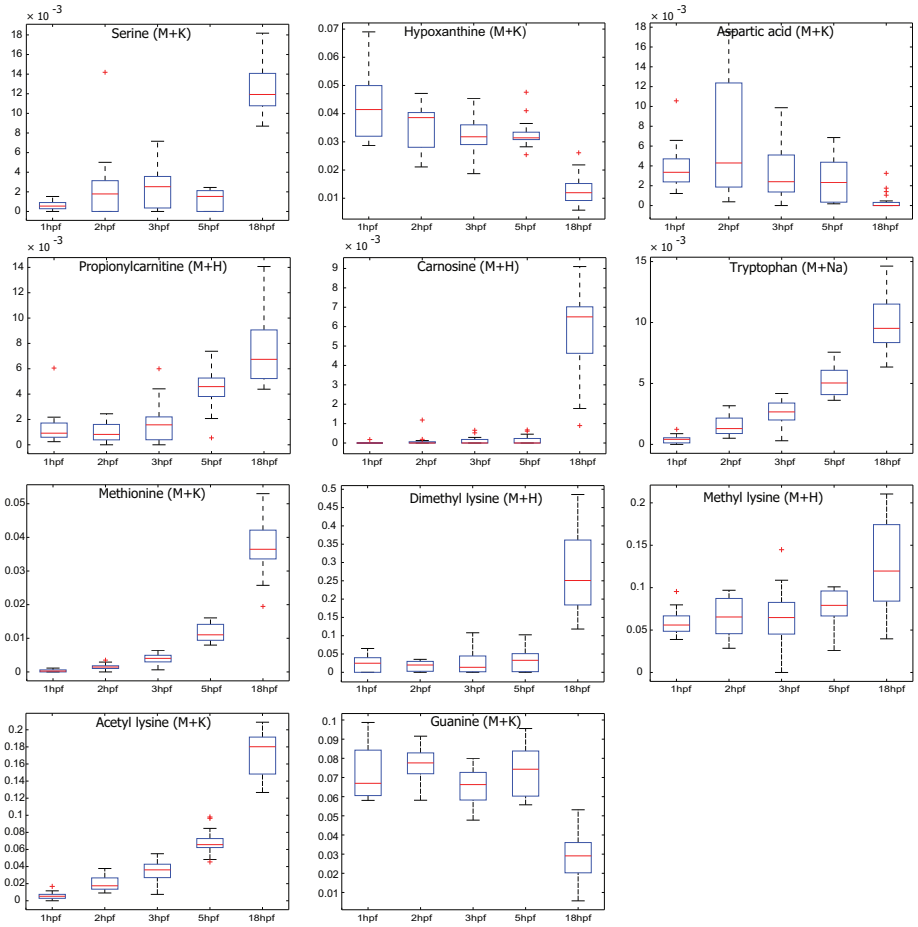


Figure S41: Boxplots of some of the discarded features due to our data cleanup steps..

5

A NEW APPROACH TO UNTARGETED INTEGRATION OF HIGH RESOLUTION LC-MS DATA

ABSTRACT

Because of its high sensitivity and specificity, hyphenated mass spectrometry has become the predominant method to detect and quantify metabolites present in bio-samples relevant for all sorts of life science studies being executed. In contrast to targeted methods that are dedicated to specific features, global profiling acquisition methods allow new unspecific metabolites to be analyzed. The challenge with these so-called untargeted methods is the proper and automated extraction and integration of features that could be of relevance. We propose a new algorithm that enables untargeted integration of samples that are measured with high resolution liquid chromatography mass spectrometry (LC-MS). In contrast to other approaches limited user interaction is needed allowing also less experienced users to integrate their data. The large amount of single features that are found within a sample is combined to a smaller list of, compound-related, grouped feature-sets representative for that sample. These feature-sets allow for easier interpretation and identification and as important, easier matching over samples. We show that the automatic obtained integration results for a set of known target metabolites match those generated with vendor software but that at least 10 times more feature-sets are extracted as well. We demonstrate our approach using high resolution LC-MS data acquired for 128 samples on a lipidomics platform. The data was also processed in a targeted manner (with a combination of automatic and manual integration) using vendor software for a set of 174 targets. As our untargeted extraction procedure is run per sample and per mass trace the implementation of it is scalable. Because of the generic approach, we envision that this data extraction lipids method will be used in a targeted as well as untargeted analysis of many different kinds of TOF-MS data, even CE- and GC-MS data or MRM. The Matlab package is available for download on request and efforts are directed towards a user-friendly Windows executable.

Frans M van der Kloet, Margriet Hendriks, Thomas Hankemeier, Theo Reijmers, A new approach to untargeted integration of high resolution LC-MS data, Anal Chim Acta. 2013 Nov 1;801:34-42. doi: 10.1016/j.aca.2013.09.028. Epub 2013 Sep 23.

5.1 INTRODUCTION

The systems biology framework aims at describing the behavior of biological systems (e.g. organisms, organs, cells) as a whole rather than the behavior of their (functional) biochemical components in isolation. During the last decade functional analysis of the transcriptome, proteome, and metabolome has increased [16, 31]. Because the metabolome is expected and found to be more sensitive to environmental (diet, drug, lifestyle) perturbations than the transcriptome or proteome, the emphasis on the phenotype at a more global systems biological level has shifted the focus towards the metabolome [16, 69, 65, 64, 36, 55, 120]. With this increasing awareness of the importance of the metabolome, the number of methods to detect and quantify metabolites is increasing. Hyphenated mass-spectrometry (GC, CE or LC-MS) has become the predominant technology for determining metabolite abundances, mainly because of its sensitivity allowing the measurement of low abundant metabolites in small sample volumes. Targeted modes of data acquisition (MRM/SRM) allow the MS to detect pre-selected compounds with an even higher sensitivity but at the same time have a limit (determined by the maximum MS/MS scan experiments possible) on the list of target compounds reported. The full scan data acquisition mode however, enables a wider, untargeted coverage of different metabolites.

Despite the limited number of compounds reported, targeted approaches are wide spread. Obvious reasons are the added advantage of data interpretation of known metabolites/compounds and the possibility to quantify them (using internal standards) often with better precision and accuracy then in untargeted modes. To a large extent the lesser use of untargeted approaches is also due to the lack of appropriate software that would enable untargeted extraction and integration without introducing artifacts and errors. As a result, integration is often limited to a set of known metabolites (targets) only and in most cases vendor software (MassLynx[138], Compass DataAnalysis[17], MassHunter[1] etc.) is used.

The lack of software that enables untargeted integration has been recognized by various academic groups and different algorithms and solutions have been suggested. For GC-MS measurements Metabolite Detector[45] or TNO-Deco[57] and Metalign[80] could be used and for high-resolution LC-MS software like XC-MS[115], Metalign[80] or MZmine[62] are available. However, these solutions do require specific user input, sometimes even sample specific, and often much user experience is needed before the data is properly extracted and integrated. All LC-MS untargeted solutions result in a huge list of features sometimes with additional putative identification (e.g. XC-MS). Several packages extract and/or report features based on differential analysis between sample groups (e.g. diseased vs. healthy). This not only limits the scalability but renders the method useless if no such grouping factor exists.

In this paper we describe a method for untargeted feature extraction and in addition we propose a new strategy that addresses the aforementioned shortcomings. The method is able to integrate untargeted data, can be incorporated in an automated environment and with only a few parameters to configure, the user interaction is kept to a minimum. Our proposed strategy is a two-step approach that in fact automates common analytical practice. The first

step after feature extraction is based on per sample grouping of single features to feature-sets according to their isotopic patterns and retention times. Here, we introduce the term feature-set as a group of two or more features in a single sample with isotopically related masses that share the same retention time. The second step of our strategy consists of matching these feature-sets over samples. This way more constraints are imposed on the search space to increase the probability for a proper match over samples. Conversely, noisy signals have a lower chance of being propagated.

We demonstrate our method using data obtained with full scan global lipidomics profiling acquired with high-resolution LC-MS (Quadrupole Time-Of-Flight (qTOF)). Lipid profiles are especially challenging for untargeted processing due to the presence of a large number of isomers. We compare our untargeted integration results for a set of known compounds to those that were obtained by vendor software (the reference set) and those obtained using XC-MS[115]. The developed Matlab package is available for download on request and efforts are directed towards a user-friendly Windows executable.

5.2 WORKFLOW (AND METHODS)

Comparable to any software package that analyzes hyphenated MS data, the basic workflow comprises reading data, detecting, extracting and integrating peaks and relating them over samples. For a list of known targets, i.e. compounds with known masses and retention times this seems a straightforward task. Integration however, is complicated by issues like retention time shifts, bad chromatographic separation of isomers, bad peak shapes, noise and small shifts in registered masses. In vendor software, to deal with these issues, the target specific mass- and/or retention time windows are manually adjusted and other, vendor specific integration settings are optimized. Even if the integration results in untargeted methods are invariant to some of these issues, mass and retention time windows still need to be defined before matching over samples can be performed. How large these shifts will be, depends on the method that was used to acquire the data; e.g. flow rate, mobile- and stationary phase changes etc.

To summarize all the steps we have taken schematically, the full workflow of our feature extraction procedure for a single sample is shown in Figure 1.

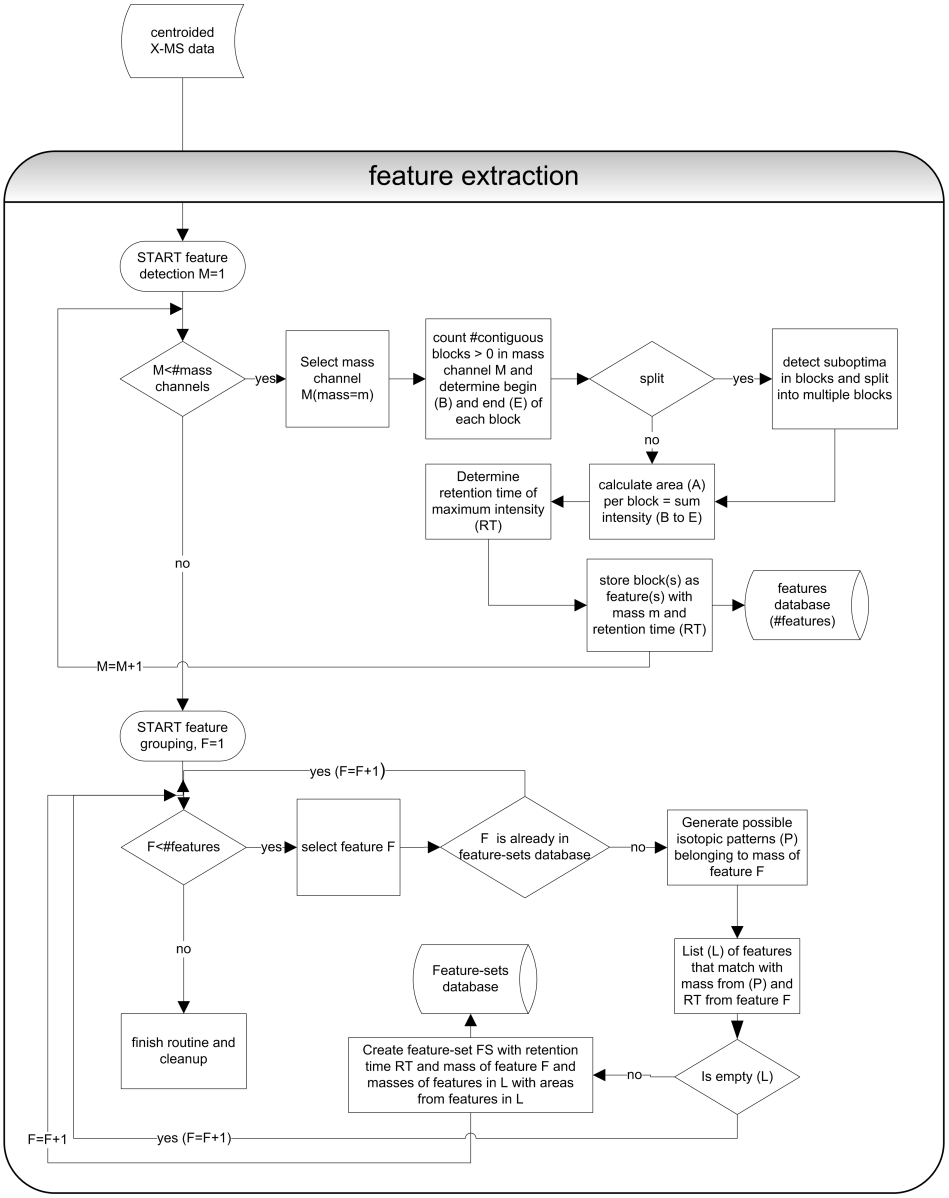


Figure 1: The workflow of per sample feature extraction and grouping; In the first stage each mass trace is scanned for the presence of one or more features. In the second part the features are combined according to their mass and retention time

5.2.1 *Detection and extraction*

The starting point and the heart of our feature extraction method is centroided hyphenated MS data. The vast amount of data stored in samples acquired in full profile at high resolution puts a huge strain on computers. To keep file sizes significantly smaller and retain as much information as possible the mass spectra are often acquired in centroid mode. In this mode the average m/z value is determined for every mass peak and only the intensity at the average m/z value is stored. Figure 2a shows the two dimensional (mass and retention time) response area of a peak that has been reconstructed from full profile data. Figure 2b shows the same peak but now after centroiding.

