Detailed Analysis of Cancer Proteomes for the Identification of Markers with Modern Liquid Chromatographic and Mass Spectrometric Methods

by

Pascal Steffen, M. Sc.

Dissertation

For the acquisition of the academic degree

Doctor rerum naturalium

Dr. rer. nat.

Joint PhD between the

University of Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Chemistry

and

Macquarie University, Faculty of Science and Engineering, Department of Molecular Sciences

August 2018

Referees:

1. Prof. Dr. Sasha Rohn

2. Prof. Dr. med. Udo Schumacher

3. Dr. Maria Riedner

4. Dr. Charlotte Uetrecht

Date of Disputation: 21.09.2018

Date of clearance for publication: 21.09.2018

This thesis was conducted as a Joint PhD thesis between the University Hamburg in Germany and the Macquarie University in Sydney, Australia. Thesis beginning was the 01.04.2015 in Hamburg with a one-year study in Sydney (01.04.2017-01.04.2018) and ended in 10.08.2018.

I. List of Publications

1. **Steffen P**, Kwiatkowski M, Robertson WD, Zarrine-Afsar A, Deterra D, Richter V, Schlüter H. Protein Species as diagnostic marker, J. Proteomics, 2016 Feb 16.

2. Kwiatkowski M, Wurlitzer M, Krutilin A, Kiani P, Nimer R, Omidi M, Mannaa A, Bussmann T, Bartkowiak K, Kruber S, Uschold S, **Steffen P**, Lübberstedt J, Küpker N, Petersen H, Knecht R, Hansen NO, Zarrine-Afsar A, Robertson WD, Miller RJD, Schlüter H. Homogenization of tissues via picosecond-infrared laser (PIRL) ablation: Giving a closer view on the in-vivo composition of protein species as compared to mechanical homogenization, J Proteomics, 2016 Feb 16.

3. Gleißner L, Kwiatkowski M, Myllynen L, **Steffen P**, Petersen C, Rothkamm K, Schlüter H, Kriegs M. Analyzing the influence of kinase inhibitors on DNA repair by differential proteomics of chromatin-interacting proteins and nuclear phosphor-proteins, Oncotarget, 2017 Nov 10.

4. **Steffen P**, Krisp C, Yi W, Yang P, Molloy MP, Schlüter H. Multi-laboratory analysis of the variability of shipped samples for proteomics following non-cooled international transport, Anal. Biochem., 2018 May 14.

5. Kwiatkowski M, Krösser D, Wurlitzer M, **Steffen P**, Barcaru A, Krisp C, Horvatovich P, Bischoff R, Schlüter H. Application of Displacement Chromatography to Online Two-Dimensional Liquid Chromatography Coupled to Tandem Mass Spectrometry Improves Peptide Separation Efiiciency and Detectability for the Analysis of Complex Proteomes, Anal. Chem., 2018 Jul 17.

II. Table of Contents

III. Abbreviations

Name	**Abbreviation**
International Society of Urological Pathology	ISUP
3-Dimensional liquid chromatography	3D-LC
Acetonitrile	ACN
Coefficient of variation	CV
Collision induced dissociation	CID
Data-Dependant Acquisition	DDA
Data-Independant Acquisition	DIA
Differential expression package	DEP
Dithithreitol	DTT
Exctracted ion chromatogram	XIC
False-discovery rate	FDR
Formalin fixed parafin-embedded	FFPE
Gene ontology	GO
Gleason-Score	GS
High pH	HpH
Higher-energy C-Trap dissociation	HCD
Iodoacetamide	IAA
Label-free Quantitation	LFQ
Linear models for Microarray data	limma
Liquid chromatography	LC
Mass Spectrometry	MS
Missing values not at random	MNAR
Multidimensional protein identification technology	Mud-PIT
Multiple-Reaction Monitoring	MRM
Phosphate buffer saline	PBS
Principle component analysis	PCA
Retention time calibration peptides	iRT

Reversed-phase	RP
Sequential window acquisition of all theoretical mass-spectra	SWATH
Strong anion exchange	SAX
Strong cation exchange	SCX
Ultra-High-Pressure liquid chromatography	UHPLC
Universitäts Klinikum Hamburg-Eppendorf	UKE

1 Zusammenfassung

Die Überlebensrate von Prostatakrebs-Patienten beträgt nahezu 100%, solang der Krebs im frühen Stadium detektiert wurde. Diese sinkt allerdings drastisch (auf 23%), wenn die Detektion in späteren Stadien erfolgt. Die größte Differenz in der Lebenserwartung wird beobachtet, wenn ein Gleason-Score von 7 bestimmt wurde, welcher entwerder als 3+4, hier werden Metastasen nur sehr selten nachgewiesen, oder 4+3 welcher eine 3-fache Erhöhung an Metastasen im Vergleich zu 3+4 aufweist, bestimmt werden kann. Das Hauptziel dieser Arbeit war es, potenzielle Proteinmarker zu identifizieren um Pathologen weitere Parameter zur Verfügung zu stellen um den Gleason-Score eines Tumors genauer zu bestimmen.

Ein Xenograft-Mausmodel wurde verwendet, um potenzielle Proteinmarker für Metastasierung zu suchen. Hierfür wurden die Quantitäten der Proteine der stark metastasierenden Zelllinie PC3 mit denen der nicht-metastasierenden Zelllinie DU145 verglichen. In PC3 Zellen wurden 32 Proteine mit signifikant höheren Mengen nachgewiesen. 25 humane Prostatagewebeproben mit unterschiedlichen Gleason-Scores (GS) wurden am UKE in Hamburg für die Proteomanalyse vorbereitet, und nach dem Transport der getrockneten tryptischen Peptide zur Macquarie-Universität in Sydney mittels quantitativer Proteomanalyse analysiert. Mittels Experimenten zur Stabilität der tryptischen Peptide wurde sichergestellt, dass diese durch den Transport keine signifikanten Änderungen erfahren. Die Proteine mit signifikanten Unterschieden in ihren Quantitäten zwischen den GS 3+4 und 4+3 wurden mit denen aus dem Xenograft-Mausmodel verglichen. Interessanterweise, wurde das Protein Agrin in erhöhter Konzentration in den 4+3 Proben sowie in den PC3 Mausproben detektiert. Zur Interpretation der Bedeutung der identifizierten potentiellen Marker-Proteine wurde ein Textmining Script entwickelt, mit welchem eine Kategorisierung in Bezug auf den Bekanntheitsgrad der Proteine im Kontext von Metastasierung und Prostatakrebs gelang. Mit diesem Werkzeug wurde erkannt, dass Agrin ein Protein sein könnte, dass eine wichtige funktionelle Rolle in der Metastasierung hat, was in zukünftigen Studien validiert werden sollte. Desweiteren wurde eine auf Displacement-Chromatographie basierte 3D-LC Methode entwickelt, mit dem Ziel eine größere Zahl niedrig abundanter Proteine zukünftig identifizeren und quantifizieren zu können. Diese Methode stellt die

erste dokumentierte Integration von Displacement-Chromatographie in einer 3D-LC Methode dar.

2 Abstract

Although when detected in early stages the survival rate of prostate cancer patients is close to 100% it drops drastically (to 23%) when diagnosed in later stages. The most severe difference in life expectancy is observed when a Gleason-Score of 7 is assigned which can be either 3+4, which shows only small rates of metastasis or 4+3 which shows a three-fold increase in metastasis compared to 3+4. The main aim of this thesis was the identification of potential protein marker to give pathologists another parameter for a more exact classification of the Gleason-Score of a tumor.

A xenograft mouse model was used to look for potential protein marker in metastasis. To achieve this, quantities of the proteins of the highly metastatic PC3 cell line were compared to those of the non-metastatic cell line DU145. This resulted in 32 statistically relevant proteins which showed higher quantities in the tumors derived from the PC3 cells. 25 human prostate cancer tissue samples representing different Gleason-Scores (GS) were prepared in the UKE in Hamburg for proteome analysis and after transport of the dried peptides to the Macquarie University in Sydney analyzed by quantitative proteome analysis. Using experiments to determine the stability of tryptic peptides it was ensured that they would not undergo significant changes. Proteins with significant differences in quantity between GS 3+4 and 4+3 were compared to those found in the xenograft mouse model. Interestingly the protein Agrin was found in higher concentration in the 4+3 samples as well as in the PC3 tumors grown in mouse. For the interpretation of the significance of the potential marker proteins a text mining script was developed with which a categorization in relation to their familiarity in context with metastasis and prostate cancer was possible. Using this tool, Agrin was identified to possibly play an important part in metastasis and should be further validated in future studies. Furthermore, a 3D-LC method based on displacement mode chromatography was developed with the aim to identify and quantify a higher number of low abundant protein in future studies. This method represents the first known integration of displacement chromatography in a 3D-LC method.

3 Introduction

3.1 Cancer

Cancer is a disease which is based on abnormal cell growth. Different disease triggers are known. Some of these triggers include exogenic factors such as the consumption of tobacco and alcohol. Cancer may also be triggered by infections such as the human papillomavirus (HPV) which acts as growth factor for the infected cell by manipulation of the DNA [1]. It is estimated that in 2018 about 1.7 million people in the US will be newly diagnosed with cancer [2].

In most cases, the primary tumor is not the cause of death and can be surgically removed. The formation of metastasis as well as the risk involved in treatment of cancer are the main reason for the death of the patient. Cancer cells show different stages of development ranging from primary tumor cells to metastasis. The progression from one stage to the next is mediated by a multitude of proteins. For the development of new diagnostics and therapies against cancer a detailed understanding of these proteins is essential. The differentiation between malign and benign tumor as well as between malign and healthy tissue in prostate cancer is difficult which is why preventive surgery is often performed. The surgery has several risks for the patient such as incontinence and impotence. Finding new marker which are specific for different tumor types may lead to a decrease in unnecessary surgery.

3.1.1 Prostate cancer

The prostate is a male gland which produces parts of the semen fluid. The fluid is secreted into the urethra and mixed with semen. 30% of the semen fluid is produced in the prostate and incorporates substances which are essential for the liquification and mobility of the sperm as well as the stimulus of the uterus. The prostate in healthy men is about the size of a walnut. It is located underneath of the urinary bladder, surrounds the urethra and is adjacent to the rectal [3]. A schematic of the male reproductive system can be seen in Figure 1.

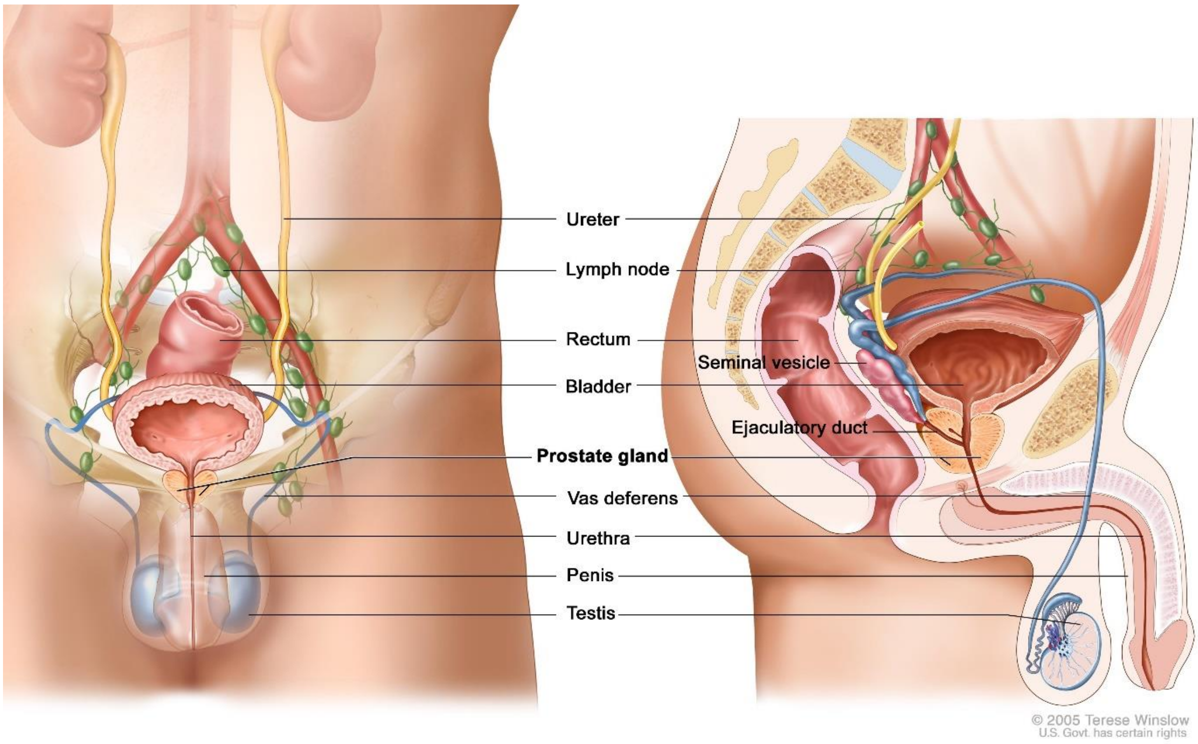


Figure 1: Schematic of the male reproductive system [3].

Different diseases relate to the prostate such as bacterial infections (Prostatitis). It is possible that the prostate increases in size with increased age of the individual which is called benign prostatic hyperplasia (BPH). These diseases often lead to false-positive diagnosis of prostate cancer.

With increased age the risk of prostate cancer increases. Prostate cancer is the most common form of cancer in the male population in America with an expected new incidence of 164.690 cases in 2018 [2] which makes it the third most commonly diagnosed cancer over all. The median age in which prostate cancer is diagnosed is 66 [4]. Because of the slow growth of this tumor it often presents no problems for the patients during their lifetime [5]. There are different methods for diagnosing prostate cancer some of which include a physical exam of the patient, digital rectal exam (DRE), a prostate-specific antigen test (PSA-Test) as well as transrectal magnetic resonance imaging (MRI) [3, 6]. These tests present first indications for prostate cancer but each of these methods has its own disadvantages and none of them is unambiguous. In case of a rectal exam the experience of the physician plays a major role. The position of the prostate itself makes it impossible for the physician to feel the whole prostate who may therefore miss tumors in the front part of the prostate [7]. All anatomic imaging techniques have the same disadvantage in that they cannot differentiate the malign tumor from surrounding healthy tissue [6]. The PSA-Test was approved by the Food and Drug Administration (FDA) in 1986 [8]. PSA is a prostate-specific glycosylated enzyme which is part of the kallikrein subfamily and is secreted from the epithelial cells of the prostate [9]. The PSA-value alone is, for a diagnosis of prostate cancer not specific enough. Elevated concentration of PSA in the blood (in the past, for patients with concentrations above 4 ng/mL a biopsy was advised) may suggest prostate cancer but could also be caused by prostatitis or other diseases [8]. Because until recently the PSA-test was used extensively for screening purposes which lead to many false-positive diagnosis many organizations question the PSA-based screening especially because there are many risks for the patients when undergoing the subsequent biopsy [10].

Treatment of prostate cancer includes radical prostatectomy, radio therapy, hormone treatment and active survey of the patient. These treatment methods have a very high success rate (5-year survival rate of close to 100% for localized cancer) but with the down-site of different side effects such as incontinence, erectile dysfunction or loss of libido. The survival rate drops drastically (to around 28%) when diagnosed at later stages [4]. To increase the accuracy and reduce overdiagnosis the development of new tests or the improvement of existing ones is necessary. A promising approach is the identification of specific markers for prostate cancer.

Diagnostic markers are defined as: “A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [11]. Different molecule classes fit this description, proteins being the ones with the highest potential as marker. The advantage of proteins is their massive diversity caused by different proteoforms and post-translational modifications [12, 13] which in turn also leads to high difficulty in analysis. To detect specific molecules in a complex mixture well established sample preparation and separation techniques as well as detectors with high sensitivity are necessary. Recent developments in mass spectrometry (such as Orbitrap mass spectrometers) make proteomics an ideal method for the identification of new marker candidates.

3.2 Mass Spectrometry of Proteins (Proteomics)

The first documented mentioning of the word “Proteome” was in a paper in 1995 by Wilkins MR *et al.* [14] and was used to describe the total set of proteins encoded by a genome. 1997 the term “proteomics” was coined by James P which he based on the term genomics [15] and was further characterized by Anderson NL *et al.* [16]. Anderson defined the proteome as “the use of quantitative protein-level measurements of gene expression to characterize biological processes (e.g. disease processes and drug effects) and decipher the mechanism of gene expression control”. Therefore, proteomics can be seen, like genomics, as a field of study which encompasses different areas of research. Pandey A *et al.* [17] defined three main areas of proteomics as “(1) protein micro-characterization of large-scale identification of proteins and their post-translational modifications; (2) ‘differential display’ proteomics for comparison of protein levels with potential application in a wide range of diseases; and (3) studies of protein-protein interactions using techniques such as mass spectrometry…”. Historically the beginning of proteomics can be seen as early as the 1970s where 2D-Gel electrophoresis was used to build protein libraries. However, one can argue that only with the completion of the human genome project in 2003 [18] proteomics, as it is seen today, emerged as a wide spread field of biological and medical research. Another factor that contributed immensely to the growing popularity of proteomics was the development of mass spectrometers with higher sensitivity, accuracy and resolving power such as the Orbitraps (the Orbitrap LTQ was presented to the public in 2005 [19]). At that time the FT-ICR analyzer was considered the state of the art mass analyzer but had major drawbacks in being very expensive (both in acquisition and maintenance), vast space requirements and requiring highly trained researchers. The Orbitrap posed a good compromise in being a benchtop instrument with less maintenance requirements although not as accurate and with less resolving power than a FT-ICR but better than a TOF, Iontrap (IT) or Quadrupole analyzer.

The simplest proteomics experiment tries to identify proteins in a mixture by identifying parts of their amino acid sequence. To achieve this, the proteins in the mixture are subjected to an enzyme catalyzed hydrolysis (Trypsin is often the enzyme of choice as it cleaves C-Terminal to an Arginine or Lysine [20] and thus resulting in peptides which carry at least one extra charge) resulting in peptide fragments. This approach is referred to as shotgun or bottom-up proteomics in contrast to top-down proteomics where the intact proteins are analyzed which as of today still poses many difficulties, mostly on the level of separation [21].

Before measuring the peptides in a mass spectrometer, the peptide mixture must be separated, typically by ultra-high-pressure liquid chromatography (UHPLC) using a reverse-phase (C18) column. The UHPLC is connected to the mass spectrometer via an ESI (Electrospray Ionization) source. Eluting peptides enter the mass spectrometer where a spectrum containing peaks for all peptides (Full-scan or MS1 spectrum) is recorded. The mass spectrometer then selects different peptides for further fragmentation, typically the most intense mass peaks are selected and subsequently subjected to a fragmentation event which results in fragment ions that are specific for each peptide (Data Dependent Acquisition, DDA). There are two major fragmentation methods the first being used mainly in TOF, IT or Quadrupole based instruments, collision induced dissociation (CID), where the peptide is fragmented by increasing the neutral gas (helium, nitrogen or argon) pressure in the TOF/IT/Quadrupole. The energy of the collisions between the peptides and gas molecules is transferred as internal energy to the peptide where the weakest bond (usually the amide bonds) breaks resulting in different fragment ions. The other fragmentation method is called higher energy C-Trap dissociation (HCD) and is exclusive to orbitrap mass analyzers. The principle of the fragmentation remains the same as in CID although a higher RF voltage is applied to the C-Trap to keep the fragments trapped. The resulting fragment ions are then transferred into the Orbitrap for detection [22]. A schematic of the possible fragment ion series can be seen in Figure 2.

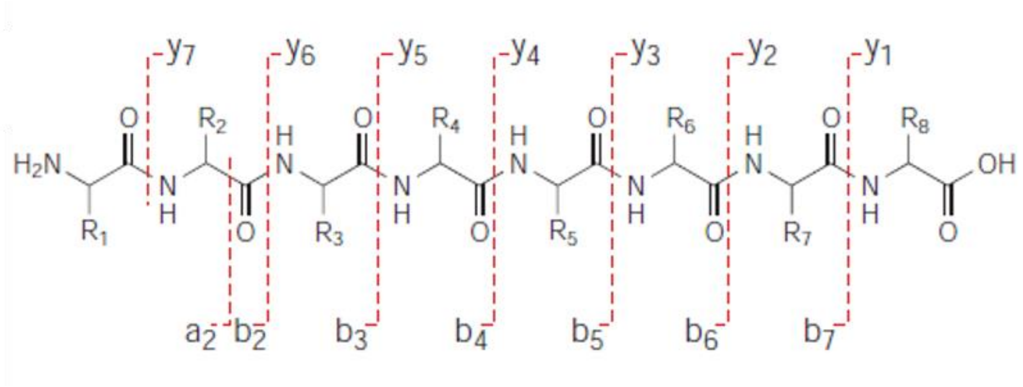


Figure 2: Schematic of the y- and b-ions nomenclature [23].

The different masses of the y- and b-ions detected in the MS2-spectrum can be allocated to their corresponding amino acids thus resulting in the amino acid sequence of the peptide. An example of such an MS2 spectrum can be seen in the results part in Figure 13. Because this method of identification is based on known sequences of the proteins a complete and curated protein database (which in case of the human genome the Human Genome Project delivered in 2003) is of utmost importance. Because of the high complexity of the peptide mixture it is not feasible to annotate and identify each MS2 spectrum by hand. Therefore, search engines such as Andromeda which is used in MaxQuant [24] or Sequest [25] which is used in Proteome Discoverer have been developed. These algorithms use the detected parent ion mass as well as the fragment ion masses of the y- and b-ions of a peptide to query a so called FASTA database which contains the amino acid sequences of proteins of a specific organism. These amino acid sequences are digested *in-silico* using the information provided by the user such as enzyme used for proteolysis and maximum allowed missed cleavages to generate theoretical peptides. For each of these theoretical peptides theoretical fragment ion masses are computed and compared against the data using user defined mass tolerances. Each identified peptide is allocated a search engine specific score which among other parameters incorporates the accuracy and number of fragment masses matched in the spectrum. Figure 3 shows a schematic of this method.

