Institute of Pharmacology and Toxycology, Vetsuisse Faculty,
University of Zürich

Director: Professor Dr. Felix Althaus

PROFILING OF BOAR TAINT BY NONTARGETED METABOLOMICS

Inaugural Dissertation

to be awarded the Doctoral Degree of the Vetsuisse Faculty, University of Zürich

submitted by

Malin Emelie Maria Olson

Veterinarian from Sweden

Approved by the Vetsuisse Faculty as inaugural dissertation on proposal from

Prof. Dr. Hanspeter Nägeli, Referee

Prof. Dr. Paul Torgerson, Co-referee

Zurich 2013

Table of content

Abstract

In many countries, male piglets are castrated shortly after birth to avoid the production of meat with an unpleasant smell and flavor known as boar taint. Extensive research has been carried out during the last 40 years to delineate compounds that are responsible for this problem. The most frequently candidates are androstenone, skatole and indole. However, other factors must be involved in causing boar taint, since a significant proportion of tainted pigs have unchanged levels of these three compounds. The aim of this thesis was to establish the conditions for a non-targeted metabolomics study and thereby identify new potential biomarkers that correlate with the appearance of boar taint. The adipose tissue of 16 nontainted and 17 strongly tainted pigs, selected by an earlier sensory panel analysis, was homogenized with methanol. After solid-phase extraction, the samples were analyzed by liquid chromatography coupled to a time-of-flight mass spectrometer using a nanoUPLC®-ESI-QTOF-HDMS™ system. By monitoring about 20'000 different masses with an accuracy of around 5 ppm, we found a metabolic pattern that is characteristic for the appearance of boar taint. A set of 16 masses can discriminate between tainted and non-tainted carcasses with a mean predictive accuracy of 90%. These results will be used to further develop a reliable test for the rapid detection of boar-tainted meat.

Abbreviations

2CID	2-Carboxyindole	MeOH	Methanol
ACN	Acetonitrile	MF	Molecular Formula
ALP	Agroscope Liebefeld-Posieux Research Station	MM	Molecular Mass
AND	Androstenone	MS	Mass Spectrometry
BIP	Base Peak Chromatogram	MS/MS	Tandem Mass Spectrometry
BW	Body Weight	MS1	Low Energy Mass Spectrometry
CV	Coefficient of Variation	MS2	High Energy Mass Spectrometry
D-crit	Maximum tolerable distance for the data	MS^E	MS/MS altering between high and low energy without precursor filtering
DDA™	Data Dependent Acquisition Mode	MSI	The Metabolomic Standard Initiative
DModX	Distance to the Model X	MTBE	Methyl-tert-butyl ether
EC	Elemental Composition	**n**	Nontainted pig
ESI	Electro Spray Ionization	NB	Naïve Bayes
FT-ICR	Fourier Transform Ion Cyclotron	NIST	National Institute For Standard and Technology
FDA	Food and Drug Administration	NMR	Nuclear Magnetic Resonance Spectroscopy
GC	Gas Chromatography	OPLS-DA	Orthogonal Projection on Latent Structure Discriminant Analysis
Glu-Fib	[Glu1]-Fibrinogen peptide B Human	PC	Principal Component
HDMS	High Definition Mass Spectrometry	PCA	Principal Component Analysis
HILIC	Hydrophilic Interaction Chromatography	PLS-DA	Partial Least Squares Discriminant Analysis
HPLC	High Performance Liquid Chromatography	Q	Quadrupole
i.e.	that is	QC	Quality Control
ID	Indole	QTOF	Quadrupole Time Of Flight
IT	Ion Trap	ROC	Receiver Operator Characteristic
KeyMix	AND+SK+ID 10ng/µl and 2CID 50 ng/µl in MEOH	RT	Retention Time
LC	Liquid Chromatography	**s**	Strong tainted pig
LD	M. longissimus dorsi	SD	Standard Deviation
Leu-Enk	Leucine Enkephaline	SK	Skatole
LIT	Linear Ion Trap	TOF	Time Of Flight
m/z	Mass to Charge Ratio	UPLC	Ultra Performance Liquid Chromatography
$\bar{x}$	Mean	XIC	Extracted Ion Chromatogram

Introduction

I. Boar Taint

An Encyclopedia (The Veenman's Agrarische Winkler Prins) from 1954 declares that "Since boars' meat is less tasty, the young boars which are intended for fattening are always castrated". The first attempt to identify the compound(s) responsible for this unpleasant taste was made by Craig and Pearson, 1959. Following this much research has been undertaken investigate the increase of unpleasant odors and flavors in some male pigs, which is now known as boar taint (see the reviews of Bonneau, 1982; Brooks and Pearson, 1986; Claus et. al., 1994). Patterson (1968) identified the sexual hormone 5α-androst-16-en-3one (androstenone) as a compound responsible for the urine-like odor associated with boar taint. Two years later 3-methylindole (skatole) and indole, which are produced from tryptophan by bacteria in the intestines (Yokoyama and Carlson, 1979; Wilkins, 1990; Deslandes et. al., 2001) were identified as additional contributors by Walstra and Maarse (1970) and Vold (1970). In addition to these findings, there are other compounds suggested to contribute to boar taint (Xue and Dial, 1997; Rius et. al., 2005) but none of them could be corroborated over time. Until today androstenone, skatole and indole are still seen as the main compounds. However, there are indications that some other factors could be involved in causing boar taint (Annor-Frempong et. al., 1998; de Kook et. al., 2001; Ampuereo and Bee, 2006; Pauly et.al., 2010). Indeed, the level of androstenone, skatole and indole correlates badly with results from classical sensory panels (Bonneau et al., 2000; Rius et. al, 2005; Ampuero and Bee, 2006). The correlation coefficient between skatole levels and the appearance of boar taint determined by a sensory panel is on order of 0.7, accounting for only 50% of the total score (Bejerholm and Barton-Gade, 1993). If androstenone content is included, about 66% of odor score can be accounted for. The magnitude of these coefficients does not exclude the contribution of other compounds.

In most European countries, male piglets intended for fattening are still castrated during their first days of life. This is to avoid the development of off-odor, as the incidence of boar taint in entire males is ranging from 18% to 64% at usual slaughter weights (Williams et al., 1963; Desmoulin et al., 1971; Malmfors and Hansson, 1974; Bonneau et. al., 2000). Numerous studies have established the advantages associated with the production of entire males (reviewed by Walstra, 1974; Walstra and Vermeer, 1993; Desmoulin et. al., 1990). One of these advantages is the substantially lower production cost for entire males than for castrates. The costs involved in performing castration are averted and possible animal losses and temporary decrease in performance following castration are avoided. Boars may also grow faster than castrates (Walstra and Kroeske, 1968; Fowler et al., 1981; Andersson et al., 1997). The smaller development of adipose tissue is another important advantage associated with entire male pigs (Prescott and Lamming, 1967; Fortin *et al.*, 1983; Hansen and Lewis, 1993). Therefore meat cuts from entire males are more appealing to the consumer (Babol and Squires, 1995) and achieve a higher grading result for the carcasses (Andersson *et al.*, 1997). This leads to a higher

income for the farmer. From the point of sustainable agriculture, entire males would also be beneficial because they require less feed per gained kg of bodyweight and have a better efficiency leading to a decreased output of nitrogen in the manure (Desmoulin et al., 1971).

With growing concern on animal welfare and aims for more sustainable production of food, a European declaration on alternatives to surgical castration of pigs was released on the 16th of December in 2010 (http://ec.europa.eu/food/animal/welfare/farm/initiatives_en.htm). Representatives of European farmers, meat industry, retailers, scientists, veterinarians and animal welfare non-governmental organizations committed to a plan to voluntarily end surgical castration of pigs in Europe by 1 January 2018. The great obstacle, however, is that meat of non-castrated boars is not popular among retailers and in international trade. Because of fear of boar taint causing consumers complains and decreasing trading profits, considerably less is paid for boars' meat. Some importers will not accept any boars' meat. As an important precondition to increase the frequency of young boar fattening, the reduction of boar taint and the development of an "on the slaughter line"-detector for tainted carcasses would be desirable. The detection of the already known compounds androstenone, skatole and indole alone will not be sufficient to meet the markets demands of a method with predictive accuracy as close to 100% as possible (Bonneau et al., 2000). With the innovative approach of metabolomics we intend to contribute with new knowledge to this area.

II. Metabolomics

Metabolomics, also known as metabonomics, is concerned with the study of low molecular weight compounds in biological samples and other complex matrixes (Nicholson et. al. 1999, Fiehn, 2002). These compounds (typically <1500 Da) make up what has been termed the "metabolome". The concept of the metabolome as the "total complement of metabolites in a cell" (Tweeddale et al., 1998) has since been broadened to include not only endogenous small molecules but also those introduced and modified by diet, environmental exposure, and coexisting organisms (Dunn, 2008). Metabolomics plays an important role in systems biology. It enables better understanding of complex interactions in biological systems, and the idea is to look for changes in metabolic activity. Metabolite amounts can change for multiple reasons. A down-regulated pathway might produce less of a particular metabolite; an up-regulated pathway might consume more. The influence of diet and environmental effects as well as genetic-factors will also affect the metabolome (Vigneau-Callahan et. al., 2001, Poste, 2011).

There are many approaches for metabolomics. They can be roughly classified according to data quality and number of metabolites that can be detected. Firstly, the "targeted metabolite analysis" or "targeted metabolomics" which refers to the detection and precise quantification of a single or small set of chemically defined compounds. Second, the "metabolic profiling" provides the identification and approximate quantification of a group of metabolites. Third is "metabolite fingerprinting" or "nontargeted metabolomics", it is used for complete

metabolome comparisons without knowledge of the compounds investigated, therefore metabolite identification is not mandatory (Krastanov, 2010).

The definition of nontargeted metabolomics can also be used when analyzing panels such as lipids, including phospholipids (i.e., lipidomics, Wenk, 2005; Castro-Perez, 2010; Bicalho, 2008), amino acids (Paik et. al., 2008; Wei, 2010), sugars (Wei, 2010), bile acids (Bobeldijk, 2008) or small molecules (Baker, 2011; Neumann, 2010). It aims to gather information on as many metabolites as possible by taking into account all information present in the data sets producing detailed information on the relative abundance of thousands of mass signals representing hundreds of metabolites. Subsequent statistics and bioinformatics tools can then be used to provide a detailed view on the differences and similarities between samples/groups of samples or to link metabolomics data to other systems biology information, genetic markers and/or specific quality parameters (Moco, 2007). The best practice stated by Poste 2011, is to test for multiple biomarkers instead of just looking for one or two candidates. He termed this approach multiplex profiling. Based on the questions asked, metabolites are selected for analysis and specific analytical methods are developed for their determination.

The presented metabolomics study is a comparative analysis of samples to inquire, "Can these samples be distinguished on the basis of their qualitative and quantitative endogenous chemical composition?" To answer this question, one must identify those differences that are a direct result of the alteration and distinguish them from normal biological variability. Ideally, differences between samples from within the same group (control or altered) will be smaller, or at least not the same as differences across groups (control versus altered). Therefore, it is important to minimize any artificially introduced variability in the samples at any step of the experiment. Following standardized and minimalistic protocols typically facilitates this. Careful consideration must be made of various parameters. The number of individuals/subjects per group, since these impact the determination of statistical significance. The diet, environment exposure, sample collection and sample storage, are also components that need to be supervised since these can directly affect the composition of the metabolome (Robertson, 2005; Scalbert, 2009).

III. Food science

Metabolomics has recently found its place in food science as reviewed by Wishart, 2008 and Cevallos-Cevallo et. al., 2009. One of the main applications is food safety or quality improvement where the analyst wants to correlate a specific property, for example taste, origin or age, to metabolite patterns using biostatistics. Taking taste as an example, the main goal is often to understand the taste of food in terms of chemical composition and physical properties or to find biomarkers that can be measured routinely and easier than the taste itself. A typical workflow of a food quality metabolomics experiment is shown in *Figure 1.* Every aspect of this workflow has to be optimized in order to make metabolomics studies a success. With respect to the analytical chemistry involved in metabolomics, sample workup, analysis of sample and data preprocessing are items that have to be dealt with.

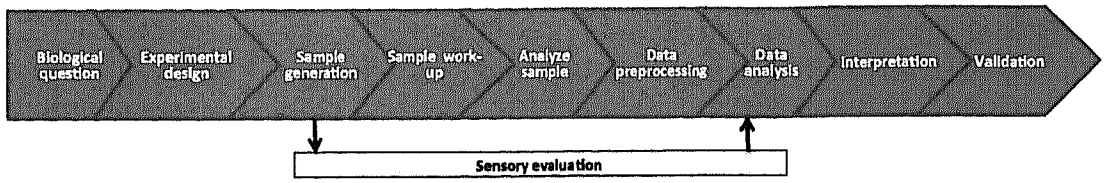


Figure 1: *Schematic diagram of the different steps in a food metabolomics*

IV. Mass spectrometry

Currently two analytical platforms are mainly being used for metabolomics applications (Scalbert, 2009), namely nuclear magnetic resonance spectroscopy (NMR) (Nicholson and Wilson, 2003; Lenz and Wilson, 2007) and mass spectrometry (MS) (Dunn, 2008; Koulman et.al., 2009). These techniques are used with direct infusion or are coupled to separation techniques. The advantages and disadvantages of these analytical approaches are listed in *Table 1*.

Table 1: *Comparison of different metabolomics technologies*

Analyzers	Advantages	Disadvantages	Technical properties
NMR	Quantitative, fast, non-destructive, minimal sample preparation, derivatization and separation not necessary, robust, allows identification of novel chemicals, good within and across lab reproducibility, easy maintenance	Not sensitive, very expensive, huge instrument specific alteration of data, requires at least 0.5 µL sample	Sensitivity dependent on the presence of a magnetically-susceptible nuclide (There are >5000 resolved lines in single pulse 750 or 800 MHz NMR spectra)
MS	Excellent sensitivity and resolution, very flexible technology, detects most organic and inorganic compounds, minimal sample size, has potential for detecting the largest portion of the metabolome, fast scan rate, data analysis automation	Sample not recoverable, not very quantitative, slow, less robust than NMR, novel compound identification difficult, molecules unable to ionize can not detectable	High resolution (~20.000), mass accuracy <5 ppm with internal calibration

NMR has the potential for high-throughput fingerprinting, as requires minimal sample preparation, and it is a non-discriminating and non-destructive technique. However, only medium to high abundance metabolites will be detected with this approach. Additionally, identification of individual metabolites based on chemical shift signals, which are causing sample clustering in multivariate analysis is challenging in complex mixtures. Also the costly price makes the accessibility of this technology highly limited (Lenz and Wilson, 2007). MS-based analysis offers relative quantitative measurements with high selectivity and sensitivity and the potential to identify compounds. With the many different MS detectors available on the market (differing in price,

resolution and accuracy), this analytical platform combines the most important features for a successful metabolomics study (Dettmer et al., 2007).

The linear Time Of Flight (TOF) is a method of mass spectrometry in which an ion's mass-to-charge ratio (m/z) is determined via a time measurement, with virtually unlimited mass range. TOF instruments offer high resolution, fast scanning capabilities (milliseconds), and accuracy on the order of 5 part per million (ppm) with internal calibration. The hybrid Quadrupole-Time Of Flight (Q-TOF) mass spectrometers combine the filtering ability, efficient transmission from low to high mass and stability of a Quadrupole (Q) analyzer with the high efficiency, sensitivity, and accuracy of a time-of-flight mass analyzer. Q-TOF mass analyzers are an obvious choice for obtaining metabolite MS and MS/MS data. The quadrupole can act as any simple quadrupole analyzer to scan across a specified m/z range, but can also be used to selectively isolate a precursor ion and direct that ion into the collision cell. Q-TOF analyzers offer significantly higher sensitivity and accuracy over tandem quadrupole (Q-Q-Q) instruments when acquiring full fragment mass spectra. Other types of mass analyzers are listed in *Table 2* summering the reviews Domon and Aebershold, 2006 and Lenz and Wilson, 2007.

Table 2: *Types of mass analyzers*

1)	Time-of-Flight (TOF)
2)	Quadrupole (Q)
3)	Ion traps (IT)/Linear ion trap (LIT)
4)	Fourier transform ion cyclotron (FT-ICR)
5)	Sector
a.	Magnetic
b.	Electric
6)	Tandem and hybrids (Q-TOF, Q-Q-Q, Q-LIT, TOF-TOF)

With tandem MS instruments, it is possible to acquire second order mass spectra (MS/MS) in data-dependent-acquisition mode (DDA™). This method obviates the need to analyze the sample in MS mode to identify the target precursor ions and then re-run the sample in MS/MS mode. The technique is particularly valuable in the analysis of unknown samples using on-line chromatography where the target precursor ions and their retention times may be different for each sample. When acquiring data with DDA™, the MS instrument switches from full-scan MS mode to full-scan MS/MS mode for any mass rising above a predefined threshold. However, DDA results both in a loss of data in the MS mode when MS/MS data are being acquired and in poor duty cycles, thus making it less than ideal for fast analysis and narrow, rapidly eluting, peaks. Both of these approaches are therefore perhaps less efficient than would be desired for the rapid analysis of complex multicomponent samples. A different approach is the acquisition of tandem mass spectra, by alternating between low and high collision energies without any precursor mass filtering, termed MS^E (Plumb et. al., 2006). Applying low collision energy in the collision cell, precursor ion information can be obtained, and with high collision energy full-scan accurate mass fragment, precursor ion and neutral loss (loss of an uncharged fragment from a molecule) information can be acquired.

V. Liquid chromatography

Due to the complex nature of biological samples, chromatographic separation is often performed before MS analysis to achieve the detection of as many metabolites as possible. Traditionally, gas chromatography (GC) was employed, as it is well known for high resolution and reproducibility. However, disadvantages of GC include cumbersome sample preparation (such as

derivatization), lengthy analysis time, and limitations as to size and type of molecules that can be separated (nonvolatile, polar- and semi-polar molecules and macromolecules are unsuitable). However, GC-MS is still widely used in plant metabolomics due in part to the nature of the metabolites being investigated (Jonsson el. al., 2004). Liquid chromatography coupled to mass spectrometry (LC-MS) has in the last couple of years become a very popular alternative. On the basis of coverage, ease-of-use, robustness to matrix and robustness in routine operation, LC was identified as the optimal platform for metabolomics experiments (Buescher et. al., 2009). With the separation of molecules using LC, a decrease of the number of competing analytes entering the mass spectrometer is achieved and this reduces ion suppression (Gangl et. al., 2001). The result of this complexity reduction is a selective approach that allows for both relative quantitation and structural elucidation, whereby sensitivities in the pg/mL range can be achieved (Plumb et. al., 2004).

Metabolite extracts contain a diversity of small molecules that differ in their physical chemical properties like size, polarity/hydrophobicity and charge. An important factor is therefore the choice of the separation column. While various LC methods are described in the literature, the most robust LC approach to small molecule separation is reversed-phase (RP) chromatography using a nonpolar stationary phase, for example C_{18} RP-LC (Idborg et.al. (1), 2005). Gradients begin with high water content, gradually adding methanol or acetonitrile to elute hydrophobic compounds. Polar molecules elute earlier and nonpolar molecules later. Although C_{18} RP-LC is a good starting point for metabolome analysis (Trauger et. al., 2008), many polar metabolites do not retain adequately, thus eluting near or within the void volume during the beginning of a chromatographic run. Another approach is to enhance retention of polar analytes using an ion-pairing agent as described by Buescher et. al., 2010. An ion-pairing reagent, for example 3-tributylamin, is a volatile charged compound that pairs with oppositely charged analytes in solution, resulting in an ion–ion complex. The ion-pairing reagent contains hydrophobic moieties that enhance binding of the ion–ion complex to the C_{18} column and is typically added to the aqueous mobile phase. Unfortunately, routine operation drastically increases maintenance and cleaning of the syringes, tubing and fittings of the liquid chromatograph to remove the contaminations caused by of the ion-pairing reagent. In addition the initial stages of the mass spectrometer are also contaminated. Therefore, such experiments can only be done on an instrument dedicated just to ion-pairing. An interesting alternative is hydrophilic interaction liquid chromatography (HILIC) as presented by Tolstikov and Fiehn, 2002. They used this approach for the analysis of highly polar compounds in plant extracts. HILIC shows a good flexibility but is still lacking robustness and reproducibility with respect to retention times (Idaborg et. al. (1), 2005).

The ability of LC to separate complex mixtures prior to mass analysis comes at a cost of speed. An alternative to traditional high performance liquid chromatography (HPLC) approaches is ultrahigh performance liquid chromatography (UPLC), which utilizes columns with much smaller particle size packing material (1.4-1.8 μm) than traditional columns (3-5 μm), thus improving separation and resolution. This technology permits pumping and injection of liquids at pressures exceeding 10.000psi (Wilson et. al., 2005; Swartz, 2005). With UPLC, narrower chromatographic peaks can be achieved (peak widths at half-height <1 s), resulting in increased peak capacity, lower ion suppression and improved signal-to-noise ratio, and thus increased sensitivity. Recent studies comparing UPLC and HPLC for their application to metabolomics studies showed that UPLC can detect more components than HPLC (Plumb et. al., 2004; Wilson et. al., 2005) with a 20% increase reported over the same chromatographic length. These studies also showed UPLC to display superior retention time reproducibility and signal-to-noise ratios over HPLC. *Figure 2* shows a state of the art UPLC-MS system.