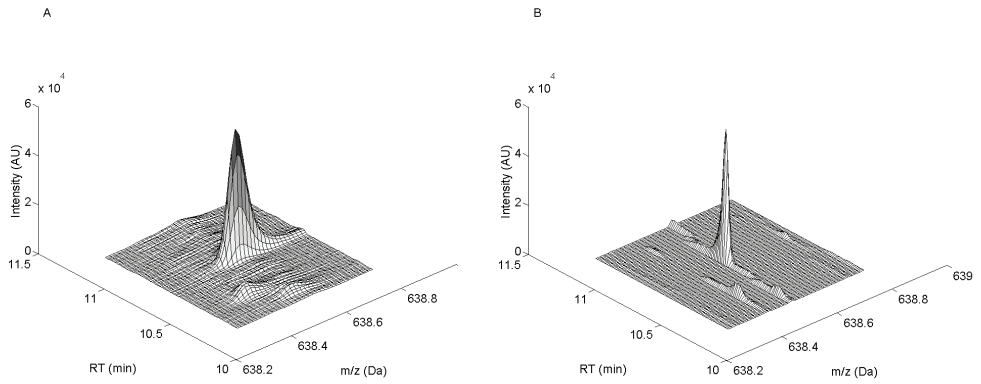


Figure 2: (A) reconstructed peak from profile data and (B) reconstruction of the same peak as (A) but now after centroiding.

One of the disadvantages of (high resolution) data acquired in centroided mode is the complexity to create chromatographic profiles of mass peaks because of mass centroids that fluctuate from scan to scan. Binning algorithms have been developed to compensate for these fluctuations [115]. To circumvent complex mass binning routines but still take advantage of the smaller data (file) sizes we re-calibrated full profile data (acquired on a TOF-MS) before applying our developed mass-centroiding step. Details can be found in the Supporting information.

The number of features per mass trace is determined by counting the number of contiguous signal containing blocks in that trace; a contiguous signal containing block was defined as showing intensity higher than the noise level in 5 or more subsequent scans. Different vendors (Waters, Agilent, Absciex and Thermo) claim that at least 8 - 15 points across a peak should be used for quantification; we therefore considered 5 to be on the safe side considering the start and end of the peak could be below the noise level. The noise level was determined per scan as **N** times the mode (most occurring intensity) of all registered intensities for that scan. This **N** is configured by the *noise-threshold-factor* parameter which is set to 3 by default. Effectively this means that if baseline separation is observed as illustrated in Figure 3a, the two peaks of the single mass trace are split into two separate features (Figure 3b and 3c). If the two peaks are not separated, only one feature is extracted by default. For the centroided data-set that contains only mass traces that contain a signal, this means that for every mass channel at least one feature is extracted. In order to split peaks that are not baseline separated, e.g. badly separated isomers, the feature extraction routine is extended by an optional to use split routine. In case of bi- or multimodality this split routine splits the feature at local minima (i.e. points where the first derivative is zero) into one or more features (Supporting Information Figure S52). The sensitivity of this parameter is defined by the so-called *split-ratio* variable. The value for this *split-ratio* is bound between 0 and 1 and the optimal value may vary per study. A certain *split-ratio* value picks up on dips between two peaks in the EIC that have an intensity at least split-ratio times the maximum intensity of the feature and splits it into multiple separate features. Higher values make this routine less sensitive to these dips. Integrated areas are obtained by accumulating the intensity values

over the retention time window for every feature. After the retention time (RT) of the maximum intensity for all features has been determined, the data is summarized as a list of features and their integrated areas, i.e. mz@RT:area.

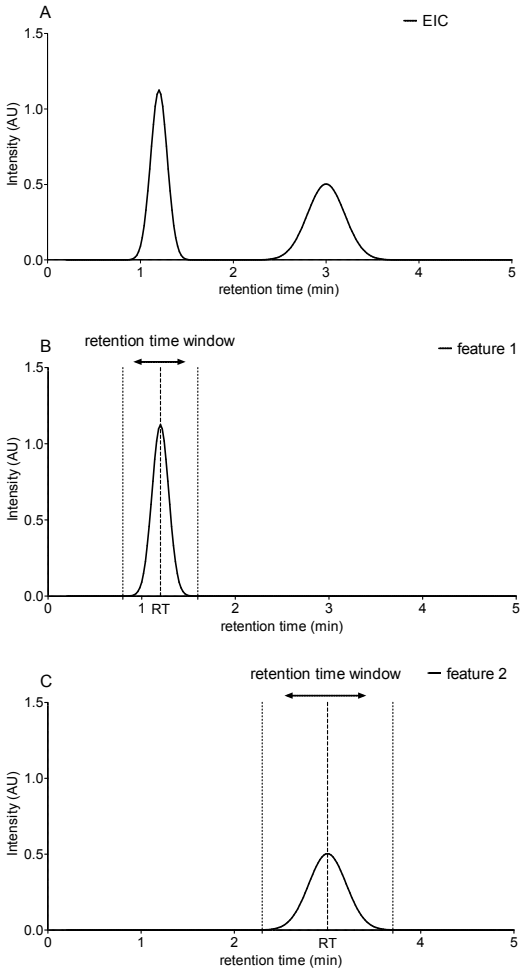


Figure 3: The transformation of the data from mass traces to features: two unimodal peaks from A form two features (B and C).

5.2.2 *Grouping of features to feature-sets*

To increase the probability of correctly matching features over samples, the features in each sample are first combined to feature-sets based on a rule set. The most common rules are based on isotopic patterns and possible adduct information. For extracting known targets one could include reference mass spectra. In this paper we group features by ^{13}C isotopic pattern matching only [125] (M+1,M+2,M+3 etc.); the isotopic masses of a chemical compound share a similar chromatographic pattern and have maximum intensity at the same (plus or minus one single scan) retention time. In contrast, features that show an irregular chromatographic profile (like noise) likely do not have an isotopic mass with the same chromatographic profile (i.e. maxima at the same position) and are therefore not grouped. The choice not to include adduct information at this stage is by design. Adducts do not necessarily share the

same chromatographic profile in the same sample as they are concentration dependent and their concentration is influenced by matrix effects (i.e. different samples), e.g. if more salts are in the sample/solvent adducts are more or less abundant. For data analysis purposes the adduct information should be analyzed separately. The lower block of Figure 1 depicts the flowchart of this grouping procedure. For every single feature the theoretical isotopic mass values are generated. By repeated comparison, features match only if the mass corresponds to one of the generated isotopic mass values. If a match is found they are grouped as a feature-set. The mass range is determined by the mass-resolution parameter which is determined by the resolution setting of the MS the data was acquired on.

5.2.3 *Comparing over samples*

After all samples have been processed in the manner as described above, comparison over multiple samples is done based on a match of the feature-sets only. When no feature-set match was found in certain cases, the search was continued in the list of single features. By selecting only feature-sets, low abundant and noisy signals are removed. Every feature-set is matched against the remaining feature-sets using repeated pairwise comparison. The matching algorithm is configured by two parameters; *retention-time-window* and *mass-resolution*. Feature-sets only match if the mass values of the feature-sets and the retention time are within range. If more than one match is found for the same sample (e.g. isomers), the match with a retention time closest to the reference retention time is selected. The script below describes the process of matching over samples in pseudo code

```
LOOP OVER SAMPLES -> S
        LOOP OVER FEATURE-SETS in SAMPLE S -> FS
                SELECT MAIN MASS in FS -> M
                SELECT RETENTION TIME in FS -> RT
                LOOP OVER OTHER SAMPLES -> OS
                        FIND feature-sets IN OS WHERE MAIN MASS is M +/- mass resolution
                           AND RETENTION TIME = RT+/- retention time window -> ID
                        IF SIZE(ID)>1
                                FIND feature-sets closest to RT -> ID
                        END
                        IF NOT FOUND
                                FIND single-features IN OS WHERE MAIN is M +/- mass resolution
                                   AND RETENTION TIME = RT+/- retention time window -> ID
                                IF SIZE(ID)>1
                                        FIND feature-sets closest to RT -> ID
                                END
                        END
                        HOUSEKEEPING: LINK FS of sample S to ID of sample OS
                END
        END
END
```

Depending on the study and type of analysis, i.e. biomarker analysis, targeted featured extraction etc. feature-sets for which no match in any of the other samples was found can optionally be removed.

The proposed approach requires little prior knowledge about the samples. This guarantees the generic applicability of the approach. However some user input cannot be avoided. Low abundant peaks can be in- or excluded depending on the *noise-threshold-factor*. Samples, where components (e.g. isomers) are not baseline separated, can be processed by an optional step which requires the setting of the *split-ratio* parameter. Matching features within and between samples not only requires a *retention-time-window* but also a *mass-resolution* to be defined. The *mass-resolution* parameter in both the feature grouping as in the

feature-set matching step is the same. Changing the values of these parameters will influence the feature grouping and feature-set matching over samples but in fact, they solely depend on the manner the data are acquired and the mass spectrometer that was used and therefore have to be defined only once.

In contrast to an absolute mass window, we prefer to use the term resolution. For a time-of-flight (TOF) detector the resolution is constant over the mass-range[145]. With a resolution of 10,000 at 200 Da, peaks that are 0.02 Da (=200/10,000) apart can be resolved, however, at 1,000 Da this changes to 0.1 Da. With only 4 parameters to configure (that are at most study specific) this untargeted processing approach requires a very low level of user input. The settings for most of these parameters follow from practical/experimental settings and need no tweaking. The *mass-resolution* is defined by the mass resolution set when acquiring mass spectra by the MS system and is as such known before the data needs to be integrated. In our case the MS system was a qTOF which was set to a resolving power of 10,000. The *noise-threshold-factor* was kept to the default value, meaning that signals having a value above 3 times the noise value are considered to be peak candidates [89]. The noise level value itself is automatically determined. The *retention-time-window* was set to +/- 3 seconds which could be deduced from the chromatographic performance of the method we used. The optimal value for *split-ratio* follows from the expected co-elution of isomeric compounds and whether or not one wishes to separate (split) them or combine them.

5.3 EXPERIMENTAL

The software was written in Matlab 2011a (64 bit) using the bioinformatics-, image processing- and statistical toolboxes. All calculations were done on a DELL workstation equipped with a 4 core Intel© Xeon© CPU X5482 @3.2 GHz processor and 16GB of memory running Windows 7 Professional 64 bit.

To demonstrate the proposed method, its functionality was tested using data of a clinical study obtained from a global lipid platform measured in positive mode[53]. The spectra obtained from this method are amongst the most complex to analyze and extract since the intensity range of the different co-eluting compounds per scan can differ in orders of magnitude (3-5 times) together with the presence of a large number of isomers. Figure 4 illustrates a typical example of a mass spectrum obtained from this LC-MS platform; it is clear that many compounds are co-eluting (Figure 4a) but also that these co-eluting compounds (with different elemental composition) have isotopic and ionic masses that are almost similar and are therefore difficult to integrate accurately and reproducibly (Figure 4b at 707.55 Da).

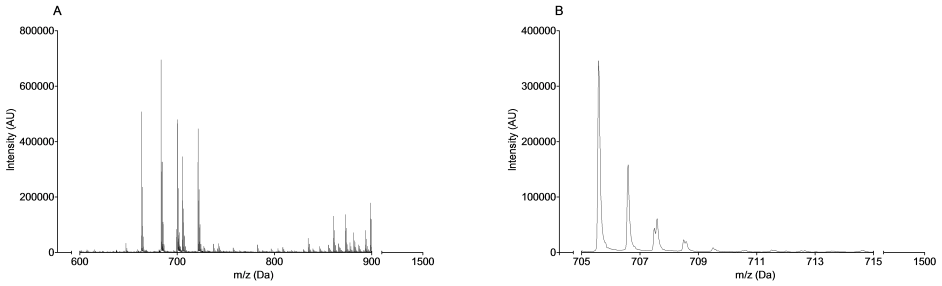


Figure 4: A typical mass spectrum obtained from LC-MS lipidomics platform. A: the full scan range shows the huge dynamic range. B: an enlargement showing interference of (isotopic) masses of co-eluting compounds at 707.55 Da(A) reconstructed peak from profile data and (B) reconstruction of the same peak (as a) but now after centroiding

In order to demonstrate the effectiveness of our integration method we compared our results in two different ways. To demonstrate the quantitative capabilities for a set of known compounds we compared our results against the reference optimized targeted and manually controlled quantification using proprietary vendor software. Secondly we compared our untargeted results against the commonly used XC-MS method for these aforementioned compounds.

The 128 samples (16 QC, 112 study samples) were processed using proprietary vendor software (Agilent, MassHunter QuantAnalysis 4.0) for a list of known targets (174, including 8 internal standards and 8 calibrants) which made them a perfect test-case (the reference set). To compare our integration results to the reference set we included all the 128 samples.

The study samples originated from subjects that underwent bariatric surgery; before and after treatment. For only a few number of patients all sample points were contained in this analytical batch. Earlier multivariate (unpublished) analysis on this sub-set of samples indicated a clear treatment effect. To demonstrate the added value of the unknown (untargeted) data we limited ourselves to this group of samples only (24 in total) for further statistical analysis.

The data were acquired on a LC-MS from Agilent (profile qTOF 6530) that was set to a mass resolving power[145] of 10,000. All the mass spectra were recorded in full scan mode and centroided mode simultaneously. Quantitation using Agilent QuantAnalysis[1] is limited to centroided data only. For the automatically generated feature extraction we used uncompressed full scan profile data. The uncompressed full scan profile LC-MS data were converted from the proprietary .d MassHunter format from Agilent to mzXML format using trapper[74] software (version 4.3.1 (build Sep 9 2009 12:29:13)) before importing in Matlab to run our integration software.

For comparison purposes the same mzXML files were also processed by XC-MS. The configuration parameters of XC-MS were optimized especially for processing high resolution LC-MS spectra from lipidomics samples. Even though these computations were performed on a machine with more internal memory (Dell Poweredge 1950, 2x quad core CPU, 40 GB memory), not all files could be processed. In order to still be able to make a fair comparison we selected 70 samples (16 QCs and 54 study samples) to be processed by XC-

MS. We assured that all metabolites from the reference dataset were present in these samples and compared the XC-MS integration results to the reference set results for these 70 samples.