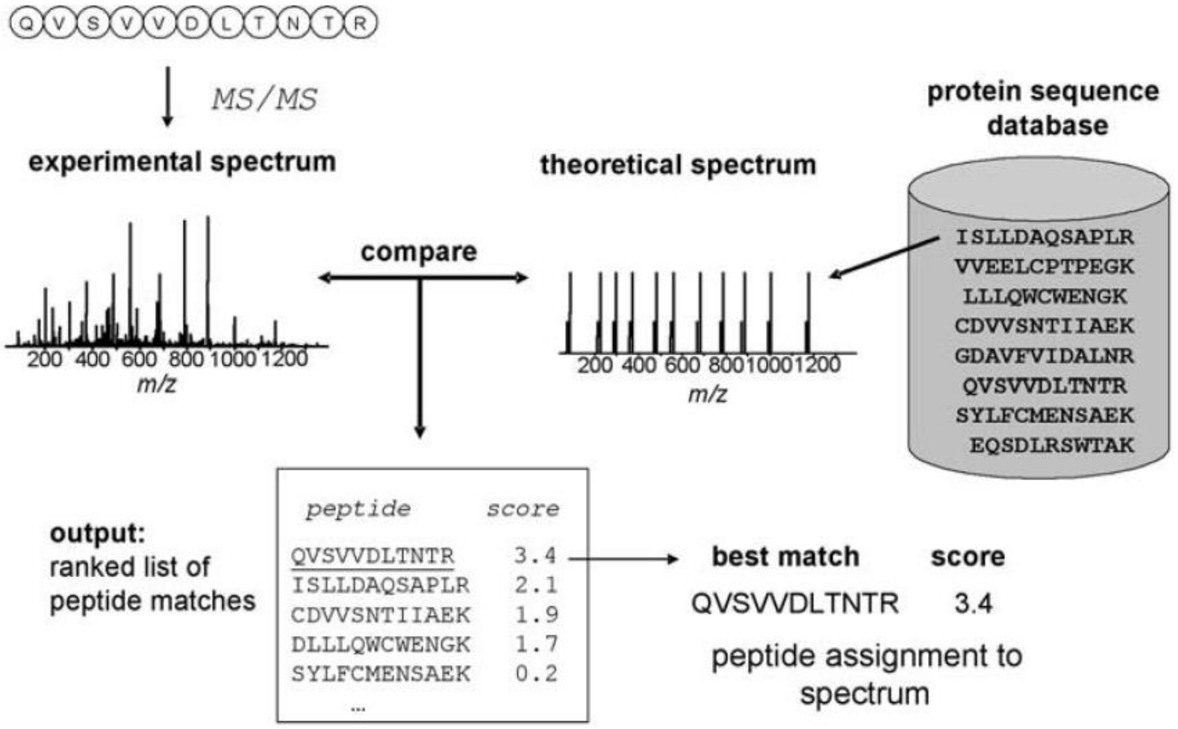
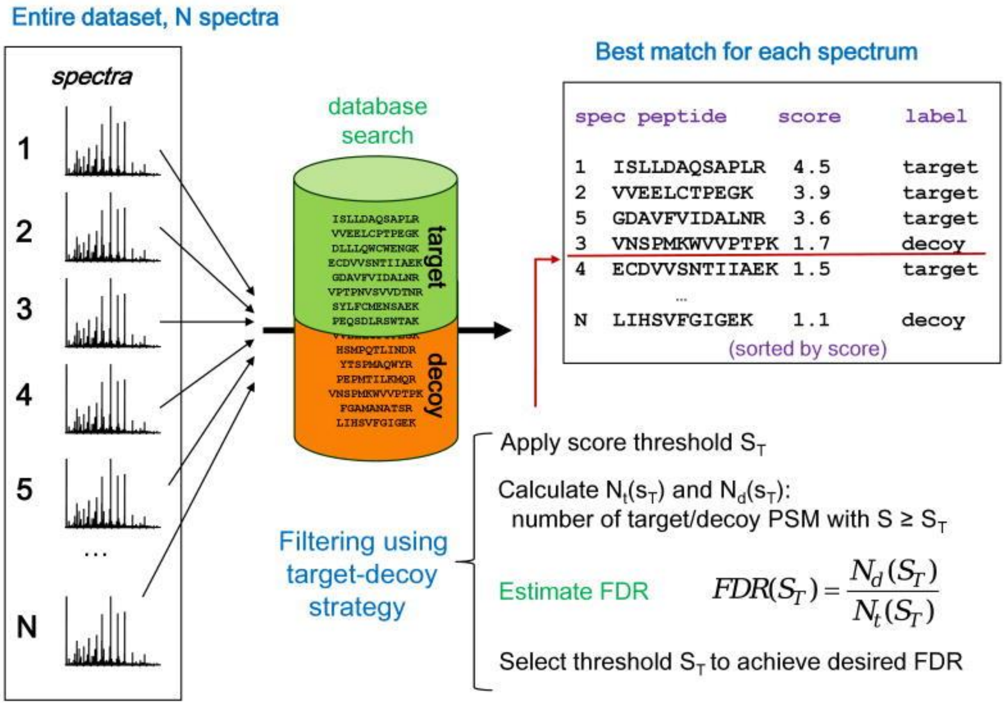


Figure 3: Schematic of a database search for MS2 spectra [26].

To account for false-positive identifications, resulting from fragmentation of multiple peptides (chimeric spectra), low MS2 quality, sequence variance and others [26] a so called false-discovery rate (FDR) is calculated. There are several approaches but the most commonly used one is to reverse all protein sequences in the FASTA database and use the amount of hits in this decoy database to compute the desired FDR. In the proteomics field an FDR of up to 1% is commonly accepted. Figure 4 shows a schematic of the FDR calculation.



spec	peptide	score	label
1	ISLLDAQSAPLR	4.5	target
2	VVEELCTPEGK	3.9	target
5	GDAVFVIDALNR	3.6	target
3	VNSPMKWVVPTPK	1.7	decoy
4	ECDVVSNTIIAEK	1.5	target
	...		
N	LIHSVFGIGEK	1.1	decoy

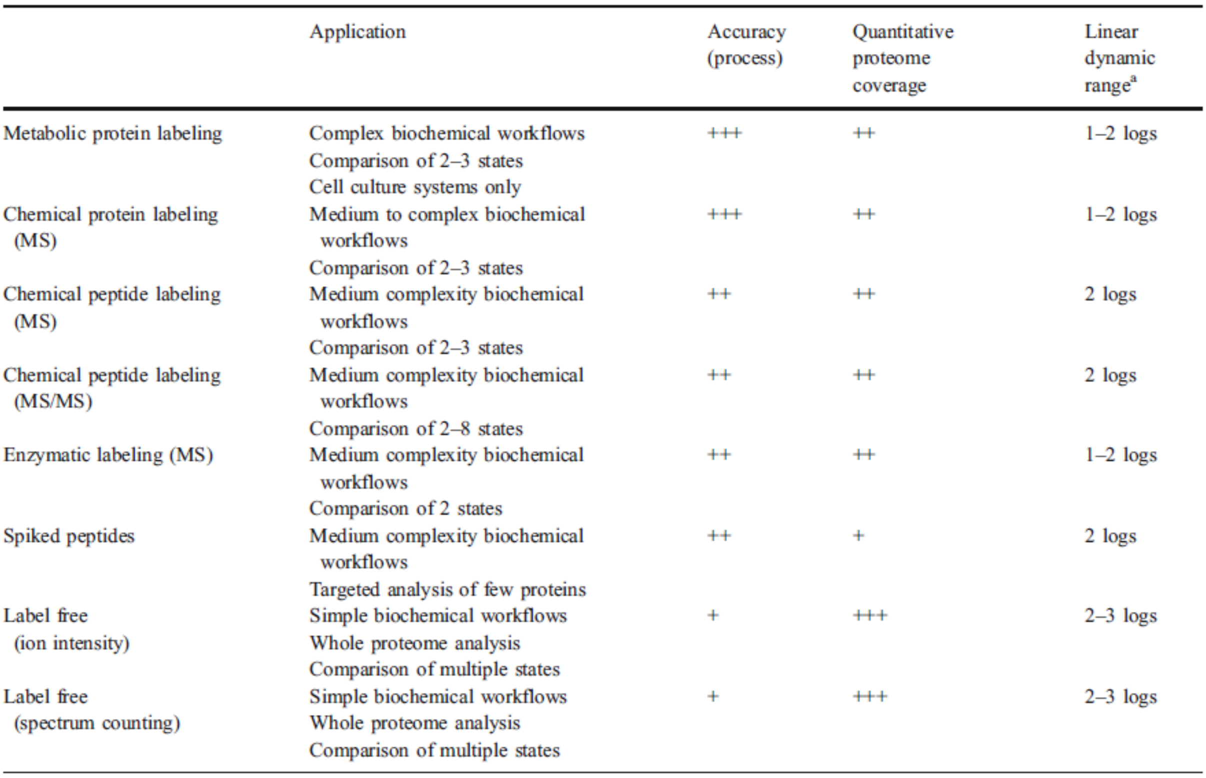
Figure 4: Schematic of the FDR calculation using a decoy database [27].

After FDR estimation the researcher is left with high confidence identifications of the proteins present in the mixture. However, to answer biological questions it is not only necessary to know which proteins are in the sample but typically also the amount of the proteins (absolute quantification) or the fold-changes (ratios) between two sample sets (e.g. healthy tissue vs. disease tissue), called relative quantification. Of specific interest in this thesis were the label-free quantification methods.

3.2.1 Label-free Quantification (LFQ)

Label-free quantification has several advantages over isotopic labeling methods. When working with biological samples it is either very expensive and laborious to label the complete organism (SILAC-Mouse [28]) or just not possible in case of human patient samples. Furthermore, all labeling methods (SILAC, Dimethyl-labelling, iTRAQ and TMT [29]) share the same problem which is incomplete labeling of the sample, which may result in a bias of the quantification. Another advantage of label-free methods is that there is theoretically no limit as to how many samples may be compared. On the other hand, labeling methods tend to be more accurate because of the use of reporter ions for quantification [30].

Figure 5 shows a comparison between the different quantification methods as well as comparing their accuracy and dynamic range.



	Application	Accuracy (process)	Quantitative proteome coverage	Linear dynamic range[a]
Metabolic protein labeling	Complex biochemical workflows Comparison of 2–3 states Cell culture systems only	+++	++	1–2 logs
Chemical protein labeling (MS)	Medium to complex biochemical workflows Comparison of 2–3 states	+++	++	1–2 logs
Chemical peptide labeling (MS)	Medium complexity biochemical workflows Comparison of 2–3 states	++	++	2 logs
Chemical peptide labeling (MS/MS)	Medium complexity biochemical workflows Comparison of 2–8 states	++	++	2 logs
Enzymatic labeling (MS)	Medium complexity biochemical workflows Comparison of 2 states	++	++	1–2 logs
Spiked peptides	Medium complexity biochemical workflows Targeted analysis of few proteins	++	+	2 logs
Label free (ion intensity)	Simple biochemical workflows Whole proteome analysis Comparison of multiple states	+	+++	2–3 logs
Label free (spectrum counting)	Simple biochemical workflows Whole proteome analysis Comparison of multiple states	+	+++	2–3 logs

Figure 5: Comparison between different quantification methods used in mass spectrometry [31].

As can be seen in Figure 5, two main label-free methods exist. One being the quantification using spectral counting [32]. This method is based around the hypothesis that the more of a protein is present in a sample the more a peptide belonging to such a protein will be identified in the sample. A benefit of using the spectral counting method lies in the use of extensive MS2 fragmentation which adds the benefit of also increasing the identification rate. However, the use of a dynamic exclusion list to minimize redundant sampling of the same peptide in order to fragment different peptides is counterproductive for this quantification method. Furthermore, it assumes a linearity between the number of spectral counts and the protein which is fundamentally flawed because of the differences in chromatographic behavior of each peptide depending on its physicochemical properties [31]. The second method is based on the ion intensity of a peptide ion over its chromatographic elution time. To this end XICs of the peptide parent ion mass is generated and the integral of this peak is taken as representative value to compare against the same peptide in a different sample (see Figure 14 in the result part). This approach benefits from a high number of full-scan spectra over the elution time of the peptide. Therefore, fast cycling mass spectrometer are advantageous but still a compromise between quantification accuracy (number of full-scans over a chromatographic peak) and peptide identification (number of MS2 fragmentation events) has to be reached. Because the quantification is based on the parent ion mass as well as on the retention time of the peptide, high reproducibility of the chromatography is essential. Therefore, algorithms have been developed to align small discrepancies in the retention time between different experiments [30]. Another influencing factor is a high mass resolution of the mass spectrometer to reduce interfering signals of peptides with similar masses.

Because of the rapid development of mass analyzers in the past years high mass resolution as well as short cycling times are generally not a bottleneck anymore [19]. However, one major issue of these quantification methods is the problem that a peptide can only be quantified after it was identified. Because of the random sampling nature of the DDA method (highest parent ion peaks are selected for fragmentation) it is possible that even using technical replicates there are differences in the number of identified peptides. Because of this the later quantification which is based on the parent ion masses is incomplete and thus contains missing values. Figure 6 shows a schematic depicting the relation between the proteins in a sample.

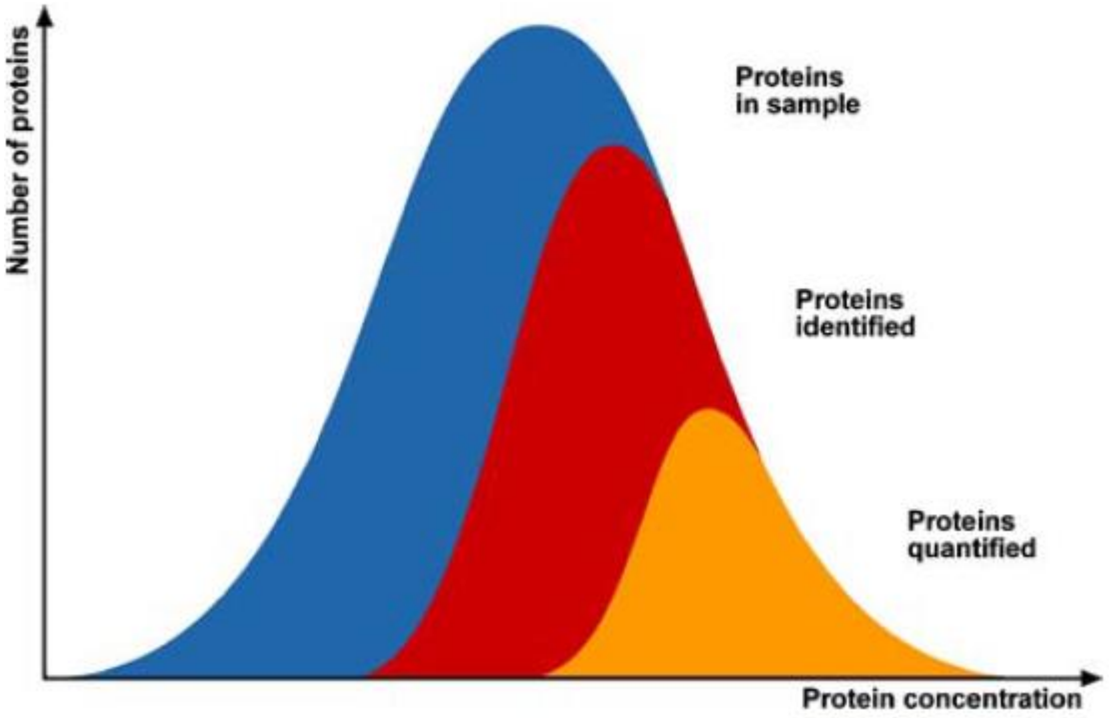


Figure 6: Schematic representation of the relation of the proteins in a sample (blue), the number of proteins identified (red), and the number of proteins quantified (yellow) [31].

These missing values pose a problem in the post-analysis of the data where statistical tests must be employed to assess the significance of a change in protein amount in different samples. There are two types of these missing values, missing at random (MAR), which are caused by the described stochastic nature of the DDA method and missing not at random (MNAR), which can be cause by real differences in protein expression in different samples resulting in peptide signals below the threshold for the LFQ algorithm to be detected or even being subject of ion suppression. Figure 17 in the result part depicts the effect of both types of missing values.

Recently, a method called Data-Independent Acquisition (DIA) which can be used synonymously with the term SWATH-MS [33] (Sequential window acquisition of all theoretical mass-spectra) has risen in popularity. This method of quantification is also based on the ion intensity over a chromatographic retention time of a given peptide but in contrast to the method described earlier it uses the fragment ion peaks of a peptide to calculate the peptide amount (MS2-level quantification). As such the DIA method can be seen as a hybrid between DDA and MRM (Multiple-Reaction Monitoring). **Figure 7** shows a schematic comparison between the MRM, DDA and DIA method.

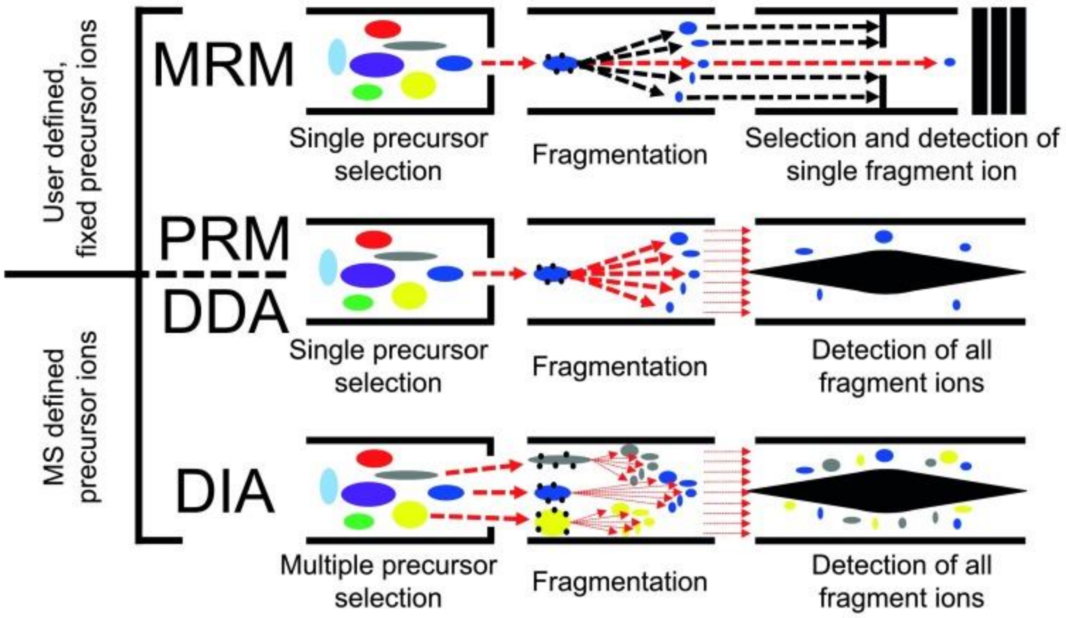

User defined, fixed precursor ions
MS defined precursor ions
MRM
PRM
DDA
DIA
Single precursor selection
Fragmentation
Selection and detection of single fragment ion
Single precursor selection
Fragmentation
Detection of all fragment ions
Multiple precursor selection
Fragmentation
Detection of all fragment ions

Figure 7: Schematic comparison between MRM, DDA and DIA acquisition methods [34].

Of the three acquisition methods MRM has the highest precision as well as the highest reproducibility because of its highly selective acquisition [34]. This however comes with the disadvantage of very high cycle times and it is therefore only feasible to monitor certain already known peptides of interest. This represents its biggest disadvantage when used in research. No discovery type experiments are possible when using MRM because of the necessity of prior knowledge of the masses of the precursor and its fragments of interest.

DDA has the lowest reproducibility but the highest rate of identification. DIA tries to find a middle ground between these two methods in being moderately precise compared to MRM and nearly as powerful as DDA in its identification rate (this however is because of the use of a spectral library generated by DDA). Because every peptide is fragmented in the DIA approach it is possible to query the collected data for peptides of interest after data collection thus giving the researcher more flexibility compared to MRM.

As the name DIA suggests this method is based on the principle to not only fragment some of the precursor peaks found in the full-scan spectrum but rather all of them in an unsupervised (independent) manner. To achieve this a mass window is selected from the full-scan rather than just a specific mass, and all of the precursor ions inside this mass window are fragmented at once. Because this leads to highly complex fragment spectra a so called spectral library is used to query the data post-acquisition. This spectral library is build using either several technical replicates in DDA mode of the desired samples or DDA data of the fractionated (often using offline high pH reversed-phase separation) pooled sample and therefore the number of identification as well as the number of possible quantification is reliant on a comprehensive DDA library. The assigned fragment ions of a peptide are then monitored over their retention time and the area under the curve is summed up to yield the peptide intensity (Figure 8).

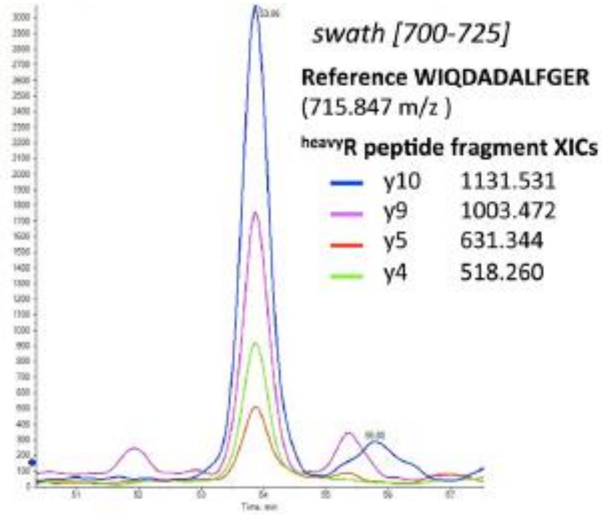


Figure 8: Fragment ion traces of a peptide precursor (adjusted from [33]). The y4, y5, y9 and y10 XICs of the heavy labelled peptide WIQDADALFGER are shown.

MS2-level quantification as seen in Figure 8 has the advantage over tradition MS1-level quantification (DDA) that the quantification is much more reliable because of the use of several specific fragment ions of the peptide instead of only using the parent ion peak envelope which might have interferences.

No matter which kind of quantification method was used to generate quantification data the need for statistical testing for significance remains the same. For robust statistical testing it is imperative that the data is as complete as possible. In case of DDA acquisition this can indeed pose a problem and lead to false-negatives or the need to use imputation methods to generate extra data points. Because of this DIA quantification data tends to be more reproducible overall.