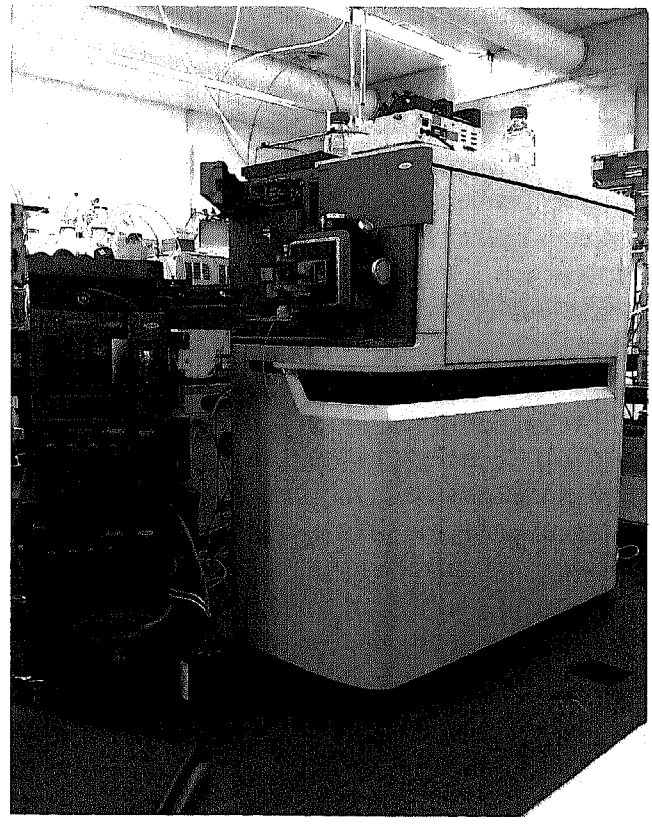


Figure 2: *Waters nanoAqcuity® and Synapt G2® (nanoUPLC-ESI-QTOF-HDMS)*

VI. Electro spray ionization

A prerequisite for a mass spectrometry analysis is that the molecule is presented as an ion (preferable as the protonated molecular ion $[M+H]^+$ in positive ionization mode). During the past 20 years, electro spray ionization (ESI) has grown to be the most popular ionization technique, whereby a strong voltage is applied to the liquid stream exiting the tip of a needle. This seemingly simple method enables efficient conversion of charged molecules from the liquid phase into gas-phase ions. The physical mechanisms of ESI remain only partially understood. Charged droplets are initially produced by electrostatic dispersion when liquid emerges from the tip of a metal needle. Solvent then evaporates from the charged droplets. As the droplets become smaller and smaller the ions within them repel due to columbic forces, eventually resulting in release of gas-phase ions (Nguyen and Fenn, 2007). A major pitfall of ESI is the competitive nature of ion formation. If too many ions are present during their expulsion from the charged capillary, ion production will not increase linearly with concentration. This results in concentration underestimation. No undisputed

method eliminating "ion suppression" exists. Instead, one needs to determine the extent of ion suppression and correct for it. This is best done by using isotopic internal standards, which will experience identical ion suppression as the analyte of interest. Unfortunately for a typical nontargeted metabolomics experiment such isotopic standards are not available or would be too costly to produce. Therefore, an absolute quantification is not possible for this novel type of experiments.

VII. Raw data extraction

Techniques with high peak capacities such as UPLC-MS will still lead to partially co-eluting peaks when analyzing complex mixtures. Moreover, low abundant compounds may not be apparent by visual inspection of chromatograms. Detection of single components from complex chromatograms is therefore performed by peak picking and mathematical deconvolution algorithms. The extraction of valuable conclusions from the raw metabolomics data is as important as performing the analytical measurements. Raw data are usually stored in sample files as series of mass spectra acquired at a given time point or scan-number. Each of these scans represents pairs of mass, m/z, and intensity vectors, counts (see *Figure 3*). It is necessary to extract information about all

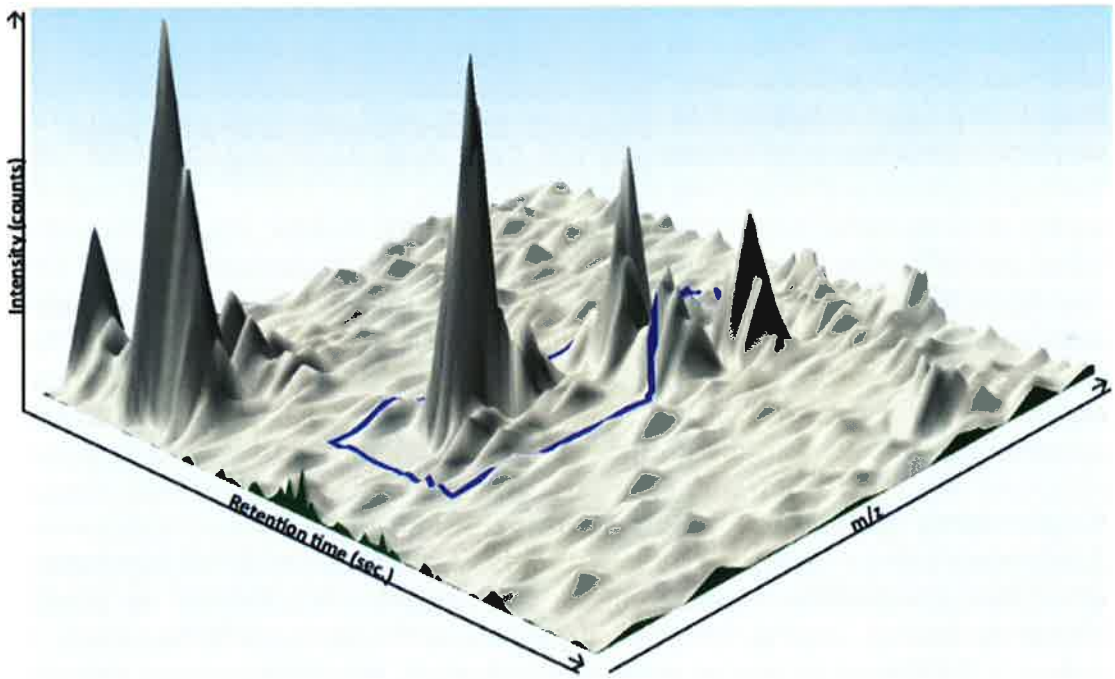


Figure 3: *Raw LC–MS data involves three dimensions including retention time, m/z, and signal intensity. The blue line is indicating co-eluting mass peaks.*

compounds, including mass and retention time (RT) as compound identifiers, and intensity as relative quantitative representation of concentration. This is followed by the combination of all data files to a uniform matrix to allow sample comparison with statistical tools.

Instrument providers have only recently become active in producing automated raw data extraction tools for metabolomics (e.g. Waters Corporation, Thermo Scientific, Agilent Technologies, Bruker Corporation) as Plug & Play Metabolomics Systems. Major disadvantages of these proprietary tools are: (1)

they only work for specific types of data and data formats and (2) they are black box systems (little is known about the underlying algorithms). At the same time more and more 3rd party software is becoming available for automated data extraction, for example GeneData (www.genedata.com), ACD-Labs (www.acdlabs.com), Rosetta (www.rosettabio.com), Non-Linear Dynamics (www.nonlinear.com) to name a few. These programs offer more flexibility with respect to the format of the input LC–MS data. This is a clear advantage when using instruments from different vendors. The "-omics" communities (proteomics, transcriptomics, metabolomics) have also been very active the last 15 years in developing their own tools, mainly because commercial software was/is not available and/or the poor performance of some of the available software. In part these open source data extraction tools are available free of charge on the web (XCMS, Smith et. al., 2006; OpenMS, Sturm et. al., 2008; MZmine, Katajamaa et. al., 2006; MetAlign, Tolstikov et. al., 2003 to name a few). They all appear to be inexpensive solutions to the data extraction problems. However, they all require training and good understanding of the software's construction and operations. Additionally, issues such as support and long-term continuity are not favorable. For an excellent overview and description of all these tools and software see Katajamaa and Oresic 2007.

The functionality and performance of all data analysis tools is a white spot on the metabolomics map, and it is widely known that all automated data extraction tools have their specific problems, which result in data with a variable amount of errors. Typical problems include:

- missed peaks
- wrongly binned signals
- integrated noise peaks
- misalignment
- integration errors

The fact that high quality raw data is being corrupted due to these errors is very alarming. Indeed, these errors become more and more dominant at low signal to noise ratios. All algorithms eventually stumble on classification problems such as: is this noise or signal? Is this peak A or B or neither of the two? Are these spectra the same? Unfortunately, many metabolites of interest happen to be present at low concentrations and with low signal to noise ratios. That's why such extraction errors have major detrimental effects on the outcome of metabolomics studies. Because of this, it is favorable to use different raw data extraction methods to be able to compare and get more confident in the results.

VIII. Chemometrics

In a LC-HDMS experiment, there are thousands of data entries per sample, complicated by a vast amount of noise, artifacts, and redundancy. In addition, the detection of minor but significant biomarkers among constitutive highly expressed compounds, a challenging analytical and statistical problem. Comparing samples has become a problem of high dimensionality (Weckwerth et. al., 2005) and chemometrics methods are needed to reduce this large number of variables. The goal is to obtain information-rich fingerprints suitable for pattern recognition and classification. Chemometrics can be broadly thought of as the application of mathematical and statistical methods to analytical

chemistry (Lavine and Workman, 2004). In the context of MS-based metabolomics, it includes any mathematical or statistical tools used for spectral processing, peak alignment, noise reduction, deconvolution, normalization and so on. However, chemical compounds are not generally identified, only their spectral patterns and intensities are recorded. Subsequently, they are statistically compared to identify relevant spectral features that are unique for sample classes (Nicholson et al., 2002; Trygg, Holmes & Lundstedt, 2007). Statistical comparison and feature identification technique usually involves unsupervised clustering, like principal component analysis (PCA). Another possibility is supervised classification like partial least squares discriminant analysis (PLS-DA) or orthogonal projection on latent structure discriminant analysis (OPLS-DA). PCA is often used for metabolomics (Choi et. al., 2004). Here it is used for the reduction of data dimensionality, to investigate a clustering trend, to detect outliers and to visualize data structures (Martens and Naes, 1991). However, PCA gives a crude representation of the information contained in spectra and cannot generally be used for additional information about the data, such as class information. Therefore, PCA is often followed by a supervised analysis technique such as PLS-DA or OPLS-DA. Lutz and colleagues showed by comparison of PCA with PLS-DA that there was a clear advantage in using a supervised model when class details are known (Lutz et. al., 2006). In such supervised two class classification cases, usually the values of the dependent variable are given 1 for one class, and 0 or -1 for the other class. A frequently used variant of PLS-DA is OPLS-DA. As here, the first components orthogonal to the dependent variable are removed from the data (Trygg and Wold, 2002). This gives a model with a single classification component (PC1). Variation that cannot be described by the first component will be described by a second principal component that is orthogonal to the class information. OPLS enhances the interpretation of PLS by forcing all classification information into a single component. The prediction power of both models is under particular conditions the same (Trygg and Wold, 2002).

The classification problem in metabolomics data analysis is complex. There are thousands of variables but often just around ten to one hundred samples. This resulting in a very "short and fat" data matrix, which makes it possible to find even with spurious data many solutions to separate the classes. This is termed overfitting and is probably today the greatest multivariate analysis problem that we observe. OPLS-DA is eager to please and thus results should be handled with great care. The problem with a multi dimensional mega-variate space is, that almost always a perfect separation between the small amounts of samples can be achieved. OPLS-DA will have no problems finding it. For example, OPLS-DA separates two groups on the base of completely random data (Westerhuis et al. 2008). Its use also becomes problematic when a high number of variables are measured. Datasets not only become larger in size and more complex, they also tend to need normalization and/or transformations. Selecting the right pretreatment method is not intuitive, in spite of its crucial influence on the outcome of a metabolomics experiment (van der Berg et. al. 2006). Moreover, methods like PLS-DA and OPLS-DA are not very well suited when addressing a typical multiple classification problem (Westerhuis et al. 2008). Here a given set of objects, each of which belongs to a known class, and each of which has a

known vector of variables, are used to construct a rule. This rule will than allow to assign future objects to a class given only the vectors of variables describing new and so far unseen objects. The construction of such rules is guided by the goal of a "high out off sample prediction accuracy". Problems of these kind, called problems of supervised classification are typical for biological research, and many methods for constructing such rules have been developed.

One very important solution is the Naïve Bayes (NB) method, also called idiot's Bayes, simple Bayes, and independence Bayes. This algorithm was identified by the IEEE International Conference on Data Mining (ICDM) and presented by Wu et. al. (2008) as one out of four of the most important data mining algorithms used to perform classification in the research community. It is very easy to construct and does not need any complicated iterative parameter estimation schemes. This means that it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making a particular classification. Finally, it often performs surprisingly well: it may not be the best possible classifier in any particular application, but it is usually the most robust one (Domingos and Pazzani, 1997). NB is a probabilistic algorithm based on Bayes Theorem. It is relying on an explicit probability model by allocating a probability to each class that corresponds to the product of the individual probabilities of every attribute value. The predicted label then corresponds to the class with the greatest probability. By invoking conditional independence assumption in Bayes rule, the likelihood term of Bayes rules can be decomposed into product terms. One can label the new predictor variable to a particular class based on highest posterior probability. NB methodology simplifies classification tasks by allowing the computation of class conditional densities for each variable separately. In effect, a multi-dimensional classification task is reduced to multiple single dimension tasks, and is therefore not depending on normalization of the data matrix (the probabilities do not need to be normalized, since their normalization constant would be the same and not affect the classification). Moreover, missing values in both design sets and new cases can be easily handled (Hand and Yu, 2001). On the other hand, one should not forget that NB, like all other classifiers, has a problem when fed with a very short and fat data matrix (Eriksson et. al., 2006) and therefor it is a favorable to filter the variables according to predefined criteria.

The receiver operator characteristic (ROC) is widely considered to be one of the best means by which to describe the utility of a variable in binary classification (Egan, 1975; Zweig and Campbell, 1993; Zhou et al., 2002; Baker, 2003; Linden, 2006 see also http://gim.unmc.edu/dxtests/ROC1.htm, http://www.anaesthetist .com/mnm/stats/roc/). To understand the ROC concept, one should have a look at the confusion matrix shown in *Figure 4*, which summarize the number of false positives, false negatives, true positives and true negatives (Broadhurst and Kell, 2006) of a classification. The case individuals, the fraction of true positives is referred to as the sensitivity while the fraction of false positives is referred to as 1-specificity. Combining the two qualifiers of a classification leads to the receiver operator characteristic. The ROC unifies two characteristics that are often used to evaluate the performance of a test or method. A ROC curve represents the sensitivity of a test as a function of the 1-specificity of a test. The

sensitivity is defined as the number of true positives found as a percentage of all positives. 1−specificity is the number of false positives as a percentage of all negatives. Sensitivities are between 0 and 1 and should be close to 1. The specificity should preferably be close to 1, and 1-specificity should be close to 0. Both, specificity and sensitivity depend on the setting of the classification boundary of the classifier used in the method. By shifting the classification boundary, more true positives may be detected, but the number of false positives also increases, and the converse also occurs. The ROC curve, therefore, is a characteristic of a method describing the sensitivity and specificity of the method for different classification boundaries. Using ROC for filtering the variables is especially attractive. It is insensitive to the nature of any underlying population distributions, i.e. it is non-parametric and independent of the prevalence of a property (Westerhuis, 2008).

Figure 4: *A so called confusion matrix describing the outcome of predictive models that cross-tabulates the observed and predicted +/- or 1/0 patterns in a binary classification system.*

	Actual +/1	**Actual -/0**
Predicted +/1	*True positive*	*False positive*
Predicted -/1	*False negative*	*True negative*

IX. Validation techniques

In order to address the already mentioned issue of overfitting, the data mining community has developed several validation techniques. Cross validation is the standard validation technique used for classification models. The major interest of cross validation lies in the universality of the data splitting heuristics. It only assumes that data are identically distributed, and training and validation samples are independent, which can even be relaxed under particular circumstances. The main idea behind cross validation is to split data, once or several times, for estimating the prediction error of each algorithm: Part of data (the training set) is used for training each algorithm, and the remaining part (the validation set) is used for estimating the prediction error of the algorithm. A single data split yields a validation estimate of the error, and averaging over several splits yields a cross-validation estimate. Cross validation selects the algorithm with the smallest estimated prediction error (Arlot and Celisse, 2010).

X. Compound identification

Once potential biomarkers have been selected and tested with a classifier, identification is desirable. Metabolite assignments using LC-MS as a tool for compound identification are usually obtained by combining accurate mass, isotopic distribution, fragmentation patterns and any other MS information available. Calculation of the chemical combinations that fit a certain accurate mass is generally one of the first steps to obtain a set of alternatives that can lead to the identity of the metabolite detected. This set of alternatives becomes less extensive if the mass spectrometer can provide a more accurate mass value (Kind and Fiehen, 2006). Using an instrument that can provide very high mass accuracies, the range of possibilities molecular formulae (MF) is limited and can, especially for lower m/z values, lead to the correct MF. Furthermore, in most cases, chemical elements can be preselected, avoiding the generation of

excessive false alternatives, which would occur if all elements of the periodic table were included. For general applications in plant, food or animal metabolomics, only the core elements are C, H, O, N, P and S are used whilst most metals can be excluded (except perhaps Na or K, which form common adduct ions in mass spectra). Another aspect to take into account when MFs are calculated from molecular masses (MM) is the algorithm used for the calculations. There are more possible mathematical combinations of elements that fit certain MM than the number of MF existing chemically. This is related to chemical rules (e.g., the octet rule, or the nitrogen rule) that dictate certain limitations on chemical bonding derived from the electronic distribution of the participating atoms. Another useful factor is the presence of rings and double bonds. As described by (Bristow, 2006), the number of rings and double bonds can be calculated from the number of C, H and N atoms that a molecule contains. One of the most powerful methods for narrowing the number of MF is to make use of the isotopic pattern of a mass signal. For most small organic molecules, the intensity of the second isotopic signal, corresponding to the ^{13}C signal, indicates the number of carbons that the molecular ion contains based on that the natural abundance of ^{13}C (1.11%). According to Kind and Fiehn (Kind and Fiehen, 2006), this strategy can remove more than 95% of false positives. The fragmentation pattern of a mass signal can provide structural information about the fragmented ion. From the fragments obtained, the structure of the molecule can be deduced, knowing that the breakages will occur at the weakest points of the ion. The possibility of isolating one ion and performing MS/MS on the successively obtained fragments can be highly informative for tracking functional groups and connectivity of fragments. In addition, the possibility of obtaining accurate mass fragments is another advantage when there is little knowledge about the possible atomic arrangements of the molecular ion. There are still only a few tools that can automatically produce a list of possible metabolites from the m/z signals at a particular retention time. Therefore, the assignment of metabolites from experimental data implies an intensive manual effort, limiting the throughput of the analytical set-up. The bridge between experimental data (MS, retention time, fragmentation pattern, chemical shift, coupling constant) and the available chemical databases is still very weak. Some identification tools (for example elemental composition calculation or MM calculation) are included in the software of different instruments, but these seldom allow spectral matching linked to a public database, as already implemented in proteomics applications. One of the few examples of spectral databases for metabolomics is the NIST database. It provides mass spectral data for some known metabolites, which can mostly be used for identifying GC-MS signals, but the newest version is also applicable to LC-MS data. Other databases are listed in *Table 3*.

Table 3: Available databases applicative for metabolomics studies

	Access	Theme	Provider/ URL
ChemSpider	Free	General	www.chemspider.com
Golm	Free	Plant, with spectral data	Max Plank Institute (Germany) www.csbdb.mpimp-golm.mpg.de
Human metabolome database	Free	Human metabolites	Department of Computing Science, University of Alberta, Canada www.hmdb.ca/extrIndex.htm
KEGG	Free	General	Kyoto University Bioinformatics Center (Japan) www.genome.jp/kegg/ligand.html
Lipidbank	Free	Lipidomics	Japanese Conference on the Biochemistry of Lipids (Japan) www.lipidbank.jp/
MassBank	Free	General, with spectral data	Kelo University, University of Tokyo, Kyoto University, RIKEN (Japan) www.massbank.jp
Metlin	Free	Human metabolites, with spectral data	Scripps Center for Mass Spectrometry www.metlin.scripps.edu
MoTo	Free	Tomato metabolites	Wageningen University (Netherlands) www.appliedbioinformatics.wur.nl
NIST	Partially free	General, with spectral data	National Institute for Standard and Technology (USA) www.nist.gov/srd/nist1a.htm
PubChem	Free	Genereal	National Center for Biotechnology Information (USA) www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pccompound
SciFinder	Licensed	General	American Chemical Society (USA)

XI. Aim of this study

Taking all the information mentioned above into account, the aim of this work was to carry out a nontargeted metabolic study of pig back-fat samples. These samples were classified as tainted or non-tainted according to the evaluation of a trained test panel. Using different chemometrics approaches we intend to evaluate the possible contribution of some other unknown metabolites to boar taint.

Experimental Section

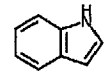
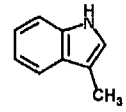
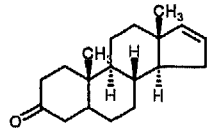
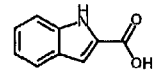
I. Chemicals

Indole (ID), skatole (SK), androstenone (AND), 2-carboxyindole (2CID), formic acid and human [Glu^1]-Fibrinogen peptide B (Glu-Fib) were obtained from Sigma-Aldrich (Steinheim, Germany), leucine enkephaline (Leu-Enk) from Waters, Milford, USA (for chemical properties see *Table 4*). High-purity HPLC grade solvents from different suppliers were used: methanol (MeOH) and acetonitrile (ACN) from Merck (Darmstadt, Germany) and HPLC-water from Scharlan S.L. (Sentmenat, Spain). Deionized pure water was prepared by using a Millipore Milli-Q system (Bedford, USA).

II. Animals, treatments and sampling

A total of 33 Swiss Large White male pigs originating from a study performed by Pauly et. al. (2009) were included in this study. All pigs had ad libitum access to the same growing and finishing diets (see *Table 5*). Individual feed intake was recorded and the body weight (BW) of all animals was determined once a week. From 80 kg BW until slaughter, pens were cleaned daily and barn ventilation was set at maximum-power. Animals were slaughtered 2 days after reaching 103 kg BW. Feed was withdrawn from the pigs 12 h before transportation to a nearby commercial abattoir. Animals were electrically stunned, exsanguinated, scalded, mechanically de-haired and eviscerated. Internal organs were removed and hot carcass weight was obtained. Thirty minutes after exsanguination, the carcasses entered air-chilling system (3°C) for 24 h. Adipose tissue samples (ca. 5 × 2 × 1 cm) consisting of the whole fat layer were collected from the right carcass side at the 10th–13th rib level the day after slaughter. The collected samples were stored at -80°C until extraction. At the same time about 1 kg of the longissimus dorsi muscle (LD) at the 13th–15th rib level and the neck (containing trapezius muscle and LD, 5th–7th rib level) were collected from the right carcass side of the selected animals for testing in the sensory panel study.

Table 4: *Name, chemical formula, structural formula, monoisotopic mass and the most abundant adducts for the test compounds.*

Name	Chemical Formula	Structural Formula	Monoisotopic Mass (Da)	$[M+H]^+$	$M+NH_4$	$M+Na^+$
Indole	C_8H_7N		117.057849	118.0657	135.0922	140.0476
Skatole *(3-methylindole)*	C_9H_9N		131.073499	132.0813	149.1079	154.0633
Androstenone *(5α-androst-16-en-3-one)*	$C_{19}H_{28}O$		272.214016	273.2218	290.2484	295.2038
2-carboxyindole	$C_9H_7NO_2$		161.0477 Da	162.0555	179.0821	184.0374
Glu-Fib				785.8421		

III. A) Selection and training of the test panel

The selection and training of the panelists, as well as the sensory evaluation conditions are described in detail by Pauly et. al. (2010). Briefly, the sensory study was carried out at the sensory laboratory of Agroscope Liebefeld-Posieux (ALP) Research Station (Posieux, Switzerland). Personnel of the research station were selected as panelists, the main selection criteria was their ability to detect androstenone. First they conducted a basic training program in sensory assessment. After that, panelists were specially trained in two sessions on boar taint. In the 1st session, the profiles of sensory attributes of boar taint were taught. In the second session, they were instructed to evaluate boar odor, boar flavor, juiciness and tenderness. The LD and neck chops from pigs with low, medium and strong AND and SK concentrations in the adipose tissue were cooked and given to the panelists with information on the concentrations of the samples. The day after the second training session the panelists retested the samples used in session two. Without receiving any information they had to score them for boar odor, boar flavor, juiciness and tenderness. Subsequently, the results were discussed in groups in order to obtain a consensus on sample evaluation.