5.4 RESULTS AND DISCUSSION

5.4.1 *Feature extraction, grouping and comparison over samples*

The mass spectra were acquired at a mass resolving power of 10,000, consequently the *mass-resolution* parameter was set to 10,000. Because we wanted to compare the results to those obtained by vendor software using highly optimized integration parameters for some compounds, the optional *split-ratio* was set to a very sensitive 0.01. This meant that if features contained two or more peaks with intensities as small as 1% of the highest peak in that feature it was split into multiple features. For every sample the automatic feature extraction method yielded approximately 100,000 single features. After the feature grouping step, approximately 13,000 combinations of 2 or more features were generated. This way, at least more than 26,000 (2x13,000) features per sample remained for comparison over samples. We allowed a *retention-time-window* of 3 seconds to match feature-sets over samples what resulted in a list of 210 thousand feature-sets over all 128 samples. The final list of feature-sets was approximately 16 times larger (210,000/13,000) than the list created per individual sample and indicated that the different samples contained sample specific/unique feature-sets. When masses and/or retention times of the same compound between samples deviate more than what is expected, it is likely that these 210,000 feature-sets include mismatches. We assume that by the right choice of the mass and retention time window the number of mismatches is limited. The use of quality control samples enables the selection of robust feature-sets (i.e. smaller than 0.2 RSD) and as such removes mismatches. Most of the feature-sets are detected in 50%-55% of all samples. 1005 feature-sets were detected in all samples. The distribution of the presence of each feature-set in all 128 samples (study + QC) is shown in Figure S53 of the Supporting information.

5.4.2 *Targeted results vs. untargeted results*

To demonstrate the capability of our extraction method to successfully extract and integrate we focused on a list of known targets and calibrants that were processed using vendor software (MassHunter Quant Analysis 4.0). By means of comparing the characteristic mass values of the feature-sets and the retention times to those of the known targets, 171 of the 174 targets from our manually optimized integration method were found in the feature-sets. The 3 missing targets could be traced back in the list of single features. Investigation confirmed that no isotopic patterns for these targets could be found. For comparison over different samples purposes the windows for mass range and retention time range were set to 10,000 and 10 seconds respectively. The allowed retention time window of 10 seconds ensured that we would definitely find the corresponding feature-sets. In Figure 5 we plotted a typical result of

peak areas that were integrated using the vendor software against those found by our untargeted method (the area of the characteristic feature only); the smaller the spread around the diagonal, the higher the similarity. Some more examples are included in the Supporting Information (Figure S54). Overall very good Pearsons correlations (>0.8) between our method and the reference method were obtained.

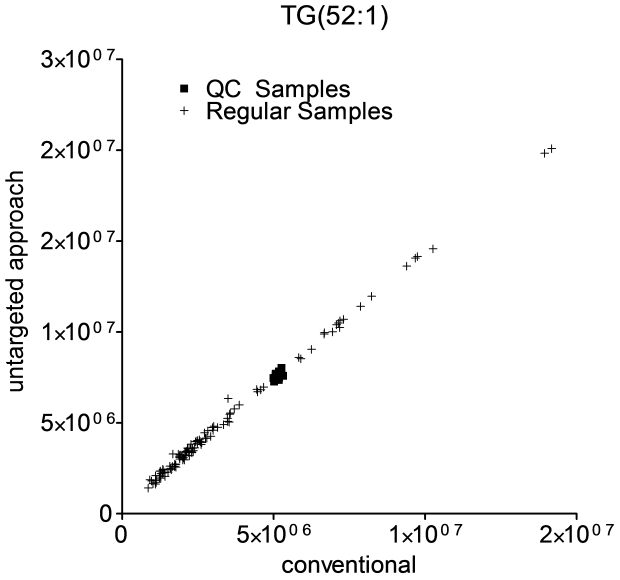


Figure 5: Comparison between untargeted integrated areas (y-axis) and areas obtained with vendor software for a known target (x-axis).

Figure 6a shows the distribution of correlations between our method and the conventional method for all targeted lipids; it was clear that most of the targets show high correlation. To reduce the obvious influence of higher intensity values, the data were first log10 transformed before calculating the correlation (Pearson) between the targeted and untargeted approach. For some feature-sets a rather low correlation was found (e.g.<0.7); further investigation revealed that in these cases the peaks were actually split into multiple features indicating bi- or multi-modality (e.g. isomers) which was not the case in the reference method where the isomers were integrated as a total. Arguably, higher split ratios would result in more comparable results to the conventional vendor software but actually our aim is to detect compounds as good as possible as opposed to what would be achievable with the vendor software. For one feature-set there was almost no correlation, however it was confirmed that this target was not detected (at all) for many samples in the reference set. Internal standards (8) and calibrants (8) were not included in this summary as they showed almost no variation over samples rendering correlation values inappropriate to compare both integration approaches. Of the remaining known targets (174-8-8-2=156) 73% (114) had a correlation higher than 0.9 and 87% (138) had a correlation higher than 0.8. Figure 6b shows the relation between the correlations found and the average peak intensity for the known targets; no clear relation between intensity and observed correlation could be observed. Low correlations were due to isomeric compounds as mentioned earlier.

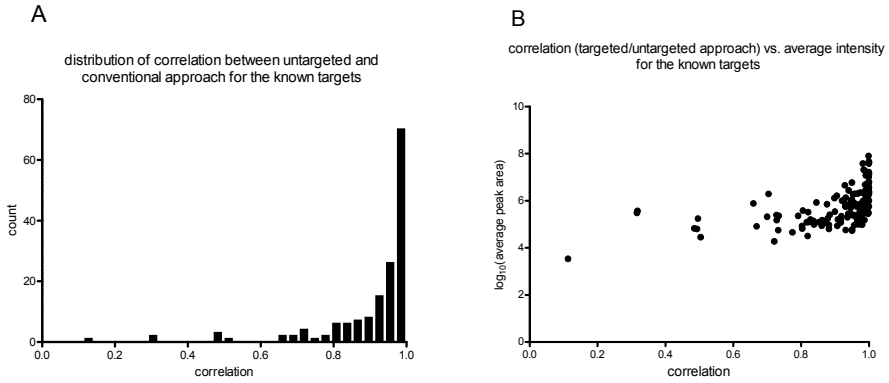


Figure 6: A: The distribution of the correlations between the untargeted and targeted results. The low correlations are the result of the sensitive split function. In 73% the correlation was higher than 0.9. B: The same correlations vs. the average intensity of the known peaks. There is no clear relation between lower intensity and lower correlation.

To compare our untargeted approach to an alternative commonly used untargeted approach, 70 samples were analyzed with XC-MS. This resulted in a list of 1558 single features. Using the same mass and retention time ranges as earlier with our software (i.e. 10,000 and 10 seconds) only 131 of the 174 target features were found. The correlation of the integrated areas for those targets that were found with those from the reference set (107 of them had a correlation higher than 0.9) was comparable to our untargeted approach but in general, XC-MS seems to have issues detecting peaks with low average peak intensity. One could argue that different settings for XC-MS should be used but the settings were already optimized for analysis of comparable lipidomics data which had been a very time consuming process by itself.

5.4.3 *Untargeted results*

Our untargeted integration method generated a lot more data than the 174 known targets and up to now the feature extraction, grouping and sample comparison did not include any study or sample information and as such could be considered really untargeted. To reduce the number of feature-sets to an easier to interpret and handle amount we included available analytical sample and study (design) information to post-process the feature-set list.

Depending on the types of samples that are measured, a biological question that needs to be answered (if any), several post-processing steps were possible using sample and study design information[13]. In our case, we focused on only those feature-sets for which a signal was detected in all quality control samples. This hugely reduced the number of feature-sets from 210,000 to approximately 3,200, a conservative estimate of at least 10 times the number of known-targets. Figure 7 shows the relation between the RSD values of the peak areas of the characteristic mass of the feature-set determined by repeated measurements of the QC samples (RSD_{QC}) vs. the average peak area for these 3,200 feature-sets. We again confirmed that all known targets from the target list were still present (Supporting Information Figure S55). In Figure 7 the ad-

ditional feature-sets are plotted from which we first removed any feature-set possibly related to a known target (e.g. other molecular adducts (NH_4, Na and K) and molecular ions in case a molecular adduct was used to quantify). For example, if for a known target with mass M there was a feature-set with the same retention time and mass M+18.033823, M+22.98922 or M+38.963158 for NH_4, Na or K respectively, it was removed. This was done for the purposes of this study only to demonstrate if the remaining feature-sets not related to known targets also contain biological information using PLS-DA modeling later. The vertical and horizontal dotted lines indicate the 20% RSD_{QC} level and the average intensity of the lowest abundant known target respectively (Ilow = 18,620 AU).

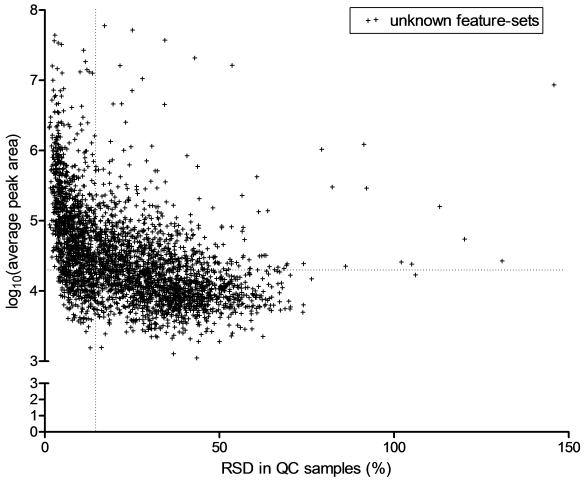


Figure 7: RSD in QC samples per feature set vs. $\log_{10}$ of average intensity of that feature set for all the unknown feature-sets that were present in all QC samples and not related to any known target.

It was observed that the average intensities of approximately 55% of the 3,200 feature-sets were above Ilow and that almost 50% had an RSD_{QC} lower than 20%, which from a data analysis point of view would be of interest at least to include in the analysis[117]. To indicate the added value of the extra feature-sets without entering the realms of a full data-analysis, we focused on four different subsets of feature-sets; the known targeted feature-sets (156 feature-sets), a sub-set of unknown feature-sets with characteristics similar to the known targets (RSD_{QC} <20% and average intensity >Ilow, 1,175 feature-sets) and unknown feature-sets that could biologically be very interesting too because of their low abundance(e.g. oxidized lipids etc.) , (RSD_{QC} <20% and average intensity <Ilow). In the fourth sub-set all feature-sets from the three other sub-sets were combined.

To compare the biological information content in either sub-set four different PLS-DA models[9], were build. For these models we focused on the sub-group of 12 subjects before and after bariatric treatment (see experimental) and object centered the data to remove inter patient variability. To limit the effect of overfitting all four models were double cross validated[114] and to get more realizations of the classification error different randomization schemes were used[41]. All four models, created with one latent variable, were highly pre-

dictive (specificity and selectivity were both 1) and performed equally with no misclassification (accuracy=1). This confirmed that relevant biological information was also contained in the new non-targeted found feature-sets.

To quantify the added value of the new feature-sets we ranked the most important variables of the PLS-DA models according to their selectivity ratio (SR)[101, 102]. To disclose the most important variable with respect to the response variable the selectivity ratios translates the PLS model regression vector to information suitable for easy (univariate) interpretation. In Figure 8 the top 20 most important feature-sets are ranked according to their SR value obtained from the combined PLS-DA model. Most (10) of the high ranking feature-sets were part of the new found feature-sets confirming their potential biological relevance, also with respect to the known targets. Furthermore 5 feature-sets originating from the low-abundant group were found in this top 20 underlining their biological potential.

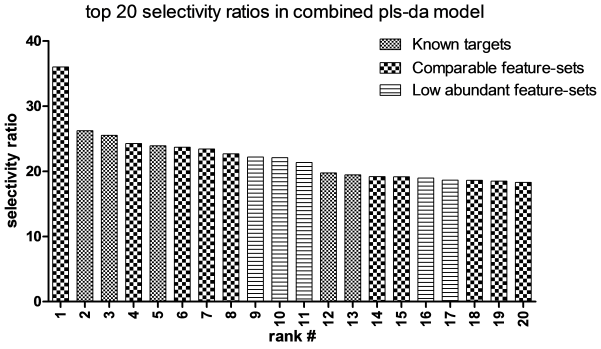


Figure 8: RSD in QC samples per feature set vs. log10 of average intensity of that feature set for all the unknown feature-sets that were present in all QC samples and not related to any known target.

5.5 DISCUSSION

As we demonstrated the effectiveness of our approach we realize that this is not the only software that is capable of doing untargeted analysis using high resolution LC-MS spectra. What makes this method different however, is that only a very limited amount of expert knowledge is required to use this method and the untargeted implementation to the very end. Subsequent matching of feature-sets instead of single features across samples arguably increases the probability of a proper match. Using only isotopic patterns we generated a lot of feature-sets and eliminated a lot of single features. In our demonstration we kept the rule-set as small as possible. The advantage of incorporating as little information as possible is that we were able to focus at feature extraction per sample which makes up-scaling for future implementations relatively easy.

The amount of feature-sets could be reduced even further by applying adduct rules or for GC-MS extended to a whole range of co-eluting features (possibly database driven). Using the framework of single sample feature extraction and grouping to feature-sets we envision that this approach can also form the base for automated CE-MS analysis which is hampered by huge migration time shifts across samples.