3.3 Chromatography in Proteomics

Chromatography is a technique to separate molecules in a mixture using their physicochemical proprieties. The molecules in the mixture which is dissolved in the mobile phase interact with the stationary phase resulting in different travel speed for different molecules along the column. Two main groups of chromatography can be defined, preparatiive and analytical. In proteomic experiments (online separation of peptides before injecting into the mass spectrometer) only analytical chromatography is used to minimize sample amount and column size as well as flow speed. Ultra-High-Pressure liquid chromatography (UHPLC) has become the standard in proteomic analysis [35]. Liquid chromatography can be classified in two major modes: Gradient and displacement mode. Every chromatographic separation technique is based on the travel time of the peptides over the column length which is directly correlated to the affinity of the peptides to the stationary phase in relation to the mobile phase (law of mass action). Different stationary phases are used for separation including reversed-phase (RP) and (strong) anion/cation exchange (SAX/SCX). RP material separates the peptides based on their Van-der-Waals (hydrophobicity) interaction which is correlated to the length and chemical composition of the peptides. SAX/SCX material separates the peptides based on ionic interaction based on the charge states of the peptides which is based on their amino acid composition and post-translational modifications.

In gradient mode, peptides are applied to the column in narrow bands and move down the column over time resulting in peak shapes resembling a gaussian curve. Longer travel times result in broader peak shapes. In this mode retention is achieved by adjusting the concentration of the mobile phase according to the stationary material used. In case of RP a polar mobile phase is used (Water) to allow the formation of the bands on the column. By decreasing the polarity of the mobile phase over time (gradient) elution of peptides with different affinity to the stationary material is achieved. When using RP acetonitrile (ACN) is commonly used for the decrease in polarity mainly because of its high volatility. When using gradient mode for separation it is important that only low concentrations of the analyte is applied so that initial band formation is achieved, and peak separation is ensured.

In contrast, when using displacement mode, it is important to load high concentrations of the analyte to ensure that nearly complete saturation of the binding sites of the stationary material is achieved. Here the high affinity peptides are retained at the head of the column thus pushing down lower affinity peptides to the bottom of the column resulting in zones of high purity. The mobile phase is chosen to ensure high affinity of the peptides to the stationary phase. To elute the peptides from the column a displacer molecule is injected in increasing concentration. This displacer has higher affinity to the stationary phase than the peptides resulting in the binding to the head of the column and pushing down the bound peptides. Each of these peptides moves further down the column replacing lower affinity peptides according to the increase in displacer amount thus forming a so-called displacement train. This results in the elution of the peptides in the shape of bands in contrast to the gaussian peak shapes when using elution mode [36, 37]. Figure 9 shows a comparison of the elution profiles between isocratic, gradient elution and displacement.

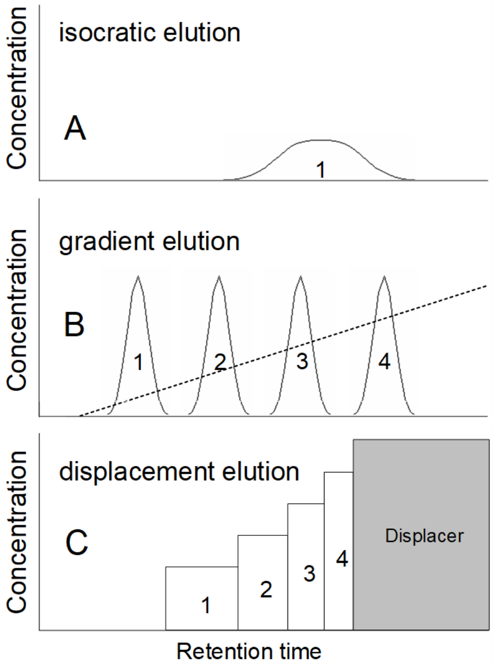


Figure 9: Schematic comparison between the elution profiles of isocratic (A), gradient (B) elution and displacement (C). The dashed line in B shows the increase in of the second component of the mobile phase over time.

4 Aim of the Thesis

Although patients diagnosed with early stage prostate cancer show a nearly 100% 5-year survival rate, treatment of later stage patients remains difficult. Early detection as well as correct categorization is therefore of utmost importance for patients. The most drastic drop in patient survival rates is observed between patients of which the tumor is classified as Gleason-Score 3+4 and 4+3 [38, 39]. The Gleason-Score is based on the classification based on the expert assessment of a pathologist by microscopy. Because of this, the classification is arguably prone to small errors especially for very similar tumor structures.

This thesis was split into one main aim which was to identify possible protein markers which could be used by pathologists for a more exact classification. To achieve this it was first hypothesized, that a single cell tumor consisting of either PC3 (highly metastatic) or DU145 (non-metastatic) cells grown subcutaneously in a xenograft mouse model would approximate the phenotype of either high or low Gleason-Scores in human patient samples. The proteins showing higher concentrations as determined by label free mass spectrometry in either tumor model should therefore be a good basis for the analysis of the human patient samples as well as providing an additional level of confidence in the analysis. Because this thesis was conducted as a Joint PhD project between the university of Hamburg and Macquarie University (Sydney) the mouse samples should be measured in Hamburg and the human patient samples prepared in Hamburg and then shipped to the laboratory at Macquarie University.

To support this aim three sub-aims were formulated, the first of those being the assessment of the stability of samples prepared for proteomics during a non-cooled transport between institutes to ensure high sample quality for the human samples measured in Sydney.

The second sub-aim being to develop a literature mining script to assist researchers in categorizing targets of interest for further analysis. This should help to give meaning to the proteins found in higher concentration between samples of the mouse and human experiments in context of the main aim.

The third sub-aim was to develop a new 3-dimensional LC method which should improve peptide identification in a reasonable time frame as well as being compatible for iRT peptides for retention time calibration for the later use as a spectral library for DIA quantification.

5 Results

The overall aim of this thesis was to develop new methods for cancer protein marker screening and to apply these methods to investigate prostate cancer. As this thesis was conducted as a Joint PhD project between the University of Hamburg (Germany) and the Macquarie University (Sydney, Australia), biological samples would have to be shipped between these countries. Two of the most hindering being the high cost of transport associated with modes that require constant cooling through dry ice addition to ensure sample stability and the regulations of the customs of the respective countries which can add delays.

A proteomic experiment using SW480 cells was conducted between the UKE, the Fudan University in Shanghai and Macquarie University in Sydney to assess if biological samples can be tryptically digested and desalted prior to drying *in vacuo* and shipped without cooling while still retaining their integrity for reliable quantitative proteomic mass spectrometry analysis.

The first set of experiments was conducted in the UKE in the laboratory of Prof. Schlüter. Primary prostate cancer tumors of two different cell lines (PC3 and DU145) were obtained by injecting the human cells subcutaneously into xenograft mice. The proteomics data of these two primary tumors were compared by label-free quantitation and statistically relevant differentially regulated proteins determined by post analysis. This set of data was used as a basis model for highly metastatic (PC3) and non-metastatic (DU145) prostate cancer phenotypes of human samples.

For interpretation of the meaning of the differentially regulated proteins a text mining algorithm was developed for classifying the proteins in three different classes, well known proteins already established as metastasis markers, proteins, which have been mentioned in association with metastasis and proteins which never have been reported to be associated with metastasis.

Human prostate cancer tissue specimens of different Gleason-scores were prepared for proteome analysis in the UKE and the resulting tryptic peptides sent to Macquarie University for label-free quantification using DIA/SWATH to find differentially regulated proteins which may serve as marker proteins for Gleason-score categorization.

Furthermore, a new 3-Dimensional peptide separation technique was developed to achieve high analysis depth at a reasonable measurement time while remaining compatible with indexed retention time peptides (iRT) for data-independent acquisition (DIA/SWATH). Its utility for building reference spectral libraries was demonstrated.

5.1 Xenograft Mouse experiments

To establish a baseline of differentially expressed proteins for different Gleason-score prostate samples, a mouse model was initially examined. Samples were prepared as described in the material and methods part (see section 7). Peptides corresponding to 1 µg of protein were injected into the LC system and measured on an Orbitrap Fusion as described in section 7.11.2.

5.1.1 Protein Identification and MS1-based Label-free quantification

Figure 10 shows a typical base peak chromatogram of tryptic peptides of a primary tumor of PC3 cells grown in an immune deficient mouse.

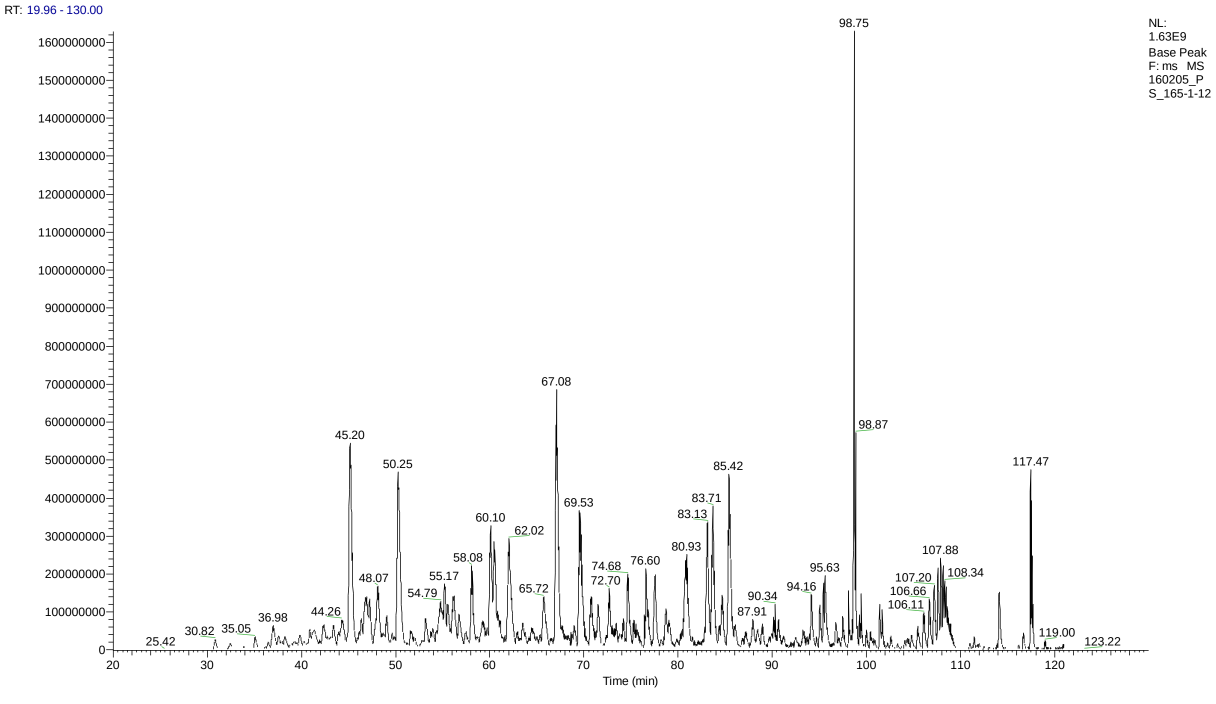


Figure 10: Base peak Chromatogram (BPC) of a tryptic peptide of a primary tumor derived from PC3 cells grown in the xenograft mouse.

The raw data was processed using MaxQuant with LFQ mode enabled (further parameters are reported in the Methods part). To identify a protein in the LC-MSMS data set the software needs to first identify a unique peptide using a FASTA database of the organism correlating to the measured sample (in this case a human database as the primary tumor was composed of human PC3 or DU145 cells). To achieve this, the software first performs an *in-silico* digestion of the whole proteome in the database and builds theoretical MS2 fragment spectra for each possible tryptic peptide as well as calculating the parent ion mass for each peptide. Now extracted ion chromatograms are computed from the theoretical MS1 mass in the specified mass thresholds. Figure 11 shows an extracted ion chromatogram of one of the peptides of Vinculin in one of the PC3 samples.

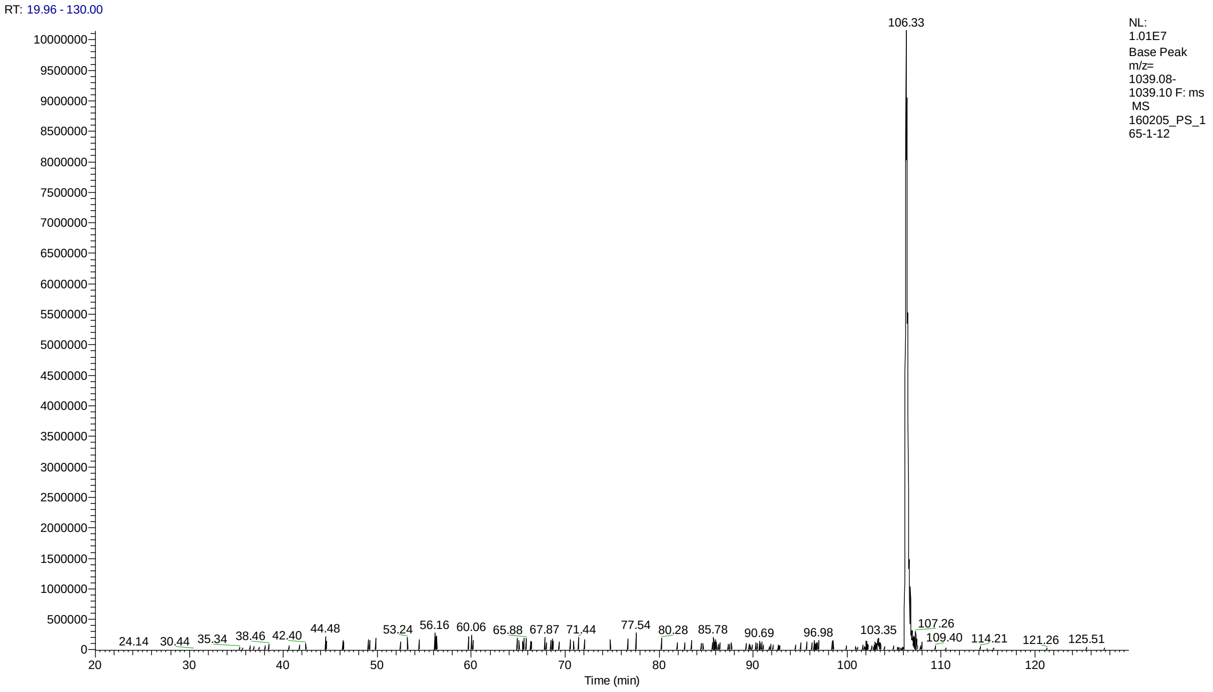


Figure 11: Extracted Ion Chromatogram (XIC) of the precursor mass m/z 1038.59 at 106.33 min in the sample of one of the PC3 mice (165-1-12).

The singular peak in Figure 11 corresponds to the precursor mass of the Vinculin peptide with a m/z of 1038.59 at 106.33 min in the gradient.

Figure 12 shows a zoomed in view of the MS1 spectrum at 106.33 min showing the isotopic envelope of a doubly charged peptide.

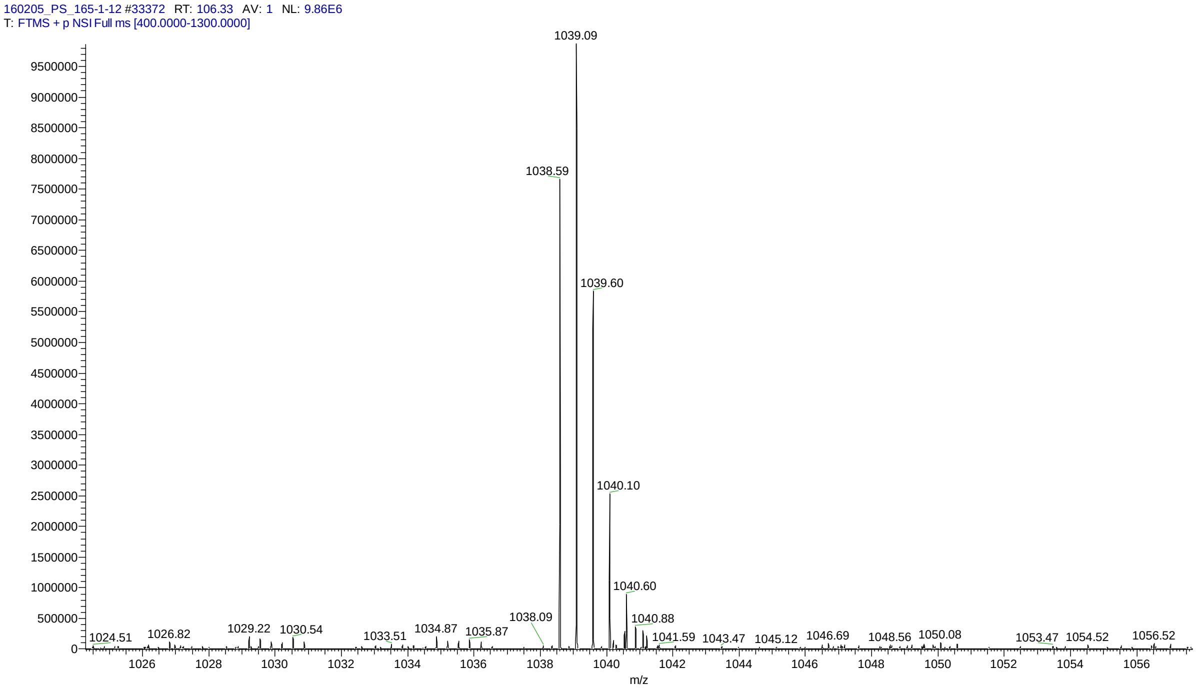


Figure 12: MS1 Spectrum at 106.33 min of a peptide of a primary tumor derived from PC3 cells grown in a xenograft mouse (165-1-12). The view is zoomed in to show the isotopic envelope of the peptide with precursor mass of 1038.59 m/z.

The isotopic envelope shown in Figure 12 shows the distribution for a doubly-charged peptide with the first peak (mono-isotopic mass) less intense than the following peak but more intense than the third. In a typical isotopic distribution for a doubly-charged peptide (mostly below 1000 Da) the mono-isotopic peak would be the most intense. In this example however, the peptide is large enough to shift the possibility of incorporating one C^{13} Isotope to be the most likely, resulting in the second peak being the most intense.

To determine if the peak corresponding to the precursor mass of 1038.59 m/z belongs to a specific peptide of Vinculin the search engine now compares the theoretical and the detected MS2 spectra. The MS2 spectrum of the peptide with 1038.95 m/z is shown in Figure 13.

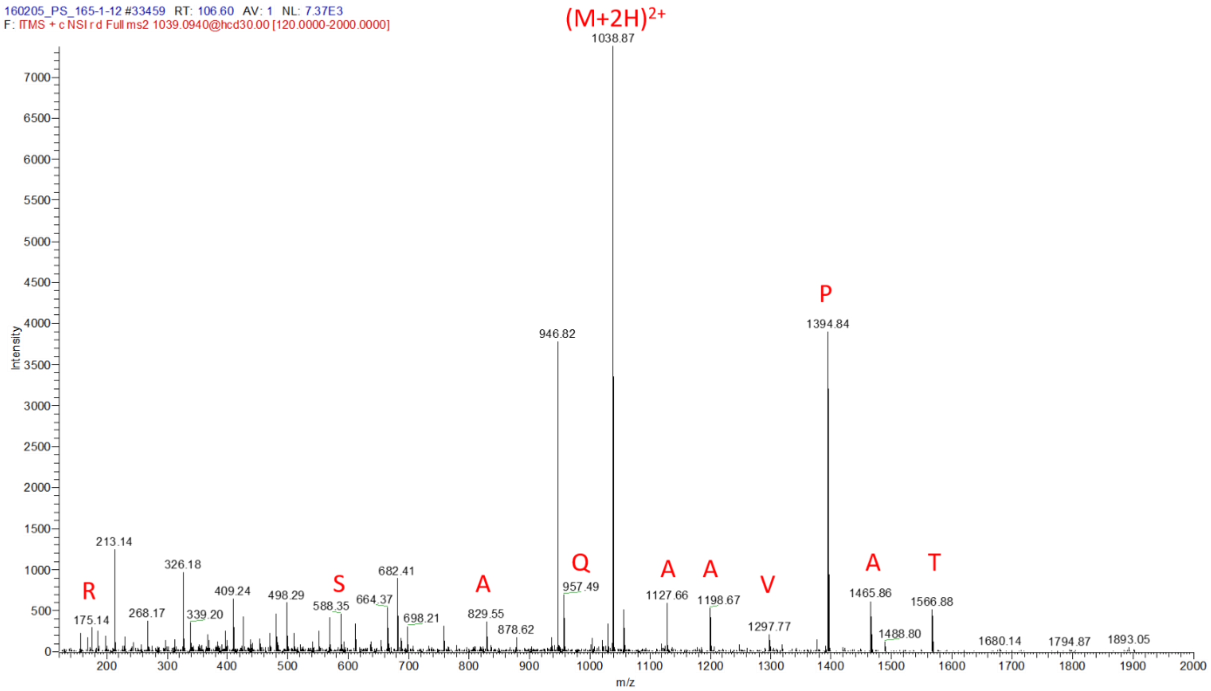


Figure 13: MS2 spectrum of the peptide with an m/z of 1038.95 (AIPDLTAPVAAVQAAVSNLVR). The *y*-ion series of the amino acid sequence for the corresponding peptide is overlaid in red.

Figure 13 shows a MS2 spectrum of the peptide with a nearly complete *y*-ion series. For additional assurance the precursor peak can also be detected. HCD spectra tend to favor the y-ion series at higher m/z.

If sufficient consensus between the theoretical and detected spectra is observed (MS2 peaks need to be inside a prior specified m/z threshold) the spectra is assigned to an amino acid sequence in this case AIPDLTAPVAAVQAAVSNLVR which is a unique peptide of Vinculin. After assigning XICs to their corresponding peptides the search engine is now able to perform label-free quantitation on the data set. In case of MaxQuant this is done using the peak areas of the MS1 peaks. Because this is a relative quantification method, meaning there is no reference standard, two datasets need to be compared directly. Figure 14 shows the MS1 peak area comparison of the Vinculin peptide of the PC3 sample 165-1-12 and the DU145 sample 202-2-12.