B) Classification of selected samples considering AND and SK concentration

The pigs were classified in three categories of SK and AND concentrations (low, medium and high). The AND and SK concentrations in the adipose tissue were measured by HPLC with a diode-array detector as described by Pauly et. al.

(2008).

The concentrations were expressed as µg/g tissue. The concentration of indole was also measured but was not considered for the classification. The cut-off levels for each compound were established according to previous studies and were: low (AND: ≤0.5 µg/g and SK: ≤0.12 µg/g), medium (AND: 1.0 or 1.1 µg/g and SK: 0.13 or 0.16 µg/g) and high (AND: ≥2.5 µg/g and SK: ≥0.33 µg/g). Results from sensory analysis and concentrations of AND, SK and ID are presented in *Table 6.*

Table 5: *Composition of the growing and finishing diet, as-fed basis*

Ingredients (%)	Diet	
	Growing	Finishing
Wheat	27.2	62.2
Barley	15.0	3.8
Corn	6.9	3.1
Wheat starch	19.7	2.8
Sugar beet pulp	2.1	10.0
Soybean cake	22.7	12.4
Sugar beet molasse	3.0	3.0
L-lysine-HCl	0.28	0.25
DL-methionine	0.13	0.06
L-threonine	0.13	0.08
L-tryptophane	0.01	
Dicalcium phosphate	1.0	0.71
Sodium chloride	0.36	0.2
Pellan†	0.3	0.3
Calcium carbonate	0.78	0.7
Vitamin–mineral-premix	0.4	0.4
Analysed composition (g/100 g DM)		
Crude protein	18.6	16.6
Lysine	11.4	9.2
Crude lipid	2.6	2.1
Crude fibre	3.4	4.5
Calcium	0.80	0.70
Phosphorus	0.61	0.54
Calculated energy content		
DE‡ (MJ/kg DM)	15.8	15.4

DM = dry matter.
†Pellan = a binder that aids in pellet formation (Mikro-Technik GmbH & Co. KG, Germany).
‡DE = digestible energy content (MJ/kg) calculated from nutrient content (expressed in g/g DM) according to ALP (2005).

IV. Sample preparation

AND, SK, ID and potential boar taint markers were extracted with pure methanol from the collected back fat samples. A 800 µL volume of methanol and 20 µL volume of 2-carboxyindole solution (2 µg/mL in methanol) used as an internal standard was added to 800 mg of back fat. Portions of 800mg back fat were extracted with 800µL methanol in the presence of 40ng of the internal standard 2-carboxyindole, added as 20µL of a solution with 2ng carboxyindole/µL). Extracts were homogenized by means of a Retsch MM300 homogenizer (F. Kurt Retsch GmbH & Co. KG, Haan, Germany) for 5 minutes at 25 Hz using stainless steel grinding balls of 5mm diameter (Schierlitz & Hausenstein AG, Zwingen, Switzerland). After homogenization, the samples were cooled for 30 min at -20 °C. After centrifugation at 1600 rcf for 10 min by 0° C, the supernatants were transferred on a reversed solid phase extraction cartige (Sep-Pak® C_{18} column, Waters, Milford, USA) previously chilled and conditioned with 2 ml -20 °C methanol. The first 300 µL of the eluate were discarded and the next 500 µL were collected. The samples wear stored at -20 °C until metabolite profiling on the nanoUPLC-QTOF-HDMS system describer below. Storage time of the extracts was no longer than one-week post extraction.

Table 6: *Identity, AND, SK, ID content and sensory classification of all the pigs analyzed with nanoUPLC®-ESI-QTOF-HDMS™.*

Pig ID	ANDRO (µg/g fat)	SCATOLE (µg/g fat)	INDOLE (µg/g fat)	HPLC class	Flaveur average	Sensory class	Gender
v26_798	0.200	0.039	0.025	**n**	2.489	**n**	IMP
v26_740	0.200	0.030	0.025	**n**	2.513	**n**	CAS
v26_790	0.200	0.056	0.025	**n**	2.854	**n**	CAS
v26_737	0.200	0.076	0.025	**n**	3.129	**n**	IMP
v26_864	0.471	0.032	0.025	**n**	3.310	**n**	ENT
v26_839	0.306	0.058	0.025	**n**	3.427	**n**	ENT
v26_834	0.200	0.030	0.025	**n**	3.507	**n**	CAS
v26_823	0.200	0.035	0.025	**n**	3.557	**n**	CAS
v26_775	0.200	0.030	0.025	**n**	3.621	**n**	CAS
v26_866	0.200	0.030	0.025	**n**	3.665	**n**	IMP
v26_842	0.200	0.030	0.025	**n**	3.695	**n**	IMP
v26_776	0.200	0.033	0.025	**n**	3.726	**n**	IMP
v22_8903	0.200	0.029	0.029	**n**	1.307	**n**	CAS
v22_8677	0.200	0.041	0.029	**n**	2.094	**n**	CAS
v22_8882	0.200	0.037	0.029	**n**	2.446	**n**	CAS
v22_8847	0.200	0.029	0.029	**n**	3.026	**n**	CAS
v26_788	0.832	0.241	0.038	**s**	5.009	**s**	ENT
v26_826	0.200	0.085	0.025	**n**	5.176	**s**	IMP
v26_659	0.200	0.048	0.025	**n**	5.211	**s**	CAS
v26_777	0.297	0.085	0.025	**n**	5.220	**s**	ENT
v26_841	0.757	0.051	0.027	**w**	5.394	**s**	ENT
v26_729	0.467	0.030	0.025	**n**	5.402	**s**	ENT
v26_726	0.200	0.035	0.025	**n**	5.409	**s**	IMP
v26_825	0.491	0.101	0.025	**n**	5.456	**s**	ENT
v26_835	1.212	0.094	0.026	**s**	6.147	**s**	ENT
v26_658	0.439	0.205	0.032	**s**	6.982	**s**	ENT
v26_836	1.091	0.343	0.025	**s**	8.010	**s**	ENT
v22_8900	0.990	0.145	0.032	**w**	5.239	**s**	ENT
v22_8894	0.752	0.137	0.031	**w**	5.580	**s**	ENT
v22_8676	1.136	0.027	0.114	**s**	6.531	**s**	ENT
v22_8849	1.937	0.044	0.039	**s**	7.194	**s**	ENT
v22_8885	0.819	0.192	0.045	**s**	7.212	**s**	ENT
v22_8883	1.669	0.124	0.121	**s**	7.734	**s**	ENT

Abbreviations: s, strong tainted pig; n, non tainted pig; w, weak tainted pig; CAS, castrated pig; IMP, immunocastrated pig; ENT, entire male pig

V. NanoUPLC®-QTOF-HDMS™ Analysis

Ultra performance liquid chromatography (UPLC) was performed on a Waters Technologies (Manchester, UK) nanoAcquity UPLC® system. All columns were packed in-house into pieces of 200µm inner diameter untreated fused silica capillaries (BGB Analytik Vertrieb Bernhard Fischer, Schlossböckelheim, Germany) with a length of 200mm. A 3µm inner diameter Atlantis® C_{18} material was used to pack the first 10 mm of the columns and a 1.8 µm inner diameter High Strength Silica (HSS) T3 C_{18} material was used to fill the column to a length of 130 mm (Waters, Milford, USA). The column was kept at room temperature during storage and measurement. The temperature of the sample manager was 4 °C and the injection volume was 0,5 µL. The mobile phases consisted of (A) 0,1% formic acid in water and (B) 0,1% formic acid in acetonitrile and a flow rate of 3 µL/min was applied. For the biological samples an isocratic period of 5 min at 50% A was followed by a linear change from 50% to 95% B in 15 min. Next, the gradient remained 15 min at 95% B (between the 25th to 32nd minute the flow was increased to 5 µL/min to achieve a higher cleaning volume) followed by a zigzag gradient for 7 min (from 95% B to 5% B, repeated three times) and returned linearly in 1 min to 50% A, remaining at this level for 8 min until the next injection. For the test mixtures (Glu-Fib 600 ng/µL in water/ACN (19:1) and a mixture of the key compounds (KeyMix) AND, SK, ID, 10 ng/µL and 2CID 50

Box 1

nanoUPLC®-QTOF-HDMS SETUP: CONDITIONING THE nanoUPLC®-SYSTEM

1. Prepare all solvents (A, B, strong needle wash, week needle wash and seal wash) and degas them for at least 5 min using an ultrasound bath. Prime A and B nanoUPLC® pump and tubing for a minimum of 10 minutes and syringes at least for seven cycles.
2. Install the column on the NanoFlowSprayer™.
3. Precondition column system by starting at 98% B with 3 µL/min for 20 min and than slowly decrease % of B until the initial gradient conditions are reached. Wait for additional 10 min for the pressure to stabilize.
4. Program the inlet file according to the gradient settings given in the text. In the standard setup, we use relatively long chromatographic runs of 55 min, including column washing and re-conditioning, with a mobile phase flow rate of 3 µL/min into the analytical column (diameter of 200 µm) resulting in a backpressure of approximately 6000 psi.
5. Check UPLC® pump for air bubbles and connections for leakage by verifying pressure stability and performing an auto zero flow transducer test. The mass spectrometer can be calibrated and checked for performance as described in Box 2.
6. Place the methanol extracts in trays inside the autosampler (4°C) during the analysis series. Program the injection system to operate in partial loop mode (with a 5 µl loop mounted on the injection valve). The injection needle is washed with 0.6 ml weak needle wash and 0.6 ml strong needle wash between injections.
7. Our nanoUPLC® system does not provide reproducible results for the first few injections for biological samples. For this reason, it was our general practice to run several (usually 5 to 10) QC samples prior to the start of the main analytical runs. In this way we were "conditioning" the system and were able to demonstrate that it has achieved stability. We have observed that there is an absolute requirement for the injection of biological samples to achieve retention time stability for the adipose tissue components, at least in the system described here. Thus, the repeated injection (up to 10 cycles) of the mixture of pure standards prior to the start of analysis of the QCs was ineffective in "conditioning" the system.

ng/µL in MeOH) an isocratic period of 5 min at 95% A was followed by a linear change from 95% A to 95% B in 5 min. Next, the gradient remained 5 min at 95% before returning linearly in 30 seconds to 95% A, remaining at this level for 4.5 min until the next injection. Between the biological samples and the test compounds, injections of the week needle wash solvent (0.1% formic acid in 95% water and 5% ACN) were performed to compensate the changes in starting and ending condition (95% A vs. 50% A). Starting at 50 % A respectively 95% A for 5 min going to 95% A respectively 50% A in 30 seconds, remaining at this condition 4.5 min. The injection system was subjected to two washing cycles with a strong (0.1% formic acid in ACN) and a weak needle wash solvent after each injection, and one cycle prior to each injection to minimize carryover. The pump seals were washed with MilliQ water/ACN (9:1 v/v) every 15 minutes. For details of the conditioning of the LC-system see *Box 1*.

The nanoUPLC® was directly interfaced with a Waters Synapt G2™ HDMS™ mass spectrometer equipped with a dual electrospray ionization probe, Zspray™-NanoLockSpray™, operating in positive electrospray ionization mode (ESI+). The source temperature was 80°C with a cone gas flow of 30 L/h and desolvation temperature of 180°C. The capillary voltage was set at 2.00 kV, with a sampling cone voltage of 30 V and an extraction cone voltage of 2.50 V. The data acquisition rate was 0.25 s, with an interscan delay of 0.024 s. Leu-Enk was employed as the lockmass compound, infused straight into the MS at a

Box 2

nanoUPLC®-QTOF-MS SETUP: CONDITIONING THE HDMS™- SYSTEM

Before starting with sample analyses, the mass spectrometer should be conditioned and calibrated to obtain a good performance in terms of mass accuracy and resolution. The procedure and settings described here are for a Waters Synapt® G2 HDMS™ with an ESI source and the TOF tube in V-mode, in combination with the nanoUPLC® conditions described above.

1. Attach the NanoFlowSprayer™ with the mounted column and an eluent flow of 3 µL/min to the stage platform of the Zspray™. Push the NanoFlowSprayer™ into the source and use the settings described in the text.
2. Adjust pump flow, capillary voltage, cone voltage, desolvation gas flow and/or collision energy (depending on room temperature and humidity these values are subjected to changes) until an ion intensity of at least 10^3 for the background is reached, and 10^5 for the lockmass compound Leu-Enk. Mass resolution is calculated by dividing the m/z value of the centered mass signal by the mass difference at half height of the Gaussian-shaped mass peak in continuum mode, and should be higher than 20.000. Combine spectra of about 50 scans during acquisition of the lockmass at optimal settings in continuum mode. Do an automatic peak detection and check the mass accuracy. The observed mass should be within 2 ppm deviation of m/z 556.2771 in positive mode. If resolution and accuracy are satisfying the basic criteria for a measurement are fulfilled; if not, a recalibration of the instrument is needed.
3. Prepare MS method file to acquire mass data from m/z 50–1.200, at a scan rate of 0.25 scan/s and an interscan delay of 0.024 s using the same settings as for the lockmass. The range of masses to be detected in sample extracts should fall within the range of calibration masses. The HDMS is programmed to switch from sample to lock spray every 20 s and to average 3 scans for lock mass correction (m/z 556.2771 ± 0.5 Da in positive mode). Adjust flow rate or concentration of the lock mass solution to obtain an intensity of about 2000 counts per scan during measurement, to enable accurate mass calculation of as many compounds in the extracts as possible. Polarity switch during the run should be avoided because it causes contamination of inner lenses and quadrupoles. Therefore, all measurements were in positive mode.

concentration of 2 ng/µL in 5% ACN and 95% water containing 0.1% formic acid at a flow rate of 0.5 µL/min. The lockmass was the monoisotopic positive ion peak observed at m/z 556.2771. All mass spectral data were acquired in the MS^E continuum mode with direct lock mass correction by scanning a m/z 50-1200 range. The collision energy in function two was ramped from 20-40 kV. All parts were controlled by the MassLynx™ Software 4.1 SCN 833 (Waters, Milford, USA). For the conditioning of the HDMS™-systems see *Box 2*.

We acquired all data in continuum mode. Here the mass signal is represented by a Gaussian curve. In comparison the mass signal in centroid mode, is the projection of an accurate m/z value by on-the-fly mathematical transformation. In this way, we do not loose relevant information on mass peak shape and purity, which can vanish during "centroiding". In addition, by using the separate lock mass spray as reference and by continuously switching between sample and reference, the MassLynx™ software can automatically correct the continuum mass values in the sample for deviations from the exact mass measurement. The procedure results in a mass accuracy higher than 5 ppm. A disadvantage is that the raw data files are markedly bigger, raging from about 600 to 700 MB, whereas a centrioded file just accounts for around 200 to 300 MB per sample (for a MS^E experiment over 55 min with 4 scans per second). However, after the acquisition a centroiding of the data is still possible. This is also desirable to achieve accurate mass and to avoid long processing time of the data by the software.

To supervise the stability of the system a standardized analysis protocol (*Figure 5*) was used. Glu-Fib 600 ng/µL in water with 5% ACN, a mixture of the key compounds AND, SK and ID (10 ng/µL) and 2CID (50 ng/µL) in MeOH (KeyMix) and a quality control (QC) sample containing 10 µL of every pig backfat extract were injected throughout the whole experimental run. A blank sample consisted of a methanol solution that had also been subjected to preparation procedures as described above, was analyzed to include any contaminating peaks that may have come from preparation steps. The pooled QC sample was injected 6 times at the beginning of the run to ensure system equilibration and then every 10 samples to further monitor the stability of the analysis. We analyzed the fat extracts in triplicate, distributing each replicate in random order in a different analysis series. In total, three series, each containing one of the 33 pigs were analyzed by nanoUPLC®-ESI+-QTOF-HDMS™ during a time span of 5 days. Transformation of continuous mass spectra into centroid spectra was than performed on the resulting 114 measurements, using Accurate Mass Measure with the function "Accurate mass measure" selected. This step is to be considered as a first to reduce spectra complexity before peak extraction.

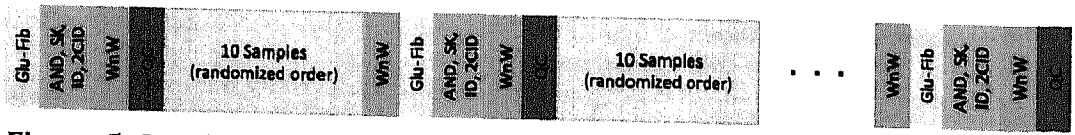


Figure 5: *Standardized analysis scheme used for metabolomics studies. Adapted and modified from Coulier et. al., 2011.*

VI. LC/MS raw data extraction

Three different methods were used for the raw data extraction. The first two were based on the software delivered together with the instruments. The metabolomics group at the FGCZ developed the third method, mainly as a part of the ongoing PhD studies of David Fischer (publication in preparation).

A) The MassLynx™ raw data files were processed using MarkerLynx™ software (Waters, Milford, USA). The MarkerLynx™ parameters were the following: initial retention time 2.00 min, final retention time 24.00 min, lowest mass 100 Da and highest mass 1150 Da with mass tolerance 0,05 Da. Peak width at 5% height 20 s, peak to peak baseline noise 500, intensity threshold 500, mass window 0.05 Da, retention time window 0.2 min, noise elimination level turned off and the deisotope data turned on. The settings for the internal standard was: Name 2CID, Mass 162.0555 Da, Mass window ± 0.050 Da, Retention time 6.40 ± 0.3 min. Masses belonging to the same peak 'cluster' are merged together and then aligned according to retention time if their masses differ by less than the specified tolerance. The data were shown as height of the peak and normalized to total marker intensity (the marker intensities are scaled such that their sum totals 10000).

B) Using MarkerLynx™ software with the same settings as mentioned above but with the noise elimination level set to 6. In the noise elimination feature within MarkerLynx™ it is assumed that the intensity distribution of the component spectra is a Gaussian distribution due to noise with signals as outliers. The noise level, N, is defined as a user defined gain of the standard deviation above the mean, or $N = X\sigma + \bar{x}$ where $\bar{x}$ is the mean, σ is the standard deviation and X is the user defined gain.

C) For the in FGCZ method, the MassLynx™ raw data files was converted into netCDF (.cdf) files using DataBridge (Waters, Milford, USA) and imported into MATLAB (MATLAB version 7.10.0. Natick, Massachusetts: The MathWorks Inc., 2010). For data matrix generation, mass bins were selected based on the total mass abundance in all files. All mass spectra (from 2-24 min) were combined into one spectrum. From this "master" mass spectrum, accurate masses and left and right locations of the full width at half height for each peak were identified using the mspeaks function. Based on this information, mass bins were selected and for each bin we combined the ion intensities from the whole LC-MS run. Each feature in the data matrix therefore corresponds to the total intensity of one respective accurate mass during the whole LC-MS run. For normalization purpose, the data matrix was $\log_2$-transformed and for each sample, values were calculated against the median feature intensity. A log transform was applied to the observed intensities for each compound because, in general, the variance increased as a function of a compound's average response. In order to test normalization success, the intensity of the internal standard was monitored.

VII. Statistical analysis

The data matrices generated with MarkerLynx™ were subjected to two different statistical methods to identify significantly altered metabolites between non-tainted pigs and tainted pigs. One based upon ROC classification and the second

one was based on OPLS-DA. The data matrix generated with the "FGCZ method" was subjected to a student's two sample t-test to identify significantly altered metabolites.

OPLS-DA: The MarkerLynx output tables were used as an input for EZinfo (Umetrics, Umeå, SE) to visualize each data matrix with PCA and OPLS-DA. Data were scaled using pareto scaling, where the weight factor is the square root of the standard deviation of each column. Pareto scaling is recommended for metabolomics data (Trygg, 2007). PCA was used to detect trends, patterns and outliers. OPLS-DA was used to do a supervised classification using the results from the sensory panel (**n** relate to non-tainted pig and **s** to strongly tainted pigs) and find potential marker candidates. The OPLS-DA loadings visualization tool "S-plot" (the modeled covariance and modeled correlation from the OPLS-DA model are combined in a scatter plot) was used to pick those marker candidates, that have high reliability and medium to high magnitude of differences between the sample classes (the p1-axis describes the magnitude, p(corr)-axis represents the reliability of each variable in the data matrix).

Students t-test: Significant features between the two group „strongly-tainted", **s**, vs „non-tainted", **n**, were calculated using the t-test function in MATLAB. Features with a p-value lower than 0.01 were considered as significant.

ROC: The MarkerLynx™ file was exported as a comma separated value file (.csv) and imported into 'R' (The R-project for statistical computing and graphics, www.cran.r-project.org/), which is a freely available open-source software package. The first step was to calculate the mean and the median of every pig. 'R' was programmed to take the triplicates of a pig for this computation. In cases where a response was not detected, it was assumed that the value was missing because the compound was below the limit of detection (due to ion-suppression, unstable ESI or other problems with the analytical platform). In those cases, the mean and the median was calculated from the responses detected, if all responses were zero the mean/median was set to zero. Each marker was than filtered separately with Receivers Operating Characteristic using 10.000 different cut-offs for intensity. Compounds achieving >80% for observed within sample statistical sensitivity and specificity was than selected. These selected marker candidates were checked on the acceptance criteria and was kept for further analysis.

Naïve Bayes: Marker candidates originated from OPLS-DA, Students t-test and ROC evaluations were crosschecked. Those entries, that were significant in three or more of the chemometric methods, were kept. The mean and the median value of these markers were than given to a naïve Bayes classifier to determine the predictive performance of unseen data. We used a 90/10 cross-validation to estimate the predictive out-of-sample accuracy and repeated this 1000 times.

VIII. Tentative metabolite identification

From accurate mass measurements, the elemental composition was determined using MarkerLynx™ Elemental Composition Method. The settings were: mass tolerance 5.0 mDa, mass mode monoisotopic; electron state, even electronic ion;

double bond equivalence, from -1.5 to 50.0; elements allowed, 0-10 C, 0-100 H, 0-5 N, 0-10 O, 0-2 S, 0-1 Na, 0-3 P and 0-1 K. The elemental composition was used to search for matching compounds within a mass window of 0.02 Da. The online open source libraries used were ChemSpider, Human Metbolome Database, KEGG, LipidBank, MassBank, Metlin and PubChem. Additionally NIST Mass Spectral Search Program 2.0 g (Standard Reference Data Program of the National Institute of Standards and Technology, USA) was used to perform an exact mass search. All hits within an accuracy of 10 ppm were listed and closely examined. Considering the results from both methods and the likelihood of biological significance tentative metabolite identification was made.