The number of unique feature-sets over all 128 samples (∼210,000) was large and therefore difficult to interpret and handle (for any statistical modeling and reporting). This reveals an intrinsic complication that untargeted analysis approaches face. This however does not have to be problematic if clever sampling schemes and study designs are used. For routine analysis one could for example select only those feature-sets that are stable in repeated measurements of the same sample or QC samples like we did in this case. If the study is focused on biomarker discovery the QC approach could be unsafe and different approaches could be followed (e.g. pooled QC samples per class).

Another aspect of our approach is the potential risk of missing out on features that do not show any discernible isotopic pattern across all samples included and therefore end up in the list of single features. The advantage of the focus on feature-sets in our opinion clearly outweighs this risk. However when small metabolites (e.g. low number of C atoms) are expected with low abundance single features should and can be included.

Even though we demonstrated the biological potential (i.e. relevance) of these additional unknown features, the outcome of our data analysis was purely driven by emphasizing the contrast between the new feature-sets and the known-features. For this reason no reference was made to the identity of the selected feature-sets.

The software was demonstrated for data from a TOF mass spectrometer. The software has been extended to allow for non-equidistant sampling points in cases the data was acquired on an FT or Orbitrap mass spectrometer. In contrast to R (runtime environment for XC-MS), Matlab is not freely available and licenses would be required to run our integration tool.

5.6 CONCLUSIONS

We introduced a new method to integrate high resolution full scan profiling LC-MS data in an untargeted manner. To demonstrate the effectiveness of our strategy of only comparing feature-sets over samples we used complex lipidomics full scan profiling LC-MS data of 128 samples. We compared the automatically integrated areas for a set of 174 known target lipids to those obtained by optimized and manually controlled quantification using vendor software. For 87% of the targets the correlation between the sample areas determined by vendor software and the untargeted approach was higher than 0.8. For 73% of the targets the correlation was higher than 0.9. Low correlations were found for isomer peaks that were integrated as one peak using vendor software or peaks that showed no discernible isotopic pattern. No clear relations between the correlation and average peak area were observed. The high correlations are impressive since the integration parameters for the different lipids (e.g. combining/splitting of isomers) in the vendor software have been highly optimized over a period of years. Furthermore, it indicates that even if the only interest is in known compounds for reasons like direct biological interpretability, our approach can still be applied. The untargeted method extracted at least 10 times more feature-sets than the known target lipids. PLS-DA models based on either the additionally found feature-sets or on the extracted known targets performed equally. Based on selectivity ratios we showed that the most important feature-sets were contained in the set

of higher abundant unknown feature-sets, confirming the potential biological relevance of these feature-sets found and thus indicating the added value of untargeted integration. The method proposed is fully automated and almost no user interaction is needed which makes it a perfect candidate for inclusion in a data pre-processing pipeline. As the extraction is on a per-sample base the method is highly scalable when more computers/processors are available. The grouping of features to feature-sets is specific to high resolution mass spectrometry data. We envision that this method can be extended to facilitate GC-MS feature extraction by using database information and automated CE-MS integration by adding the necessary alignment routines. A copy of the Matlab package is available on request from the corresponding author.

ACKNOWLEDGEMENT

We would like to thank Professor W. van Dijk at Leiden University Medical Center for providing the samples. We acknowledge Professor A.H.C. van Kampen and Mia Pras-Raves at the Academic Medical Center, Amsterdam for processing the samples with XC-MS. We further would like to thank Jorne Troost for his feedback on the analytical issues we faced and Adrie Dane for constructive discussions. This project was (co)financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research.

S5

SUPPLEMENT TOA NEW APPROACH TO UNTARGETED INTEGRATION OF HIGH RESOLUTION LC-MS DATA

S5.1 RE-CALIBRATION OF PROFILE DATA

To explain why a specific mass for a compound has a bandwidth and not one exact mass it is convenient to recall some of the basics of time of flight mass spectrometry; depending on the mass over charge ratio (m/z) it takes longer to reach the detector (reflectron). Peak broadening occurs because not all of the (identical) molecules get desorbed at the same time in the same place (in the reflector). In Figure S51a this peak broadening is demonstrated with a mass spectrum of a molecular ion and its (2) isotopic masses. The numbers above the peaks indicate the m/z value that was determined. Figure S51b displays the same part of the mass spectrum but now a number of scans later. Even though the masses look very alike they are not the same.

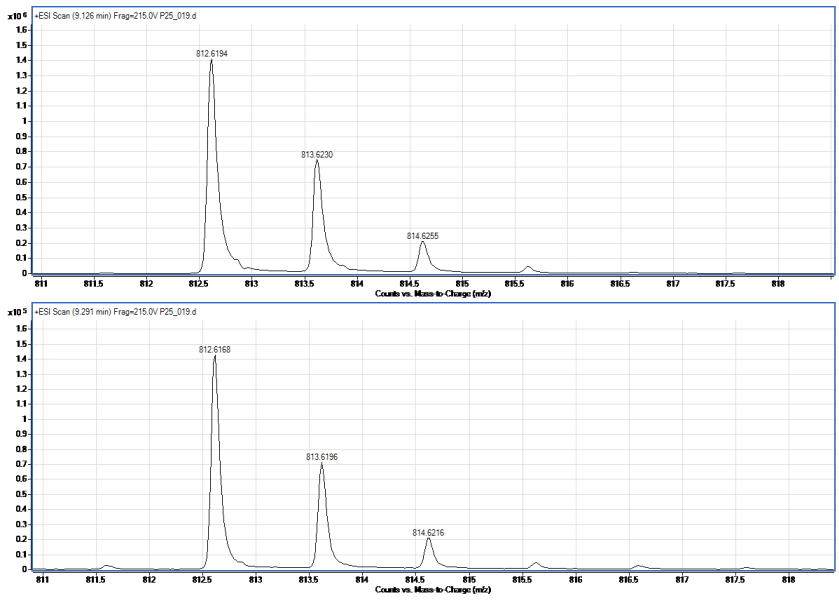


Figure S51: A: part of a profile mass spectrum at point x. B: the same part of the mass spectrum some scans later. The masses are not exactly the same for each scan

The masses were determined from the time that it took to reach the detector (Time of Flight). Consequently this meant that the bulk of the compound in the different scans reached the detector at a slightly different time with each scan. Since a lock mass calibration fluid (922.0098 Da) was post-column infused to

the sample, every scan contained this absolute reference mass. We recalibrated the mass values using this lock mass.

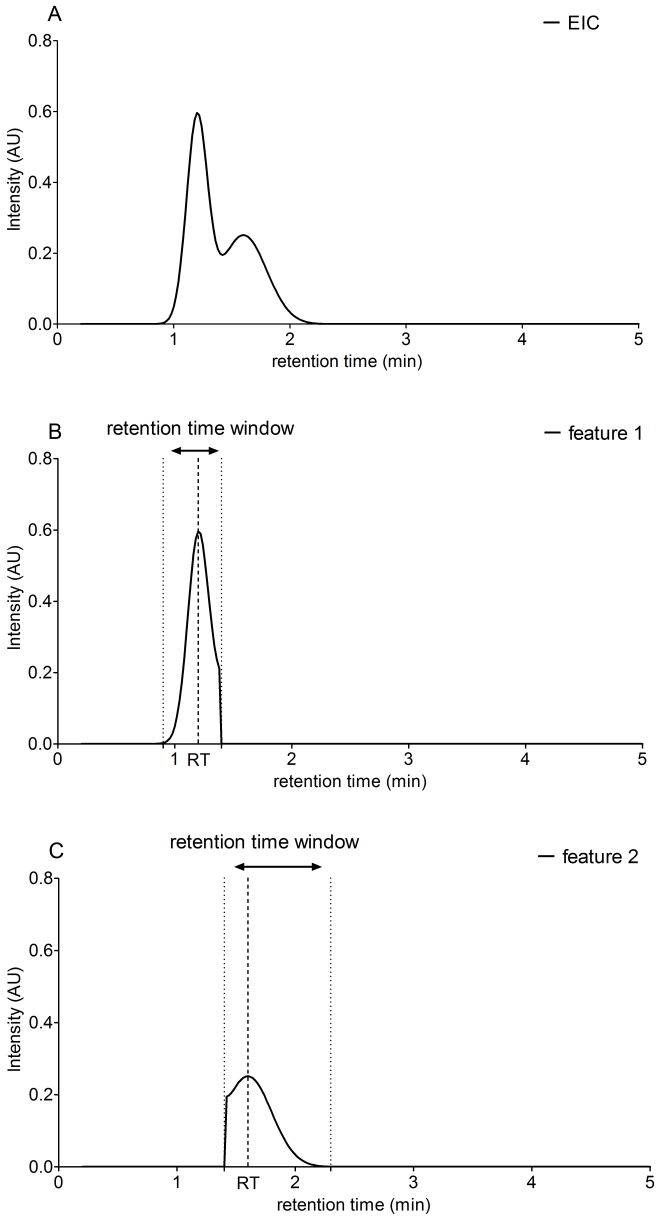


Figure S52: a bimodal peak, e.g. no separation at baseline level, results in 2 separate features after using the optional split function (b and c).

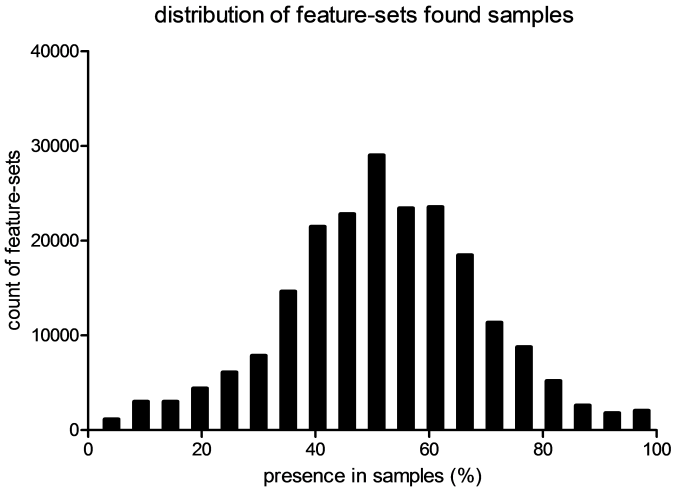


Figure S53: The distribution of the feature-sets found in study and QC samples, most of the feature-sets are found in 50-55% of all samples.

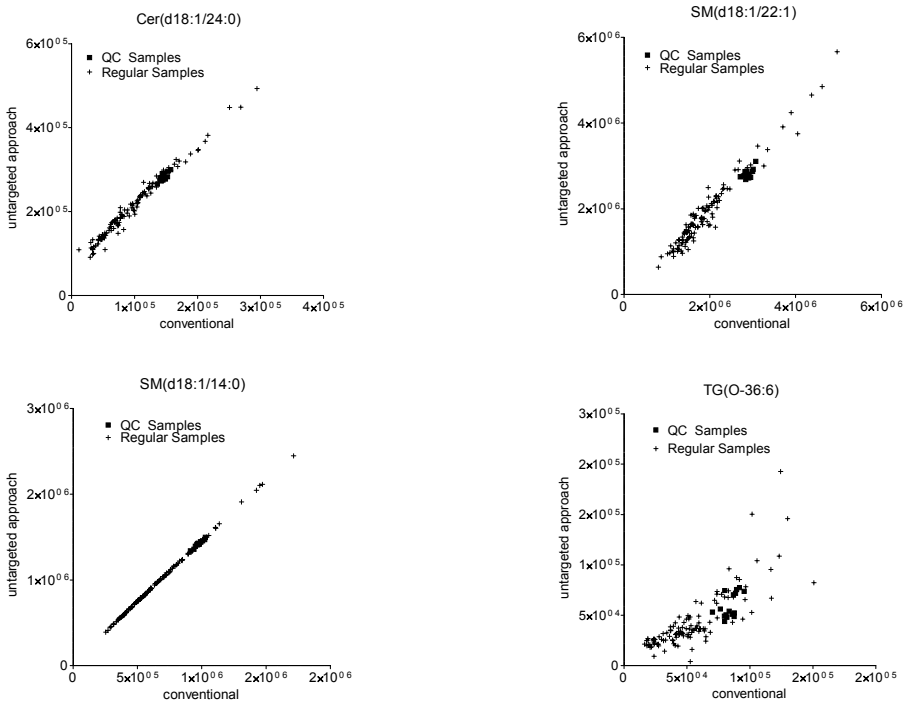


Figure S54: comparison between untargeted integrated areas (y-axis) and areas obtained using a target list with vendor software (x-asis). The squares indicate QC samples, the crosses the regular samples.