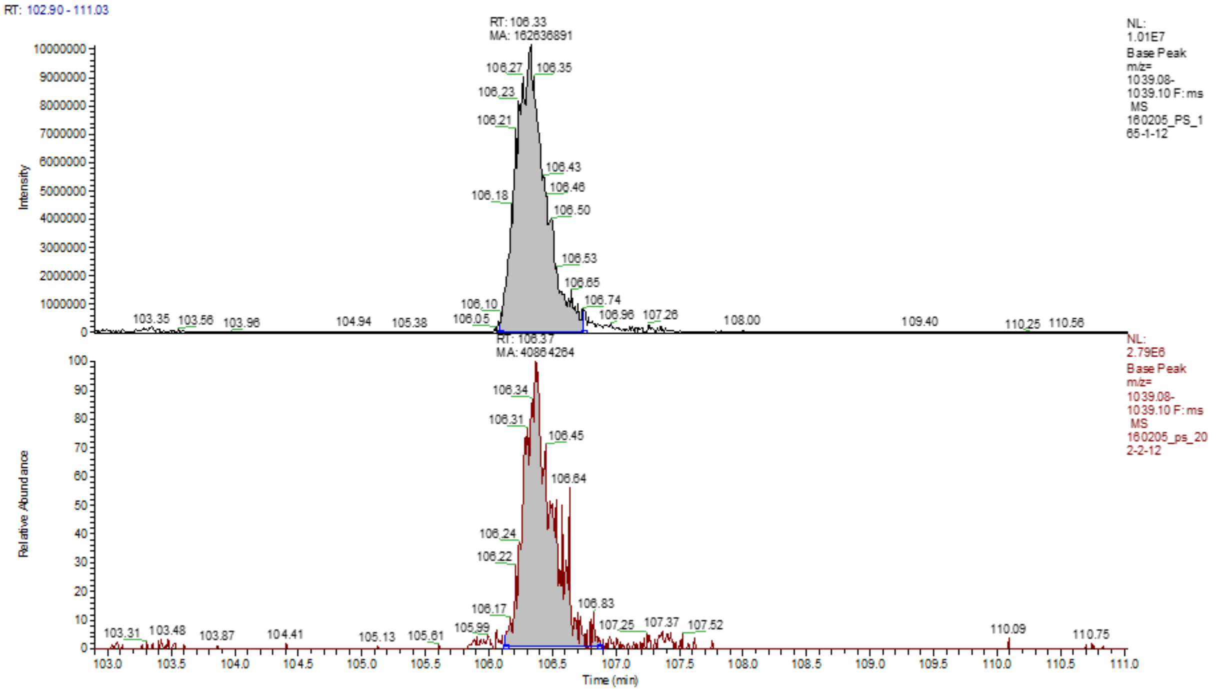


Figure 14: MS1 peak area comparison of the peak at 106.33 min with an m/z of 1038.95 corresponding to the peptide of Vinculin with amino acid sequence of AIPDLTAPVAAVQAAVSNLVR. Shown are the integration boundaries in blue as set manually in Xcalibur for illustration of the LFQ process in MaxQuant.

Figure 14 illustrates the LFQ process in MaxQuant using the raw data viewer Xcalibur and manually set integration boundaries which do not correspond to those chosen by MaxQuant. In this example, the peak area of the peptide in the PC3 sample is determined to be 1.626e^{8} and for the DU145 sample 0.408e^{8} (arbitrary units). The difference in abundance for this peptide for these two samples is therefore around 4-fold higher in the PC3 sample compared to the DU145 sample. MaxQuant normalizes these values to minimize the effect of different injection concentrations of the different samples and sums the peptide peak areas of one protein and reports the LFQ intensities for the whole protein rather than every single peptide.

5.1.2 Post-processing of the data

After identification and quantification of the data by MaxQuant, the resulting data needs to be further processed for statistical relevance. Here, a combination of the R package DEP from Bioconductor as well as Perseus is used for data analysis. Because the raw data are searched against an additional contaminant database (including human keratin, BSA and other proteins that might be introduced externally to the sample) and a false discovery rate (FDR) is computed using the inverse amino acid sequences these hits need to be filtered out. The data are then log2 transformed to reduce variability and conform more closely to a normal distribution. As a next step the data should be examined more closely, Figure 15 shows the protein identification overlap of the different samples.

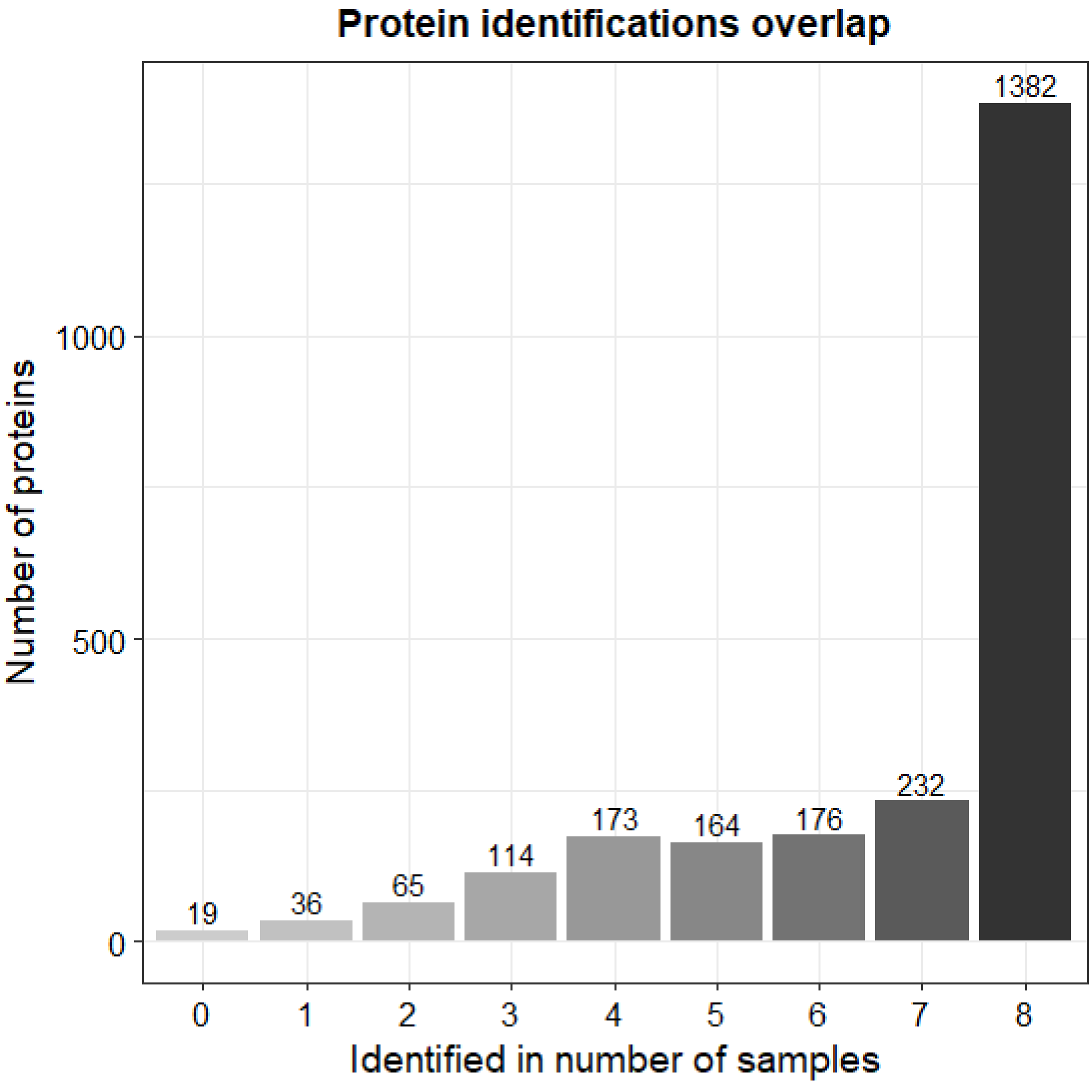


Figure 15: Protein overlap between the 8 different samples.

Figure 15 shows that most of the proteins are quantified in all 8 samples with some proteins only quantified in a subset of the samples. This is caused by missing values introduced into the dataset because of the random sampling of the DDA method of the mass spectrometer. The 19 Proteins that were identified in 0 of the 8 samples represent artifacts of the search engine where a peptide for a protein could be identified by an MS2 spectrum but corresponding MS1 peak of the peptide could not be detected by the algorithm in either of the 8 samples.

Label-free quantitation data based on MS1 peak areas acquired by DDA has to deal with missing values in the data set originating in the inherent random sampling of the MS method. To further process the data, these missing values need to be imputed. However, too many missing values lead to a bias or false imputation and therefore some of these missing values need to be filtered out. Here proteins were filtered out that had less than 3 quantitative values across one sample group (PC3 or DU145) resulting in the reduction of the total number of proteins from 2371 to 2156. Figure 16 shows the number of quantified proteins after filtering.

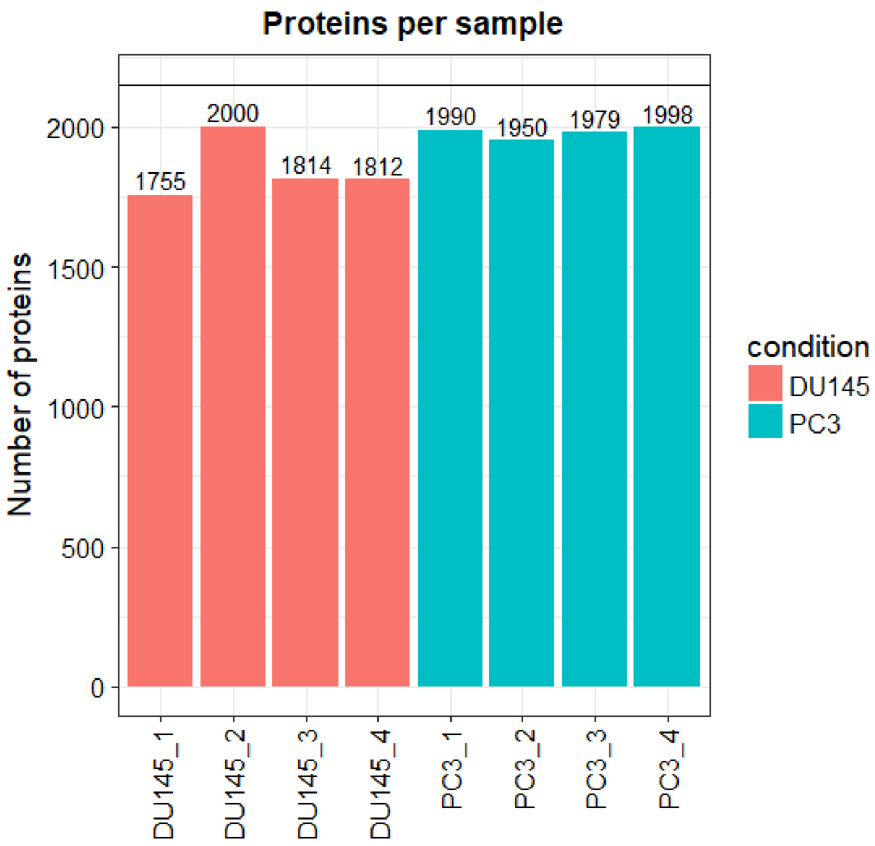


Figure 16: Number of quantified proteins per sample after filtering.

In a next step the data are normalized to reduce variance across the samples as well as variance stabilized. Before the imputation method can be chosen the type of missing value prevalent in the data should be determined. Figure 17 depicts the missing value distribution across the samples.

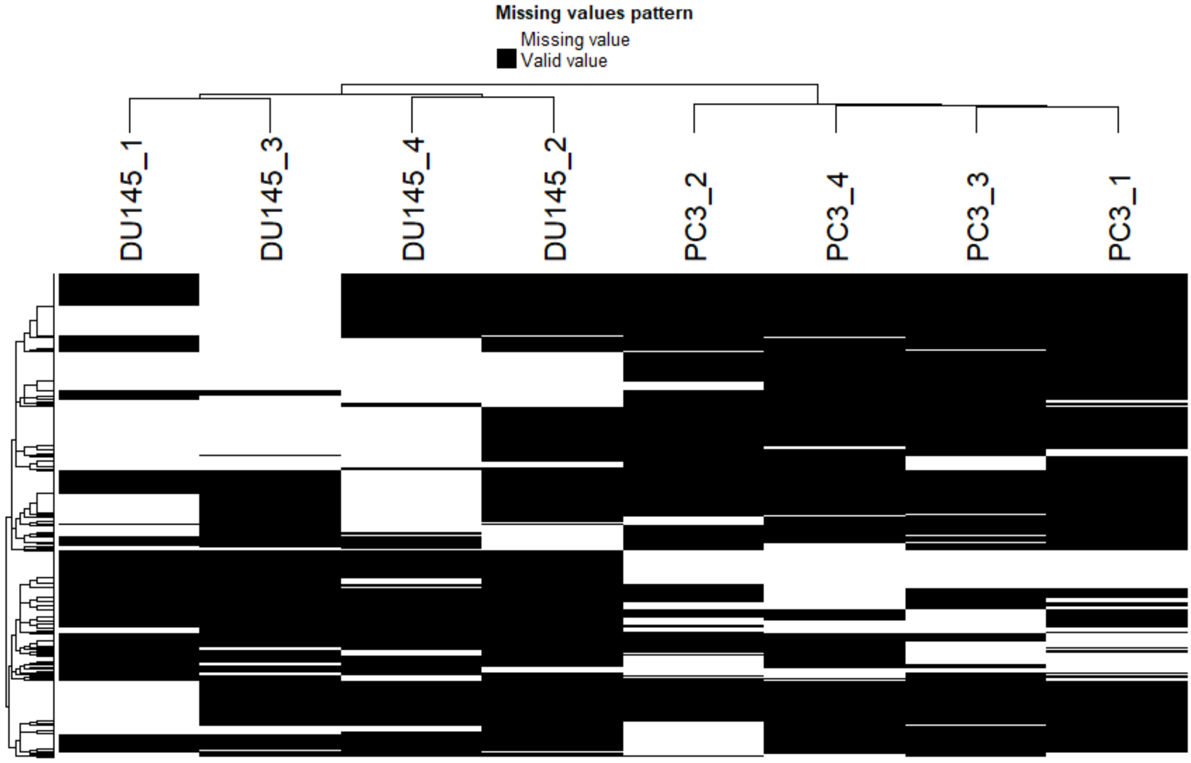


Figure 17: Missing value distribution across the 8 samples. White spaces indicate missing values while black spaces indicate valid values.

There are two types of missing values, the first one being missing at random (MAR) which occurs because of the random nature of the DDA method and does not correlate with the sample composition. The second type is missing not at random (MNAR) which occurs when a certain protein has a low expression and is therefore beneath the limit of quantification or detection and MaxQuant's peak picking algorithm could not find the corresponding MS1 peak. MNAR values can be seen in Figure 17 in the top left corner where one cluster of proteins was not quantified in all of the DU145 samples as well as on the right in the middle where the same can be observed for a cluster of proteins which were not quantified in all of the PC3 samples. Single white space in a single column of the heatmap correspond to MAR values. Because this data contains many MNAR values, random draws from a left shifted normal distribution was chosen as imputation method. Of the 2156 quantified proteins, 773 protein quantification values across all samples were imputed (36%). Figure 18 illustrates the effect of the imputation on the data.

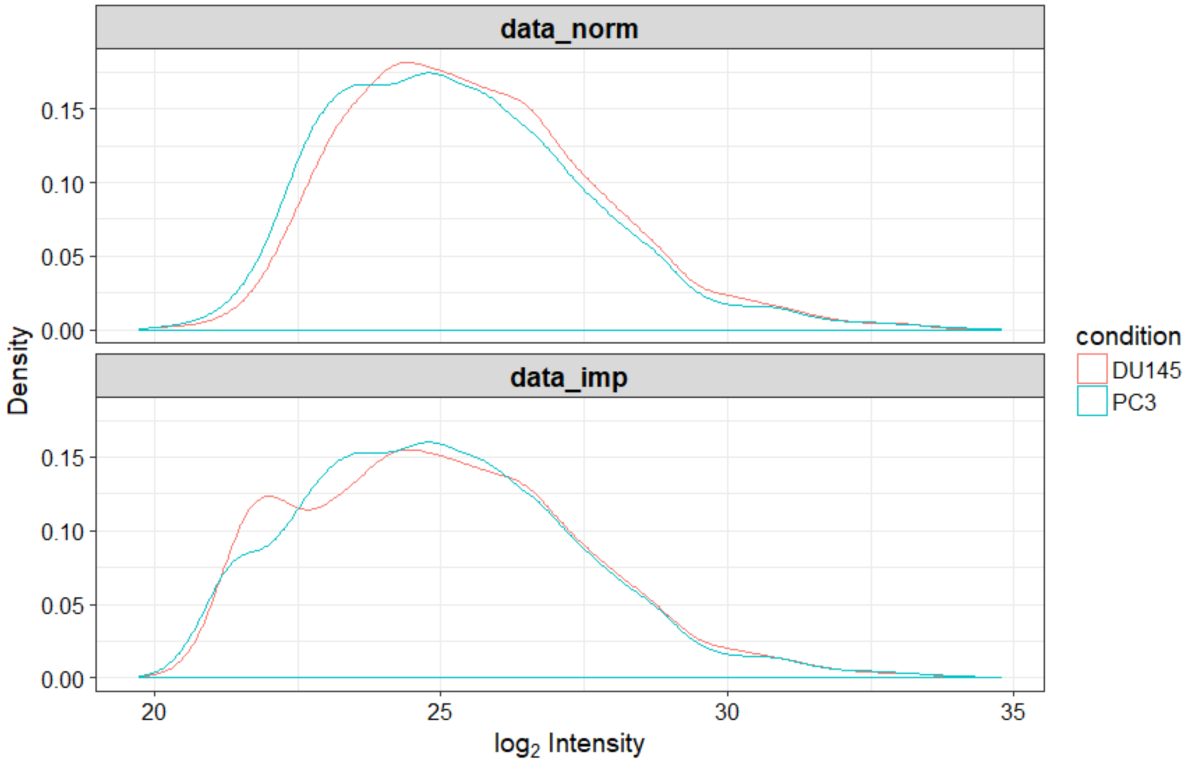


Figure 18: Density distribution of the data before (data_norm) and after imputation (data_imp) using random draws from a left shifted normal distribution.

As expected, after imputation a local maximum of the density distribution for the DU145 sample is observed indicating imputation with low intensity values. The more values needed to be imputed the lower the reliability of the following statistical tests as these values are completely artificially generated.

5.1.3 Statistical analysis

After filtering and imputation, the data is ready for statistical testing and processing. To obtain an overview of the data, PCA is computed (Figure 19).

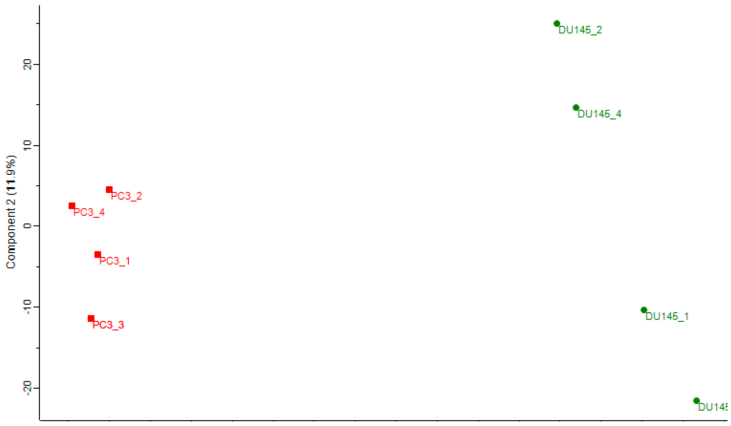


Figure 19: Principle component analysis (PCA) of the data. Component 1 and 2 are shown for each sample. PC3 sample are shown as red squares and DU145 samples are shown as green circles.

Figure 19 shows a good clustering between PC3 and DU145 samples with a clear distinction between the two sample groups. This was to be expected because these two prostate cell lines had shown two different phenotypes with PC3 being mesenchymal and DU145 being epithelial in previous experiments performed by Prof. Tobias Lange (Anatomical Institute, UKE).

Figure 20 shows the protein projection for the PCA.

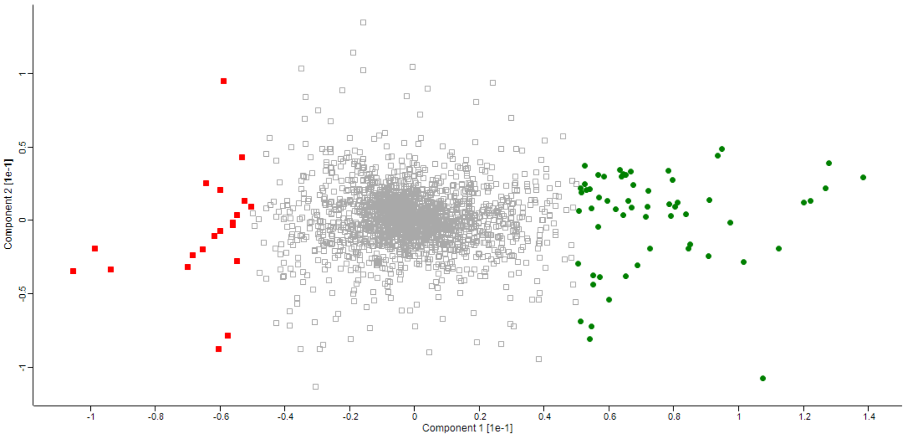


Figure 20: Protein projection of the PCA. Proteins responsible for the left shift of the PC3 samples are marked as red squares and proteins responsible for the right shift of the DU145 samples are marked as green circles. An arbitrary cutoff of < -0.5 and > 0.5 was chosen.

Figure 20 shows the 79 proteins which are responsible for the left/right shift of the PC3/DU145 samples (Supplement table 1). These proteins can be considered as the first potential marker proteins distinguishing between the PC3 (20 proteins) and DU145 (59 proteins) cells but as a PCA does not involve any statistical testing for variance and the chosen threshold of |0.5| is arbitrarily chosen, these proteins would need to be further evaluated.