Results

I. Experimental setup of sample extraction

When entering the terms "metabolomics and adipose tissue" in the search field of PubMed, we received 29 hits. From these 29 hits, two concerning metabolomics approaches applied to adipose tissue (Mattila et. al. 2008; Zyromsky et. al., 2009). None of these publications were discussing how to extract adipose tissue for measuring the metabolome. Two older protocols from Folch (Folch et. al. 1957) or Bligh and Dyer (Bligh, E. G., & Dyer, W. J.,1959) were than our first choice of extraction method. Lipids of all major classes are recovered via chloroform/methanol extraction, where they are mostly enriched in the chloroform phase. The second phase, containing water/methanol and more polar metabolites, could also be subjected to analysis if desired. However using these two protocols, we were not able to get extracts suitable for measurements. The lipid phase was adversely affected by fat precipitation and the polar fraction was hardly containing any metabolites. In addition, it is not recommended to use chloroform when working with the nanoUPLC® equipment.

Another interesting approach was described by Masson et. al. (2010). Here, they compared six different extraction protocols for nontargeted metabolic profiling of liver samples. They concluded that an aqueous extraction with methanol/water followed by an organic extraction with dichloromethane/methanol and reconstitution of both dried extracts in methanol/water was the best method. This protocol adopted to our fat samples led to a very nice and clear aqueous extract but the organic extract was suffering from micelle formation after reconstitution. We also got followed the suggestions of Matyash et. al. (2008) to use methyl-tert-butyl ether (MTBE) instead of chloroform. Here, lipids are recovered into the MTBE phase, that, because of its lower density, is the upper phase of the two-phase solvent system. In contrast to the Folch (Folch et. al. 1957) method, non-extractable matrix residuals are in the aqueous phase at the bottom of the extraction vial. The organic phase enriched with lipids is easily accessible by a micropipette from the top. The author finally states that the MTBE extraction procedure allows faster and cleaner recovery of most of the major lipid classes. Additionally it is also well suited for shotgun profiling, in which total extracts are infused directly into a mass spectrometer with no prior cleanup. Unfortunately, implementing this protocol did not lead to any improvement because the extracts still contained too many suffering impurities.

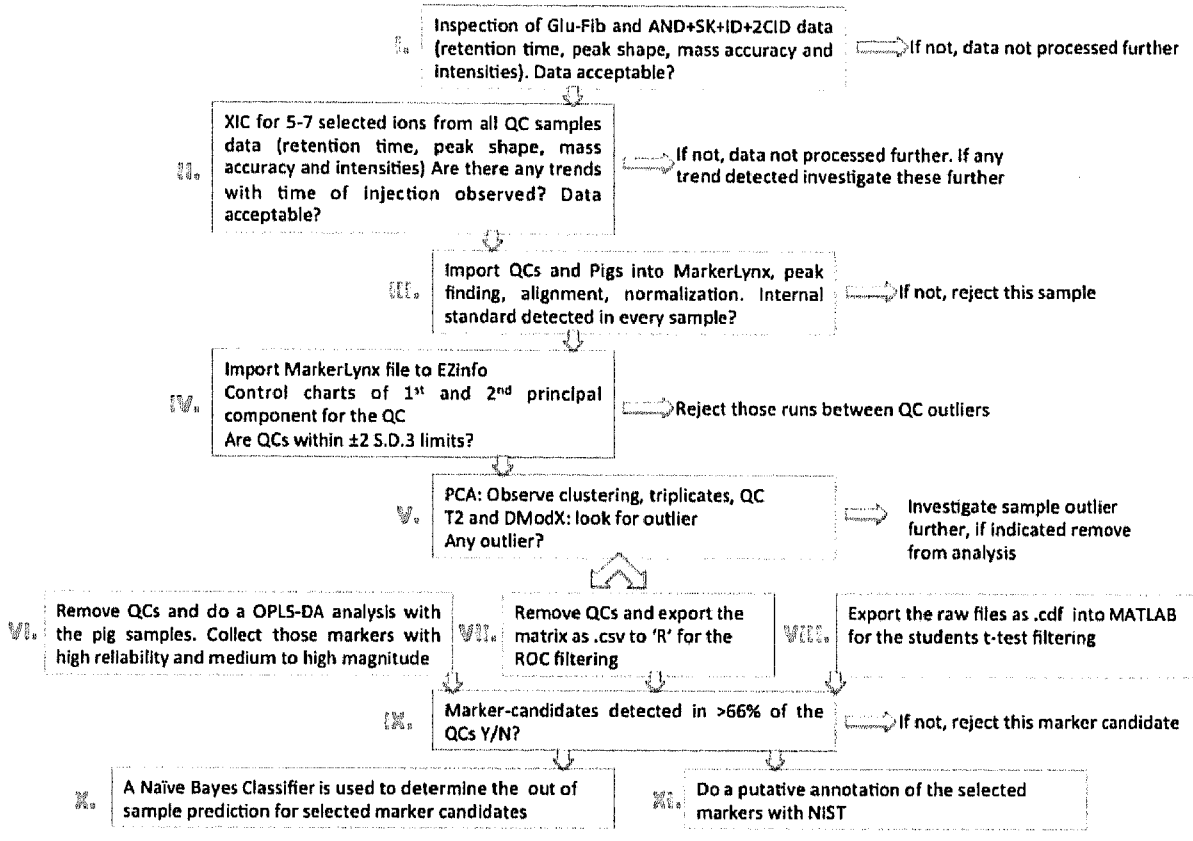
The literature on fat extractions for the detection of boar taint we include four main methods (Mortensen and Sorensen, 1984; Garcia-Regureiro and Diaz, 1989; Hansen-Møller, 1994; Toumola et. al., 1996). In all these publications the scientists had been working with adipose tissue and recognized the problem with the high content of lipids in the extracts. Due to the fact that chromatographic systems are disturbed by very lipidrich samples, extracts must be cleaned prior to injection. The removal of as many lipids as possible, without loosing the lipophilic boar taint compounds, is therefore a critical step in the sample preparation. In the method described by Toumola et. al. (1996) the fat is liquefied in a microwave prior to extraction with methanol. Bearing in mind that we were searching for unknown compounds, this idea seemed to be fraught with the risk of loosing interesting compounds. Hansen-Møller (1994) presenting another approach where they kept the samples chilled throughout the entire extraction procedure, using a solid phase extraction column to eliminate lipids. He states that chilling of the solid-phase extraction columns prior to application of tissue homogenates and the application of cold homogenate is essential. If homogenates were at ambient temperature, the lipids would melt, with the risk of excessive amount of lipids passing through the column. Consequently, lipids would be injected on to the analytical column leading to an increase of backpressure, which could even damage the column. The risk of contaminating the injection system of the LC and the source of the MS is also not to be underestimated. With this method the interfering fat and cell debris is retained by the stationary C_{18} phase of the column. We therefor adopted the method of Hansen-Møller (1994) and transformed it in to a minimalistic metabolomics protocol.

II. LC-MS analysis

The chromatographic conditions applied are always a compromise to achieve best chromatographic resolution, retention time stability and sample throughput. The settings in this protocol were selected after testing different types of solvent systems (different concentrations of acetic acid, water and ACN in A, different concentrations of acetic acid, water and ACN in the week needle wash) gradients and columns (HELIC and BEH C18 both from Waters, Milford, USA) for their ability to retain and separate compounds of our prime interest (AND, SK, ID and 2CID). A long cleaning and re-equilibration time was necessary to avoid carry-over and ensure stability of the chromatographic separation. The reliable multicomponent analysis of complex biological samples via LC-MS-based methods provides a number of challenges. The limitations of the technique must be kept in mind and controlled at all times (potential for drift in both chromatographic and mass spectrometer performance, for example decreased detector response, altered ionization efficiency, decreased mass-accuracy and shifting retention times). To eliminate the bias due to a gradual change in the performance of the system, the samples must be analyzed in a random order. In addition, quality control samples should be used to rigorously monitor the performance of the platform. We have used standard mixtures of test compounds (Glu-Fib, KeyMix).

These test mixtures of pure standards provided an initial screen. A very rapid visual examination of the performance of the system was obtained by overlay and comparison of the 10 runs performed during the experiment. A pooled fat

Figure 6: *Data acceptance and analysis workflow*



extract sample (QC) enabled us to demonstrate that the nanoUPLC-HDMS system was providing useful and reliable data also for biological samples. Fulfilling the first acceptance level with pure standards, the QC samples are than evaluated against a set of predefined criteria to enable acceptance or rejection of the batch. Random selections of masses are monitored against predetermined acceptance criteria for peak shape, intensity, mass accuracy and retention time. If the QC samples pass this preliminary screen, multivariate statistical analysis can be performed to determine if the QC data show no time related trends and cluster closely together. Highly variable QC data would mean that the run failed while close QC data do not automatically mean that the run was successful, but justify further data analysis. When potential biomarkers have been identified, it is necessary to reexamine the QC data. For variability of the QCs results obtained with selected marker candidates. Every step in the "Data acceptance and analysis workflow" is described in *Figure 6,* and the results are shown according to this scheme.

i. Inspection of KeyMix and Glu-Fib

Any change or deterioration, in either chromatographic or detector performance would be evident as alteration in retention time, peak shape and mass accuracy.

Peak shape was unchanged and the maximum deviation in retention time for all of the test compounds was 0.021 min. The mass accuracy never exceeded 5.5 ppm, which was considered to be acceptable. All standard compounds are shown in *Figure 7* as overlaid extracted ion chromatogram (XIC, ± 50 mDa). Signal intensity was the most significant source of variability (coefficient of variance (CV) ranging from 9.4% to 69%) rather than retention time or changes in mass accuracy. This fact is not surprising, because the data is not yet normalized to accommodate variations across experiments. In addition, the same two samples were measured throughout the whole experimental run. After the first measurement, the top cover is damaged and MeOH will start to evaporate, leading to an increase of concentration. However, low abundant compounds are more biased than high abundant compounds.

ii. Quality control base peak chromatogram and extracted ion chromatogram for 7 selected ions

Next, the internal standard 2CID and a small subset of six peaks present in the QC samples, covering a range of retention times and signal intensities, were examined. We looked at extracted ion chromatograms as a further means of screening the QC raw data prior to processing with the peak finding algorithm. This enabled us to determine whether retention time, detector response, and the mass accuracy over the course of a biological sample run were altered (for example due to matrix effects). The deviation from accurate mass of the internal standard was 0.56 ppm. The deviation from standard retention time was 0.013 min (*Figure 8*). *Figure 9* shows a section of the whole chromatographic run, were the overall profile can be assessed. *Figure 10* shows the extracted ion chromatograms intensities in all the 13 QCs for a peak eluting shortly after the void volume at 4.19 min (m/z: 146.0599), one of the most intense peak eluting at 12.80 min (m/z: 522.354), a peak eluting towards the very end of the gradient at 22.90 min (m/z: 326.306) and two intermediate eluting peaks with different intensities (m/z 277.215 at 8.62 min and m/z 324.293 at 16.30 min). This showed that the stability of the retention times for the five components over the 120 h run was good (standard deviation 0.021 min to 0.038 min from the mean) and measured mass also acceptable (CV ranging from 3.1 ppm to 5.7 ppm). Thus, once the system had come to equilibrium, the main cause of variability over the 120 h of the experiment was also here the intensity.

iii. Quality control and Pigs: Inspection of internal standard 2CID

After centroiding all raw data, the entire set of 114 samples was processed with MarkerLynx™. All analytical information in the raw profiles is first transformed into coordinates on the basis of mass, retention time and signal amplitude (Idborg et. al. (2), 2005). MarkerLynx™ extracts *m/z* chromatograms (XIC), if a peak is detected above the given threshold peak top mass (single scan). A retention time is used and assigned to an extracted mass retention time bin (predefined time- and m/z-slot) accordingly. The *m/z* reported in the marker table is the intensity weighted mean *m/z* of the detected markers in all samples and not the *m/z* of any single sample, i.e. it is the *m/z* of the bin. These coordinates are then aligned across all samples and presented as a matrix in the MarkerLynx™ result window. The first matrix resulted in 67569 marker candidates (referred to as A). The second matrix, using the noise elimination

algorithm set to 6, resulted in 8286 marker candidates (referred to as B). The matrices were normalized to the total marker intensity. Subsequently the intensity of the internal standard in the QC samples was investigated. The coefficient of variance was computed to 11.3% ($\bar{x}$ 138.78, SD ±24.6), which means that the normalization led to decrease of CV of almost 10%. This confirms the efficiency of normalization of acquired UPLC-HDMS data. The internal standard was also detected in all of the pig samples with a SD for retention time of just 0.012 min. The CV for intensity was computed to 17.6% ($\bar{x}$ =140.8, SD ±24.8). This means that the extraction accounts for additional 6% of the intensity variance.

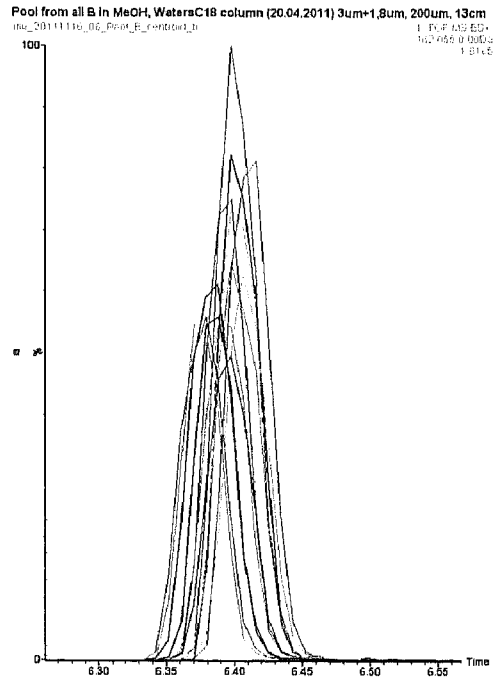


Figure 8: *2CID exact mass 162.0555, CV 3.2 ppm ($\bar{x}$ 162.0556, SD 0.52 mDa); Intensity CV 18.45% ($\bar{x}$ 447656, SD 82612); Retention time CV 0.2% ($\bar{x}$ 6.397 min SD 0.013 min)*

iv. Quality controls within ±2SD limit

PCA was performed on the thirteen QC samples separately, to determine trends and shifts depending upon time. Their scores represent weighted average trajectories of the original variables. *Figure 11* shows the time series properties of the first PCA component. The second PCA component is describing how the QC samples behaved as the run progressed. This type of result gives some confidence that the analysis was stable for the duration of the run considering all of the detected marker candidates. Thus, it provides a pragmatic means of assessing the quality of the data and deciding if it is sufficient to warrant further statistical analysis of the results to detect biomarkers. In PCA scores plots, multiple injections of the QC samples should cluster tightly together and ideally show random variation without any drift over time.

v. PCA with all Markers: trends, outliers, QCs

PCA was performed on matrix A as a preliminary step of data examination to assess the samples' distribution with respect to their class labels. Samples were not clearly separated on the first principal plane (PC1 10.8% vs. PC2 7.0%). Indeed, principal components correspond to the directions that maximize the variance and there is no guarantee that these dimensions are discriminant (Hotelling, 1933). Additionally an underlying structure was discovered within total experimental run. *Figure 12* shows the samples colored by measurement date (16th, 17th, 18th, 19th and 20th of November 2011).

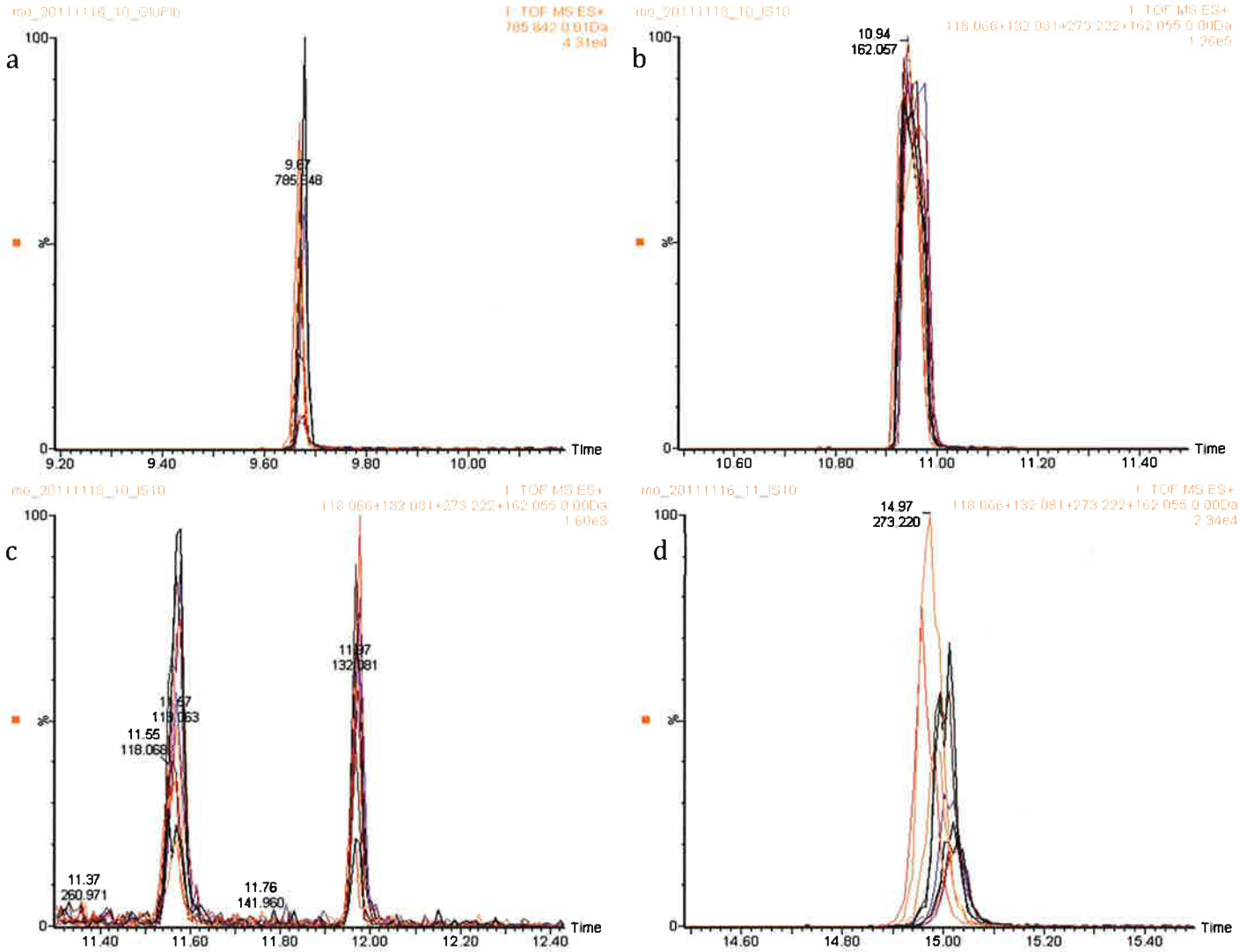


Figure 7: a) *Glu-Fib: Exact mass 785.8421, CV 1.4 ppm ($\bar{x}$ m/z 785.8417, SD 1.11 mDa); Intensity CV 69% ($\bar{x}$ 67704, SD 46802.7); Retention time CV 0.06% ($\bar{x}$ 9.671 min, SD 0.056)*
b) *2CID: Exact mass 162.0555, CV 2.7 ppm ($\bar{x}$ m/z 162.0557, SD 0.43 mDa); Intensity CV 9.4% ($\bar{x}$ 914167, SD 89206); Retention time CV 0.05% ($\bar{x}$ 10.945 min, SD 0.005)*
c): *ID: Exact mass 118.0657, CV 2.9 ppm ($\bar{x}$ m/z 118.0662, SD 0.34 mDa); Intensity CV 12.3% ($\bar{x}$ 177708, SD 21802.7); Retention time CV 0.07% ($\bar{x}$ 11.567 min, SD 0.0675)*
SK: Exact mass 132.0813, CV 5.5 ppm ($\bar{x}$ 132.08172 SD 0.72 mDa) Intensity CV 49.5% ($\bar{x}$ 4903, SD 2426); Retention time CV 0.42% ($\bar{x}$ 11.975 min, SD 0.005)
d) *AND: Exact mass 273.2220, CV 1.1 ppm ($\bar{x}$ m/z 273.2216, SD 0.31 mDa); Intensity CV 47.7 % ($\bar{x}$ 124328 SD 59314); Retention time CV 0.14% ($\bar{x}$ 15.002 min, SD 0.139)*

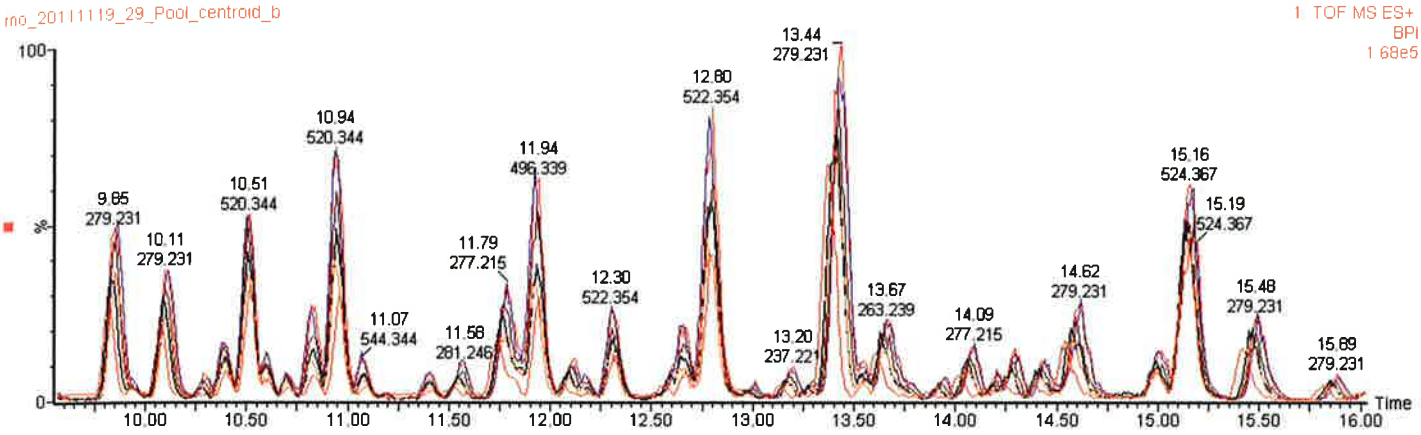


***Figure 9**: Section of the complete run of all of the 13 QC as overlaid graphs with linked vertical axes. Visualizing the reproducibility within the body of the runs.*