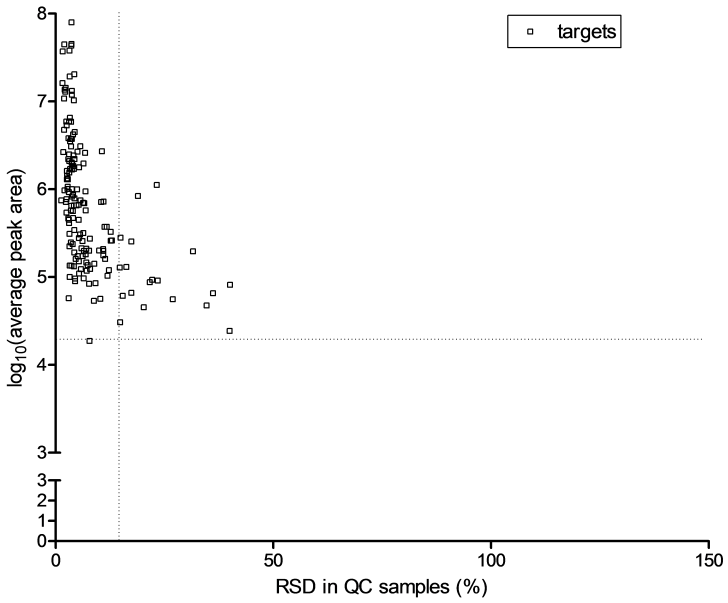


Figure S55: the RSD in QC samples for the feature-sets belonging to the known targets vs log10 average intensity of that feature-set

6

SUMMARY AND FUTURE PERSPECTIVES

The aim of this thesis was to develop concepts and methods to extract qualitative and quantitative information about metabolites from untargeted mass spectrometric data of biological samples. Several typical challenges in data handling were addressed that prevent a straightforward interpretation (data analysis) of the data acquired with different types of mass spectrometric-based metabolomics methods (GC-MS, LC-MS, CE-MS or DI-MS) methods. The critical parameters causing variation in quantitative results were identified and studied at different stages in the metabolomics workflow such as data acquisition (**Chapter 2**), data pre-processing (**Chapters 2**, **4** and **5**) and data analysis (**Chapters 3** and **5**). Different methods and concepts were developed to address these and to improve the quantitation of metabolites and the comparison between metabolite data in different samples of the same study measured at different moments or between studies. The methods developed focused on improved normalization, data pre-processing of untargeted analysis and data pre-processing of high resolution direct infusion mass spectrometry data. Furthermore it was demonstrated that even for metabolomic studies with few samples cross-validation of multivariate models can be very time consuming and parallel implementation on a (large) cluster of computers is the way to make such computations feasible.

All methods were developed in such a way that they can be used as an automated module within the data processing pipeline. In **Chapter 2** it was shown that quantification of metabolites in metabolomics studies greatly can be improved using single point calibration. The abundance of each metabolite is related to a predefined amount of internal standard added to each sample. This one-point calibration is indispensable when large-scale metabolomics studies are performed, in which both within and between batch differences become a problem due to instrumental and environmental changes during the measurements of the large number of samples. By repeated measurements of a representative sample for the study, usually pooled Quality Control (QC) samples, we demonstrated that the relative standard deviation (RSD) of individual metabolites in these QC samples before and after internal standard correction is a good measure to find the best match between a given metabolite and a set of internal standards that were spiked in the sample. This is a practical alternative to using a separate (isotope labeled) internal standard for each metabolite, which is often not feasible due to high costs and/or limited availability of such standards. It was shown that two types of QC samples are required whereby the first type, the calibration QC samples, is used to perform a one-point calibration, and the second type, validation QC samples, is used

to assess how well the calibration procedure improves the data quality. As a result, the biological variation in the study samples becomes more apparent, and more meaningful statistical models can be built.

In **Chapter 3** it was demonstrated that based on untargeted LC-MS measurements of urine samples a statistically significant multivariate model could be constructed to distinguish between progressive and non-progressive subjects within the normal urinary albumin excretion rate (AER) group with 75% accuracy. The metabolic profile defined by the multivariate model included all of the measured compounds that showed a univariate significant difference between the two groups. The profile, however, also included metabolites that did not show a univariate significant difference and emphasizes the additional benefit of multivariate statistics over univariate statistics alone in preventing overlooking candidate biomarkers. Key for multivariate modeling however is proper model validation and permutation testing.

We demonstrated the use of rapid metabolic fingerprinting for rapid determination of metabolic changes in **Chapter 4** and we were able to distinguish metabolic profiles of early developmental stages of zebrafish embryos using High Resolution Direct Infusion Mass Spectrometry (HR-DI-MS). Actually, in this project data preprocessing was not the focus of the project at the beginning, but the lack of automated data pre-processing of this type of data initiated the development of the method described in this chapter. The huge number of features that were generated in a single mass spectrum made clear that some kind of unique (virtual) reference point (e.g. a feature appearing in all samples) over the samples was needed. After careful data preprocessing and analysis, we were able to isolate 102 features that showed consistent behavior within each developmental stage. Principal Component Analysis revealed that early development stages of zebrafish embryos could be differentiated from each other. In total 27 out of the 102 features were (putatively) identified. Several observed trends of these putatively identified metabolites were supported by previous publications. But more importantly, with our method several new features were discovered as being relevant during early development of zebrafish embryos. It could be concluded that HR-DI-MS is suitable for rapid metabolic profiling on zebrafish embryos. However, to improve robustness and obtain more high-quality features better, fast but appropriate sample preparation methods are required.

In **Chapter 5** a new method to integrate high resolution full scan profiling LC-MS data in an untargeted manner is introduced. To demonstrate the effectiveness of the strategy of only comparing feature-sets over samples we used complex lipidomics full scan profiling LC-MS data. The automatically integrated areas for a set of known target lipids were compared to those obtained by optimized and manually controlled quantification using vendor software. i.e. which were considered as the reference data. Very high correlations with the reference data were observed which was impressive since the integration parameters for the different lipids (e.g. combining/splitting of isomers) in the vendor software have been highly optimized over a period of years. The untargeted method extracted at least 10 times more feature-sets than the known

target lipids and PLS-DA models based on either the new found feature-sets or on the extracted known targets performed equally well. Selectivity ratios however, showed that the most important discriminating feature-sets were contained in the set of higher abundant unknown feature-sets, confirming the potential biological relevance of these feature-sets found and thus indicating the added value of untargeted integration, as these allow also the detection of metabolites and lipids at low concentrations, and include also so far unknown metabolites. The proposed integration method requires only a very limited amount of expert knowledge and is fully automated with almost no user interaction which makes it a perfect candidate for inclusion in a data pre-processing pipeline. Because the extraction and integration method is implemented on a per-sample base, the method is highly scalable when more computers/processors are available. We envision this method to be extended to also facilitate GC-MS feature extraction but also enable automated CE-MS data analysis that suffers from huge migration time shifts between samples.

With respect to the title of this thesis, one may ask the question whether quantification of untargeted mass spectrometry data was significantly improved by the methods described in this thesis, and are these methods suitable for metabolomics research project? The answer is complex as the developed methods indeed improved the quantification of metabolites and allowed to identify and deal with the sources of analytical variation, but more developments are required.

The inclusion of multiple internal standards in untargeted profiling methods is essential and it should always be investigated which internal standards to use. Often there is a high correlation between the added internal standards. This is partly due to the way the samples are prepared (i.e. adding a mix of internal standards as one solution). Ion suppression can jeopardize quantification. The experimental setup to qualitatively screen for regions that suffer from matrix effects using post column infusion is very interesting. Infusing a mixture of reference signals could then act as calibration points and correct for the amount of suppression at that point in time, i.e. acting as internal standards.

The huge number of additional features that was generated using the untargeted integration method also introduces a new challenge: what is the right strategy to analyze these features, which feature-set is good, which is not? To test this, often, like we did, the reproducibility in repeated measurement of pooled quality control (QC) samples is used. If however the focus is on biomarker discovery such a strategy is riskful. The nature of a pool of QC samples is that it is an average of all metabolites that can be expected. But what if the (unknown) biomarker that we are looking for is too much diluted in this average sample? If the concentration of a biomarker is close to the quantification limit, and if it is only present in one of several classes of a (clinical) study, the concentration in the pooled QC sample may be so low that the RSD values of repeated measurements in the QC samples are unacceptable high and the biomarker will be excluded for further data analysis. In cases like these, a pooled reference sample per sample class group would be beneficial.

RSDs can then be determined per sample class and data-analysis would be enriched with metabolites that otherwise would not have passed the criteria. Consequently the standard data acquisition protocol should be adjusted accordingly for future metabolomic studies by including multiple class-specific QC samples. Experiments should quantify/qualify this dilution effect and assess whether pooled QC samples per control group are more representative than overall pooled QC samples for all metabolites in both groups. In these experiments the response of the QC pool per group(s) should be compared to the response of the combined (regular) pooled QC sample over a large number of measurements to mimic the experimental variations of a large metabolomics study as much as possible.

The importance of the QC samples becomes more and more apparent but the limited amount of material often prohibits re-measurement in different studies but sometimes there is not even enough material available to create an adequate QC pool for the original study. In such cases it would be of much interest to evaluate the response of different types of reference samples. Do these reference samples behave adequately for QC monitoring and QC correction and/or would they act as good validation samples to test the QC correction and IS correction steps? It would be even more interesting to evaluate if these references samples could be used as transfer samples to create so-called transfer models and enable direct comparisons between different studies.

In Chapter 5 feature-sets like molecular adducts that were related to known lipids were explicitly left out not to influence the model prediction. But what if they were also included? It is known that in some matrices (e.g. different samples, experimental conditions) adducts are more easily created then in others. It could be expected that the sum of all compound related masses could improve the quantitative comparison between the different samples and minimize matrix effects, another hypothesis worthwhile to test in the future.

In conclusion, significant progress in the data processing of MS-based metabolomics data was achieved in this thesis, but much progress still has to be and will be made in this field. The ever increasing computing power enables ever more complex data extraction, data processing and data analysis procedures. The popularity of untargeted methods is likely to increase as a consequence of the development of new software, complex deconvolution algorithms, higher mass accuracies and reference databases. And such untargeted methods allow the detection of a target list of metabolites. But even though fewer compounds are measured and/or reported, targeted methods will always have its advantages (e.g. less processing time), especially when high throughput of samples is of concern. The success of an analytical platform hugely depends on the ability of the subsequent data (pre)processing steps to obtain high quality data. This dependency will always be there but new data (pre)processing techniques may also inspire technical changes and/or different analytical setups. In conclusion, metabolomics research should be approached as an integrative effort combining knowledge on the biological question, sample processing and data acquisition and data processing, and subsequent data analysis, to ultimately answer biological questions.

7

BIBLIOGRAPHY

[1] Agilent. MassHunter, 2013.

[2] Shaukat Ali, Danielle L. Champagne, Herman P. Spaink, and Michael K. Richardson. Zebrafish embryos and larvae: A new generation of disease models and drug screens. *Birth Defects Research Part C: Embryo Today: Reviews*, 93(2):115–133, 2011.

[3] J L Anderson, J D Carten, and S A Farber. Zebrafish lipid metabolism: from mediating early patterning to the metabolism of dietary fat and cholesterol. *Methods in cell biology*, 101:111–141, 2011.

[4] Gregor Arh, Leo Klasinc, Marjan Veber, and Matev Pompe. Modeling of the Mass Spectrometric Response Factors in Non-target Analysis. *Acta Chim. Slov.*, pages 581–585, 2010.

[5] No authors listed. Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy. The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group. *New England Journal of Medicine*, pages 381–389, 2000.

[6] Michael J Avery. Quantitative characterization of differential ion suppression on liquid chromatography / atmospheric pressure ionization mass spectrometric bioanalytical methods. *Rapid Communications in Mass Spectrometry*, pages 197–201, 2003.

[7] G Bakker. http://clinicaltrials.gov/ct2/show/nct00655798, 4 2008.

[8] Gertruud C M Bakker, Marjan J Van Erk, Linette Pellis, Suzan Wopereis, Carina M Rubingh, Nicole H P Cnubben, Teake Kooistra, Ben Van Ommen, and Henk F J Hendriks. An antiinflammatory dietary mix modulates inflammation and oxidative and metabolic stress in overweight men : a nutrigenomics approach 1 4. *American Society for Nutrition*, pages 1044–1059, 2010.

[9] Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, March 2003.

[10] Daniel G Beach and Wojciech Gabryelski. Linear and Nonlinear Regimes of Electrospray Signal Response in Analysis of Urine by Electrospray Ionization-High Field Asymmetric Waveform Ion Mobility Spectrometry-MS and Implications for Nontarget Quantification. *Analytical chemistry*, 85(4):2127–34, February 2013.

[11] Francis Beaudry and Pascal Vachon. Electrospray ionization suppression, a physical or a chemical phenomenon? *Biomedical Chromatography*, 20(2):200–205, 2005.

[12] Manfred Beckmann, David Parker, David P Enot, Emilie Duval, and John Draper. High-throughput, nontargeted metabolite fingerprinting using nominal mass flow injection electrospray mass spectrometry. *Nature protocols*, 3(3):486–504, January 2008.

[13] Sabina Bijlsma, Ivana Bobeldijk, Elwin R. Verheij, Raymond Ramaker, Sunil Kochhar, Ian A Macdonald, Ben Van Ommen, and Age K Smilde. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Analytical Chemistry*, 78:567–574, 2006.

[14] I. Bobeldijk, M. Hekman, J. de Vries-van der Weij, L. Coulier, R. Ramaker, R. Kleemann, T. Kooistra, C. Rubingh, A. Freidig, and E. Verheij. Quantitative profiling of bile acids in biofluids and tissues based on accurate mass high resolution lc-ft-ms: compound class targeting in a metabolomics workflow. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 871(2):306–13, 2008.

[15] Luisa Bonafe, Heinz Troxler, Thomas Kuster, and Claus W Heizmann. Evaluation of Urinary Acylglycines by Electrospray Tandem Mass Spectrometry in Mitochondrial Energy Metabolism Defects and Organic Acidurias. *Molecular Genetics and Metabolism*, 311:302–311, 2000.

[16] Rainer Breitling. What is systems biology? *Frontiers in physiology*, 1(May):9, January 2010.

[17] Bruker. Compass Analysis, 2013.

[18] L. Burton, G. Ivosev, S. Tate, G. Impey, J. Wingate, and R. Bonner. Instrumental and experimental effects in lc-ms-based metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 871(2):227–35, 2008.

[19] E Businge, K Brackmann, T Moritz, and U Egertsdotter. Metabolite profiling reveals clear metabolic changes during somatic embryo development of Norway spruce (Picea abies). *Tree physiology*, 32(2):232–244, 2012.

[20] M Luiza Caramori, Paola Fioretto, and Michael Mauer. Enhancing the Predictive Value of Urinary Albumin for Diabetic Nephropathy. *Journal of the American Society of Nephrology*, pages 339–352, 2006.