For a more thorough analysis of differentially expressed proteins, statistical tests need to be performed. There are two groups in this dataset: the PC3 and the DU145 group. Because of this it is possible to test this dataset using a pair-wise model. This model is combined with an empirical Bayes statistic called LIMMA (linear models for microarray data) which is routinely used in microarray assays but can be used for any type of data. Using this t-Test a total of 103 significantly different proteins were determined, 32 showing higher concentration in the PC3 group and 71 showing higher concentration in the DU145 group (the list of all 103 significantly different proteins can be found in supplement table 2). As expected these numbers differ from those found using the PCA because a PCA does not compute any p-values and thus lacks a value for statistical relevance. The significantly different proteins found by the t-Test are shown as a volcano-plot in Figure 21.

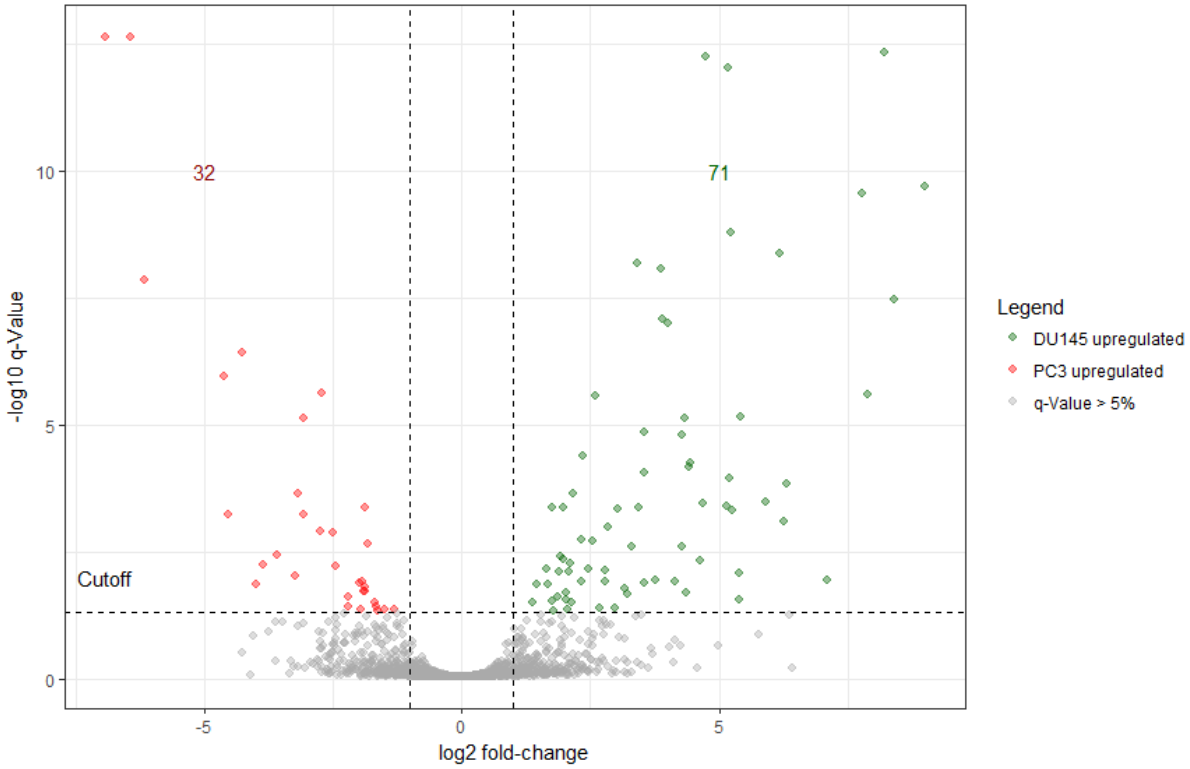


Figure 21: Volcano-Plot of the differentially expressed proteins using a pair-wise comparison. The dashed lines represent the log2 fold-change cutoff of 1 and the adj. p-Value cutoff of max 5%. Proteins showing higher concentration in the PC3 group fulfilling these criteria are marked in red and proteins showing higher concentrations in the DU145 group are marked in green.

A list of all proteins with their log2 fold-change as well as adjusted p-Values (q-Values) can be found in supplement table 2. When comparing the proteins found by the PCA and the significantly different proteins determined by t-Test (see supplement table 1) it is evident, that in the PCA list only proteins with a log2 fold-change of more than |3.0| are present. This was expected as the PCA only takes the log2 intensities of each sample for each protein into account. This leads to a bias for high abundant proteins even though they were measured with high variance between samples (see supplement table 1, ACTBL2 which shows a log2 fold change of -4.09 but an adjusted p-Value of 0.867) as well as a loss of hits for lower abundant proteins but with very low variance (see supplement table 2, CD44 which has a log2 ratio of 1.87 and an adjusted p-Value of 0.0155).

5.2 Text Mining

5.2.1 Network analysis and Gene Ontology enrichment

The most common way to further analyze differential -omics data is the so-called network analysis. This kind of analysis tries to connect the proteins showing differential concentrations via their Gene Ontology (GO) terms or other literature-based criteria. Figure 22 shows such a network for these proteins found in the mouse experiment using the ReactomeFI app in Cytoscape software.

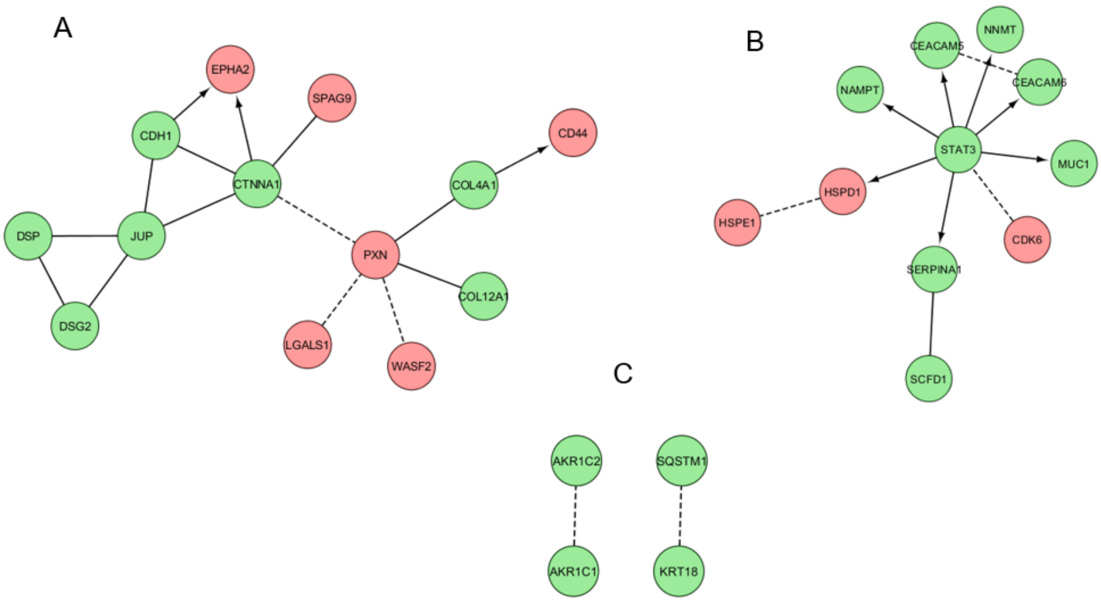


Figure 22: Protein networks derived from the 103 significantly different proteins. Proteins showing higher concentration in the PC3 group are shown in red and proteins showing higher concentrations in the DU145 group are shown in green. Connections with arrows indicate an activation, connections with complete lines indicate protein-protein interactions and connections with dashed lines indicate predicted interactions. For legibility, only proteins with one or more interaction partners are shown.

Figure 22 shows three distinct interaction networks after GLay clustering. Even though these kinds of representations are well established in the -omics community their significance is highly debatable. Here, the GLay clustering algorithm was used to identify cluster in the network, when using a different algorithm different clusters are obtained diminishing the meaningfulness and exacerbate the interpretation of such a network for the researcher. Another difficulty are the different databases, in this case the ReactomeFI annotations were used for network calculation. Another prominent tool is the STRING database. When using STRING different networks are obtained, again diminishing the usefulness of such networks. In the end the researcher must have either complete background knowledge about these proteins and their interactions or has to do a thorough literature search. Such networks should only be used for visualization of data, when a tabular form would be too complex or confusing for the reader.

In Figure 22 CD44 shows higher concentrations in the PC3 group, when looking into the literature CD44 seems to be overexpressed in highly tumorigenic cells, which is in good correlation with this study (PC3 cells being highly metastatic in comparison to DU145 cells).

Another, maybe more meaningful interpretation of the data is the so-called GO enrichment. In this case the GO-Terms of the significantly different proteins are queried against a larger dataset (i.e. the whole human proteome/genome) and overrepresented biological pathways, molecular functions or cell compartments can be displayed. Table 1 shows the top 10 enriched biological pathways for the 32 proteins showing higher concentration in the PC3 group and table 2 shows the top 10 biological pathways for the 71 proteins showing higher concentration in the DU145 group.

Table 1: GO-Enrichment of the biological pathways in the 32 proteins showing higher concentration in the PC3 group. Shown are the pathways, the corrected p-Value the number of proteins in the pathway (x), the number of total proteins in the pathway in the human proteome (n) and the total number of proteins in the human proteome set used for enrichment (N) as well as the gene names.

Description	corr p-value	x	n	N	Genes in test set
response to axon injury	6.46E-03	3	27	17788	LGALS1|MAP1B|NEFL
cellular component assembly	6.46E-03	9	913	17788	PSMD10|MAP1B|PXN|HIST1H1D|HMGA1|NEFL|WASF2|HSPD1|PLEC
axon regeneration in the peripheral nervous system	6.46E-03	2	4	17788	MAP1B|NEFL
cellular macromolecular complex subunit organization	8.19E-03	6	356	17788	PSMD10|PXN|HIST1H1D|HMGA1|NEFL|HSPD1
negative regulation of DNA damage response, signal transduction by	8.53E-03	2	6	17788	PSMD10|CD44

p53 class mediator					
cellular component biogenesis	8.53E-03	9	1035	17788	PSMD10|MAP1B|PXN|HIST1H1D|HMGA1|NEFL|WASF2|HSPD1|PLEC
nervous system development	1.71E-02	9	1155	17788	LGALS1|CDK6|AHNAK|MAP1B|NEFL|PHGDH|NES|CD44|EPHA2
cellular protein complex assembly	1.81E-02	4	151	17788	PSMD10|PXN|NEFL|HSPD1
regulation of DNA damage response, signal transduction by p53 class mediator	1.96E-02	2	11	17788	PSMD10|CD44
axon regeneration	2.00E-02	2	12	17788	MAP1B|NEFL

Table 1 shows significant enrichment of axon and nervous system pathways in the PC3 upregulated proteins. Axon development is typically associated with dendritic and fast cell growth. Using this enrichment, a neuronal phenotype for the PC3 cells can be deducted which correlates with the high metastatic potential of these cells.

Table 2: GO-Enrichment of the biological pathways in the 71 proteins showing higher concentration in the DU145 group. Shown are the pathways, the corrected p-Value the number of proteins in the pathway (x), the number of total proteins in the pathway in the human proteome (n) and the total number of proteins in the human proteome set used for enrichment (N) as well as the gene names.

Description	corr p-value	x	n	N	Genes in test set
cellular ketone metabolic process	3.39E-07	17	577	14299	CBR1|LYPLA1|AKR1C1|AKR1C3|AKR1C2|PGD|PTGR1|ASRGL1|ASS1|UGDH|GRHPR|GCLC|KYNU|CKB|PRODH|FBP1|CBR3
oxidation reduction	9.50E-07	17	646	14299	CBR1|CYB5A|NQO1|GPX2|AKR1C1|GSR|AKR1C3|AKR1C2|SQRDL|PGD|PTGR1|UGDH|GRHPR|AKR1B10|ALDH16A1|PRODH|CBR3
small molecule metabolic process	2.88E-06	23	1369	14299	CBR1|BCHE|LYPLA1|AKR1C1|GSR|AKR1C3|AKR1C2|NUDT5|PGD|PTGR1|ASRGL1|ASS1|TYMP|UGDH|GRHPR|GCLC|KYNU|NAMPT|CKB|PRODH|FBP1|SULT1A4|CBR3

oxoacid metabolic process	3.36E-06	15	563	14299	LYPLA1|AKR1C1|AKR1C3|AKR1C2|PGD|PTGR1|ASRGL1|ASS1|UGDH|GRHPR|GCLC|KYNU|CKB|PRODH|FBP1
carboxylic acid metabolic process	3.36E-06	15	563	14299	LYPLA1|AKR1C1|AKR1C3|AKR1C2|PGD|PTGR1|ASRGL1|ASS1|UGDH|GRHPR|GCLC|KYNU|CKB|PRODH|FBP1
organic acid metabolic process	3.36E-06	15	570	14299	LYPLA1|AKR1C1|AKR1C3|AKR1C2|PGD|PTGR1|ASRGL1|ASS1|UGDH|GRHPR|GCLC|KYNU|CKB|PRODH|FBP1
response to xenobiotic stimulus	7.77E-05	5	37	14299	NQO1|GCLC|KYNU|AKR1C1|SULT1A4
cellular response to indole-3-methanol	8.59E-05	3	5	14299	JUP|CDH1|CTNNA1
response to indole-3-methanol	8.59E-05	3	5	14299	JUP|CDH1|CTNNA1
amine metabolic process	1.59E-04	11	409	14299	UGDH|BCHE|GCLC|KYNU|CKB|GUSB|PGD|PRODH|SULT1A4|ASRGL1|ASS1

Table 2 shows significant enrichment in metabolic processes. This enrichment shows a typical cancer phenotype where proteins connected to metabolic processes are highly upregulated which in turn can lead to oxygen radicals which lead to oncogenic mutations [40].

5.2.2 Development of a text mining Script for differential proteomics

In cooperation with Dr. Jemma Wu (Macquarie University, APAF), who was responsible for the coding of the script while I provided the conceptual design and testing, an R script was developed to examine differential proteomics data in context with a specific process or keyword. To achieve this, the first iteration of the script used a list of gene names of the proteins of interest, the taxonomy number and a keyword of the desired condition or process and looks up the iHOP database to access all known synonyms for each gene name in the list. After retrieving all names, the script starts a literature search using www.pubmed.org to find all scientific articles which have the gene name or its synonyms plus the desired condition or process in their title to return hits with high confidence. These hits are then given a value between 1 and 3 with 1 being hits that include reviews, 2 being hits that do not include reviews and 3 no hits at all for a specific gene and keyword combination. The Excel output file contains information about the author, publication year, abstract, allocated value between 1 and 3 and several other criteria. Figure 23 shows a schematic of the script.

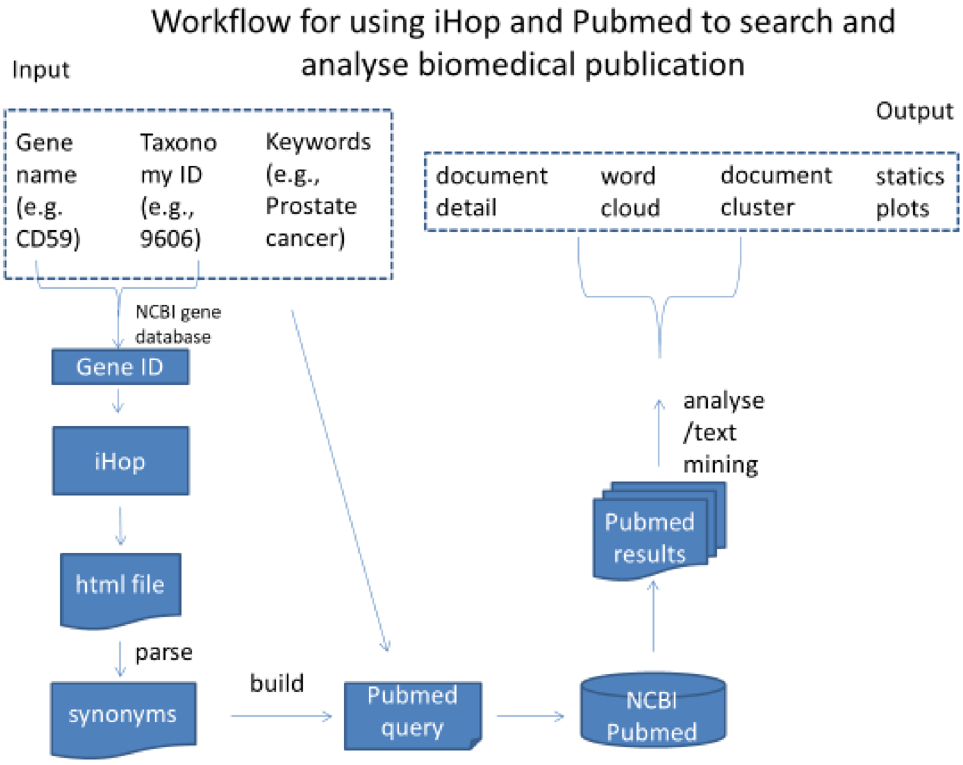


Figure 23: Schematic of the workflow of the first iteration of the text mining script written in R.

As of the second quarter of 2018 the iHOP server was taken off the internet and is therefore not accessible for the script to use. Because of this a second iteration of the script was developed using the www.Uniprot.org database for retrieval of gene synonyms instead of iHOP. Because Uniprot is used, the input was changed to include Uniprot identifiers instead of gene names thus forgoing the need to access the gene database. Other than that, the script operates the same as the first iteration. A schematic of the second iteration of the script can be seen in Figure 24.

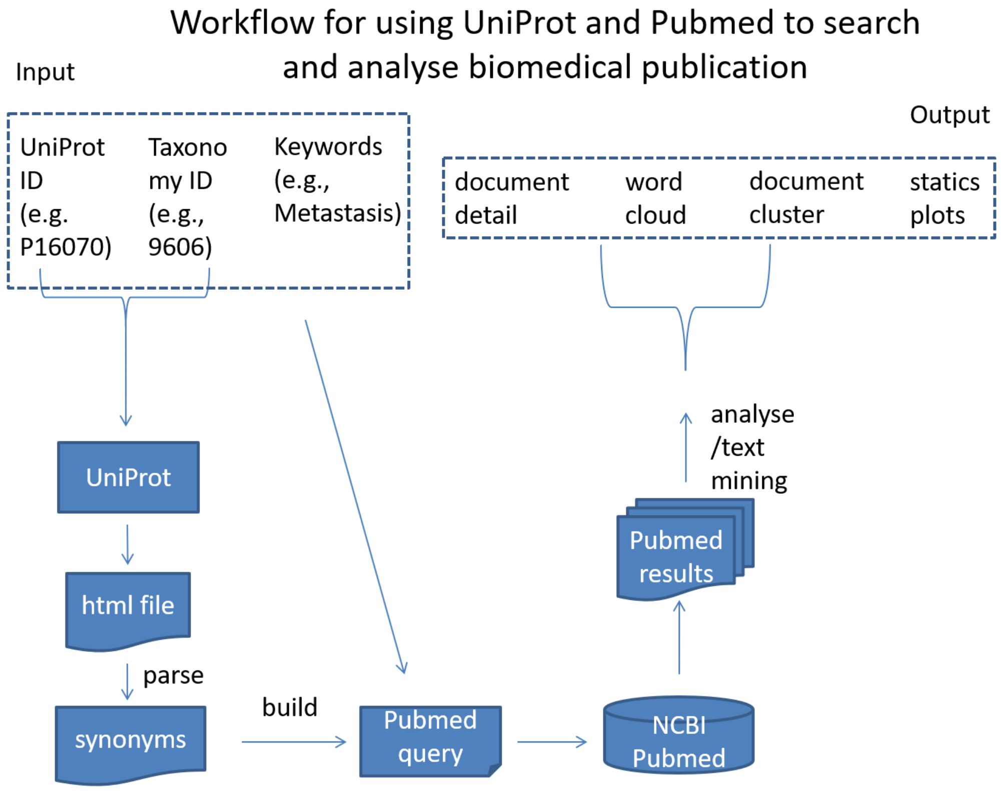


Figure 24: Schematic of the workflow of the second iteration of the text mining script written in R.

Using this new script, the significantly different proteins of the PC3 and DU145 groups were used to query a search in conjunction with the keyword metastasis. Table 3 shows the first 5 lines of the summary tab of the query output of the script.

Table 3: First 5 lines of the summary tab of the excel output file generated by the text mining script using UniProt.

N	UniProtID	TaxID	Synonyms	Keyword	KeywordInTitleOnly	Results	Category	Other	PubmedQuery
1	P48681	9606	NES,Nbla00170	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("NES"[TI] OR "Nbla00170"[TI])
2	Q15847	9606	ADIRF,AFRO,APM2,C10orf116	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("ADIRF"[TI] OR "AFRO"[TI] OR "APM2"[TI] OR "C10orf116"[TI])
3	Q1L6U9	9606	MSMP,PSMP	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("MSMP"[TI] OR "PSMP"[TI])
4	P29034	9606	S100A2,S100L	Metastasis	Yes	2	2	0	"Metastasis"[TI] AND ("S100A2"[TI] OR "S100L"[TI])
5	P46821	9606	MAP1B	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("MAP1B"[TI])

Table 3 shows part of the first tab in the generated Excel file for the search. This version of the script only uses gene name synonyms for the search, it is planned to also include the protein names in the future to have a more comprehensive query for the Pubmed search. 3 out of the 5 displayed queries resulted in no hits for this specific combination of gene names and keyword, meaning that these proteins currently are not associated with metastasis. These would be interesting although probably challenging targets for further research. The second and fourth query resulted in 1/2 hits respectively. Table 4 shows the result tab of the second query and table 5 shows the result tab of the fourth query.

Table 4: Result tab of the second query of generated excel file by the R script. The abstract and MeSH terms have been shortened for greater legibility.