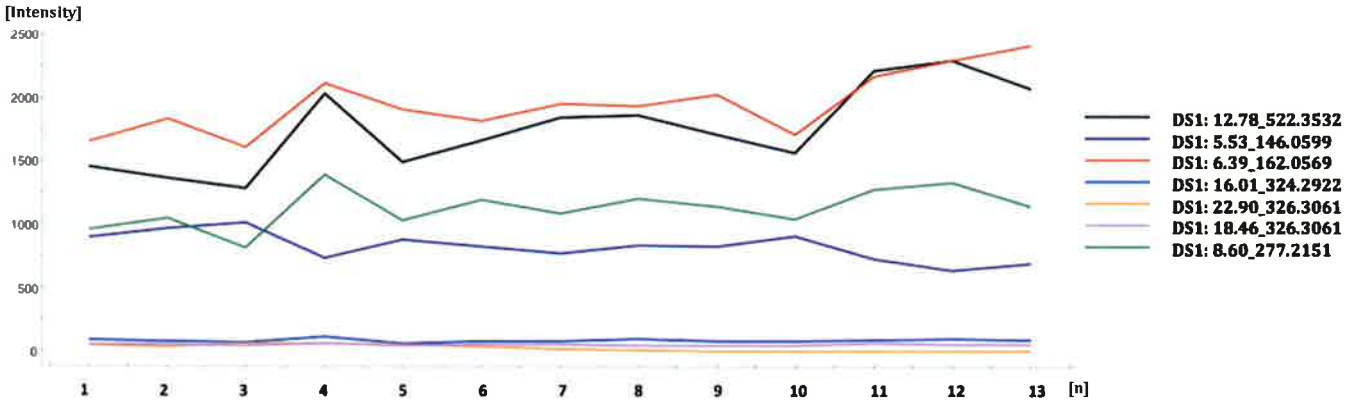


***Figure 10**: Intensity fluctuation of the internal standard (6.39_162.0569) and six randomly selected peaks (RT_m/z) collected by MassLynx™ software in the thirteen QC samples analyzed during five days with an interval of approximately ten hours*

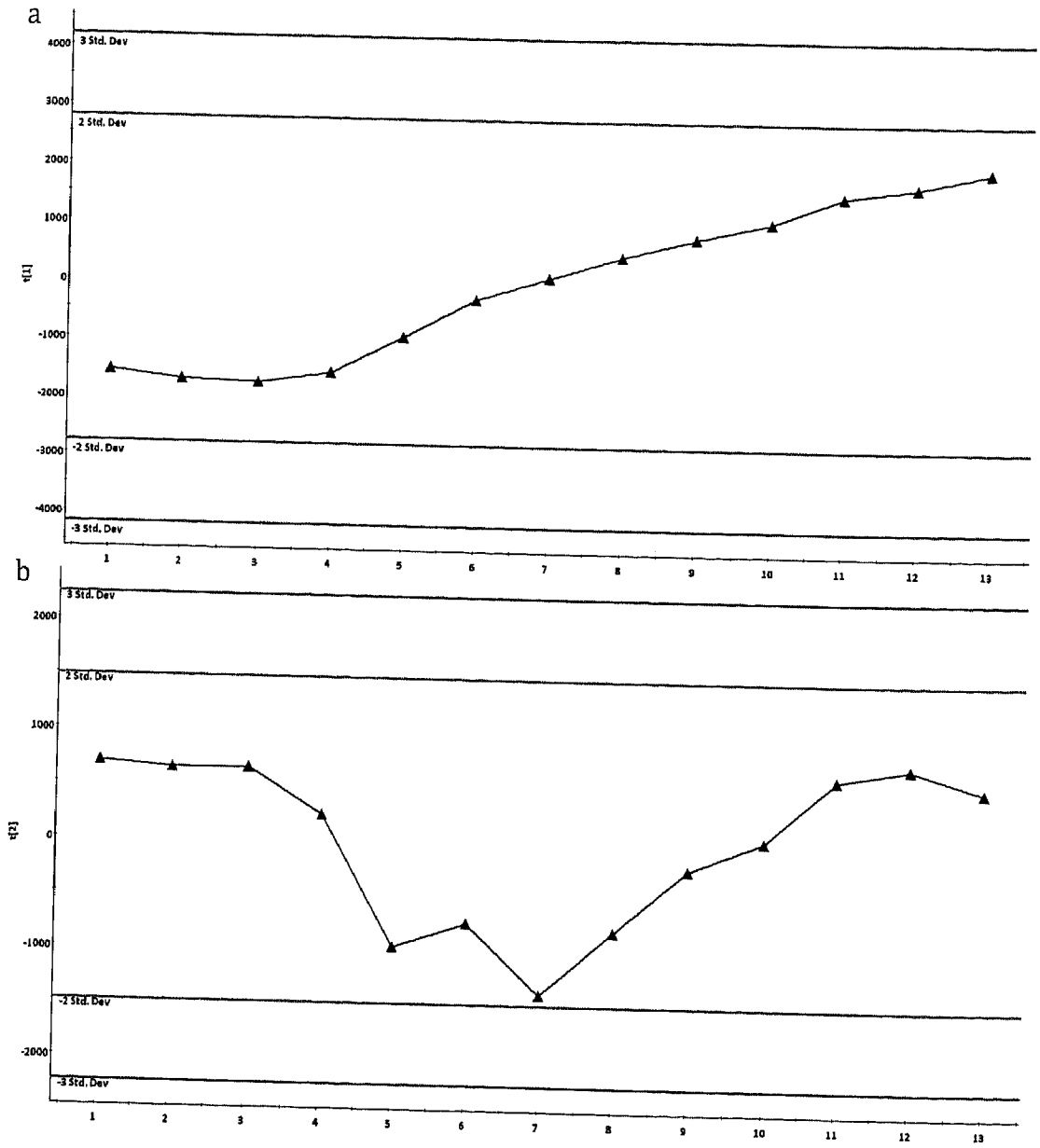


Figure 11: *PCA first (a) and second (b) component for the QC samples versus time. Component 1 R2X [1] = 0.4461, Component 2 R2X [2] = 0.09747*

It is clearly visible that the first component describes a time-dependent variation. However, looking at PC2 and PC3 (6.8%) we can conclude that the measurements are reproducible, as the QCs are clustering tightly together in the middle of the Score-Plot (*Figure 13*). As already stated, signal intensity was the most significant source of variability rather than retention time or changes in mass accuracy. With a matrix containing all of the chromatographic noise, it is not surprising that the PC1 describes the changes over time. On the basis of these results, it seems reasonable to suggest that for the type of samples studied here, a fairly strict acceptance criterion must be applied to candidate markers. Evidence of high variability within the body of the run, would constitute a significant reason for concern. Sample data obtained either side of that QC

cannot be considered to be reliable, and might indeed require the whole run to be repeated. Here, we don't find any indication which would ratify that this is the case in our experiment. However, the matrix contains a lot of noise that could be responsible for a high false discovery rate. Therefore, we concluded that the use of matrix B is more appropriate for the subsequent selection of marker candidates. Matrix A is used as a safety net to ensure that we don't miss any relevant markers present in the noise.

Hotelling's T^2 Range is a multivariate generalization of Student's t-distribution and by defining the "normal" area in the score plot, simplifies the identification of strong outliers. *Figure 14* shows the summary of the 12 scores. The larger the distance, the more extreme is the sample. If the sample is above the green or the red horizontal line, the probability that the sample is similar to others is less than 5% and 1% respectively. The first run of sample n866, s726, s729 and the third run of sample s841 and s8883 are significantly different from the other two runs. Therefore, they were excluded from any further analysis. The third run of sample s8676 was kept as it is not significantly different from the other two runs.

The distance to the model X (DModX) indicates how well an observation fits the PCA model. It is a summary of the unexplained variance in the model space (i.e. the noise in each sample). A value for DModX can be calculated for each observation; based on considering the elements of the residual Matrix and summarizing these row by row. These values can be plotted in a control chart where the maximum tolerable distance (Dcrit) for the data set is given. Moderate outliers have values larger than Dcrit. Samples well above the red line are significantly different from the others (Eriksson et. al., 2006). The 114 measurements are plotted in *Figure 15*. It was concluded that the noise of the samples being outliers in the Hotelling's T^2 Range are not the cause of the differences, and therefore it is correct to exclude them.

vi. OPLS-DA

The first component includes all variation that differentiates the two groups. To simplify the visualization of OPLS-DA, we used the S-plot. The S-plot combines the modeled covariance and modeled correlation from the OPLS-DA model and displays it as a scatter plot (*Figure 16*). The p(corr)1P-axis represent the reliability of each variable and varies between -1 and 1. Ideal marker candidates have high magnitude and high reliability. Unlikely cases have high magnitude with low reliability. Using the S-plot it is possible to get more information on the selected marker candidate, and subsequently look at the raw data, to ensure that the marker candidate is selected correctly. Because of this uncertainty it is not possible to specify a cut off, and the marker candidates were picked by hand. We collected those marker candidates with high contribution and high confidence. Altogether 107 marker candidates were selected, 39 for tainted pigs and 68 for non-tainted pigs. The procedure was repeated with matrix B and here 30 marker candidates for tainted pigs and 13 marker candidates for non-tainted pigs were selected. Doing this process, we ensured that we were not missing any relevant markers, which may be of low abundance. Indeed, they could be removed inadvertent when using the Noise elimination algorithm.

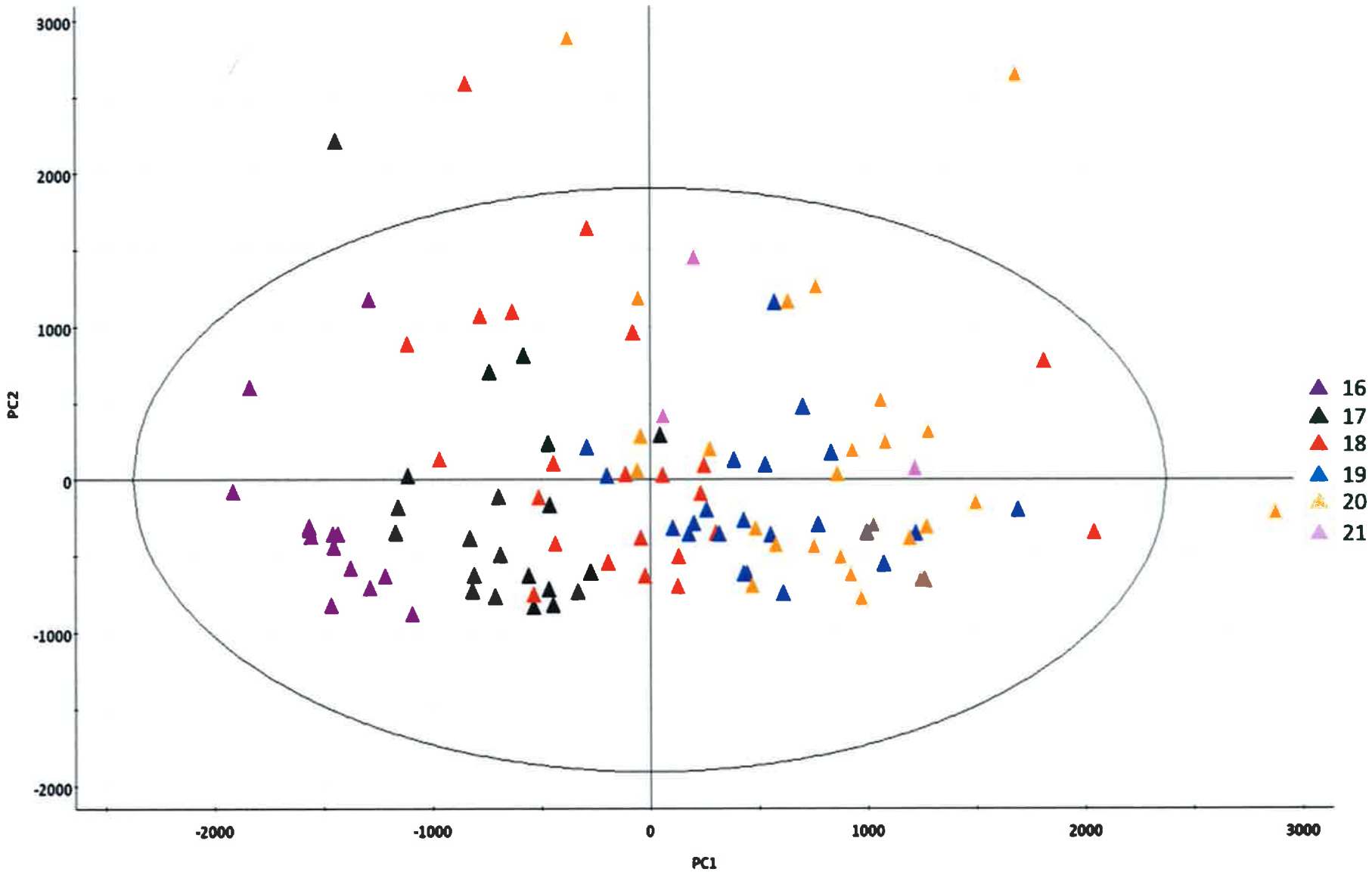


Figure 12: *Matrix A scaled using Pareto. The samples are colored according to dates of measurement (16th, 17th, 18th, 19th, 20th and 21st of November 2011) visualizing the time depending factor on PC1 (10,7%). After the first two days the system is stabilized.*

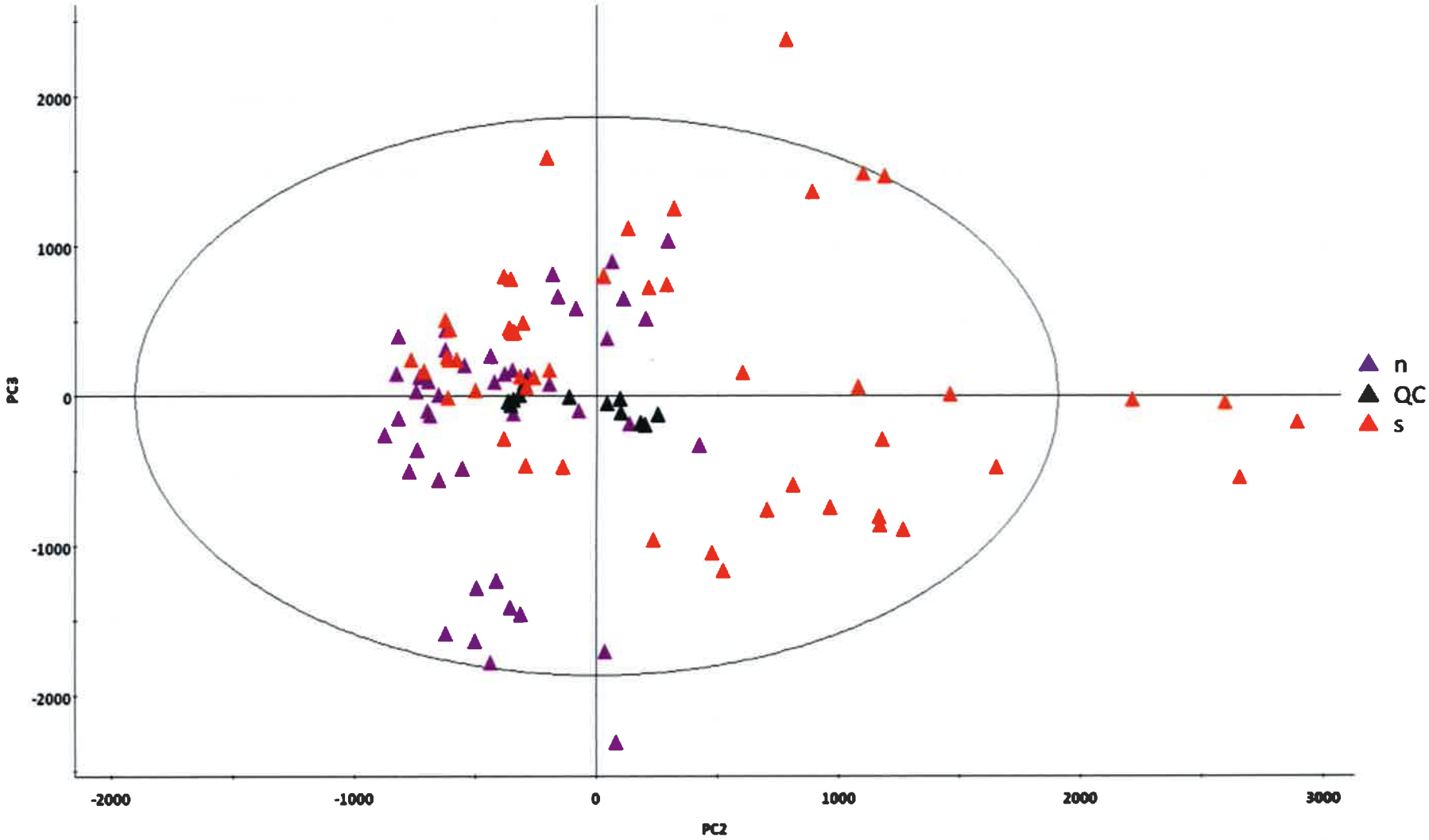


Figure 13: *Matrix A scaled using Pareto. Samples are colored according to sensory groups. The QCs are clustering in the middle of the score plot when plotting on PC2 and PC3.*

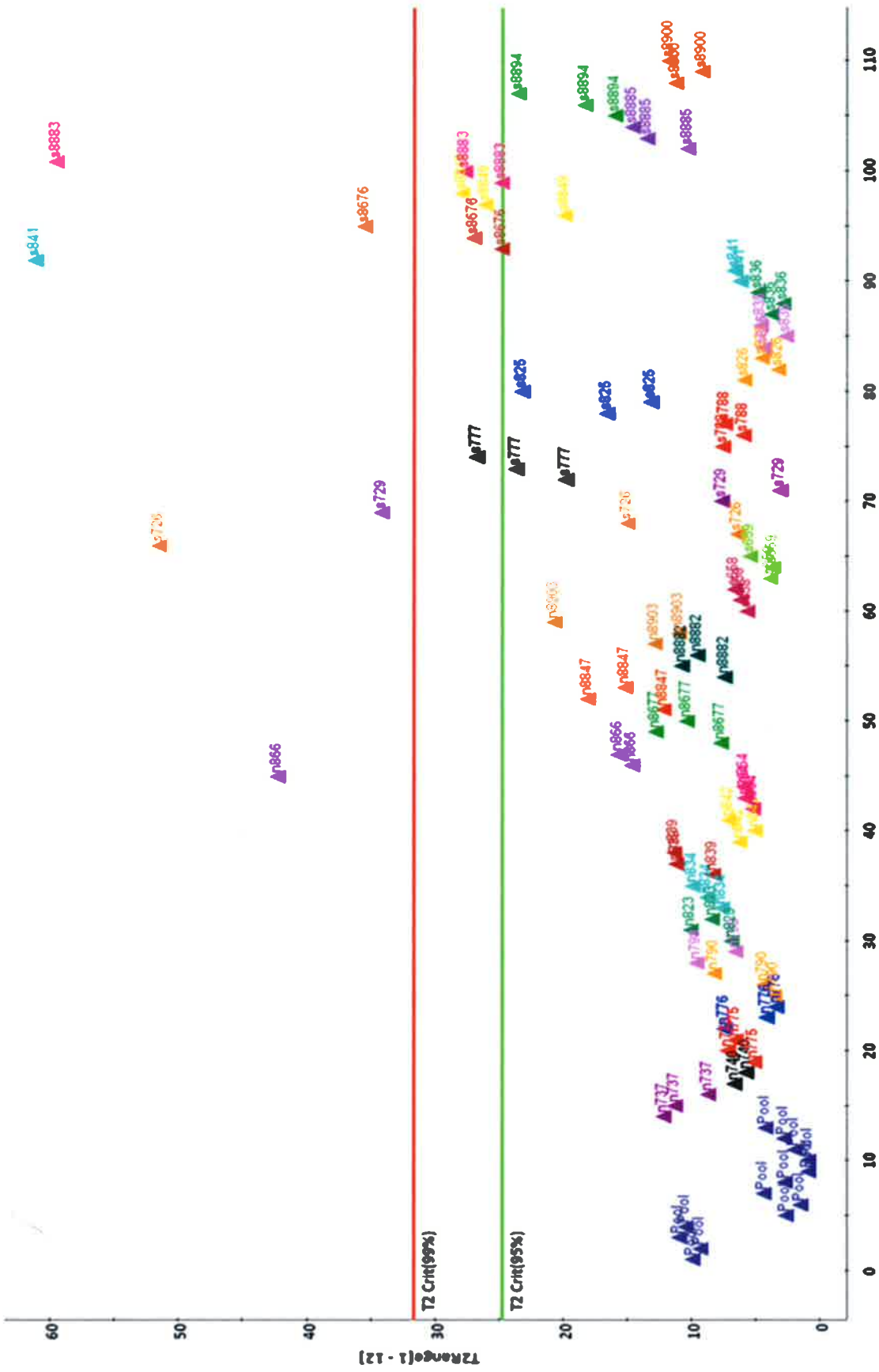


Figure 14: Hotellings T² Range [Component 1-12] colored by pig identity number.