[21] R Carvalho, J de Sonneville, O W Stockhammer, N D Savage, W J Veneman, T H Ottenhoff, R P Dirks, A H Meijer, and H P Spaink. A high-throughput screen for tuberculosis progression. *PloS one*, 6(2):e16779, 2011.

[22] R.A. Chalmers, C.R. Roe, T.E. Stacey, and C.L. Hoppel. Urinary excretion of l-Carnitine and Acylcarnitines by patients with disorders of organic acid metabolism: Evidence for secondary insufficiency of l-Carnitine. *Pediatric Research*, 18(12):1325–1328, 1984.

[23] Erin Chambers, Diane M Wagrowski-Diehl, Ziling Lu, and Jeffrey R Mazzeo. Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, 852(1-2):22–34, June 2007.

[24] W. Chang, U. Thissen, K. A. Ehlert, M. M. Koek, R. H. Jellema, T. Hankemeier, J. v. d. Greef, and M. Wang. Effects of Growth Conditions and Processing on Rehmannia glutinosa using Fingerprint Strategy. *Planta Med*, 72:458–67, 2006.

[25] S. Drake, R. Bowen, Remaley. A., and Hortin G. Potential interferences from blood collection tubes in mass spectrometric analyses of serum polypeptides. *Clinical Chemistry*, 50(12):2398–2401, 2004.

[26] John Draper, Amanda J. Lloyd, Royston Goodacre, and Manfred Beckmann. Flow infusion electrospray ionisation mass spectrometry for high throughput, non-targeted metabolite fingerprinting: a review. *Metabolomics*, July 2012.

[27] W B Dunn, D I Broadhurst, H J Atherton, R Goodacre, and J L Griffin. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, 40(1):387–426, 2011.

[28] W.B. Dunn, D. Broadhurst, M. Brown, P.N. Baker, C.W. Redman, L.C. Kenny, and D.B. Kell. Metabolic profiling of serum using ultra performance liquid chromatography and the ltq-orbitrap mass spectrometry system. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 871(2):288–98, 2008.

[29] Paul H C Eilers. A Perfect Smoother. *Analytical Chemistry*, 75(14):3631–3636, 2003.

[30] A El-Faramawy, K W Siu, and B A Thomson. Efficiency of nanoelectrospray ionization. *Journal of the American Society for Mass Spectrometry*, 16(10):1702–1707, 2005.

[31] David I Ellis, Warwick B Dunn, Julian L Griffin, J William Allwood, and Royston Goodacre. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*, 8(9):1243–66, September 2007.

[32] Alisdair R Fernie, Asaph Aharoni, Lothar Willmitzer, Mark Stitt, Takayuki Tohge, Joachim Kopka, Adam J Carroll, Kazuki Saito, Paul D Fraser, and Vincenzo DeLuca. Recommendations for reporting metabolite data. *The Plant cell*, 23(7):2477–82, July 2011.

[33] Patrik Finne, Antti Reunanen, Svante Stenman, Per-Henrik Groop, and Carola Grönhagen-Riska. Incidence of end-stage renal disease in patients with type 1 diabetes. *JAMA : the journal of the American Medical Association*, 294(14):1782–7, October 2005.

[34] Tobias Fuhrer, Dominik Heer, Boris Begemann, and Nicola Zamboni. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Analytical chemistry*, 83(18):7074–80, September 2011.

[35] Michael J Gibney, Marianne Walsh, Lorraine Brennan, Helen M Roche, Bruce German, and Ben Van Ommen. Metabolomics in human nutrition : opportunities and challenges 1 3. *American Journal of Clinical Nutrition*, pages 497–503, 2005.

[36] Christian Gieger, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé de Angelis, Florian Kronenberg, Thomas Meitinger, Hans-Werner Mewes, H-Erich Wichmann, Klaus M Weinberger, Jerzy Adamski, Thomas Illig, and Karsten Suhre. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics*, 4(11):e1000282, November 2008.

[37] H.G. Gika, G.A. Theodoridis, J.E. Wingate, and I.D. Wilson. Within-day reproducibility of an hplc-ms-based method for metabonomic analysis: application to human urine. *Journal of proteome research*, 6(8):3291–303, 2007 Aug.

[38] W Goessling, T E North, and L I Zon. New waves of discovery: modeling cancer in zebrafish. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(17):2473–2479, 2007.

[39] Per-henrik Groop, Merlin C Thomas, John L Moran, Johan Wade, Lena M Thorn, Milla Rosengå rd ba, Markku Saraheimo, and Kustaa Hietala. The Presence and Severity of Chronic Kidney Disease Predicts All-Cause Mortality in Type 1 Diabetes. *Baseline*, 58(July), 2009.

[40] J.L. Gross, M.J. de Azevedo, S.P. Silveiro, L.H. Canani, M.L. Caramori, and T Zelmanovitz. Diabetic Nephropathy : Diagnosis , Prevention , and Treatment. *Diabetes Care*, 28:176–188, 2005.

[41] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2nd edition, 2008.

[42] S Hayashi, S Akiyama, Y Tamaru, Y Takeda, T Fujiwara, K Inoue, A Kobayashi, S Maegawa, and E Fukusaki. A novel application of metabolomics in vertebrate development. *Biochemical and biophysical research communications*, 386(1):268–272, 2009.

[43] S Hayashi, M Yoshida, T Fujiwara, S Maegawa, and E Fukusaki. Single-embryo metabolomics and systematic prediction of developmental stage in zebrafish. *Zeitschrift fur Naturforschung. C, Journal of biosciences*, 66(3-4):191–198, 2011.

[44] Diana M Hendrickx, Margriet M W B Hendriks, Paul H C Eilers, Age K Smilde, and Huub C J Hoefsloot. Reverse engineering of metabolic networks, a critical assessment. *Molecular bioSystems*, 7(2):511–20, February 2011.

[45] Karsten Hiller, Jasper Hangebrauk, Christian Jäger, Jana Spura, Kerstin Schreiber, and Dietmar Schomburg. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Analytical chemistry*, 81(9):3429–39, May 2009.

[46] Olga Hrydziuszk and Mark R. Viant. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, 8(1 Supplement):161–174, 2012.

[47] Oxford Dictionaries http://oxforddictionaries.com, 2013.

[48] Nutritional Phenotype Database http://www.dbnp.org. Nutritional Phenotype Database.

[49] Eigenvector Research Inc. http://www.eigenvector.com, 2008.

[50] The Mathworks http://www.mathworks.com, 2007.

[51] The Mathworks http://www.mathworks.com, 2008.

[52] SAS http://www.sas.com, 2007.

[53] Chunxiu Hu, Judith Van Dommelen, Rob Van Der Heijden, Gerwin Spijksma, Theo H Reijmers, Mei Wang, Elizabeth Slee, Xin Lu, Guowang Xu, Jan Van Der Greef, and Thomas Hankemeier. RPLC-Ion-Trap-FTMS Method for Lipid Profiling of Plasma : Method Validation and Application to p53 Mutant Mouse Model research articles. *Journal of proteome research*, pages 4982–4991, 2008.

[54] IAEA. Development and use of reference materials and quality control materials. Technical Report April, Agency, International Atomic Energy, 2003.

[55] Thomas Illig, Christian Gieger, Guangju Zhai, Werner Römisch-Margl, Rui Wang-Sattler, Cornelia Prehn, Elisabeth Altmaier, Gabi Kastenmüller, Bernet S Kato, Hans-Werner Mewes, Thomas Meitinger, Martin Hrabé de Angelis, Florian Kronenberg, Nicole Soranzo, H-Erich Wichmann, Tim D Spector, Jerzy Adamski, and Karsten Suhre. A genome-wide perspective of genetic variation in human metabolism. *Nature genetics*, 42(2):137–41, February 2010.

[56] Andris Jankevics, Edvards Liepinsh, Edgars Liepinsh, Reinis Vilskersts, Solveiga Grinberga, Osvalds Pugovics, and Maija Dambrova. Chemometrics and Intelligent Laboratory Systems Metabolomic studies of experimental diabetic urine samples by 1 H NMR spectroscopy and LC / MS method. *Chemometrics and Intelligent Laboratory Systems*, 97(1):11–17, 2009.

[57] Renger H. Jellema, Shaji Krishnan, Margriet M.W.B. Hendriks, Bas Muilwijk, and Jack T.W.E. Vogels. Deconvolution using signal segmentation. *Chemometrics and Intelligent Laboratory Systems*, 104(1):132–139, November 2010.

[58] Qiqin Yin-Goen John H. Phan and May D. Wang Andrew N. Young. Improving The Efficiency Of Biomarker Identification Using Biological Knowledge. In *Pacific Symposium on Biocomputing*, pages 14:427–438 (2009), 2009.

[59] Peter F Kador, Yukio Takahashi, Yoshio Akagi, Heike Neuenschwander, William Greentree, Petra Lackner, Karen Blessing, and Milton Wyman. Effect of Galactose Diet Removal on the Progression of Retinal Vessel Changes in Galactose-Fed Dogs. *Investigative Ophthalmology*, pages 1916–1921, 2002.

[60] D A Kane and C B Kimmel. The zebrafish midblastula transition. *Development*, 119(2):447–456, 1993.

[61] Piotr T. Kasper, Miguel Rojas-Chertó, Robert Mistrik, Theo Reijmers, Thomas Hankemeier, and Rob J. Vreeken. Fragmentation trees for the structural characterisation of metabolites. *Rapid Communications in Mass Spectrometry*, 26(19):2275–2286, October 2012.

[62] Mikko Katajamaa, Jarkko Miettinen, and Matej Oresic. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics (Oxford, England)*, 22(5):634–6, March 2006.

[63] Douglas B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discovery Today*, 11(2324):1085 – 1092, 2006.

[64] Douglas B Kell. Metabolomic biomarkers: search, discovery and validation. *Expert review of molecular diagnostics*, 7(4):329–33, July 2007.

[65] Douglas B Kell, Marie Brown, Hazel M Davey, Warwick B Dunn, Irena Spasic, and Stephen G Oliver. Metabolic footprinting and systems biology: the medium is the message. *Nature reviews. Microbiology*, 3(7):557–65, July 2005.

[66] Ramses F J Kemperman, Peter L Horvatovich, Berend Hoekman, Theo H Reijmers, Frits a J Muskiet, and Rainer Bischoff. Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: method development, evaluation, and application to proteinuria. *Journal of proteome research*, 6(1):194–206, January 2007.

[67] Carol Kilkenny, Nick Parsons, Ed Kadyszewski, Michael F W Festing, Innes C Cuthill, Derek Fry, Jane Hutton, and Douglas G Altman. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS one*, 4(11):e7824, January 2009.

[68] P Kimmelstiell and C Wilson. Benign and malignant hypertension and nephrosclerosis. *Am J Pathol*, 12:4548, 1936.

[69] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–10, November 2002.

[70] Frans M Van Der Kloet, Ivana Bobeldijk, Elwin R Verheij, and Renger H Jellema. Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. *Journal of Proteome Research*, 2009.

[71] Kevin J Knecht, Milton S Feather, John W Baynes, and South Carolina. Detection of 3-Deoxyfructose and 3-Deoxyglucosone Human Urine and Plasma : Evidence for Intermediate Stages of the Maillard Reaction in viva . *Review Literature And Arts Of The Americas*, 294(1):130–137, 1992.

[72] M. M. Koek, B. Muilwijk, M.J. van der Werf, and T. Hankemeier. Microbial metabolomics with gas chromatography/mass spectrometry. *Anal. Chem.*, 78 (4):1272–1281, 2006.

[73] Maud M Koek, Bas Muilwijk, J Van Der Werf, and Thomas Hankemeier. Microbial Metabolomics with Gas Chromatography / Mass Spectrometry. *Science*, 78(4):1272–1281, 2006.

[74] Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, and Marc Sturm. TOPP–the OpenMS proteomics pipeline. *Bioinformatics (Oxford, England)*, 23(2):e191–7, January 2007.

[75] C Lawrence. Advances in zebrafish husbandry and management. *Methods in cell biology*, 104:429–451, 2011.

[76] Alexander M Lawson, Ronald A Chalmers, and Richard W E Watts. Urinary Organic Acids in Man . I . Normal Patterns. *Clin. Chem*, 22:1283–1287, 1976.

[77] Wai-nang Paul Lee, Laszlo G Boros, and Vay-liang W Go. Metabolic Pathways as Targets for Drug Screening. *Metabolomics*, 2:31–9, 2006.

[78] V Lesage, Y Morin, È Rioux, C Pomerleau, Sh Ferguson, and É Pelletier. Stable isotopes and trace elements as indicators of diet and habitat use in cetaceans: predicting errors related to preservation, lipid extraction, and lipid normalization. *Marine Ecology Progress Series*, 419:249–265, November 2010.

[79] Guowen Liu, Qin C Ji, and Mark E Arnold. Identifying, evaluating, and controlling bioanalytical risks resulting from nonuniform matrix ion suppression/enhancement and nonlinear liquid chromatography-mass spectrometry assay response. *Analytical chemistry*, 82(23):9671–7, December 2010.

[80] Arjen Lommen. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Analytical chemistry*, 81(8):3079–86, April 2009.

[81] G.L Lou, C. Pinsky, and D. S. Sitar. Kynurenic acid distribution into brain and peripheral tissues of mice. *Canadian Journal of Physiology and Pharmacology,*, 72(2):161–167, 1994.

[82] Geoffrey Madalinski, Emmanuel Godat, Sandra Alves, Denis Lesage, Eric Genin, Philippe Levi, Jean Labarre, Jean-Claude Tabet, Eric Ezan, and Christophe Junot. Direct introduction of biological samples into a LTQ-Orbitrap hybrid mass spectrometer as a tool for fast metabolome analysis. *Analytical chemistry*, 80(9):3291–303, May 2008.