UniProtID	GeneID	TaxID	Synonyms	Keywords	KeywordInTitleOnly	Results	Category	Other	PubmedQuery
Q15847		9606	ADIRF, AFRO, APM2, C10orf116	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("ADIRF"[TI] OR "AFRO"[TI] OR "APM2"[TI] OR "C10orf116"[TI])
Title	**Abstract**	**Year**	**Author**	**Country**	**Link**	**ConditionInTitle**	**IsReview**	**MeSH**	
High incidence of prostate cancer metastasis in Afro-Brazilian men with low educational levels: a retrospective observational study.	BACKGROUND: This study investigated factors related to ethnicity and educational level, their correlation with tumor stage at the time of diagnosis, and their influence on treatment outcomes in patients with prostate cancer.…	2013	AB de Souza,HG Guedes,VC Oliveira,FA de Araújo,CC Ramos,KC Medeiros,RF Araújo	England	https://www.ncbi.nlm.nih.gov/pubmed/23734601	TRUE	FALSE	Aged,Aged, 80 and over,Brazil…	

Table 5: Result tab of the second query of generated excel file by the R script. The abstract and MeSH terms have been shortened for greater legibility.

UniProtID	GeneID	TaxID	Synonyms	Keywords	KeywordInTitleOnly	TotalResults	Category	Other	PubmedQuery
P29034		9606	S100A2,S100L	Metastasis	Yes	2	2	0	"Metastasis"[TI] AND ("S100A2"[TI] OR "S100L"[TI])
Title	Abstract	Year	Author	Country	Link	ConditionInTitle	IsReview	MeSH	
S100A2 induces metastasis in non-small cell lung cancer.	PURPOSE: S100 proteins are implicated in metastasis development in several cancers...	2009	E Bulk,B Sargin,U Krug,A Hascher,Y Jun,M Knop,C Kerkhoff,V Gerke,R Liersch,RM Mesters,M Hotfilder,A Marra,S Koschmieder,M Dugas,WE Berdel,H Serve,C Müller-Tidow	United States	https://www.ncbi.nlm.nih.gov/pubmed/19118029	TRUE	FALSE	Animals, Carcinoma…	
S100A2 expression as a predictive marker for late cervical metastasis in stage I and II invasive squamous cell carcinoma of the oral cavity.	The purpose of this study was to discover whether S100A2 expression is associated with late cervical metastasis in patients with stage I and II invasive squamous cell carcinoma of the oral cavity...	2005	F Suzuki,N Oridate,A Homma,Y Nakamaru,T Nagahashi,K Yagi,S Yamaguchi,Y Furuta,S Fukuda	Greece	https://www.ncbi.nlm.nih.gov/pubmed/16273244	TRUE	FALSE	Adult, Aged…	

Table 4 shows a case where the script returns a false-positive hit. The problem here was that one of the gene name synonyms is “AFRO” which was found in the title of the article in the word “Afro-Brazilian” and therefore has another meaning as intended. Currently there is no way to adjust the script to exclude false-positives such as these which are caused by synonyms which are words or part of words in the English language and therefore must be excluded manually. Table 5 on the other hand shows the intended results of the script. The gene name of the query as well as the keyword were found in 2 articles. Both articles associate S100A2 expression with an increase in metastasis in different cancers which correlates nicely with the data (S100A2 being upregulated in the PC3 samples). Supplement table 3 shows the complete summary tab for the 32 proteins showing higher concentration in the PC3 samples. 14 (after elimination of the false-positive result for query 2) of the proteins showed at least one literature hit in conjunction with the keyword metastasis with one of them being a category one hit (CD44 which showed 156 hits including 14 review articles). When looking through the results of all hits (data not shown) every article associates the correlating protein directly or indirectly with an increase of metastasis. These text mining results support the statistical data of the experiment and provide additional confidence in these possible protein markers.

5.3 Shipping Test

To assess the influence of non-cooled transport of tryptically digested, desalted and dried peptides from a biological sample an international multi-laboratory experiment was conducted. Briefly, SW480 cells were lysed and tryptically digested, desalted, aliquoted and dried *in vacuo* in Australia. One aliquot was shipped by air to Shanghai to the laboratory of Prof. Yang (Fudan University) and two were sent to Germany to the laboratory of Prof. Schlüter (one of which was sent back to Australia to the lab of Prof. Molloy). Another aliquot was kept in Australia serving as control sample. All samples were kept at -20 °C at the respective sites until measurement. To reduce bias, the samples were measured on a Q Exactive in every laboratory using reversed phase columns of the same batch distributed by the laboratory of Prof. Schlüter. Furthermore, the same LC and MS methods were used. The data was collected at one site and analyzed as one set using MaxQuant version 1.5.8.3. The data from this part of the thesis was published in the journal Analytical Biochemistry with the title “Multi-laboratory analysis of the variability of shipped samples for proteomics following non-cooled international transport” in May 2018.

5.3.1 Reproducibility across laboratories

Figure 25A shows a bar graph with the identified proteins (with at least two unique peptides) in every laboratory. Of special interest were the similarities and differences between the “MQ-Control” and the “MQ-Shipped” samples as those were measured in the same laboratory with the same analysis setup and the same operator thus being the least prone to external bias.

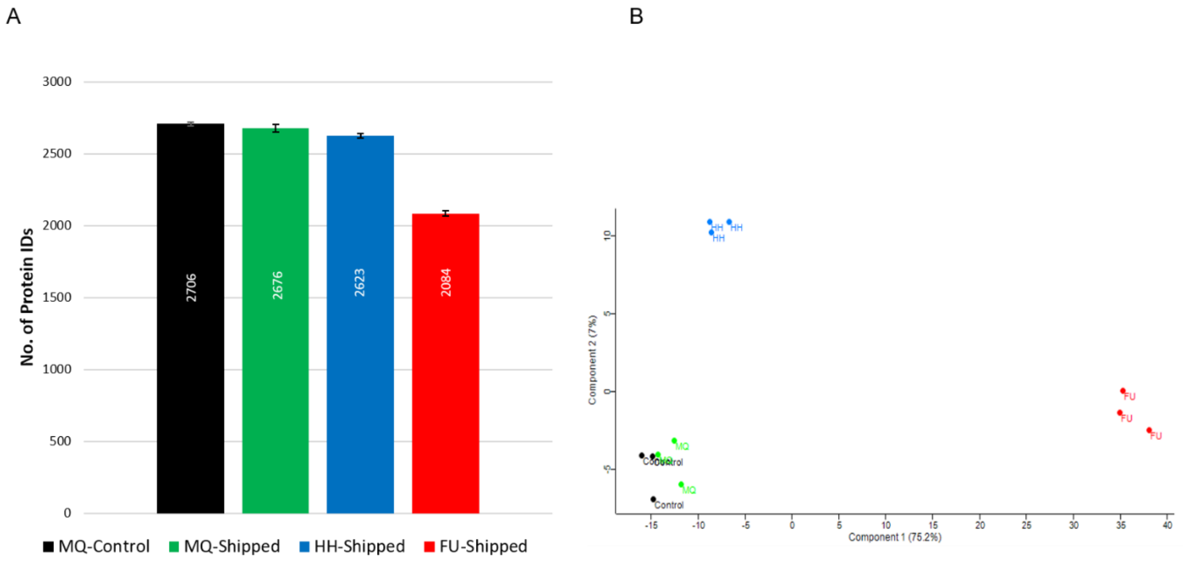


Figure 25: A) The mean number of proteins identified with at least two unique peptides of the technical replicates (n=3) in every laboratory with error bars representing their respective standard deviation. B) Principle component analysis (PCA) of the sample groups.

As shown in Figure 25A, the highest number of proteins was detected in MQ-Control (2706 proteins) followed by MQ-Shipped (2676 proteins), HH-Shipped (2623 proteins) and FU-Shipped (2084 proteins). Additionally, a PCA was conducted (Figure 25B) which shows very high comparability between MQ-Control and MQ-Shipped while clearly separating the other sample groups but showing very low intra laboratory variance. To further assess the reproducibility of the protein identification across the sample groups, the median CVs were computed as shown in Figure 26.

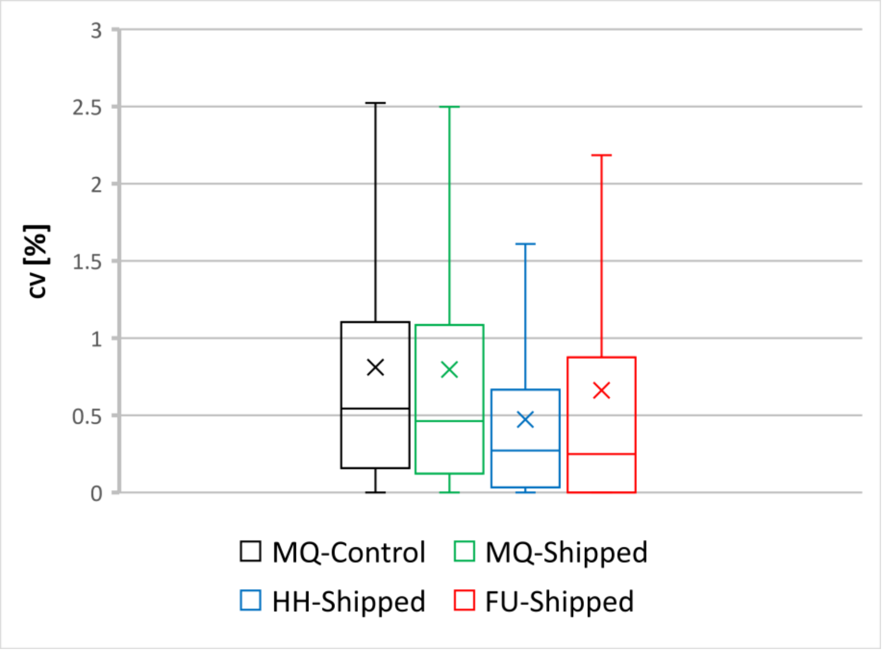


Figure 26: CVs of the different sample groups shown in %. The x in the boxplot marks the median, while the line marks the mean.

Figure 26 shows that the CVs of all samples groups are well below 1% indicating very high intra laboratory measurement reproducibility for the technical replicates.

The overlap between the identified proteins and peptides between the different sample groups were examined using Venn-diagrams (Figure 27).

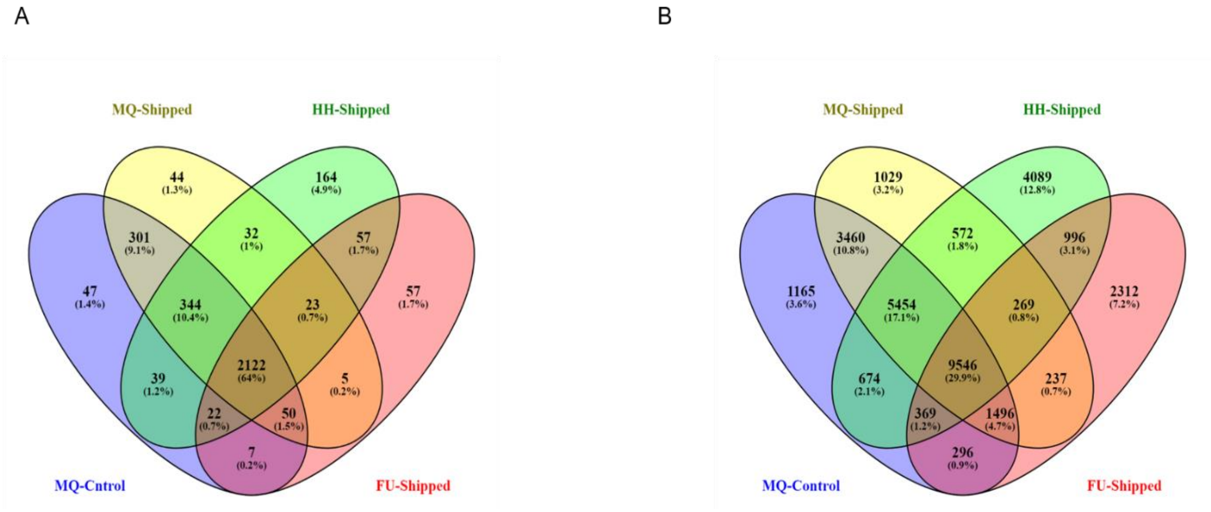


Figure 27: A) Venn-Diagram of the overlap of the identified proteins in each sample group. B) Venn-Diagram of the overlap of the identified peptides in each sample group. Colour code: MQ-Control (blue), MQ-Shipped (yellow), HH-Shipped (green) and FU-Shipped (red).

2122 proteins (64%) were identified in all four sample groups, 78.7% protein identity was detected between the HH-Shipped and MQ-Shipped sample and 92.8% protein identity was observed between MQ-Shipped and MQ-Control again showing very high comparability for these two groups. Figure 27B shows the peptide overlap between the sample groups. Here 9546 peptides (29.9%) were identified in all four groups. 56.2% peptide identity was observed between the HH-Shipped and MQ-Shipped group and 81.2% peptide identity between MQ-Shipped and MQ-Control.

The number of unique peptides per protein for those uniquely identified in only one sample group (Figure 27A) were plotted in Figure 28 to exclude possible bias for low peptide/protein identification for certain groups.

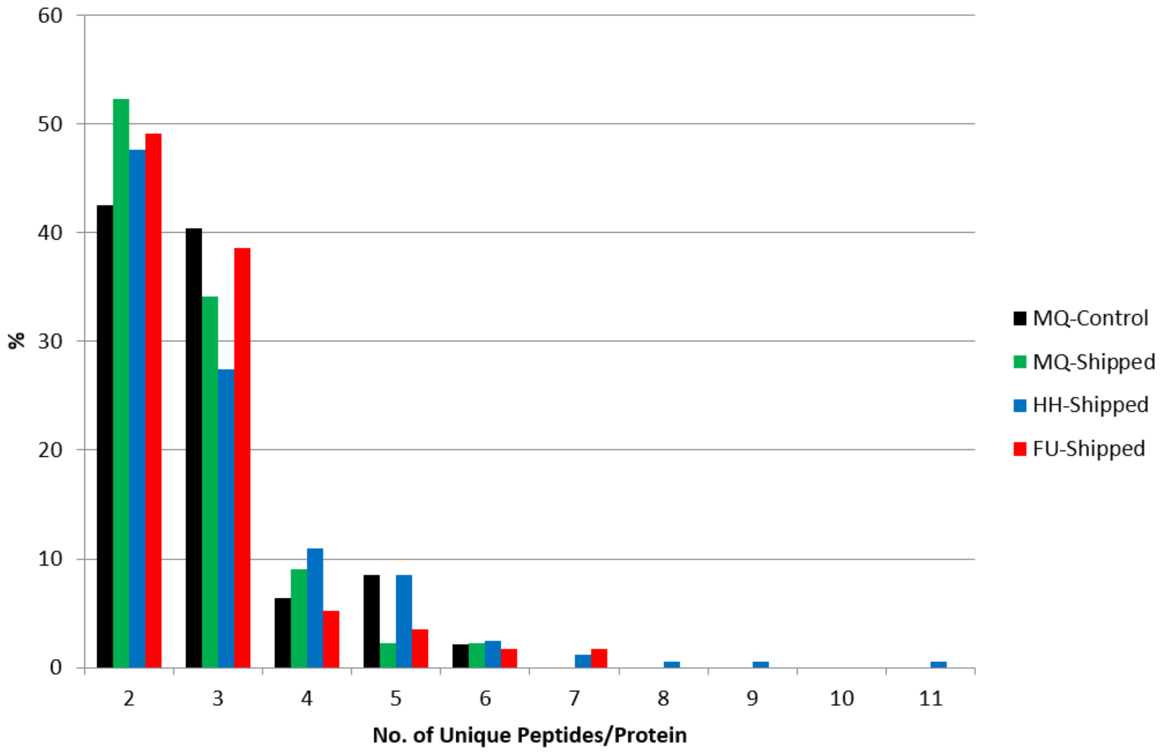


Figure 28: Number of unique peptides/protein in % for each of the sample groups. Here only those proteins were considered which were uniquely identified in only one of the groups.

No apparent bias can be seen in the distribution of unique peptides/protein in Figure 28.

To examine possible sample loss, the protein abundances were compared using label-free quantitation in MaxQuant (Figure 29).

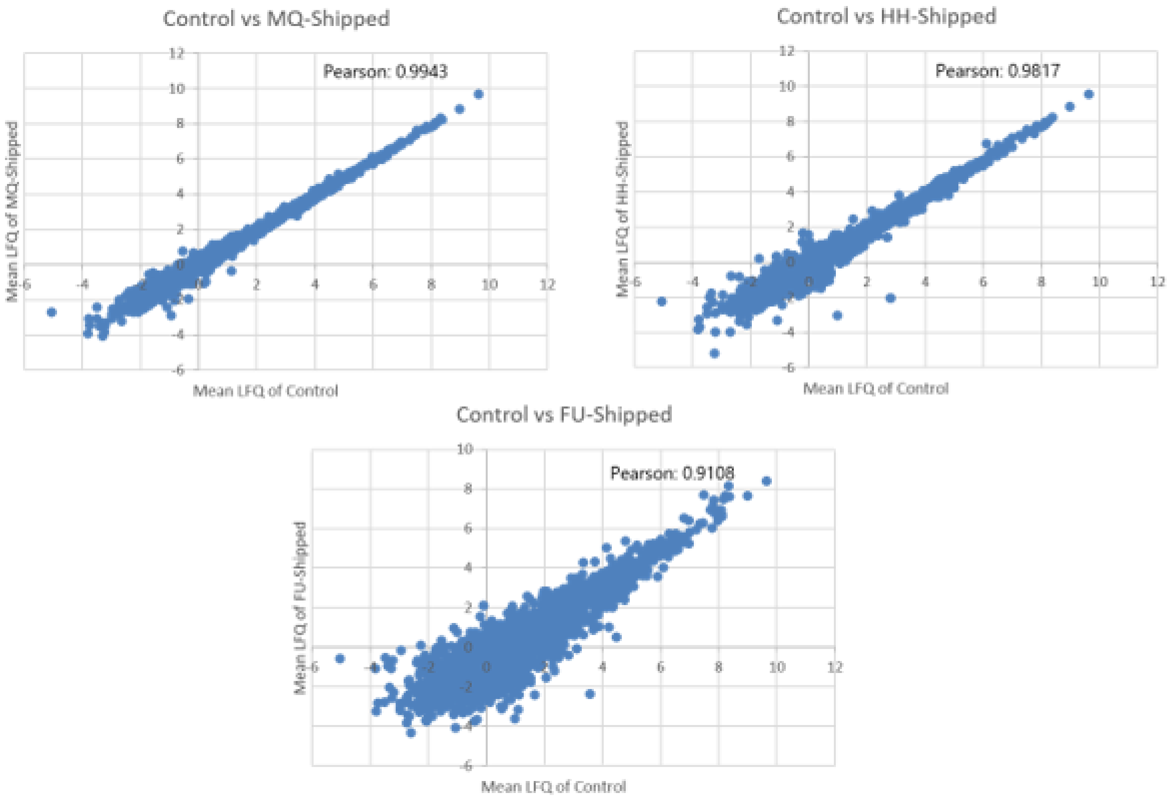


Figure 29: Scatter-plot of the protein abundances of the Control group against all other groups with calculated Pearson-Correlation.

Figure 29 shows scatter-plots of the protein abundances (summed peak areas of the peptides for each protein) of the control group against each of the other sample groups. A very high correlation can be seen between the control and MQ-Shipped with a Pearson-correlation of 99.43% and the control and HH-Shipped with a Pearson-correlation of 98.17% and slightly less correlation between the control and the FU-Shipped group with a Pearson-Score of 91.08%, overall suggesting no apparent loss in protein amount caused by the transport.

5.3.2 Chemical modification

Possible chemical modification which might occur during heating of the sample such as cyclization of glutamine to pyro-glutamine and differences in oxidation of methionine as well as N-terminal acetylation were examined on peptide level in Figure 30.

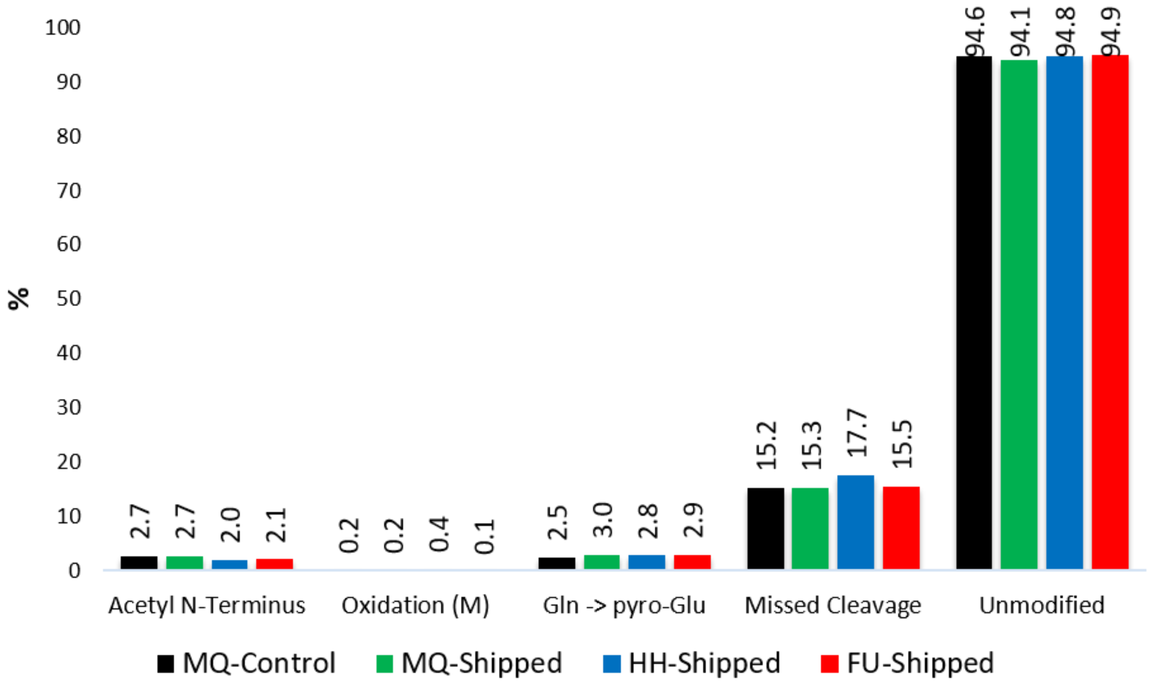


Figure 30: Amount of all modified peptides as well as the amount of missed cleavages in % in each sample group.

As can be seen in Figure 30 nearly all of the detected peptides in all sample groups were unmodified (94.1%-94.9%). No bias of one sample group towards a specific heat induced modification can be observed.