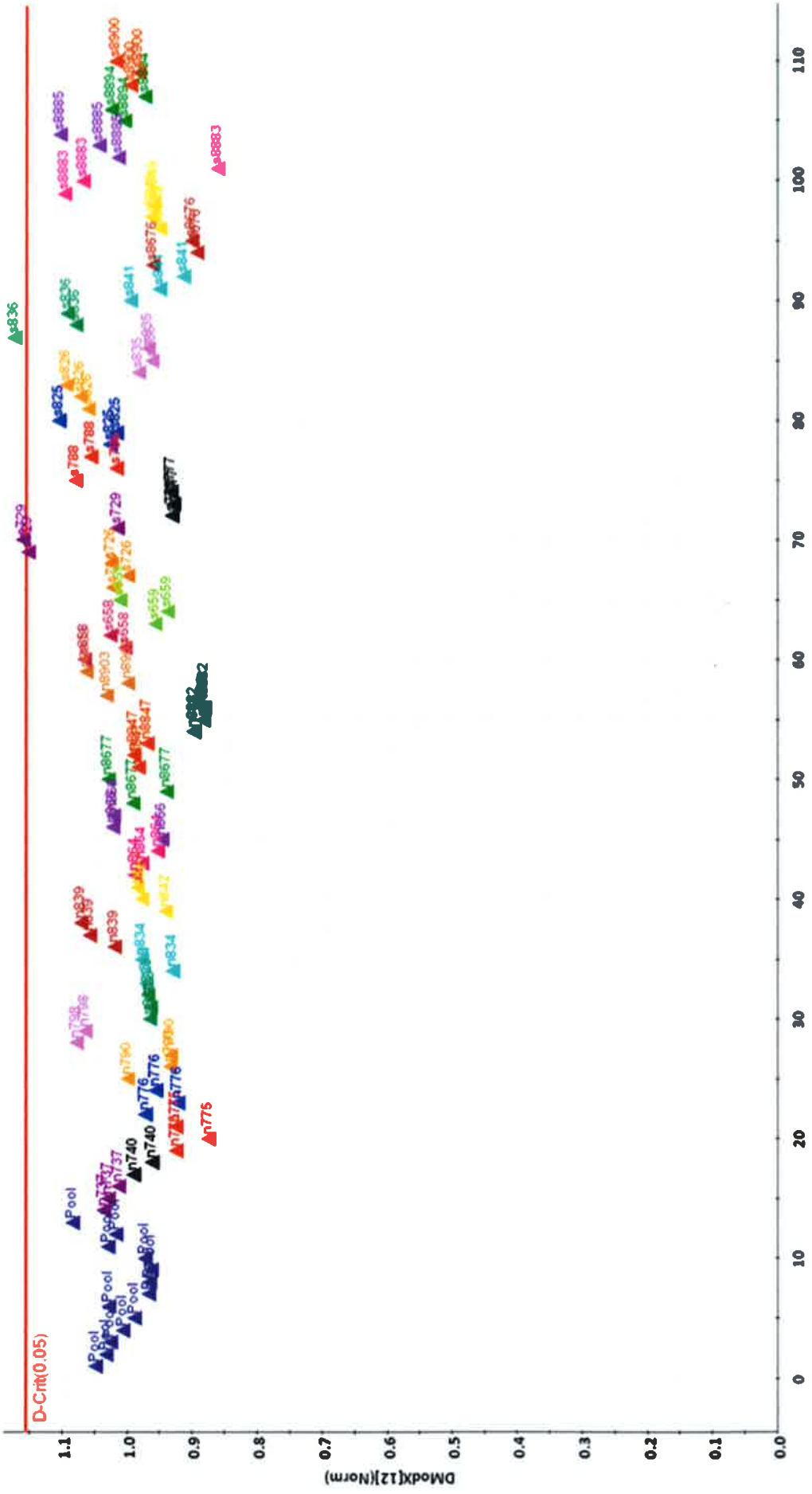


Figure 15: *Distance to model X colored following pig identification numbers*

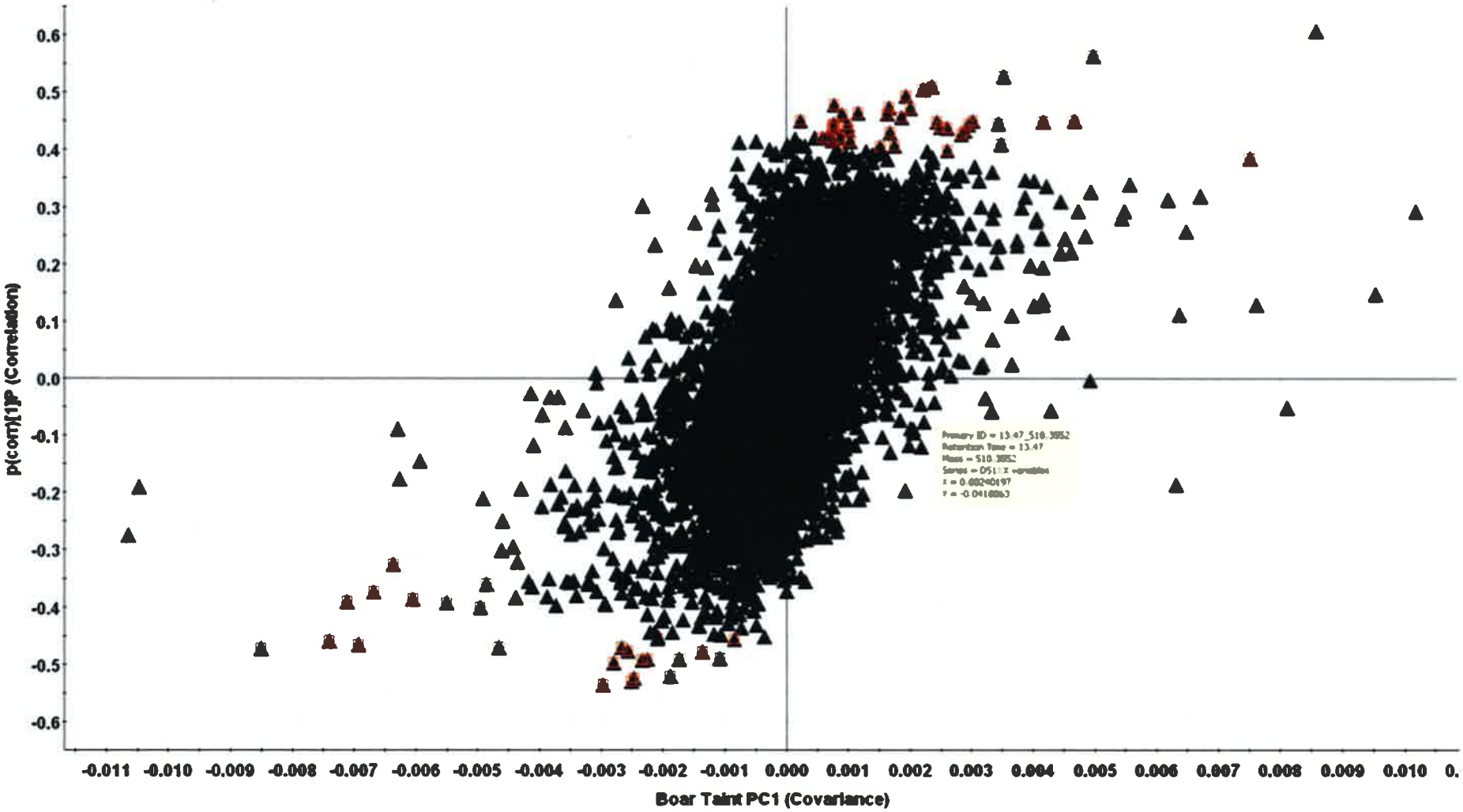


Figure 16: *S-plot showing the distribution of the 67569 markers. The selected markers are marked with a red square*

vii. ROC

Prior to filtering with the ROC algorithm we calculated the mean and the median of the three measurements for each pig originating from matrix A and B. Taking the mean and the median, we ensured that the time dependent variation was accounted for. This resulted in four new matrices:

a: mean of the sample without noise elimination (67569 variables, 33 subjects)
α: median of the sample without noise elimination (67569 variables, 33 subjects)
b: mean of the sample with noise elimination(8286 variables, 33 subjects)
β: median of the sample with noise elimination (8286 variables, 33 subjects)

Compounds achieving >80% for "observed within sample statistical sensitivity and specificity" were then selected using two different decision rules: "> than cutoff means tainted", collecting those marker which are significant for pigs achieving high score in the sensory panel, and "> than cutoff means non-tainted", collecting markers which are highly abundant in pigs without boar taint. The results are shown in *Table 7.*

Table 7: *Number of marker candidates within the different matrices using two different decision rules*

	Selected marker candidates	*Number of markers present in a/α and b/β (i.e. duplicates)*	*Sum of markers (mean+median after duplicate was removed)*
a *= mean per sample, only compounds included which had >=80 within sample accuracy, RULE: >=cutoff is non tainted*	*41*	*25*	*A: 67*
α *= median per sample, only compounds included which had >=80 within sample accuracy, RULE: >=cutoff is non tainted*	*51*	*25*	
b *= mean per sample, only compounds included which had >=80 within sample accuracy, RULE: >=cutoff is non tainted*	*9*	*3*	*B: 12*
β *= median per sample, only compounds included which had >=80 within sample accuracy, RULE: >=cutoff is non tainted*	*6*	*3*	
a *= mean per sample, only compounds included which had >=80 within sample accuracy, RULE: >cutoff is boar tainted*	*36*	*33*	*A: 40*
α *= median per sample, only compounds included which had >=80 within sample accuracy, RULE: >cutoff is boar tainted*	*37*	*33*	
b *= mean per sample, only compounds included which had >=80 within sample accuracy, RULE: >cutoff is boar tainted*	*17*	*15*	*B: 18*
β *= median per sample, only compounds included which had >=80 within sample accuracy, RULE: >cutoff is boar tainted*	*16*	*15*	

viii. FGCZ methode

All the mass spectra are added up to yield one single "master spectrum" where there is no misalignment and the sensitivity is increased. However, by this procedure the information about retention time is lost and isomers will be added up as one single m/z.

Detected masses are actually present and therefore there will be no zeros in this data matrix. Another benefit is the decrease of the process time with almost 66%. This method generated 14264 m/z values. 97 of these marker candidates have a p-value lower than 0.01 and were therefor considered as significant and kept for further analysis.

ix. Validating candidate markers with acceptance criteria

The Food and Drug Administration (FDA) published guidance on analytical method validation for bio-analytical methods in the industry (2001). In this guidance they are listing some criteria on acceptable degree of reproducibility for a particular maker candidate. Their criteria allow 5 QC samples out of 13 QC (i.e. 33%) to fall outside the acceptance criteria. The CV should not exceed 20%,

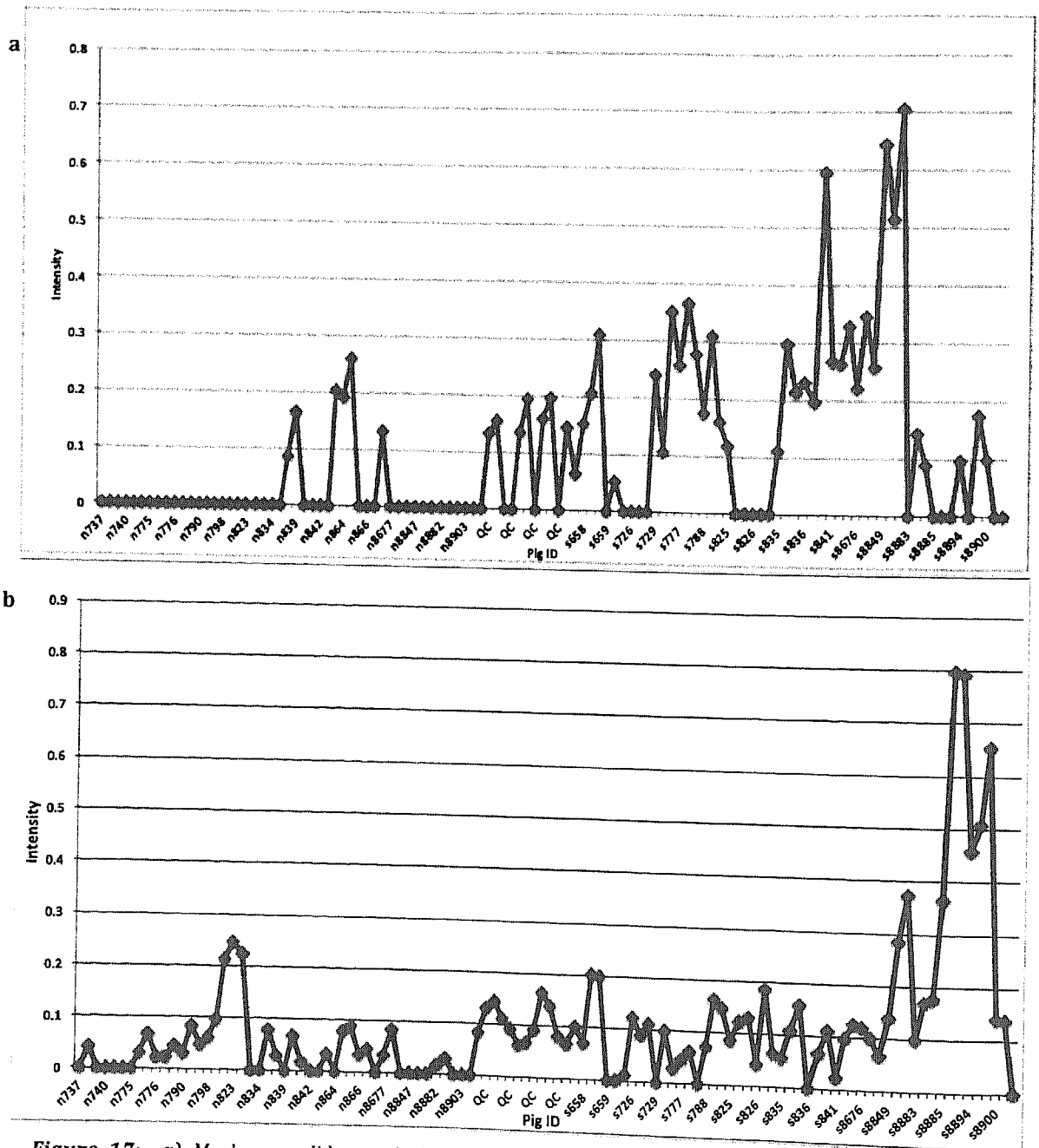


Figure 17: a) Marker candidate m/z 299.2379 at 7.88 min was selected by the OPLS-DA and ROC approach but invalidated because it was detected in just 8 of the 13 QC.
b) Marker candidate m/z 271. 2062 at 17.5617 was selected by OPLS-DA, ROC and the in House method and was detected in all of the QC. Eliminating 3 QC gave a CV of 19%.

which represents an acceptable degree of reproducibility for a particular maker candidate. Therefore, we examined the subset of marker candidates, in total 323, originated from our three different chemometrics methods, using such guidelines. Practical examples of how the QC data is used to validate the results are shown in *Figure 17*. In *"a"* the plot profiles of one mass selected to be significant for a strong tainted pig is shown. It is clearly seen that this mass is not a reliable marker, because it is just detected in 60% of the QC. Therefore it is not valid and should be excluded. 283 marker candidates were omitted because they did not fulfill the acceptance criteria. Over 90% of the marker candidates generated from the A matrix were not passing the acceptance criteria. However, we eliminated the risk to miss any important low abundant marker. The remaining 40 markers are presented in the *Table 8,* together with the information with which methods they were detected. We decided to do a tentative annotation of those markers listed in three or more of the result tables. The 16 candidate markers fulfilling this criterion were tested on out of sample accuracy with the naïve Bayes classifier.

x. Naïve Bayes

Naïve Bayes is an algorithm relying on an explicit probability model by allocating a probability to each class that corresponds to the product of the individual probabilities of every attribute value. The predicted class label then corresponds to the class with the greatest probability. The 16 selected marker candidates were tested with a naïve Bayes classifier to determine the predictive performance of unseen data (*Figure 18*). We used a 90/10 cross-validation to estimate the predictive out-of-sample accuracy and repeated this 1000 times. Naïve Bayes predicts with 90% accuracy if the pig is going to be tainted or not. Identical results were obtained looking at A mean/median or B mean/median.

xi. Marker annotation /Marker Candidates

The Metabolomics Standards Initiative (MSI) has published several guidelines (http://msi-workgroups.sourceforge.net/) for the publication of metabolomics experiments. One of these covers the "proposed minimum reporting standards for chemical analysis" that define confidence levels for the identification of compounds, ranging from unidentified signals at level 4, to level 1 for a rigorous identification based on independent measurements of authentic standards.

4) Unknown compounds: Although unidentified or unclassified these metabolites can still be differentiated based upon spectral data, thus enabling relative quantifications.

3) Putatively characterized compound classes: Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class.

2) Putatively annotated compounds: Without chemical reference standards, based solely upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries.

1) Identified compounds: A minimum of two independent and orthogonal types of data relative to an authentic standard analyzed under identical experimental conditions. In MS-based techniques this could include: retention time/index and mass spectrum, or accurate mass and tandem MS.

Table 8: *A list of 40 markers passing the acceptance criteria stated by the FDA.*

RT	m/z	Marker for	A ROC	B ROC	A OPLS-DA	B OPLS-DA	"FGCZ" t-Test
13.6368	111.1181	n			X	X	
5.5302	146.0599	n			X		X
13.6399	185.1539	n			X	X	
13.3974	187.149	n			X	X	
13.6232	193.16	n			X	X	
8.603	219.1751	s			X	X	
13.9004	225.2229	n			X	X	
8.603	235.1687	s	X	X	X	X	
5.5221	244.1185	s		X		X	
13.6282	245.2257	n			X	X	
5.5195	261.1462	s			X	X	
13.6314	263.2392	n			X	X	
17.5613	271.2062	s	X	X	X	X	X
20.1881	273.2215	s	X	X	X	X	X
8.6047	277.2151	s			X	X	
13.6326	281.2462	n			X	X	
8.8091	287.2024	s	X		X	X	X
10.6721	287.2053	s	X		X		X
7.812	289.2169	s	X	X	X	X	X
8.605	295.2268	s			X	X	
8.6038	313.2396	s			X	X	
6.3871	317.212	n			X	X	
5.9234	322.2005	s				X	X
8.6033	335.2179	s	X		X	X	
5.9269	340.2125	s				X	X
6.396	350.2342	s			X	X	X
8.6013	351.1914	s			X	X	
5.9271	358.2226	s	X			X	X
6.3974	368.241	s				X	X
10.6415	374.3166	s			X		X
5.9239	380.206	s	X	X			X
6.3992	386.2546	s			X	X	
5.926	396.1774	s	X	X			X
5.7805	410.2505	s	X	X	X	X	X
7.51	410.3282	s	X	X	X		X
6.2767	414.3287	s			X		X
5.782	424.2302	s	X	X			X
5.8096	426.247	s	X	X			X
12.6437	462.2974	n	X	X		X	
5.6004	528.2956	n			X	X	

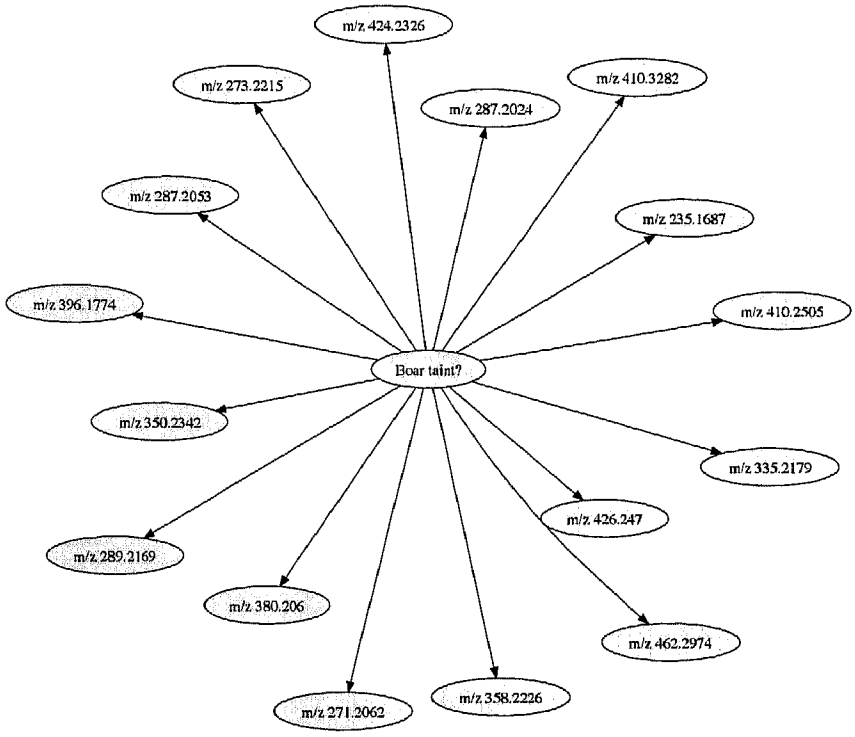


Figure 18: *The naïve Bayes classifier allocate a probability to each class that corresponds to the product of the individual probabilities of every value. Here the selected m/z values are displayed.*

With modern high-resolution mass spectrometers, the determination of the elemental composition for low to medium weight metabolites from accurate measurements is clearly feasible. However, this is only the first step of compound identification. Appropriate tools have long been a part of most vendor software's, and many of today's algorithms are known to perform well in practice (Bocker et. al., 2009). Yet, the American Society for Mass Spectrometry (ASMS) presented a survey in 2009, were the 600 participants revealed that the identification of compounds was still perceived as the bottleneck in the interpretation of metabolomics data (http://metabolomicssurvey.com/). This shows the difficulty still present in the annotation workflow. It is often stated that mass accuracy is the most important parameter for the determination of elemental compositions. Nevertheless, the number of possible elemental compositions increases exponentially with increasing ion mass, even with ultra-high resolution instruments (for example 33 different molecular formulas for mass 200 Da within a 1-ppm window). Therefore, restrictive criteria based on physicochemical rules and spectral information, as found in mass spectrometry textbooks (i.e. nitrogen rules, valence considerations, isotopic patterns), are required to remove irrelevant proposals. The most popular of these chemical rules is the nitrogen rule that states that odd nominal molecular mass compounds contain an odd number of nitrogen atoms (McLafferty, 1993).

The MarkerLynx™ elemental composition (EC) function yielded up to twenty-five possible elemental compositions for the 16 m/z values, sorted after their accuracy in mDa. The nitrogen rule has been automatically applied. Isotope peaks were eliminated with the MarkerLynx™ Software. Adducts and fragments were treated as separate features. Adduct formation occurs during ionization and each analyte present in the samples may generate multiple adduct ions. Accordingly, different adducts of the same metabolite co-elute chromatographically. In positive ion mode LC–MS, quantitation is typically based on $[M+H]^+$; however, one may also see $[M+Na]^+$, $[M+K]^+$ and $[M+NH_4]^+$. For a more comprehensive list of ionization adducts, see Crutchfield et. al. (2010). The proposed EC was crosschecked with the isotopic pattern in the raw data. *Figure 19* shows the isotopic distribution of $C_{19}H_{28}O$. The best fit was subsequently used to carry out an online library search. We also performed an exact mass search in the NIST library (m/z ± 10 ppm), where also the fragment pattern is considered. The results are presented in *Table 9.*

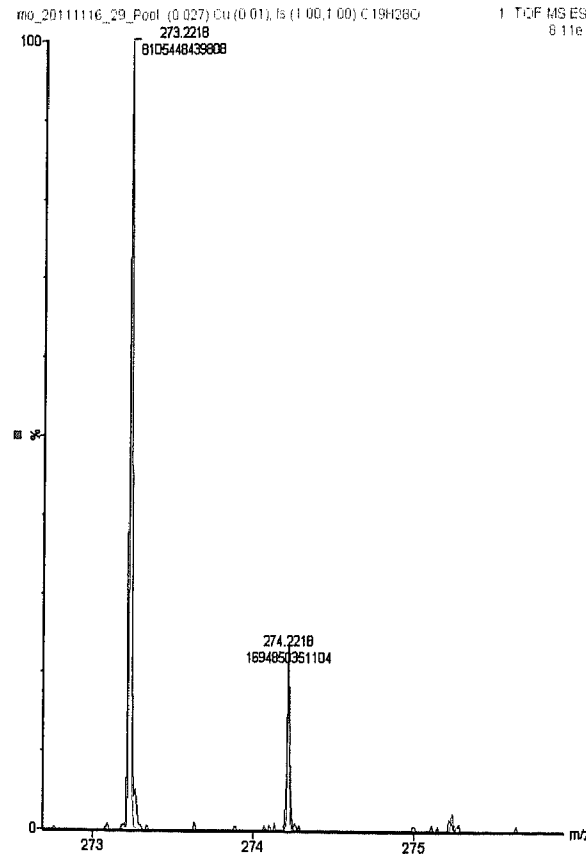


Figure 19: *Isotopic distribution of $C_{19}H_{28}O$. The red line is the distribution in the raw data. The purple line is the calculated theoretical distribution.*

Once a molecular structure is proposed and the isotopic distribution inspected, its likely hood can be further evaluated. This is possible by an interpretation of fragments found at the MS^E stage. A difficulty is the superposition of fragments from all co-eluting compounds, including background ions. The assignment to precursor and product ions was therefor possible just for those compounds for which the fragments were already known. *Figure 20* shows the $[M+H]^+$ adduct in the low (MS1) and high (MS2) energy spectrum and the most abundant fragment from testosterone. Taking its isotopic and fragmentation pattern into account, compound reached the confidence level 2 for identification.