[83] VP Mäkinen, Carol Forsblom, Lena M Thorn, Johan Wade, Daniel Gordin, Milla Rosengå rd ba, Kustaa Hietala, Laura Kyllo, Markku Saraheimo, Nina Tolonen, Maija Parkkonen, Kimmo Kaski, and Mika Alakorpela. Metabolic Phenotypes, Vascular Complications, and Premature Deaths in a Population of 4,197 Patients With Type 1 Diabetes. *DIABETES*, 57(September):2480–2487, 2008.

[84] VP Mäkinen, Pasi Soininen, Carol Forsblom, Maija Parkkonen, Petri Ingman, Kimmo Kaski, and Per-henrik Groop. H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Molecular Systems Biology*, 4(167):1–12, 2008.

[85] VP Mäkinen, Pasi Soininen, Carol Forsblom, Maija Parkkonen, Petri Ingman, Kimmo Kaski, Per-henrik Groop, Mika Ala-korpela, and Study Group. Diagnosing diabetic nephropathy by 1 H NMR metabonomics of serum. *Magnetic Resonance Materials In Physics Biology And Medicine*, 19(6):281–296, 2006.

[86] Luc Massart and Bernard Vandeginste. *Chemometrix: a textbook*. Elsevier Science, 1991.

[87] B K Matuszewski and M L Constanzer. Strategies for the Assessment of Matrix Effect in Quantitative Bioanalytical Methods Based on HPLC-MS / MS. *Analytical chemistry*, 75(13):3019–3030, 2003.

[88] A H Meijer and H P Spaink. Host-pathogen interactions made transparent with the zebrafish model. *Current drug targets*, 12(7):1000–1017, 2011.

[89] James N Miller and Jane C Miller. *Chemometrics for Analytical Chemistry*. Trans-Atlantic Pubns, 5th edition, 2005.

[90] Sofia Moco, Jacques Vervoort, Raoul J. Bino, Ric C.H. De Vos, and Raoul Bino. Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry*, 26(9):855–866, October 2007.

[91] John a Morgan and David Rhodes. Mathematical modeling of plant metabolic pathways. *Metabolic engineering*, 4(1):80–9, January 2002.

[92] J.E Mrocheck and W T Jr. Rainey. Identification and Biochemical Significance of Substituted Furans in Human Urine. *Clinical Chemistry*, 18(8):821–828, 1972.

[93] Jeremy K Nicholson, Elaine Holmes, James M Kinross, Ara W Darzi, Zoltan Takats, and John C Lindon. Metabolic phenotyping in clinical and surgical environments. *Nature*, 491(7424):384–92, November 2012.

[94] E S Ong, C F Chor, L Zou, and C N Ong. A multi-analytical approach for metabolomic profiling of zebrafish (Danio rerio) livers. *Molecular bioSystems*, 5(3):288–298, 2009.

[95] Fred C Pampel. *Logistic Regression: A primer*. Sage Publications, Thousand Oaks, London, New Delhi, 2000.

[96] C Papan and L Chen. Metabolic fingerprinting reveals developmental regulation of metabolites during early zebrafish embryogenesis. *Omics : a journal of integrative biology*, 13(5):397–405, 2009.

[97] C Pardo-Martin, T Y Chang, B K Koo, C L Gilleland, S C Wasserman, and M F Yanik. High-throughput in vivo vertebrate screening. *Nature methods*, 7(8):634–636, 2010.

[98] D Pawlak, A Tankiewicz, P Mysliwiec, and W Buczko. Tryptophan Metabolism via the Kynurenine Pathway in Experimental. *Nephron*, pages 328–335, 2002.

[99] A E Pegg. Spermidine/spermine-N(1)-acetyltransferase: a key metabolic regulator. *American journal of physiology. Endocrinology and metabolism*, 294(6):E995–1010, 2008.

[100] Yang Qiu, Dilip Rajagopalan, Susan C Connor, Doris Damian, Lei Zhu, Amir Handzel, Hu Guanghui, Arshad Amanullah, Steve Bao, Nathaniel Woody, David MacLean, Kwan Lee, Dana Vanderwall, and Terence Ryan. Multivariate classification analysis of metabolomic data for candidate biomarker discovery in type 2 diabetes mellitus. *Cytokine*, pages 337–346, 2008.

[101] Tarja Rajalahti, Reidar Arneberg, Ann C Kroksveen, Magnus Berle, Kjell-Morten Myhr, and Olav M Kvalheim. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical chemistry*, 81(7):2581–90, April 2009.

[102] Tarja Rajalahti and Olav M Kvalheim. Multivariate data analysis in pharmaceutics: a tutorial review. *International journal of pharmaceutics*, 417(1-2):280–90, September 2011.

[103] Charles J Rebouche and Hermann Seim. Carnitine metabolism and its regulation in microorganisms and mammals. *Structure*, pages 39–61, 1998.

[104] Daniela Remane, Markus R Meyer, Dirk K Wissenbach, and Hans H Maurer. Ion suppression and enhancement effects of co-eluting analytes in multi-analyte approaches : systematic investigation using ultra-high-performance liquid chromatography / mass spectrometry with atmospheric- pressure chemical ionization or electrospray ion. *Rapid Communications in Mass Spectrometry*, pages 3103–3108, 2010.

[105] Daniela Remane, Dirk K Wissenbach, Markus R Meyer, and Hans H Maurer. Systematic investigation of ion suppression and enhancement effects of fourteen stable-isotope-labeled internal standards by their native

analogues using atmospheric-pressure chemical ionization and electrospray ionization and the relevance for multi-anal. *Rapid Communications in Mass Spectrometry*, pages 859–867, 2010.

[106] Miguel Rojas-Chertó, Piotr T Kasper, Egon L Willighagen, Rob J Vreeken, Thomas Hankemeier, and Theo H Reijmers. Elemental composition determination based on MSn. *Bioinformatics (Oxford, England)*, 27(17):2376–83, September 2011.

[107] S S Rubakhin, E V Romanova, P Nemes, and J V Sweedler. Profiling metabolites and peptides in single cells. *Nature methods*, 8(4 Suppl):S20–9, 2011.

[108] L Saint-Amant and P Drapeau. Time course of the development of motor behaviors in the zebrafish embryo. *Journal of neurobiology*, 37(4):622–632, 1998.

[109] Erik J. Saude, Darryl Adamko, Brian H. Rowe, Tom Marrie, and Brian D. Sykes. Variation of metabolites in normal human urine. *Metabolomics*, 3(4):439–451, September 2007.

[110] M J Schaaf, W J Koopmans, T Meckel, J van Noort, B E Snaar-Jagalska, T S Schmidt, and H P Spaink. Single-molecule microscopy reveals membrane microdomain organization of cells in a living vertebrate. *Biophysical Journal*, 97(4):1206–1214, 2009.

[111] S Schaub, J. Wilkins, T. Weiler, K. Sangster, D. Rush, and P. Nickerson. Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney International*, 65:323–332, 2004.

[112] Mark Schmidt, Glenn Fung, and Romer Rosaless. Optimization Methods for. *Solutions*, (x), 2009.

[113] A Shilatifard. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annual review of biochemistry*, 75:243–269, 2006.

[114] Suzanne Smit, J Van Breemen, Huub C J Hoefsloot, Age K Smilde, Johannes M F G Aerts, and Chris G De Koster. Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta*, 592:210–217, 2007.

[115] Colin a Smith, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, 78(3):779–87, February 2006.

[116] K H Soanes, J C Achenbach, I W Burton, J P Hui, S L Penny, and T K Karakach. Molecular characterization of zebrafish embryogenesis via DNA microarrays and multiplatform time course metabolomics studies. *Journal of Proteome Research*, 10(11):5102–5117, 2011.

[117] Chem Soc, Warwick B Dunn, David I Broadhurst, Helen J Atherton, Royston Goodacre, and Julian L Griffin. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*, pages 387–426, 2011.

[118] S.S. Soedamah-Muthu, N. Chaturvedi, M. Toeller, B. Ferriss, P. Reboldi, G. Michel, C. Manes, and J.H. Fuller. Risk Factors for Coronary Heart Disease in Type 1 Diabetic Patients in Europe The EURODIAB Prospective Complications Study. *Diabetes Care*, 24(2):530–537, 2004.

[119] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

[120] Karsten Suhre, So-Youn Shin, Ann-Kristin Petersen, Robert P. Mohney, David Meredith, Brigitte Wägele, Elisabeth Altmaier, Panos Deloukas, Jeanette Erdmann, Elin Grundberg, Christopher J. Hammond, Martin Hrabé de Angelis, Gabi Kastenmüller, Anna Köttgen, Florian Kronenberg, Massimo Mangino, Christa Meisinger, Thomas Meitinger, Hans-Werner Mewes, Michael V. Milburn, Cornelia Prehn, Johannes Raffler, Janina S. Ried, Werner Römisch-Margl, Nilesh J. Samani, Kerrin S. Small, H. -Erich Wichmann, Guangju Zhai, Thomas Illig, Tim D. Spector, Jerzy Adamski, Nicole Soranzo, and Christian Gieger. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, August 2011.

[121] H Sukardi, C Y Ung, Z Gong, and S H Lam. Incorporating zebrafish omics into chemical biology and toxicology. *Zebrafish*, 7(1):41–52, 2010.

[122] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, and M. Orešič. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, page 8:93, 2007 Mar 15.

[123] David W Johnson T. Contemporary clinical usage of LC / MS : Analysis of biologically important carboxylic acids. *Clinical Biochemistry*, 38:351 – 361, 2005.

[124] Merlin C Thomas and Robert C Atkins. Blood Pressure Lowering for the Kidney Disease. *Drugs*, 66(17):2213–2234, 2006.

[125] E Michael Thurman and Imma Ferrer. The isotopic mass defect: a tool for limiting molecular formulas by accurate mass. *Analytical and bioanalytical chemistry*, 397(7):2807–16, August 2010.

[126] Ewa Urbanczyk-Wochniak, Alexander Luedemann, Joachim Kopka, Joachim Selbig, Ute Roessner-Tunali, Lothar Willmitzer, and Alisdair R Fernie. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports*, 4(10):989–93, October 2003.

[127] Robert a van den Berg, Huub C J Hoefsloot, Johan a Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7:142, January 2006.

[128] Robert A. van den Berg, Huub C.J. Hoefsloot, Johan A. Westerhuis, A.K. Smilde, and Mariet J. Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7:142, 2006.

[129] J. van der Greef, S. Martin, P. Juhasz, A. Adourian, T. Plasterer, E.R. Verheij, and R.N. McBurney. The art and practice of systems biology in medicine: mapping patterns of relationships. *Journal of proteome research*, 6(4):1540–59, 2007 Apr.

[130] Jan van der Greef, Thomas Hankemeier, and Robert N McBurney. Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics*, 7(7):1087–94, October 2006.

[131] M. van der Werf, K. M. overkamp, b. Muilwijk, M. M. Koek, B. J. C. van der Werff-van der Vat, R. H. Jellema, L. Coulier, and T Hankemeier. Comprehensive analysis of the metabolome of Pseudomonas putida S12 grown on different carbon sources. *Molecular Biosystems*, 4:315–27, 2008.

[132] M. J. van der Werf, R. Takors, Jørn Smedsgaard, Jens Nielsen, T. Ferenci, J. C. Portais, C. Wittmann, M. Hooks, A. Tomassini, M. Oldiges, J. Fostel, and U. Sauer. Standard reporting requirements for biological samples in metabolomics experiments: Microbial and in vitro biology experiments. *Metabolomics*, 3:189–194, 2007.

[133] Laurence Vernez. *Analysis of carnitine and acylcarnitines in biological fluids and application to a clinical study*. PhD thesis, Universität Basel, 2005.

[134] Siegfried Wagner, Thomas Deufel, and Walter G Guder C. Carnitine Metabolism in Isolated Rat Kidney Cortex Tubules Summary : Renal carnitine metabolism was studied in isolated kidney cortex tubules from fed rats . The tubular distribution of free carni- in AcC was paralleled by a decrease in C , result- ing in. *Metabolism Clinical And Experimental*, 367(January):75–79, 1986.

[135] Marianne C Walsh, Lorraine Brennan, J Paul G Malthouse, Helen M Roche, and Michael J Gibney. Effect of acute dietary standardization on the urinary , plasma , and salivary metabolomic profiles of healthy humans. *American journal of clinical nutrition*, 84(3), 2006.

[136] Y C Wang and C Li. Evolutionarily conserved protein arginine methyltransferases in non-mammalian animal systems. *The FEBS journal*, 279(6):932–945, 2012.

[137] M Watanabe, D Watanabe, and S Kondo. Polyamine sensitivity of gap junctions is required for skin pattern formation in zebrafish. *Scientific reports*, 2:473, 2012.

[138] Waters. MassLynx, 2013.

[139] Wolfram Weckwerth and Katja Morgenthal. Metabolomics: from pattern recognition to biological interpretation. *Drug discovery today*, 10(22):1551–8, 2005.

[140] Gerald Weissmann. Monumental revolutions: scientific, sanitary and 'omic. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 23(11):3639–43, November 2009.

[141] Marit J. van der Werf, Renger H. Jellema, and Thomas Hankemeier. Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *Journal of Industrial Microbiology and Biotechnology*, 32(6):234–252, 2005.

[142] Johan a. Westerhuis, Huub C. J. Hoefsloot, Suzanne Smit, Daniel J. Vis, Age K. Smilde, Ewoud J. J. Velzen, John P. M. Duijnhoven, and Ferdi a. Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, January 2008.