To exclude bias of the LC setup towards hydrophilic peptides caused by slight differences in the setup (in MQ no trapping column was used) in each laboratory again only those proteins which were uniquely identified in only one of the sample groups (Figure 27A) were chosen and the amount of acid and basic amino acids present in the peptides plotted (Figure 31 and Figure 32).

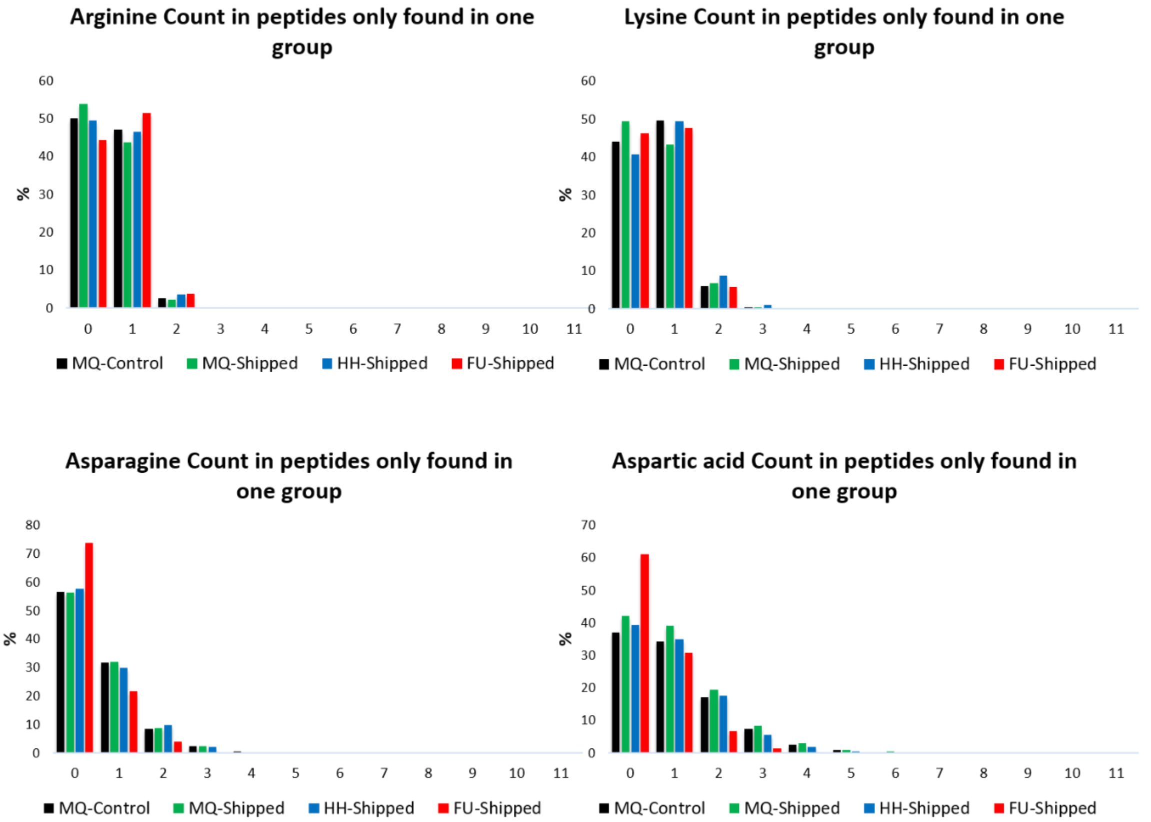


Figure 31: Amount of hydrophilic amino acids present in the peptides of the uniquely identified proteins in each sample group.

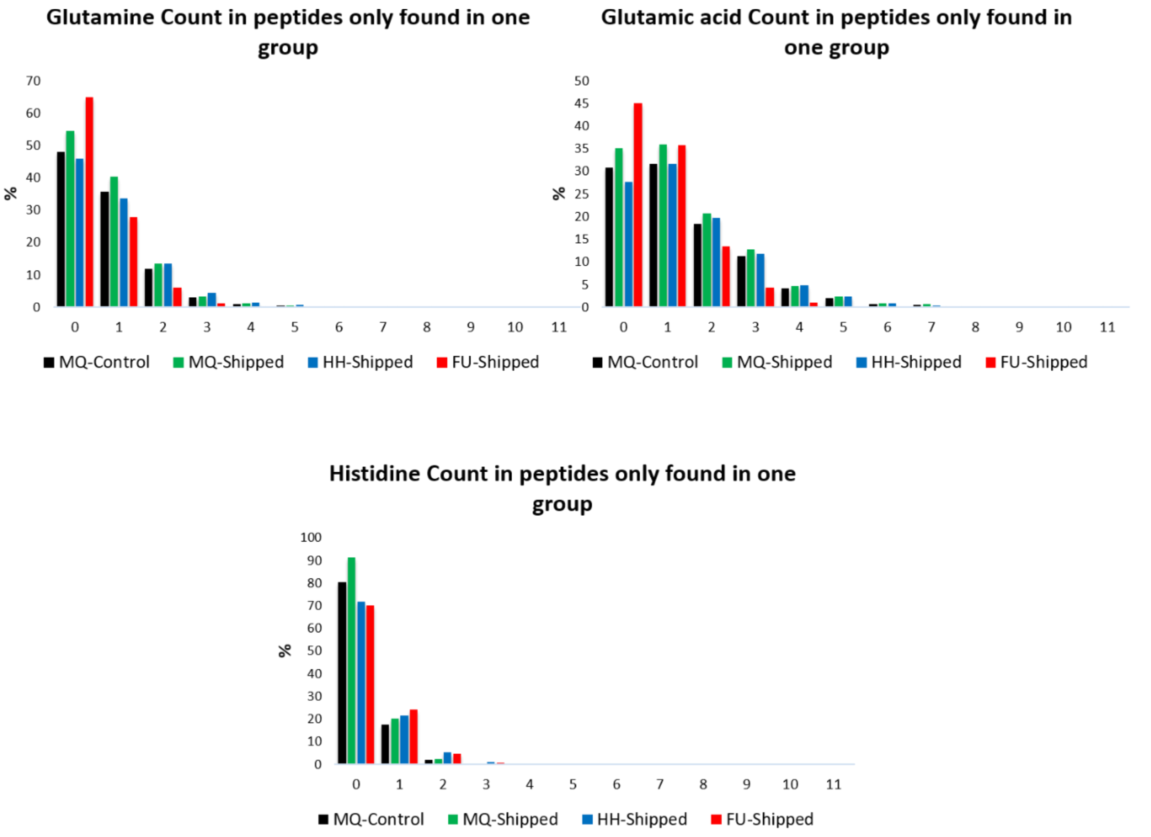


Figure 32: Amount of hydrophobic amino acids present in the peptides of the uniquely identified proteins in each sample group.

No bias towards hydrophilic peptides can be seen for the MQ-Control or MQ-Shipped sample. Interestingly the peptides of the uniquely identified proteins in the FU-Shipped sample show slightly higher hydrophilicity compared to the other sample groups.

5.4 Human pCa experiments

25 frozen human prostate tissue specimens were prepared for proteomic analysis (see section 7.4). 5 of these were from normal prostate tissue (adjacent to cancer tissue), the other 20 samples consisted of prostate cancer tissue of different Gleason-Scores (3x 3+3, 5x 3+4, 5x 4+3, 3x 4+5, 4x 5+4). Normal tissue as well as prostate cancer tissue with 3+4 and 4+3 had the most biological replicates assigned to them as these were considered the most clinically important samples among the sample cohort. The normal tissue were used as control and the two Gleason-Score 7 tissue were selected because of the drastic difference in the metastatic incidence between these two as well as there being no good differentiation parameter as of yet. The tissue samples were prepared as described in section 7. The dried samples were taken to the laboratory of Prof. Molloy in Sydney in the Macquarie University where the following experiment was conducted.

5.4.1 Protein identification and MS2-based Label-free quantification (DIA)

Because SWATH MS (DIA) was used as a quantification method, a spectral library first had to be generated. This was done as described in section 7. 28051 peptides were identified corresponding to 2559 proteins with an FDR of 1% or less. This is a rather small number of proteins considering that an offline fractionation was performed. A possible explanation for this might be that the provided samples were not frozen, but formalin fixed. Table 6 shows selected proteins with a formyl group which would support this hypothesis in total 2367 peptides with a formyl modification at the peptide N-Terminus were detected.

Table 6: Results of the peptide identification using ProteinPilot. Shown are four peptides of the Protein Myosin-11 which were detected to have a formyl-group at the N-Terminus.

Names	Start Position	Best Conf (Peptide)	Sequence	Modifications
Myosin-11	34	71	RLVWVPSEK	Formyl@N-term
Myosin-11	43	99	QGFEAASIKEEK	Formyl@N-term
Myosin-11	68	99	KVTVGKDDIQK	Formyl@N-term
Myosin-11	69	99	VTVGKDDIQK	Formyl@N-term

Because there was no possible way to acquire new samples on site in Sydney it was decided to use these for further analysis.

All 25 samples were measured in DIA mode as described in section 7.11.5 and data analysis was performed as described in section 7.13.2 and 7.13.3. After processing 467 proteins were quantified over all 25 samples.

5.4.2 Post-processing of the Data and statistical tests

The quantification data obtained through PeakView were further analyzed using the DEP R Script as well as Perseus. DIA quantification has the big advantage over MS1-based quantification that there are nearly no missing values in the data. In this case no missing values were detected over all samples and imputation of values was not necessary. To get a first look at the distribution of the data a PCA was computed (Figure 33).

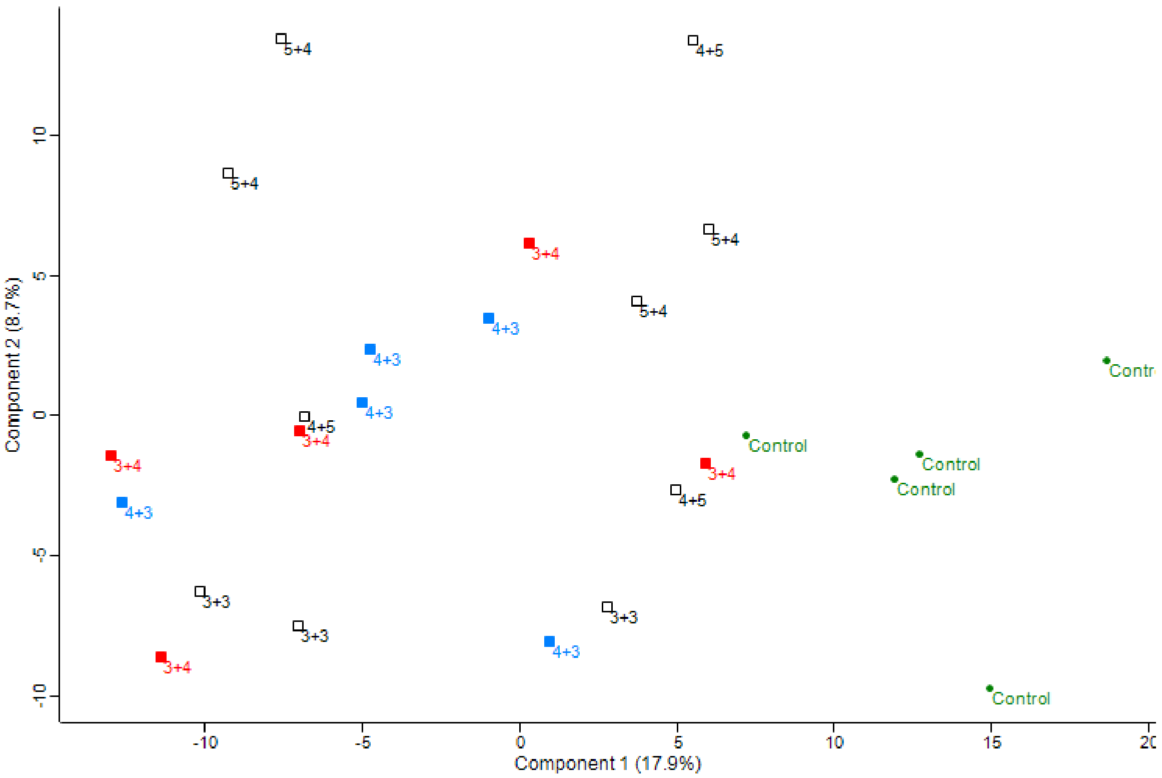


Figure 33: Principle component analysis (PCA) of the data. Component 1 and 2 are shown for each sample. Control samples are shown as green circles, Pearson score 3+4 and 4+3 are shown as filled squares in red and blue respectively and the rest of the samples are shown as black squares.

Figure 33 shows that there is no possible distinction possible between the prostate cancer samples. The only samples that show clear clustering are the controls. Our hypothesis was that the samples with higher Gleason-scores could be distinguished from the lower ones using proteome expression. A possible reason for this might be the rather small dataset of 467 quantified proteins compared to the mouse experiment where over 2000 proteins where used for post-processing.

For statistical analysis a t-Test using LIMMA was performed in R. Here each sample group was tested against each other resulting in 15 pair-wise tests. Most of these comparisons showed no significantly different proteins, which was surprising because the control samples should at least differ to each of the cancer tissues. Again, a possible explanation could be the small dataset but also the large variance in each of the biological samples. This high variance might be caused by the difference in each individual patient but is more likely due to either the heterogenous nature of the cancer tissue, difficulty in determining the actual Gleason-Score or the possibility of formalin fixed tissue (which would have required a different lysis method to the one used here) instead of frozen tissue. Nevertheless, when comparing the 3+4 and 4+3 samples 15 proteins were detected that were significantly different and showed a log2 ratio of more than |1.0|. A list of these proteins can be seen in table 7.

Table 7: Significantly different proteins as computed with the t-Test and showing log2 ratios of more than |1.0|. The log2 ratios as well as the adjusted p-Values are shown.

name	3+4_vs_4+3_diff	3+4_vs_4+3_p.adj
AGRN	1.519787153	0.021148801
CCAR1	1.424766573	0.014159194
CCAR2	-3.180468028	2.36E-05
COPG1	-1.122970712	0.017506009
DC1L2	1.563557239	0.037746391
EFHD2	1.559693781	0.0460721
FSCN1	-1.482510422	0.03332054
GLOD4	-1.310595427	0.007168668
GRP78	1.402438314	0.000205063
IST1	1.508426172	0.024711101
ITPR1	1.338528441	0.001952039
KPRB	-2.996459682	4.02E-06
PRDX1	1.049789674	0.048922106
RINI	1.055297308	0.039841289
RSMN	1.054817083	0.010541207

When comparing this list to the significantly different proteins in the mouse experiments no overlap can be observed, but when compared with the list of possible marker candidates using only the PCA one protein matches which is AGRN (Agrin). It was expected that the overlap between these two datasets would be minimal as there is much more variance when using patient samples compared to mouse samples. Furthermore, the cancer samples from the mouse experiment consisted only of a single prostate cancer cell type whereas human prostate cancer is an amalgamation of several different types and therefore hard to compare. Even so this one hit seems to be a very promising one. Although Agrin was not statistically relevant in the mouse experiment it showed a rather high log2 fold change of 4.0.

5.4.3 Pathway analysis and Gene ontology enrichment

No protein network could be generated using ReactomeFI in Cytoscape as no connection between these 15 proteins could be found. No biological pathway or molecular function was found to be enriched. These kinds of analysis typically require larger lists of proteins to work properly so this result was to be expected.

Nevertheless, Agrin was subjected to a more thorough literature search. Agrin is a heparan sulfate basal lamina glycoprotein of which 6 isoforms are described in UniProt. Interestingly, most of these isoforms are found in neuronal tissue with isoform 1 having a major role in the maintenance of neuromuscular junction [41].

Looking back at the mouse experiments a significant number of axon pathways was enriched in the highly metastatic PC3 samples which was absent in the DU145 samples. In this context it makes sense for Agrin to be upregulated in 4+3 Gleason-score samples which exhibit a much more aggressive phenotype than the 3+4 Gleason-score samples. As such Agrin would be an interesting protein marker for further testing in this context for example by performing immunohistology on tissue microarrays.

5.4.4 Analysis using the R-Script

To assess if any of the identified possible markers for the distinction between GS 3+4 and GS 4+3 have already been investigated the text mining script was used this time with the keyword “Prostate cancer”. Table 8 shows the summary tabel of the results from the query for all 15 proteins.

Table 8: Summary tab of the results from the query for the 15 differentially regulated proteins between the GS 3+4 and GS 4+3 samples.

N	UniProtID	TaxID	Synonyms	Keywords	KeywordInTitleOnly	Results	Category	Other	PubmedQuery
1	P11021	9606	HSPA5,GRP78	Prostate cancer	Yes	22	1	1	"Prostate cancer"[TI] AND ("HSPA5"[TI] OR "GRP78"[TI])
2	Q9Y678	9606	COPG1,COPG	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("COPG1"[TI] OR "COPG"[TI])
3	P13489	9606	RNH1,PRI,RNH	Prostate cancer	Yes	1	2	0	"Prostate cancer"[TI] AND ("RNH1"[TI] OR "PRI"[TI] OR "RNH"[TI])
4	Q06830	9606	PRDX1,PAGA,PAGB,TDPX2	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("PRDX1"[TI] OR "PAGA"[TI] OR "PAGB"[TI] OR "TDPX2"[TI])
5	O00468	9606	AGRN,AGRIN	Prostate cancer	Yes	1	2	0	"Prostate cancer"[TI] AND ("AGRN"[TI] OR "AGRIN"[TI])
6	Q9HC38	9606	GLOD4,C17orf25,CGI-150, My027	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("GLOD4"[TI] OR "C17orf25"[TI] OR "CGI-150"[TI] OR "My027"[TI])
7	Q8N163	9606	CCAR2,DBC1,KIAA1967	Prostate cancer	Yes	1	2	0	"Prostate cancer"[TI] AND ("CCAR2"[TI] OR "DBC1"[TI] OR "KIAA1967"[TI])
8	Q14643	9606	ITPR1,INSP3R1	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("ITPR1"[TI] OR "INSP3R1"[TI])
9	Q16658	9606	FSCN1,FAN1,HSN,SNL	Prostate cancer	Yes	2	2	0	"Prostate cancer"[TI] AND ("FSCN1"[TI] OR "FAN1"[TI] OR "HSN"[TI] OR "SNL"[TI])
10	O43237	9606	DYNC1LI2,DNCLI2,LIC2	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("DYNC1LI2"[TI] OR "DNCLI2"[TI] OR "LIC2"[TI])
11	Q96C19	9606	EFHD2,SWS1	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("EFHD2"[TI] OR "SWS1"[TI])
12	P53990	9606	IST1,KIAA0174	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("IST1"[TI] OR "KIAA0174"[TI])
13	P63162	9606	SNRPN,HCERN3,SMN	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("SNRPN"[TI] OR "HCERN3"[TI] OR "SMN"[TI])
14	O60256	9606	PRPSAP2	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("PRPSAP2"[TI])
15	Q8IX12	9606	CCAR1,CARP1,DIS	Prostate cancer	Yes	0	3	0	"Prostate cancer"[TI] AND ("CCAR1"[TI] OR "CARP1"[TI] OR "DIS"[TI])

5 out of the 15 proteins where found to have been connected to prostate cancer in the literature with HSPA5 being a category one hit with one review article. Interestingly, one article was found that associates Agrin with prostate cancer [42]. The data presented in this article suggests that downregulation of NEAT1, which has been reported to play a major role in cancer related cellular activities such as apoptosis, can be correlated to downregulation of Agrin via the NEAT1-CDCL5-AGRN circuit. This supports the findings in the mouse as well as the human sample experiments which makes Agrin a strong candidate for future studies.

5.5 Development of a 3-Dimension LC method

A new semi-online 3-Dimensional LC peptide fractionation method was developed to enable deeper analysis of proteomes. In the first dimension an offline HpH reversed phase chromatography is performed (see methods part section 7.11.3). Fractions over the whole gradient are conjoint to obtain 13 fractions. These 13 fractions are separated another 3 times using an online displacement cartridge (SCX-Cartridge) placed into a second valve before the reversed phase trapping column. The flowthrough, the fraction of the displacer pulse and the regeneration fraction are subsequently trapped and analyzed for all 13 HpH fractions resulting in a total of 39 fractions (see methods part section 7.12.3).

The maximum loading capacity and the amount of displacer needed to elute the desired fractions had to be determined before testing the method (for an in-depth description of these experiments, see the methods part section 7.12).

5.5.1 Determination of key parameter for the 3D-LC method

The maximum loading capacity of the displacement cartridge was tested using sequential injections of 1 µg of a peptide mixture of PC3 and DU145 cell lysate. Figure 34 shows the charge-plot of the capacity test.

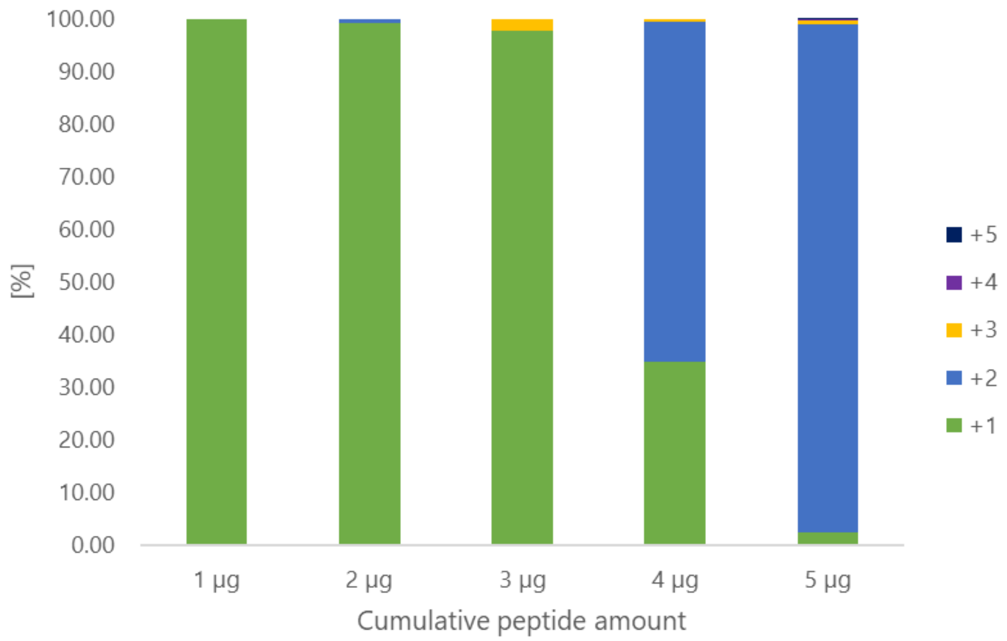


Figure 34: Charge-Plot of the capacity test of the displacement cartridge. The sequential injections are plotted on the x-axis, the color code of the charge states of the identified peptides in solution are shown on the right side and the percentage of each charge state per run is shown on the y-axis.