Table 9: *EC generated with MarkerLynx™ and through exact mass search within NIST. If EC is inconsistent both results are listed. Different annotation possibilities are listed starting with the best hit.*

m/z	Elemental Composition	MarkerLynx (Systematic name)	NIST (Systematic name)
235.1687	C15H22O	1) 4-methylphenyl octanoate 2 2) 2-Methyl-4-phenyl-2-butanyl 2-methylpropanoate 3) 3-Phenylpropyl hexanoate	1) Benzyl octanoate 2)(3R,3a'R,4'S,7a'R)-3a',4'-Dimethyl-4-methylenedecahydrospiro[furan-3,2'-inden]-2-one 3) Hexyl 3-phenylpropanoate
271.2062	C19H26O	1) 10,17-Dimethylgona-4,13(17)-dien-3-one 2) Androsta-4,16-dien-3-one 3) Androsta-3,5-dien-7-one	No hits
273.2215	C19H28O	1)(5α)-Androst-16-en-3-one (*Androstenone*)	1)(5α)-Androst-16-en-3-one (*Androstenone*)
287.2024	C19H26O2	1) Androst-4-ene-3,17-dione 2)(17β)-17-Hydroxyandrosta-1,4-dien-3-one	1) Androst-4-ene-3,17-dione
287.2053	C19H26O2	1) Androst-4-ene-3,17-dione 2) (17β)-17-Hydroxyandrosta-1,4-dien-3-one	1) Androst-4-ene-3,17-dione
289.2169	C19H28O2	1)(17β)-17-Hydroxyandrost-4-en-3-one (*Testosterone*)	1)(17β)-17-Hydroxyandrost-4-en-3-one (*Testosterone*)
335.2179	C20H30O4	1)(2E,4E,6E,8E)-2,4,6,8-Icosatetraenedioic acid	1) (Z)-7-({1R,2S)-2-((S,E)-3-hydroxyoct-1-enyl)-5-oxocyclopent-3-enyl)hept-5-enoic acid (*Prostaglandin A2*) 2)(5Z)-9,15-Dioxoprosta-5,10-dien-1-oic acid 2)(5Z,13E,15S)-15-Hydroxy-9-oxoprosta-5,10,13-trien-1-oic acid (*dhk-PGA2*)
350.2342	C20H31NO4	No hits	1) L-proline,N-furoyl-2)-,decyl ester 2) L-valine, N-(2-methoxybenzoyl)-, heptyl ester
358.2226	C4H31O4 C15H27N5O5	1) 7a-(2-Hydroxy-2-propanyl)-3-(2-methyl-2-propanyl)-1-oxotetrahydro-1H-pyrrolo[1,2-c][1,3]oxazol-6-yl 3-hydroxy-3-methylbutanoate	1) 6-amino-2-[[1-(2,4-diamino-4-oxo-butanoyl)pyrrolidine-2-carbonyl] amino]hexanoic acid 2) 6-amino-2-[[4-amino-4-oxo-2-(pyrrolidine-2-carbonylamino)butanoyl]amino]hex anoic acid
380.206	C20H30NO6	No hits	No hits
396.1774	C27H26NS	No hits	No hits
410.2505	C29H32NO	No hits	No hits
410.3282	C24H44NO4	1) 1-Allyl 2-pentadecyl 1,2-pyrrolidinedicarboxylate 2) 1-Allyl 2-pentadecyl 1,2-pyrrolidinedicarboxylate	1) 1-Allyl 2-pentadecyl 1,2-pyrrolidinedicarboxylate
424.2302	C19H38NO7P C22H33NO7	1) (2R)-1-{[(2-Aminoethoxy) (hydroxy) phosphoryl]oxy}-3-hydroxy-2-propanyl (9Z)-9-tetradecenoate	1) 3'-acetylechiupinine/3'-acetylmyoscorpine
426.247	C19H40NO7P	1) 1-tetradecanoyl-sn-glycero-3-phosphoethanolamine	No hits
462.2974	C20H39N5O7	No hits	No hits

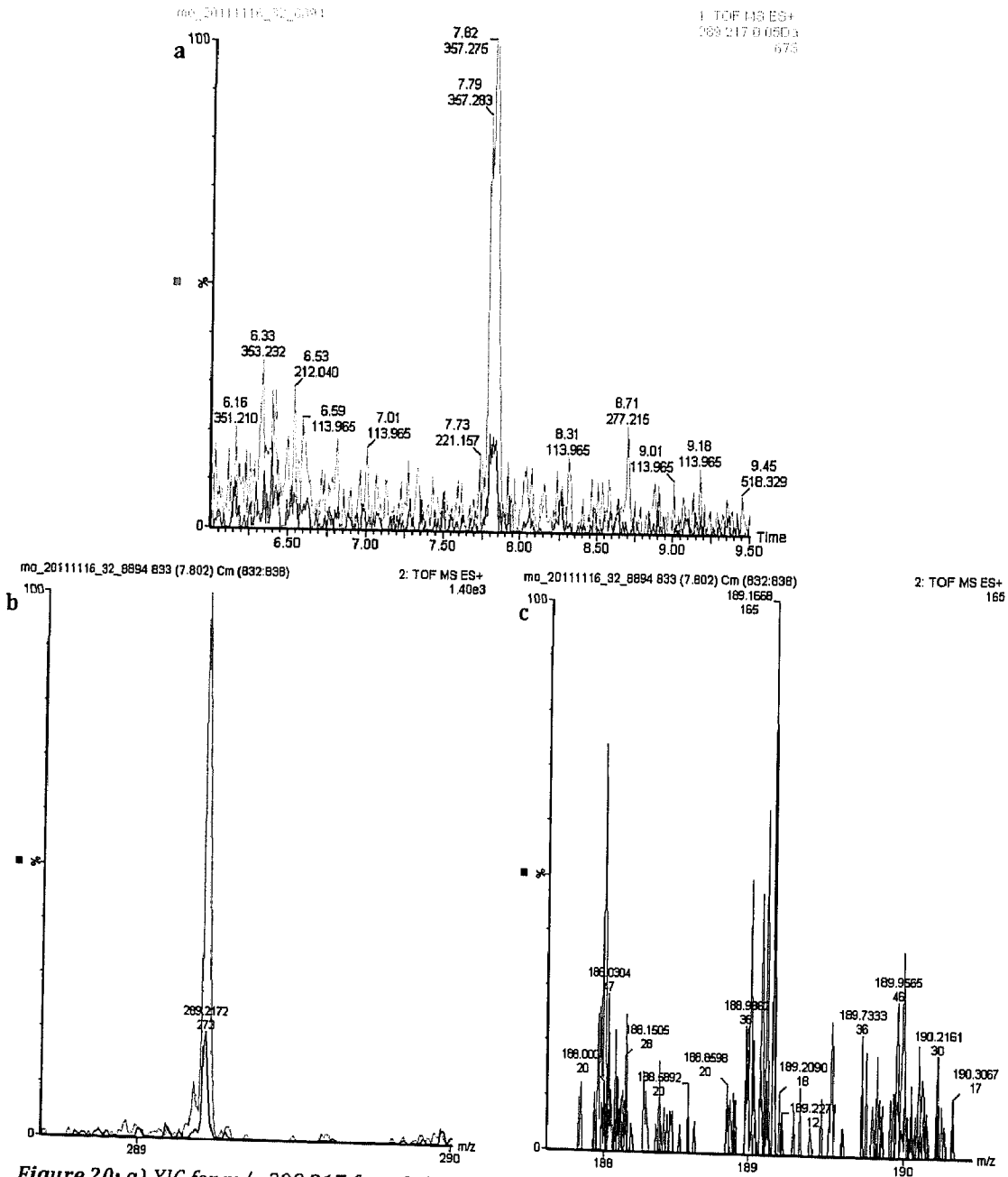


Figure 20: *a) XIC for m/z 298.217 from MS1 (yellow) and MS2 (purple)*
b) Spectra of M+H⁺ adduct of testosterone m/z 289.217 in MS1 (purple) and MS2 (black)
c) Spectra of the main fragment of testosterone m/z 189.1668 in MS1 (purple) and MS2 (black)

Another very well known candidate was AND, also reaching the MSI confidence level 2 for identification. The relative quantification of AND was comparable with the measurements at ALP. These results prove the feasibility of our method to detect relevant biological markers for boar taint. Another interesting suggestion, for the mass 335.2179 was prostaglandin A2. Prostaglandins are a group of lipid compounds that are derived enzymatically from fatty acids. Prostaglandin A2 reached the confidence level 3 for identification. A fragmentation pattern for Prostaglandin A2 was available in ESI^-, which can't be used for ESI^+ data. For the

most suggested annotations, a search within the NIST library revealed no spectral hits for ESI^+. Therefore, annotations were not possible with the data obtained within this study. Further experiments with a targeted approach, where the precursor is filtered and subsequently fragmented in an MS/MS data accusation mode, is necessary. This in order to obtain fragments from the particular m/z, and the same time avoiding superposition of co-eluting compounds and background ions.

Discussion

The pigs used in this study are representing the male pig population in Switzerland, fattening entire males (ENT), surgical castrated males (CAS) and immunocastrated males (IMP) (Improvac® is approved for the Swiss market since 2007). A classification based on AND, SK and ID concentrations in the adipose tissue, resulted in 27% of misclassified pigs (9 out of 33 pigs were misclassified, see *Table 6*). Using the method presented within this work, we reached an out of sample predictive performance of 90%, which means a decrease of misclassification of 14%. *Figure 21* shows a PCA-plot with the mean of the 16 selected marker candidates. The nontainted pigs are clustering very tightly with two exceptions. The strong tainted pigs seem to have a bigger diversity in the intensities of the above-mentioned markers. This leads to a broader distribution in the PCA-plot. Another possibility to visualize the data, is coloring the samples according to pig gender group (i.e. ENT, CAS, IMP) as shown in *Figure 22*. This representation unveils that also found markers correlate to some extent with pig gender. We conclude that for future metabolomics studies it is recommendable to consider only entire mal pigs. Most notably representatives of European farmers, the meat industry, retailers, scientists, veterinarians and animal welfare non-governmental organizations committed to a plan to voluntarily end surgical castration of pigs in Europe by 1 January 2018.

NanoUPLC®-HDMS™ nontargeted metabolomics in positive ionization mode has been applied to fat samples for the assessment of the presence or the absence of some metabolites in tainted pig carcasses. The comprehensive metabolomics approach described in this thesis enabled us to examine small molecules in fat extracts. Through chemometrics models, a selection of 16 compounds could be identified and putatively annotated. These markers also showed a respectable out off samples accuracy of 90%. All the different parts of the metabolomics workflow should be of high quality in order to be successful. Via a combination of test mixtures of known compounds and a pooled fat sample "QC", it has been possible to demonstrate that the nanoUPLC®-HDMS™ system is suitable for sample analysis. Using the data from the QC samples we were able to identify the factors that contributed to non-reproducibility between runs and to put control measurements in place. Variability in both mass accuracy and retention time could be neglected. However, signal intensity had a major effect on reproducibility. In particular, the variability of lower intensity peaks was significantly higher than those of higher intensity.

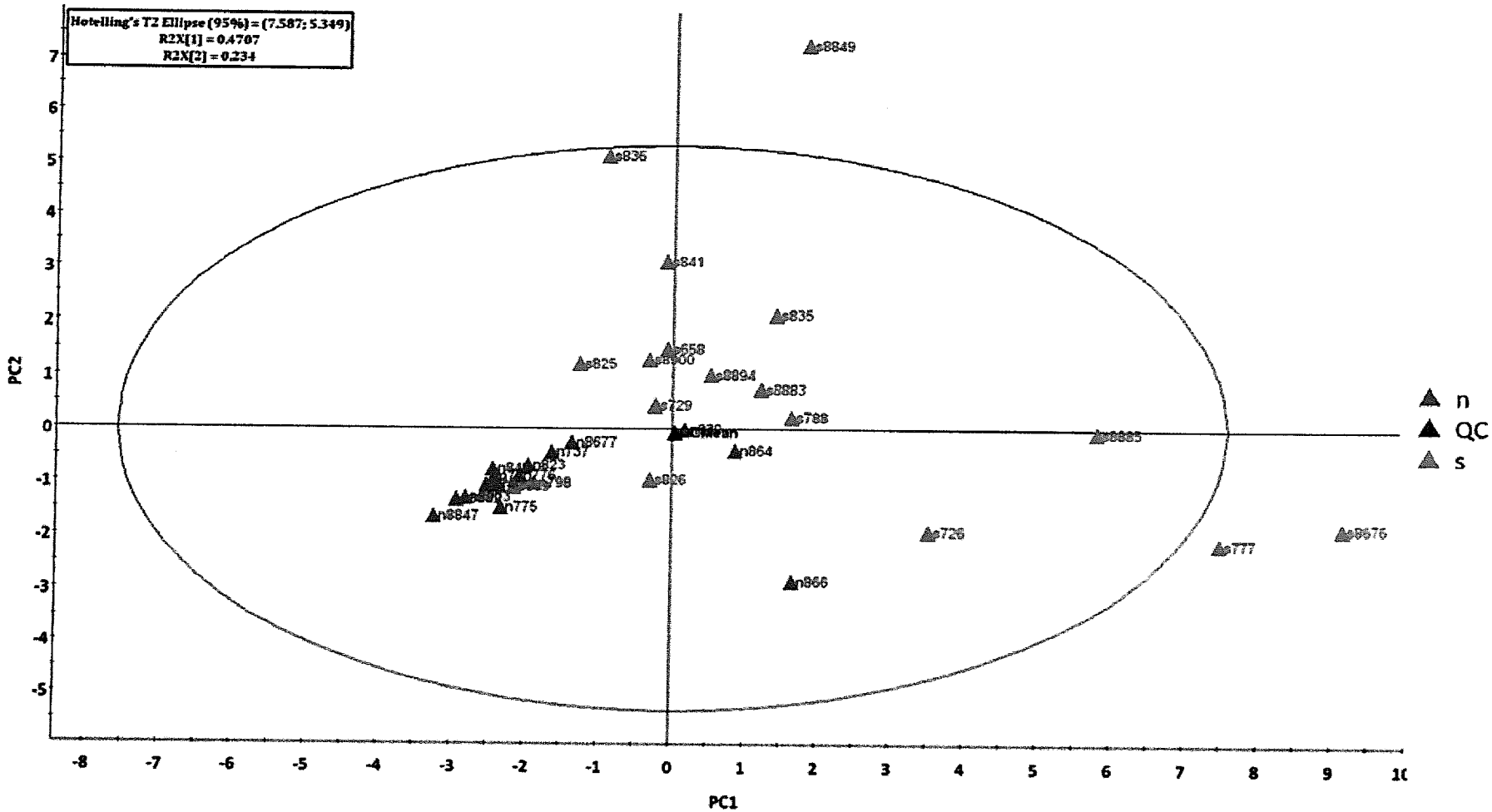


Figure 21: *By taking the mean of all pigs and just consider the 16 selected markers the PCA shows clustering (UV scaling). The* ***n*** *pigs are very homogenous, were as the* ***s*** *pigs shows more diversity.*

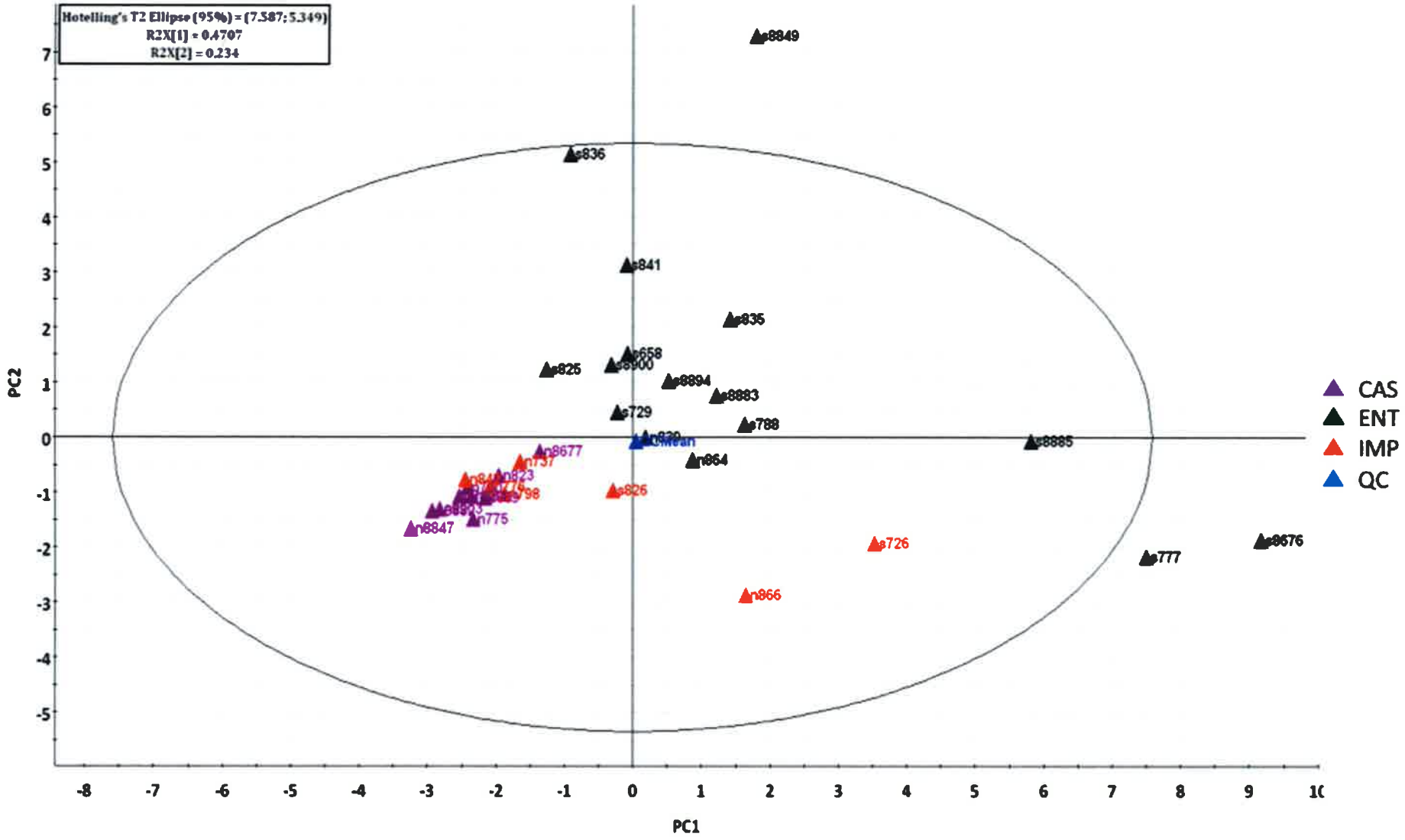


Figure 20: *By taking the mean of all pigs and just consider the 16 selected markers the PCA shows a clear clustering (UV scaling). The CAS pigs are very homogenous, whereas the ENT pigs showing more diversity. The IMP animals are an intermediate group sharing a metabolic pattern with both CAS and ENT.*

It is practically impossible to measure simultaneously the levels of all metabolites in the biological sample with a single analytical platform. The reason is that metabolites are biochemically diverse and can cover a dynamic range of over 10 orders of magnitude in concentration. Therefore, a single extraction and detection method for all metabolites from biological matrices is impracticable. Yet it is a first step to gain more knowledge and set up a workflow. An interesting complement to the present metabolomics study would be measurement in ESI-mode or a chromatographic method suitable for polar metabolites, which would add more knowledge as to the metabolites present within the adipose tissue of a pig. Using this platform, we have encountered many promising marker candidates. However, identification was often not possible. The reasons being very low intensity peaks, co-fragmentation and ambiguous spectra, as well as in some cases complex spectra likely corresponding to modified or uncommon adducts. As of now, availability of exhaustive metabolites libraries reflecting the whole diverse spectrum of metabolites and their modifications is far from reality. As a result, assignment of metabolites to all or even most metabolite profiles in a nontargeted screening continues to be a challenging task (Baker, 2011). The appearance of unidentified data is therefore a common observation in such experiments. Therefore, when exact identifications are not available, functional label annotations for unidentified peaks may be a helpful intermediate step, both as part of data analysis, as well as a guide towards the further analytical steps to identify the compounds (Broadhurst and Kell, 2006). Here, we address this challenge by creating a putative elemental composition and comparing its isotopic distribution with the isotopic distribution in the raw data of the selected marker.

The increasing emphasis on data quality in systems biology research presents great practical difficulties: the requirement for the simultaneous measurement of multiple variables in complex samples in which the identity of many of the components is unknown. This makes nontargeted metabolomics studies often laborious, tedious and costly. However, in case where targeted approaches using existing knowledge are not sufficient, the nontargeted approach is a valuable tool to gain new knowledge. New technological development in analytical chemistry, chemometrics and extension of MS libraries will increase the value of metabolomics studies even further (Dettmer et. al., 2006). Generalizability, sometimes called 'external validity', is a separate problem. It concerns the results of the comparison of two groups. The generalizability of a study depends on the characteristics of the subjects and how they are selected, regarding age, gender, morbidity, diet, environment, etc. Initial studies have limited generalizability but are satisfactory to establish a 'proof of principle' and provide the basis for larger and potentially more expensive studies that assess broader generalizability. Strong internal validity is critically important for initial studies, to avoid wasted effort and costs in follow up research (De Vos et. al., 2007).

In future we propose that this nontargeted metabolomics approach, may provide the basis of a population-screening tool to select pigs according to their suitability as meet for the food industry. Furthermore, this approach could allow a breeding selection depending on the phenotype. In particular, the nontargeted metabolomics approach has the potential to provide new biomarkers that are predictive of individual responses

Literature

Ampuero, S., & Bee, G. (2006). The potential to detect boar tainted carcasses by using an electronic nose based on mass spectrometry. *Acta Veterinaria Scandinavica, 48.*

Andersson, K., Schaub, A., Lundstrom, K., Thomke, S., & Hansson, I. (1997). The effects of feeding system, lysine level and gilt contact on performance, skatole levels and economy of entire male pigs. *Livestock Production Science, 51*(1-3), 131-140.