[143] William R Wikoff, Andrew T Anfora, Jun Liu, Peter G Schultz, Scott a Lesley, Eric C Peters, and Gary Siuzdak. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3698–703, March 2009.

[144] S. Wopereis, C. Rubingh, M. van Erk, E.R. Verheij, T. van Vliet, N.H.P Cnubben, A.K. Smilde, J. van der Greef, B. van Ommen, and H.F.J Hendriks. Metabolic profiling of the response to an oral glucose tolerance test detects subtle metabolic changes. *PLoS ONE*, 4(2):e4525, February 2009.

[145] Feng Xian, Christopher L Hendrickson, and Alan G Marshall. High resolution mass spectrometry. *Analytical chemistry*, 84(2):708–19, January 2012.

[146] Jie Zhang, Lijuan Yan, Wengui Chen, Lin Lin, Xiuyu Song, Xiaomei Yan, Wei Hang, and Benli Huang. Analytica Chimica Acta Metabonomics research of diabetic nephropathy and type 2 diabetes mellitus based on UPLC oaTOF-MS system. *Analytica Chimica Acta*, 650:16–22, 2009.

[147] Zhen Zhang and Daniel W Chan. Cancer Proteomics : In Pursuit of True Biomarker Discovery. *Biomarkers*, 14(October):2283–2286, 2005.

8

SAMENVATTING

Daar waar genomics zich bezighoudt met het genoom, proteomics met het proteoom, houdt men zich in het vakgebied metabolomics bezig met het metaboloom. Het metaboloom beschrijft het geheel van metabolieten; tussen- of eindproducten die ontstaan nadat een chemische stof in een biologisch systeem een chemische omzetting (stoffwisseling) heeft ondergaan. Metabolieten worden gekenmerkt door hun (kleine) molecuulgrootte. De diversiteit aan chemische eigenschappen van de metabolieten is essentieel voor de rol die metabolieten spelen bij de verschillende biologische reacties in een biologisch systeem. Karakterisatie en biologische interpretatie van het metaboloom zijn daardoor een zeer grote uitdaging.

Vanwege de gevoeligheid en het vermogen om veel verschillende chemische verbindingen te kunnen detecteren is massa spectrometrie een detectie methode die uitermate geschikt en veel gebruikt wordt binnen metabolomics. Met een massa spectrometer (MS) is het mogelijk om op basis van de massa en intensiteit van karakteristike ionen verkregen van een molecuul de verschillende verbindingen in het metaboloom te identificeren en kwantificeren. Wanneer deze methode vooraf gegaan wordt door een scheidingsmethode (bijv. gas chromatografie, liquid chromatografie (LC) of capillaire elektroforese kunnen verbindingen met gelijke massas van karakteristieke ionen op basis van (andere) chemische eigenschappen toch van elkaar gescheiden worden.

Het nadeel van het gebruik van een massa spectrometer is dat de kwantificatie van de metabolieten zeer gevoelig is voor de sterk variërende samenstelling van de te meten monsters (maar ook voor verschillen in experimentele meetcondities (bijv. de temperatuur tijdens het meten). Om toch de concentratie van verbindingen in een monster betrouwbaar te kunnen bepalen worden één of meerdere referentie componenten met bekende concentraties toegevoegd. Dit referentie component is idealiter een natuurlijk isotoop (^{13}C of ^{15}N gelabeld) van de component in kwestie die zich in chemisch opzicht gelijk zal gedragen als de ongelabelde component maar op basis van hun verschillende massa onderscheiden kan worden. Om metingen tussen verschillende apparaten te kunnen vergelijken worden referentie monsters, d.w.z. monsters van gelijke (en soms volledig bekende) samenstelling, gebruikt.

Het aantal gedetecteerde unieke massa's voor één enkel monster varieert maar detectie van meer dan 10.000 massa's is niet uitzonderlijk. Het absoluut kwantificeren van zoveel componenten door het toevoegen van de juiste natuurlijke isotopen gelabelde standarden zou niet alleen praktisch een onhaalbare situatie opleveren maar ook erg prijzig zijn. Bovendien is de identiteit van alle unieke massas op voorhand vaak niet bekend. In dit geval wordt de concentratie van geïdentificeerde massas vaak uitgedrukt in relatieve concen-

traties t.o.v. een referentie component die representatief is voor een bepaalde chemische klasse van verbindingen. De focus op kwantificering (of detectie) van bekende componenten wordt targeted genoemd. Vanwege de grote chemische verscheidenheid aan mogelijke metabolieten zijn er veel verschillende analytische platforms nodig om een zo volledig mogelijk metaboloom te meten. Hoewel targeted de meest wijd verspreide methode is binnen metabolomics, sluit deze aanpak het ontdekken van nieuwe metabolieten juist uit. Voor onderzoek naar nieuwe, nog onbekende metabolieten wordt de untargeted methode gebruikt. In dit geval is er in het algemeen geen sprake van een voorkeur voor specifieke componenten en wordt de groep van te detecteren metabolieten juist zo groot mogelijk gehouden. Dit gaat dat vaak ten koste van het vermogen om de metabolieten goed te kunnen kwantificeren.

In dit proefschrift wordt de ontwikkeling van methoden en concepten beschreven om zowel kwalitatieve als kwantitatieve metabolomics gegevens te extraheren uit untargeted massa spectra data verkregen van biologische monsters. De biologische monsters kunnen variëren van kleine individuele cellen tot clusters van cellen of een coupe van weefsel of verschillende lichaamsvloeistoffen zoals urine, bloed of cerebrospinale vloeistof. Een algemene inleiding van dit onderzoek is gegeven in hoofdstuk 1.

In hoofdstuk 2 wordt beschreven hoe de kwantificatie van de metabolieten significant verbeterd kan worden wanneer naast de te meten biologische monsters bepaalde referentie monsters worden meegenomen tijdens de analyse. Systematische variaties die optreden tijdens analyses van grote series van monsters (die soms weken tot wel maanden kunnen duren) kunnen niet alleen in kaart worden gebracht maar ook worden gecorrigeerd door slim gebruik te maken van deze referentie monsters.

In hoofdstuk 3 tonen we aan dat multivariate statistiek gebruikt kan worden om predictieve modellen te maken van untargeted metabolomics data. In dit geval beschrijft het model de kans, op basis van het metabole profiel van de gemeten urine, dat de persoon in kwestie een progressieve vorm van nierfalen heeft. De multivariate modellen bevatten informatie over de metabolieten die ook in de univariate modellen als statistisch significant naar voren komen maar ook metabolieten die univariaat juist niet significant bleken te zijn. Het valideren van de multivariate modellen is echter niet eenvoudig en kan zeer tijdrovend zijn maar door de ontwikkelde parallelle implementatie van het predictie model en gebruikmakend van de supercomputer faciliteiten bij SARA niet onmogelijk.

De kracht van een snelle untargeted metabole vingerprint wordt aangetoond in hoofdstuk 4. Door het ontbreken van een scheidingsmethode is de identificatie alleen gebaseerd op massa (spectra). In dit hoofdstuk worden massa spectra van zebravisjes in verschillende beginstadia van hun ontwikkeling na bevruchting met elkaar vergeleken. In dit hoofdstuk laten we zien hoe en dat de verschillende embryonale stadia van elkaar onderscheiden kunnen worden d.m.v. metabolomics data.

Het feit dat targeted methoden in metabolomics wijdverbreid zijn is onder andere het gevolg van het ontbreken van automatische methoden om untargeted data te kunnen verwerken. In hoofdstuk 5 wordt een nieuwe methode geïntroduceerd waarmee hoge resolutie LC-MS data geïntegreerd kan worden

op een volledig automatische manier. De resultaten van deze nieuwe methode komen sterk overeen met de huidige gebruikte semi-automatische standaard (targeted) aanpak. Naast de subset van bekende targets levert de nieuwe methode echter ook een scala aan onbekende componenten op die ook biologisch relevante informatie bevat.

In de hiervoor genoemde hoofdstukken is significante progressie beschreven in het verwerken van untargeted, op MS gebaseerde, metabolomics data maar tegelijkertijd wordt ook duidelijk dat er nog veel verbeterd kan worden. In hoofdstuk 6 staan algemene conclusies en perspectieven voor toekomstig onderzoek.

.

9

DANKWOORD

Dit proefschrift was nooit tot stand komen zonder de hulp van anderen. Daarom wil ik van de gelegenheid gebruik maken een aantal mensen te bedanken.

Allereerst wil ik iedereen van de afdeling Analytical Biosciences bedanken voor de inspirerende en gezellige sfeer. De bereidheid om elkaar waar mogelijk te helpen heb ik altijd erg gewaardeerd. Thomas, je brede kennis, enthousiasme en energie hebben me altijd zeer gemotiveerd. Theo, je hebt me altijd eerst zelf ideeën laten uitwerken, maar je hebt gelukkig op tijd ook bijgestuurd. Mijn kamergenoten Adrie en Margriet wil ik bedanken voor hun samenwerking maar zeker ook voor het delen van hun expertise. Voor het toelichten van de voor mij in het begin soms bijna onbegrijpelijke massa spectra wil ik Jorne bedanken. Robert-Jan wil ik bedanken voor de uitdaging om directe infusie data te verwerken. Gerwin wil ik bedanken voor zijn inzet bij alle experimenten die ik bedacht had. Jammer genoeg hebben we ze nooit allemaal kunnen uitvoeren. Loes, Bea en Coby, ik wil jullie bedanken voor het op zich nemen van alle tijdrovende en lastige administratieve taken.

Voor de tijd bij TNO wil ik al mijn toenmalige collega's bedanken en in het bijzonder Ivana, Elwin en Renger. Het QC paper was een zet in de goede richting.

Tenslotte wil ik mijn familie en in het bijzonder mijn ouders bedanken die mij hebben geleerd mijn eigen weg te kiezen en voor mij klaar stonden wanneer ik ze nodig had.

Ik heb geprobeerd zo volledig mogelijk te zijn, maar er zal altijd iemand zijn die ik vergeet. Bij deze alsnog bedankt.

10

CURRICULUM VITAE

Frans Meindert van der Kloet (1970) was born and grew up in Joure, Friesland, The Netherlands. He obtained his bachelor at the Noordelijke Hogeschool (NHL) te Leeuwarden specializing in analytical chemistry. During this period he developed an interest in computers and he continued his study to obtain his masters in Computational Chemistry at the Radboud University in Nijmegen. After some part-time jobs he started his first full-time job (1997-1999) at the Radboud University and was responsible for implementing an IT-environment to facilitate remote learning for students at the chemistry faculty. He then started working for a small marketing consultancy company (BrandmarC) where he was responsible for the development of statistical tools to analyze and quantify marketing data for customers like P&G, Philips and Unilver (1999-2006). He then became project assistant at TNO Quality of Life in Zeist where he was involved in integrating and interpreting data from biological origin using mass spectrometry or near infrared spectra. Inspired by his work at TNO he decided to return to university, and started in 2009 as biostatistician at the Division of Analytical BioSciences at Leiden University as part of a large European project (FinnDiaNa). In 2010 he started to work on his PhD research to develop better methods for quantification in metabolomics. This was a collaborative project within the Netherlands Metabolomics Centre. As from June 2014 Frans will be working as a post-doctoral researcher at the University of Amsterdam (UvA) under supervision of Dr. J.A. Westerhuis.

11

PUBLICATION LIST

Frans M van der Kloet, Ivana Bobeldijk, Elwin R Verheij, and Renger H Jellema. Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. *Journal of Proteome Research*, 2009.

Harmen H M Draisma, Theo H Reijmers, Frans van der Kloet, Ivana Bobeldijk-Pastorova, Elly Spies-Faber, Jack T W E Vogels, Jacqueline J Meulman, Dorret I Boomsma, Jan van der Greef, Thomas Hankemeier, Frans Van Der Kloet, and Jan Van Der Greef. Equating, or correction for between-block effects with application to body fluid LC-MS and NMR metabolomics data sets. *Analytical chemistry*, 82(3):1039–46, February 2010.

Maud M. Koek, Frans M. Kloet van der, Robert Kleemann, Teake Kooistra, Elwin R. Verheij, and Thomas Hankemeier. Semi-automated non-target processing in GC GCMS metabolomics analysis: applicability for biomedical studies. *Metabolomics*, 7(1):1–14, July 2010.

F. M. van der Kloet, F. W. a. Tempels, N. Ismail, R. Heijden, P. T. Kasper, M. Rojas-Cherto, R. Doorn, G. Spijksma, M. Koek, J. Greef, V. P. Mäkinen, C. Forsblom, H. Holthöfer, P. H. Groop, T. H. Reijmers, and T. Hankemeier. Discovery of early-stage biomarkers for diabetic kidney disease using ms-based metabolomics (FinnDiane study). *Metabolomics*, pages 109–119, February 2011.

Robert-Jan Raterink, Frans Meindert van der Kloet, Jiajie Li, Niels Abraham Wattel, Marcel Johannes Maria Schaaf, Herman Peter Spaink, Ruud Berger, Robert Jan Vreeken, and Thomas Hankemeier. Rapid metabolic screening of early zebrafish embryogenesis based on direct infusion-nanoESI-FTMS. *Metabolomics*, January 2013.

Vanessa Gonzalez-Covarrubias, Marian Beekman, Hae-Won Uh, Adrie Dane, Jorne Troost, Iryna Paliukhovich, Frans M van der Kloet, Jeanine Houwing-Duistermaat, Rob J Vreeken, Thomas Hankemeier, and Eline P Slagboom. Lipidomics of familial longevity. *Aging cell*, 12(3):426–34, 2013.

Frans M van der Kloet, Margriet Hendriks, Thomas Hankemeier, and Theo Reijmers. A new approach to untargeted integration of high resolution liquid chromatography-mass spectrometry data. *Analytica chimica acta*, 801:34–42, 2013.