Annor-Frempong, I. E., Nute, G. R., Wood, J. D., Whittington, F. W., & West, A. (1998). The measurement of the responses to different odour intensities of 'boar taint' using a sensory panel and an electronic nose. *Meat Science, 50*(2), 139-151.

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

Babol, J., & Squires, E. J. (1995). QUALITY OF MEAT FROM ENTIRE MALE PIGS. *Food Research International, 28*(3), 201-212.

Baker, M. (2011). Metabolomics: from small molecules to big ideas. *Nature Methods, 8*(2), 117-121.

Baker, S. G. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute, 95*(7), 511-515.

Bejerholm, C., & Gade, P. B. (1993). THE RELATIONSHIP BETWEEN SKATOLE ANDROSTENONE AND ODOR FLAVOR OF MEAT FROM ENTIRE MALE PIGS. In: M. Bonneau, *Measurement and Prevention of Boar Taint in Entire Male Pigs*, vol. 60 (pp. 75-79). Paris: Inst Natl Recherche Agronomique.

Bicalho, B., David, F., Rumplel, K., Kindt, E., & Sandra, P. (2008). Creating a fatty acid methyl ester database for lipid profiling in a single drop of human blood using high resolution capillary gas chromatography and mass spectrometry. *Journal of Chromatography A, 1211*(1-2), 120-128.

Bligh, E. G., & Dyer, W. J. (1959). A RAPID METHOD OF TOTAL LIPID EXTRACTION AND PURIFICATION. *Canadian Journal of Biochemistry and Physiology, 37*(8), 911-917.

Bobeldijk, I., Hekman, M., de Vries-van der Weij, J., Coulier, L., Ramaker, R., Kleemann, R., Kooistra, T., Rubingh, C., Freidig, A., & Verheij, E. (2008). Quantitative profiling of bile acids in biofluids and tissues based on accurate mass high resolution LC-FF-MS: Compound class targeting in a metabolomics workflow. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 871*(2), 306-313.

Bocker, S., Letzel, M. C., Liptak, Z., & Pervukhin, A. (2009). SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics, 25*(2), 218-224.

Bonneau, M. (1982). Compounds responsible for boar taint, with special emphasis on androstenone - a review. *Livestock Production Science, 9*(6), 687-705.

Bonneau, M., Walstra, P., Claudi-Magnussen, C., Kempster, A. J., Tornberg, E., Fischer, K., Diestre, A., Siret, F., Chevillon, P., Claus, R., Dijksterhuis, G., Punter, P., Matthews, K. R., Agerhem, H., Beague, M. P., Oliver, M. A., Gispert, M., Weiler, U., von Seth, G., Leask, H., Furnols, M. F. I., Homer, D. B., & Cook, G. L. (2000). An international study on the importance of androstenone and skatole for boar taint: IV. Simulation studies on consumer dissatisfaction with entire male pork and the effect of sorting carcasses on the slaughter line, main conclusions and recommendations. *Meat Science, 54*(3), 285-295.

Bristow, A. W. T. (2006). Accurate mass measurement for the determination of elemental formula - A tutorial. *Mass Spectrometry Reviews, 25*(1), 99-111.
Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics, 2*(4), 171-196.
Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. New York Academic Press.

Brooks, R. I., & Pearson, A. M. (1986). Steroid Hormone Pathways in the Pig, with Special Emphasis on Boar Odor: A Review. *Journal of Animal Science, 62*(3), 632-645.

Buescher, J. M., Czernik, D., Ewald, J. C., Sauer, U., & Zamboni, N. (2009). Cross-Platform Comparison of Methods for Quantitative Metabolomics of Primary Metabolism. *Analytical Chemistry, 81*(6), 2135-2143.

Buescher, J. M., Moco, S., Sauer, U., & Zamboni, N. (2010). Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectrometry Method for Fast and Robust Quantification of Anionic and Aromatic Metabolites. *Analytical Chemistry, 82*(11), 4403-4412.

Castro-Perez, J. M., Kamphorst, J., DeGroot, J., Lafeber, F., Goshawk, J., Yu, K., Shockcor, J. P., Vreeken, R. J., & Hankemeier, T. (2010). Comprehensive LC/MSE Lipidomic Analysis using a Shotgun Approach and Its Application to Biomarker Detection and Identification in Osteoarthritis Patients. *Journal of Proteome Research, 9*(5), 2377-2389.

Cevallos-Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science & Technology, 20*(11-12), 557-566.

Choi, H. K., Choi, Y. H., Verberne, M., Lefeber, A. W. M., Erkelens, C., & Verpoorte, R. (2004). Metabolic fingerprinting of wild type and transgenic tobacco plants by H-1 NMR and multivariate analysis technique. *Phytochemistry, 65*(7), 857-864.

Claus, R., Weiler, U., & Herzog, A. (1994). Physiological-aspects of androstenone and skatole formation in the boar- a review with experimental-data. *Meat Science, 38*(2), 289-305

Coulier, L., Tas, A., & Thissen, U. (2011). Food Metabolomics: Fact or Fiction? *Lc Gc Europe, 24*(2), 60-71

Craig, H. B., & Pearson, A. M. (1959). Some preliminary studies on sex odor in pork. *Journal of Animal Science, 18: 1557.*

Crutchfield, C. A., Lu, W. Y., Melamud, E., & Rabinowitz, J. D. (2010). MASS SPECTROMETRY-BASED METABOLOMICS OF YEAST. In: J. Weissman, C. Guthrie, & G. R. Fink, *Methods in Enzymology, Vol 470: Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis, 2nd Edition,* vol. 470 (pp. 393-426). San Diego: Elsevier Academic Press Inc.

de Kock, H. L., Heinze, P. H., Potgieter, C. M., Dijksterhuis, G. B., & Minnaar, A. (2001). Temporal aspects related to the perception of skatole and androstenone, the major boar odour compounds. *Meat Science, 57*(1), 61-70.

Deslandes, B., Gariépy, C., & Houde, A. (2001). Review of microbiological and biochemical effects of skatole on animal production. *Livestock Production Science, 71*(2-3), 193-200

Desmoulin, B., Dumont, B. L. and Jacquet, B. (1971) Le port male de race Large-White: aptitudes a la production de viande. Journees de la Recherche Porcine en France 3, 187-195.

Desmoulin, B., Bonneau, M. and Bourdon, D. (1974) Etude en bilan azote et composition corpor- elle des ports males entiers ou castrb de race Large White. Journees de la Recherche Porcine en France 6, 247-255.

Desmoulin, B., Aumaitre, A., & Peiniau, J. (1990). INFLUENCE OF WEIGHT AT 10 D AND AGE AT CASTRATION IN MALE PIGLETS ON GROWTH-RATE AND CARCASS QUALITY. *Annales De Zootechnie, 39*(3-4), 219-227

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews, 26*(1), 51-78.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*(2-3), 103-130.

Domon, B., & Aebersold, R. (2006). Mass Spectrometry and Protein Analysis. *Science, 312*(5771), 212-217.

Dunn, W. B. (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology, 5*(1).

Eriksson, L., Johansson, E., Ketteneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2006). *Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications.* Umeå: Umetrix Academy.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology, 48*(1-2), 155-171.

Folch, J., Lees, M., & Stanley, G. H. S. (1957). A SIMPLE METHOD FOR THE ISOLATION AND PURIFICATION OF TOTAL LIPIDES FROM ANIMAL TISSUES. *Journal of Biological Chemistry, 226*(1), 497-509.

Fortin, A., Friend, D. W., & Sarkar, N. K. (1983). A NOTE ON THE CARCASS COMPOSITION OF YORKSHIRE BOARS AND BARROWS. *Canadian Journal of Animal Science, 63*(3), 711-714.

Fowler, V. R., McWilliam, R., & Aitken, R. (1981). VOLUNTARY FEED-INTAKE OF BOARS, CASTRATES AND GILTS GIVEN DIETS OF DIFFERENT NUTRIENT DENSITY. Animal Production, 32(JUN), 357-357.

Gangl, E. T., Annan, M., Spooner, N., & Vouros, P. (2001). Reduction of Signal Suppression Effects in ESI-MS Using a Nanosplitting Device. *Analytical Chemistry, 73*(23), 5635-5644.

Garcia-Regueiro, J. A., & Diaz, I. (1989). EVALUATION OF THE CONTRIBUTION OF SKATOLE, INDOLE, ANDROSTENONE AND ANDROSTENOLS TO BOAR-TAINT IN BACK FAT OF PIGS BY HPLC AND CAPILLARY GAS-CHROMATOGRAPHY (CGC). *Meat Science, 25*(4), 307-316.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—Not So Stupid After All? *International Statistical Review, 69*(3), 385-398.

Hansen-Møller, J. (1994). RAPID HIGH-PERFORMANCE LIQUID-CHROMATOGRAPHIC METHOD FOR SIMULTANEOUS DETERMINATION OF ANDROSTENONE, SKATOLE AND INDOLE IN BACK FAT FROM PIGS. *Journal of Chromatography B-Biomedical Applications, 661*(2), 219-230.

Hansen, B. C., & Lewis, A. J. (1993). EFFECTS OF DIETARY-PROTEIN CONCENTRATION (CORN-SOYBEAN MEAL RATIO) ON THE PERFORMANCE AND CARCASS CHARACTERISTICS OF GROWING BOARS, BARROWS, AND GILTS - MATHEMATICAL DESCRIPTIONS. *Journal of Animal Science, 71*(8), 2122-2132.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(7), 498-520.

Idborg, H., Zamani, L., Edlund, P. O., Schuppe-Koistinen, I., & Jacobsson, S. P. (2005). Metabolic fingerprinting of rat urine by LC/MS Part 1. Analysis by hydrophilic interaction liquid chromatography-electrospray ionization mass spectrometry. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 828*(1-2), 9-13.

Idborg, H., Zamani, L., Edlund, P. O., Schuppe-Koistinen, I., & Jacobsson, S. P. (2005). Metabolic fingerprinting of rat urine by LC/MS Part 2. Data pretreatment methods for handling of complex data. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 828*(1-2), 14-20.

Jonsson, P., Gullberg, J., Nordstrom, A., Kusano, M., Kowalczyk, M., Sjostrom, M., & Moritz, T. (2004). A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry, 76*(6), 1738-1745.

Katajamaa, M., Miettinen, J., & Oresic, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics, 22*(5), 634-636.

Katajamaa, M., & Oresic, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A, 1158*(1-2), 318-328.

Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics, 7*.

Koulman, A., Woffendin, G., Narayana, V. K., Welchman, H., Crone, C., & Volmer, D. A. (2009). High-resolution extracted ion chromatography, a new tool for metabolomics and lipidomics using a second-generation orbitrap mass spectrometer. *Rapid Communications in Mass Spectrometry, 23*(10), 1411-1418.

Krastanov, A. (2010). METABOLOMICS - THE STATE OF ART. *Biotechnology & Biotechnological Equipment, 24*(1), 1537-1543.

Lavine, B., & Workman, J. J. (2004). Chemometrics. *Analytical Chemistry, 76*(12), 3365-3371.

Lenz, E. M., & Wilson, I. D. (2007). Analytical strategies in metabonomics. *Journal of Proteome Research, 6*(2), 443-458.

Lutz, U., Lutz, R. W., & Lutz, W. K. (2006). Metabolic Profiling of Glucuronides in Human Urine by LC-MS/MS and Partial Least-Squares Discriminant Analysis for Classification and Prediction of Gender. *Analytical Chemistry, 78*(13), 4564-4571.

Malmfors, B., & Hansson, I. (1974). Incidence of boar taint in Swedish Landrace and Yorkshire boars. Livestock Production Science, 1(4), 411-420.

Martens, H., & Naes, T. (1991). *Multivariate Calibration*. New York: John Wiely & Sons Inc.

Masson, P., Alves, A. C., Ebbels, T. M. D., Nicholson, J. K., & Want, E. J. (2010). Optimization and Evaluation of Metabolite Extraction Protocols for Untargeted Metabolic Profiling of Liver Samples by UPLC-MS. *Analytical Chemistry, 82*(18), 7779-7786.

Mattila, I., Seppanen-Laakso, T., Suortti, T., & Oresic, M. (2008). Application of Lipidomics and Metabolomics to the Study of Adipose Tissue. In: K. Yang, *Methods in Molecular Biology*, vol. 456 (pp. 123-130).

Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A., & Schwudke, D. (2008). Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of Lipid Research, 49*(5), 1137-1146.

McLafferty, F. W. (1993). *Interpretion of Mass Spectra 4th ed.* New York: Wiely University Science Book

Moco, S., Bino, R. J., De Vos, R. C. H., & Vervoort, J. (2007). Metabolomics technologies and metabolite identification. *Trac-Trends in Analytical Chemistry, 26*(9), 855-866.

Mortensen, A. B., & Sorensen, S. E. (1984). Relationship between boar taint and skatole determined with a new analysis method *30th European Meeting of Meat Research Workers, Bristol.*

Neumann, S., & Bocker, S. (2010). Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules. *Analytical and Bioanalytical Chemistry, 398*(7-8), 2779-2788.

Nicholson, J. K., & Wilson, I. D. (2003). Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery, 2*(8), 668-676.

Nicholson, J. K., Connelly, J., Lindon, J. C., & Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery, 1*(2), 153-161.

Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). 'Metabonomics': understanding metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of NMR spectroscopic data. *Xenobiotica, 29*(11), 1181-1189.

Paik, M.-J., Moon, S.-M., Kim, K.-R., Choi, S., Ahn, Y.-H., & Lee, G. (2008). Target metabolic profiling analysis of free amino acids in plasma as EOC/TBDMS derivatives by GC-SIM-MS. *Biomedical Chromatography, 22*(4), 339-342.

Patterson, R. L. S. (1968). 5α-androst-16-ene-3-one: Compound responsible for taint in boar fat. *Journal of the Science of Food and Agriculture, 19*(1), 31-38.

Pauly, C., Spring, P., O'Doherty, J. V., Kragten, S. A., & Bee, G. (2008). Performances, meat quality and boar taint of castrates and entire male pigs fed a standard and a raw potato starch-enriched diet. *Animal, 2*(11), 1707-1715.

Pauly, C., Spring, P., O'Doherty, J. V., Kragten, S. A., & Bee, G. (2009). Growth performance, carcass characteristics and meat quality of group-penned surgically castrated, immunocastrated (Improvac (R)) and entire male pigs and individually penned entire male pigs. *Animal, 3*(7), 1057-1066.

Pauly, C., Spring-Staehli, P., O'Doherty, J. V., Kragten, S. A., Dubois, S., Messadène, J., & Bee, G. (2010). The effects of method of castration, rearing condition and diet on sensory quality of pork assessed by a trained panel. *Meat Science, 86*(2), 498-504.

Plumb, R., Castro-Perez, J., Granger, J., Beattie, I., Joncour, K., & Wright, A. (2004). Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry, 18*(19), 2331-2337.

Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., & Nicholson, J. K. (2006). UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry, 20*(13), 1989-1994.

Poste, G. (2011). Bring on the biomarkers. *Nature, 469*(7329), 156-157.

Prescott, J. H. D., & Lamming, G. E. (1967). The influence of castration on the growth of male pigs in relation to high levels of dietary protein. *Animal Science, 9*(04), 535-545.

Rius, M. A., Hortos, M., & Garcia-Regueiro, J. A. (2005). Influence of volatile compounds on the development of off-flavours in pig back fat samples classified with boar taint by a test panel. *Meat Science, 71*(4), 595-602.

Robertson, D. G. (2005). Metabonomics in toxicology: A review. *Toxicological Sciences, 85*(2), 809-822.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B. S., van Ommen, B., Pujos-Guillot, E., Verheij, E., Wishart, D., & Wopereis, S. (2009). Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics, 5*(4), 435-458.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry, 78*(3), 779-787.

Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., & Kohlbacher, O. (2008). OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinformatics, 9*(1), 163.

Swartz, M. E. (2005). UPLC™: An Introduction and Review. *Journal of Liquid Chromatography & Related Technologies, 28*(7-8), 1253-1263

Tolstikov, V. V., & Fiehn, O. (2002). Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry, 301*(2), 298-307.

Trauger, S. A., Kalisak, E., Kalisiak, J., Morita, H., Weinberg, M. V., Menon, A. L., Poole, F. L., II, Adams, M. W. W., & Siuzdak, G. (2008). Correlating the transcriptome, proteome, and metabolome in the environmental adaptation of a hyperthermophile. *Journal of Proteome Research, 7*(3), 1027-1035.

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

Trygg, J., Holmes, E., & Lundstedt, T. (2007). Chemometrics in metabonomics. *Journal of Proteome Research, 6*(2), 469-479.

Tuomola, M., Vahva, M., & Kallio, H. (1996). High-performance liquid chromatography determination of skatole and indole levels in pig serum, subcutaneous fat, and submaxillary salivary glands. *Journal of Agricultural and Food Chemistry, 44*(5), 1265-1270.

Tweeddale, H., Notley-McRobb, L., & Ferenci, T. (1998). Effect of slow growth on metabolism of Escherichia coli, as revealed by global metabolite pool ("Metabolome") analysis. *Journal of Bacteriology, 180*(19), 5109-5116.

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

Vigneau-Callahan, K. E., Shestopalov, A. I., Milbury, P. E., Matson, W. R., & Kristal, B. S. (2001). Characterization of diet-dependent metabolic serotypes: Analytical and biological variability issues in rats. *Journal of Nutrition, 131*(3), 924S-932S.

Vold, E., (1970). Meat production from boars and castrates. IV. Organoleptic and gas chromatographic studies on the steam distillate of back fat from boars. Report No. 238. Vollabekk, Norway: Institute of Animal Genetics and Breeding, N.L.H.

Walstra, P. and Kroeske, D. (1968) The effect of castration on meat production in male pigs. *Worfd Review of Animal Production 4, 59-64.*

Walstra, P., Maarse, H., (1970). IVO-report no. 2. Researchgroep Vlees en Vleeswaren TNO, Zeist.

Walstra, P. (1974). Fattening of young boars: Quantification of negative and positive aspects. *Livestock Production Science, 1*(2), 187-196.

Walstra, P., & Vermeer, A. W. (1993). Aspects of micro and macro economics in the production of young boars. *44th annual meeting of the EAAP 1993*, vol. 2 (2) (p. 325). Aarhus, Denmark.

Weckwerth, W., & Morgenthal, K. (2005). Metabolomics: from pattern recognition to biological interpretation. *Drug Discovery Today, 10*(22), 1551-1558.

Wei, R., Li, G., & Seymour, A. B. (2010). High-Throughput and Multiplexed LC/MS/MRM Method for Targeted Metabolomics. *Analytical Chemistry, 82*(13), 5527-5533.

Wenk, M. R. (2005). The emerging field of lipidomics. *Nature Reviews Drug Discovery, 4*(7), 594-610.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., van Duijnhoven, J. P. M., & van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81-89.

Wilkins, C. K. (1990). Analysis of indole and skatole in porcine gut contents. *International Journal of Food Science & Technology, 25*(3), 313-317.

Williams, L. D., Webb, N. B., & Pearson, A. M. (1963). INCIDENCE OF SEX ODOR IN BOARS, SOWS, BARROWS AND GILTS. *Journal of Animal Science, 22*(1), 166-&.

Wilson, I. D., Nicholson, J. K., Castro-Perez, J., Granger, J. H., Johnson, K. A., Smith, B. W., & Plumb, R. S. (2005). High resolution "Ultra performance" liquid chromatography coupled to a TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *Journal of Proteome Research, 4*(2), 591-598.

Wishart, D. S. (2008). Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology, 19*(9), 482-493.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1-37.

Xue JL, Dial GD. Raising intact male pigs for meat: Detecting and preventing boar taint. *Swine Health and Production.* 1997;5(4):151–158.

Yokoyama, M. T., & Carlson, J. R. (1979). MICROBIAL METABOLITES OF TRYPTOPHAN IN THE INTESTINAL-TRACT WITH SPECIAL REFERENCE TO SKATOLE. *American Journal of Clinical Nutrition, 32*(1), 173-178.

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicin*. New York: Wiley.

Zyromski, N. J., Mathur, A., Gowda, G. A. N., Murphy, C., Swartz-Basile, D. A., Wade, T. E., Pitt, H. A., & Raftery, D. (2009). Nuclear Magnetic Resonance Spectroscopy-Based Metabolomics of the Fatty Pancreas: Implicating Fat in Pancreatic Pathology. *Pancreatology, 9*(4), 410-419.

Acknowledgments

I am sincerely grateful to my advisors, Prof. Nägeli and Dr. Laczko, for the support and guidance he showed me throughout my dissertation writing. I am sure it hadn't been possible without their help.

A very special thanks goes out to David Fischer, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate his vast knowledge and skill in many areas (e.g. vision, aging, ethics, interaction). He provided me with direction, technical support and became a mentor and friend. It was though his, persistence, understanding and kindness that I completed my medical doctor.

Appreciation also goes out to the entire crew of FGCZ and to the staff of the Institute of Pharmacology and Toxicology for all the instances in which their assistance helped me along the way. The support from Richard Lock I am also acknowledging. Even thou he had I very tight schedule, he always took time to listen to my problems and to find a solution.

I would like to thank Dr. Philippe Wyrsch and Sandro Imhasly from the Faculty of Pharmacology and Toxicology for taking time to serve as my external reader.

In conclusion, I recognize that this research would not have been possible without the financial assistance of Bundesamt für Landwirtschaft (CH), Bundesamt für Veterinärwesen (CH) and University of Zurich.

I would also like to thank my family for the support they provided me through my entire life and in particular, I must acknowledge my parents Anneli and Sören, without whose love, encouragement and assistance I would not have finished this thesis.

Finally, I thank my friends Anna Layer, Susanne Berger, Nina Kazmareck and Sabrina Schäffle for instilling in me confidence and a drive for pursuing my doctor of veterinary medicine, and of course for being there whenever I needed them.

Curriculum Vitae

Name	Malin Emelie Maria Olson
Date of birth	26. Juli 1983
Place of birth	Örebro, Sweden
Citizenship	Swedish
08/1991 – 06/1997	Primary School Vasaskolan / Örebro, Sweden
08/1997 – 06/1999	Middle School Vasaskolan / Örebro, Sweden
08/1999 – 06/2002	High School for Natural Sciences Rudbecksskolan / Örebro, Sweden
10.06.2002	University entrance diploma
10/2004 – 03/2010	Veterinary medicine Ludwig-Maximilians University / Munich, Germany
26.03.2010	Veterinary medicine diploma
04/2010 – 04/2012	Doctoral Student Supervisor Prof. Hanspeter Nägeli Institute of Veterinary Pharmacology and Toxicology, Vetsuisse-Faculty University of Zurich Director Prof. Felix Althaus
06/2012 – current	Account manager at Waters AG Baden-Dättwil / Switzerland