The maximum capacity of the displacement cartridge can be determined to be between 3 and 4 µg of peptides as seen in Figure 34, which is consistent with previous experiments with these types of cartridges [37]. The amount of displacer needed to elute the desired fraction was determined using 50 sequential injections of 50 ng Spermine. Figure 35 shows the charge-plot of the spermine test.

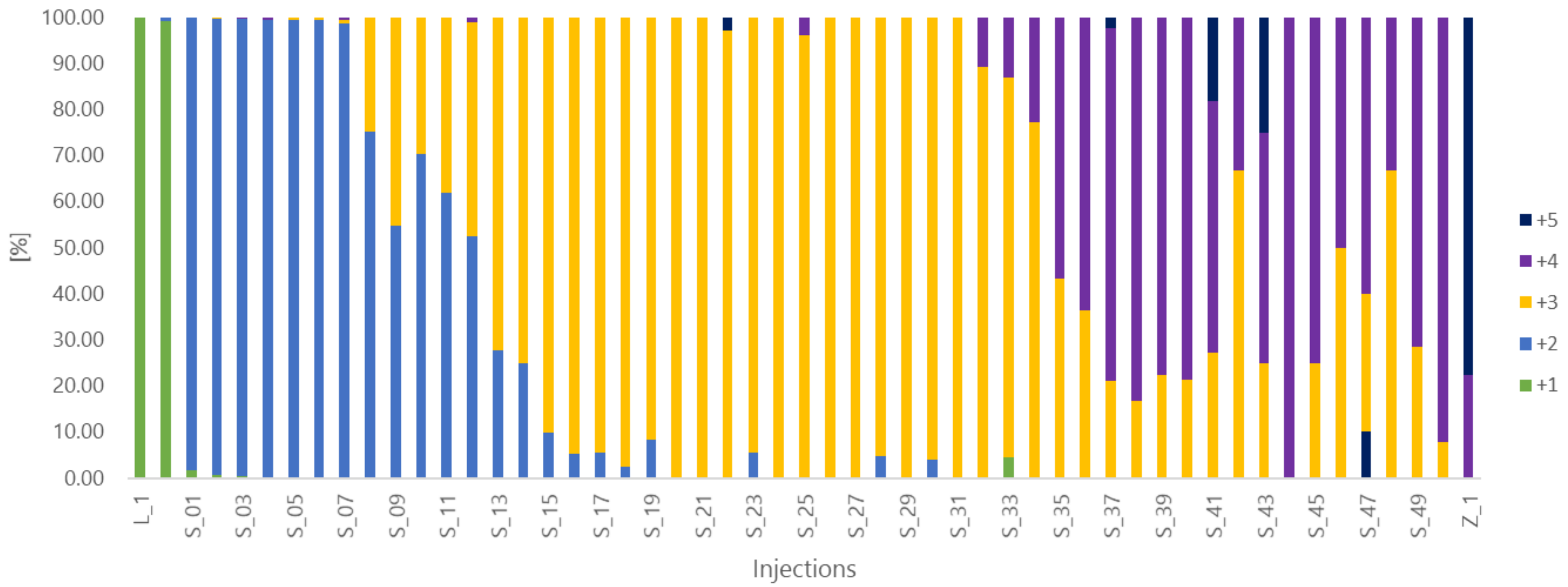


Figure 35: Charge-Plot of the spermine test of the displacement cartridge. The two peptide loading injections (L_1, L_2) the sequential injections of sperime (S_01-S_50) and the regeneration injection (Z_1) are plotted on the x-axis, the color code of the charge states of the identified peptides in solution are shown on the right side and the percentage of each charge state per run is shown on the y-axis.

A characteristic charge distribution for SCX-Columns can be seen in Figure 35 with nearly all singly charged peptides being in the flowthrough fractions. With the first injection of the displacer molecule doubly charged peptides elute and with increasing displacer amount absorbed on the cartridge peptides with higher charge states in solution begin to elute forming sharp bands.

To keep the measurement time at a reasonable level the decision was made to have one fraction of singly charged peptides (flowthrough) a second fraction of doubly charged peptides (spermine fraction) using 600 ng Spermine and a third fraction of triply, quadruply and quintuply charged peptides (regeneration fraction) using 1M ammonium acetate.

Figure 36 shows a schematic of the LC-setup for the 3D-Method.

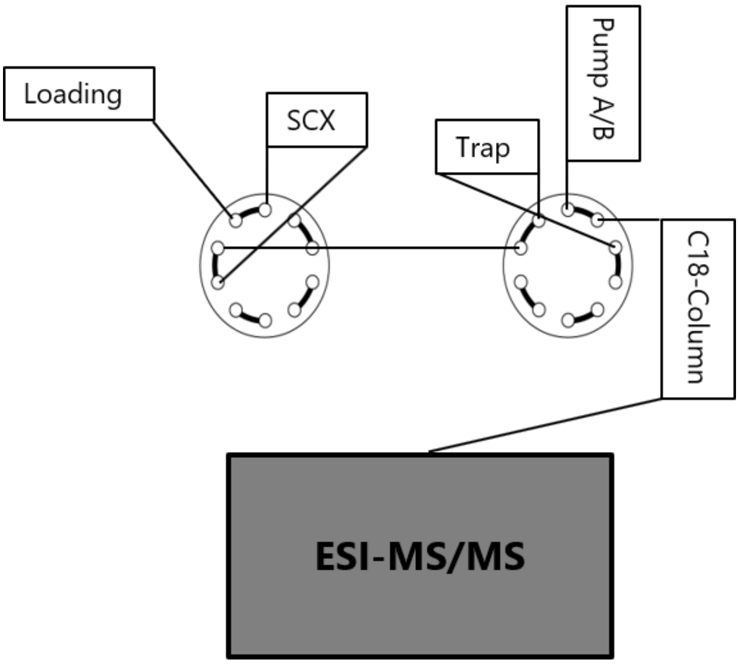


Figure 36: Schematic of the LC-Setup used for the 3D-Method. The loading in the top left refers to the autosampler and loading pump which is connected through a valve to the SCX cartridge. The first valve with the SCX cartridge is connected to the second valve using a standard proteomics setup with a C18-reversed phase trapping column coupled to an analytical C18-reveresed phase column. Peptides are eluted directly into the mass spectrometer.

5.5.2 Testing of the 3D-LC method against a standard 2D offline HpH method

The 3D-LC setup was tested against a common 2D workflow containing just the offline HpH reversed phase separation followed by a standard low pH analysis. As biological material a 1:1 mix of PC3 and DU145 cell lysate was used. The peptide mixture was fractionated using the offline HpH setup and a portion from the resulting 13 fractions was used for the standard 2D workflow with a 2-hour gradient. Another portion (approximately 4 µg of protein of each fraction) was used for the 3D analysis. The BPC of the flowthrough, spermine and regeneration fraction for the first HpH fraction of the 3D-LC is shown in Figure 37.

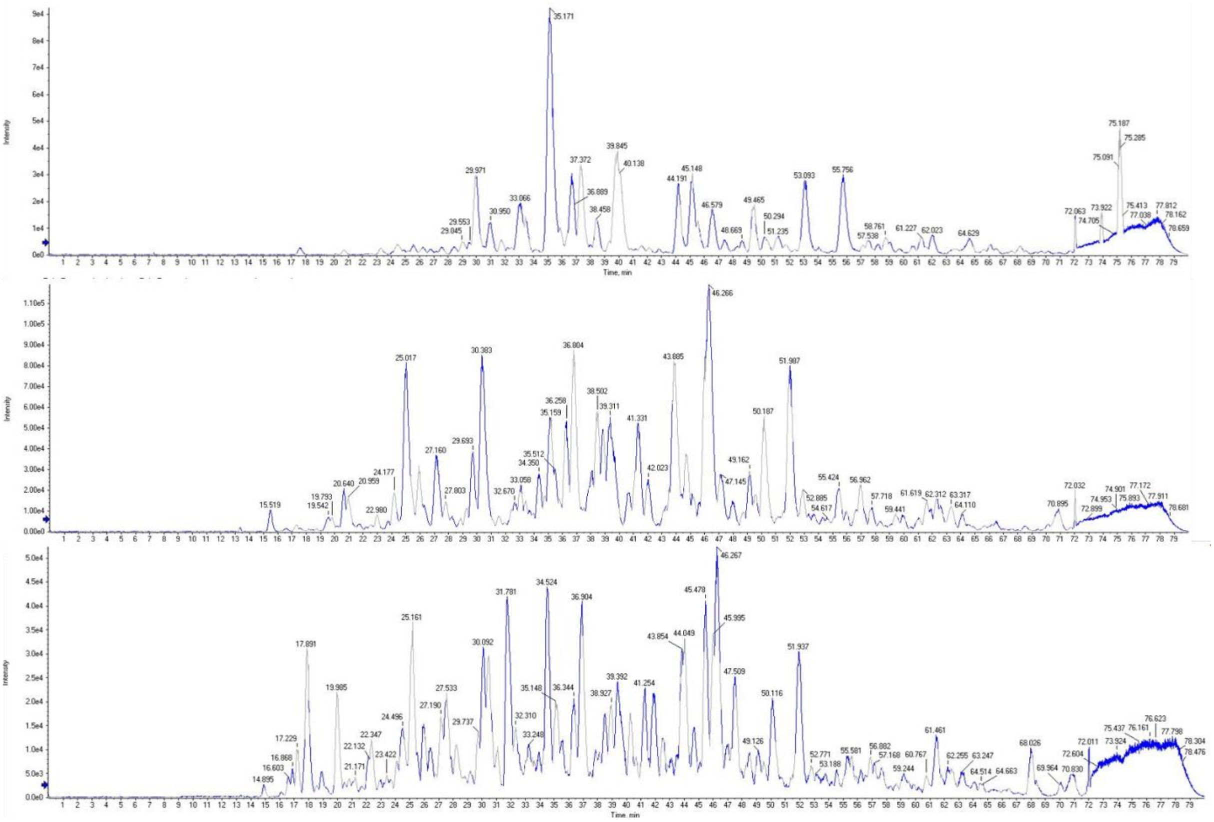


Figure 37: BPC of the flowthrough (top, 01A), spermine (middle, 01B) and regeneration (bottom, 01C) fraction of the first HpH fraction.

In Figure 37 an increase in signal intensity can be observed. This was to be expected as only a very small fraction of the sample consists of singly charged peptides (mostly acetylated N-Termini) with the main component being the doubly charged peptides in the spermine fraction (Figure 37 middle). To further assess the separation efficiency of the 3D-LC method the charge-plot for the three fractions in Figure 37 was computed (Figure 38).

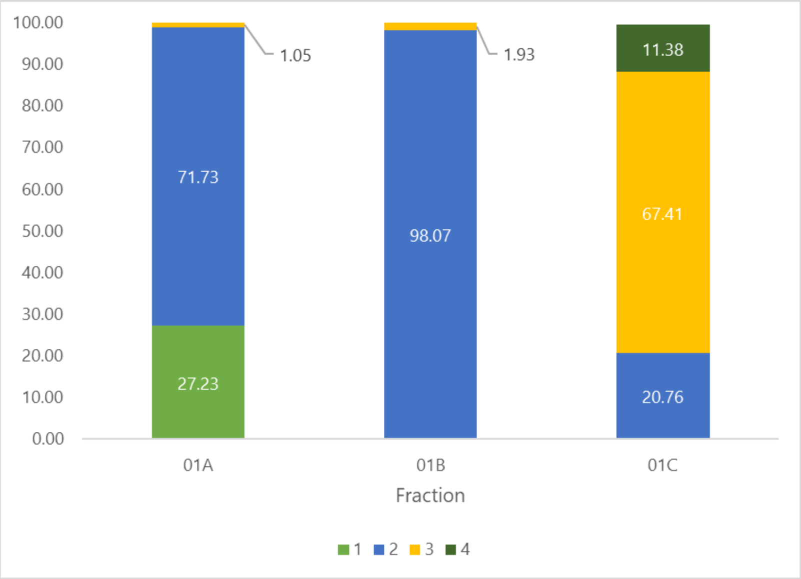


Figure 38: Charge-Plot of the three fractions shown in figure 27. The fractions are plotted on the x-axis and the number of charged peptides in % is plotted on the y-axis. Singly charged peptides are shown in light green, doubly charged peptides in blue, triply charged peptides in yellow and quadruply charged peptides in dark green.

Based on the charge-plot it is clear that the loading capacity was slightly exceeded mainly because of the difficulty to assess the right amount of peptides in each fraction after HpH fractionation. However, fraction 01B (the spermine fraction) consists nearly purely of doubly charged peptides which was the main aim using this 3-fraction split. The regeneration fraction consists of doubly, triply and quadruply charged peptides. This split allows for the highest depth of separation in the least amount of time possible.

Table 9 shows the differences between the new 3D-LC method and the standard 2D HpH method.

Table 9: Comparison of important parameter between the 3D-LC method and the standard 2D HpH method.

All data corresponds to 1% FDR		
	3D-LC	HpH
Protein ID	4047	2989
Unique Peptides	40731	21314
Acetylation	635 (1.6%)	332 (1.6%)
Mean Sequence Coverage (%)	22.6	17.7
Median Peptide/Protein	4	3

Benefit of 3D-LC to normal HpH		
	No.	%
Protein ID	1058	35.4
Distinct Peptides	19417	91.1
Acetylation	303	91.3
Mean Sequence Coverage	-	4.9
Median Peptide/Protein	1	33.3

A 35.4% increase in protein IDs could be achieved using the new method. The more important value in the table however, is the increase of 91.1% of peptide IDs resulting in a potentially much more comprehensive spectral library for later use in DIA/SWATH experiments. An increase in identification of 91.3% of peptides containing N-terminal acetylation could also be observed. Most of these peptides were found in the flowthrough fractions.

To further investigate the differences on protein and peptide level, two Venn-Diagrams were computed (Figure 39).

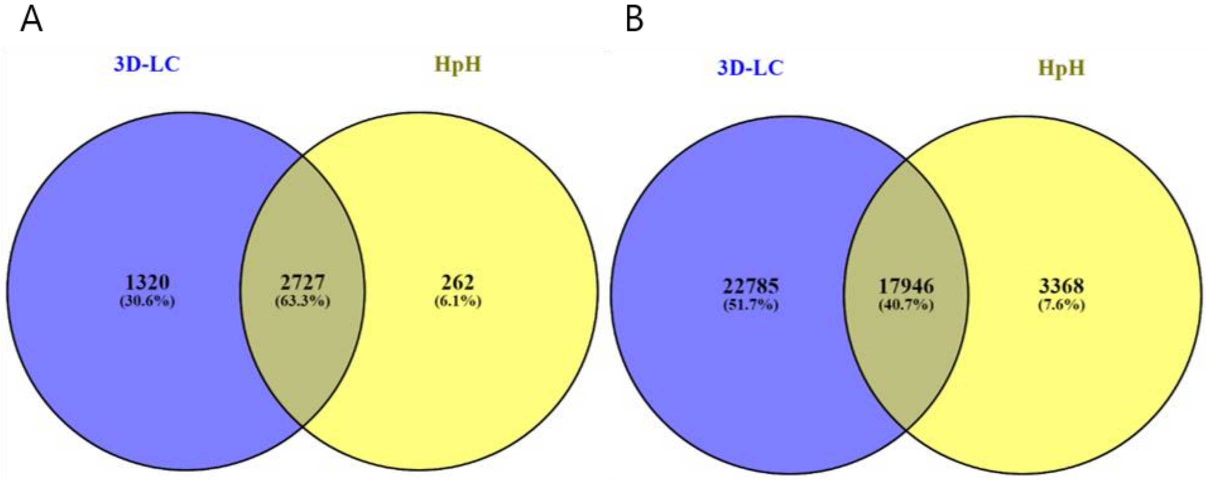


Figure 39: A) Venn-Diagram of the proteins identified in both methods. B) Venn-Diagram of the peptides identified in both methods.

Nearly all proteins identified using the standard HpH method could also be identified using the 3D-LC method (identity of 93.9%) including an additional 30.6% more proteins identified. On peptide level (Figure 39B) this trend is even more apparent with the 3D-LC identifying 92.4% of the peptides of the HpH method and an additional 51.7% more peptide IDs.

Using this semi-3D-LC method it is possible to get a very high depth of information of the sample in a reasonable amount of time without having to change the LC setup compared to normal proteomic workflows. In the future this method will also be implemented in the laboratory in the UKE and tested using samples with biological relevant background. One of the major benefits of using this 2-valve-setup is the possibility to use this method for spectral library generation even when using iRT peptides for retention time calibration.

6 Discussion

The overall aim of this thesis was to identify possible protein marker for different stages of prostate cancer (Gleason-Scores). This kind of investigation would typically involve a large clinical trial with numerous patients showing different stages of prostate cancer. Because such a study is very time and money intensive it was not in the scope of this thesis. Therefore, it was decided to conduct this research using patient tissue readily available in the biobank of the pathology department in the UKE. Furthermore, because such tissue samples are very valuable a xenograft mouse model was used to get a first baseline of proteins showing differences in concentration in primary tumors derived from single-cell cultures of two different human prostate cancer cell lines. Lange *et al.* [43] developed a clinically relevant mouse model (pfp$^{-/-}$/rag2$^{-/-}$) to monitor the metastasis spread of different prostate cancer cell lines. For this study the same xenograft model was used. PC3 and DU145 cells were chosen as representative cell lines as these showed high respectively no metastatic potential in previous experiments. The reasoning behind using these two cell lines was that they represent the most extreme phenotypes of prostate cancer. Another reason was the differentiation of the two possible stages of GS 7. Recently, the International Society of Urological Pathology (ISUP) decided to assign a higher risk to the 4+3 score based on several reports in literature [39]. The hypothesis in this thesis for using PC3 and DU145 cells was that there should be a significant difference in metastatic potential between the GS 3+4 and 4+3 and thus would show a rather similar protein profile to the mouse experiments. Indeed, a recently published paper from Kamel MH *et al.* [38] showed that there seems to be a 3-fold increase in incidence of metastasis for patients diagnosed with GS 4+3 compared to 3+4.

The experiments with the primary tumors from the mouse xenograft model were analyzed in the UKE. Here, MS1 quantification using DDA experiments was used to identify proteins significantly different between the PC3 and DU145 samples. MS1 quantification is the most commonly used relative quantification method in proteomics mostly because DDA shotgun methods are the standard approach for protein identification and thus the method is known to the majority of the proteomics community [23]. With the help of software like Maxquant [24] it is possible to adjust small retention time differences across the data and quantify the area under the curve for XICs of identified peptides in the sample. The main problem with this technique is the stochastic variability between measurements based in the random sampling of the DDA method resulting in missing values and reducing the confidence of the quantification. Tabb *et al.* reported a repeatability and reproducibility between 70% and 80% for technical replicates, with Orbitrap instruments performing more consistent than TOF instruments [44]. In case of the data in the mouse experiment (see Figure 17) 36% of the quantification values across all samples represented missing values and had to be imputed resulting in an overall repeatability of 64% which is slightly below the reported values of Tabb *et al. C*onsidering that in this thesis biological replicates (4 mice for each cell line) were used as well as two different cell lines (PC3 and DU145) which resulted in a larger amount of MNAR, it is safe to say that the repeatability and amount of missing values is well in the range of typical proteomic experiments.

To analyze significantly different proteins across different sample sets, the data have to be tested for significance. The most commonly used statistical test between two sample groups is the t-Test. An ordinary t-Test compares the mean of the replicates between two sample groups for each protein and assesses the variance. A drawback of using an ordinary t-Test is when only a small set of replicates per sample is available which results in unstable variance estimations. This is because the degrees of freedom for this test are calculated as *m+n-2*, where *n* and *m* are the sample sizes of both sample groups. When dealing with small sample sizes the degrees of freedom are close to 0 which results in an unstable variance estimation. In this study for the mouse as well as the human experiments the number of biological replicates per group was rather small (a maximum of 5 replicates in case of the human experiments). To account for the variance of the ordinary t-Test a so called limma statistic was chosen. Here, a moderated t-Test is used integrating a linear empirical Bayes model for variance estimation [45]. This approach is widely used in the genomics field to analyze microarray data but can also be used for proteomic data [46]. Using this test, 103 significantly different proteins with a log2 fold-change higher or equal to |1.0| between the PC3 and DU145 samples were identified. As a comparison, 339 significantly different proteins with a log2 fold-change higher or equal to |1.0| were identified using an ordinary t-Test (data not shown).

Recently, the analysis of proteomics data using network analysis has become more popular. As many other innovations in proteomics this also originated in the genomics field. The aim of such an analysis is to use the proteomics data to fill a network of nodes and edges with information to establish a network which displays interactions between the differentially expressed proteins [47]. Figure 22 shows such a network using the 103 significantly different proteins found in the PC3 and DU145 samples. To build such a network the data is correlated to different databases and depending on the database used may present different messages to the reader. Herein lies the main problem when using such networks: it is not easy for the reader to assess which kind of interactions of the proteins/genes are displayed. In case of Figure 22 the Reactome database was used which is an open access, manually curated, peer-reviewed pathway database [48]. In this network activation of proteins as well as protein-protein interaction and predicted interaction are displayed. Other popular databases used for creating such a networks include STRING-DB which only displays protein-protein interactions [49]. Another problem apart from using different databases is the use of different cluster algorithms for analysis of the networks. ClusterMaker2 [50], an app in the Cytoscape environment offers 9 different algorithms for network clustering each resulting in slightly different interaction cluster. Yet another problem of these kinds of network analysis is that the interactions are based on literature. Although the source for each suggested interaction is present for the researcher it is not feasible to display such information in a graphic and as such is not visible for the reader.

A more reasonable approach to analyze the data is a pathway analysis based on GO-Enrichment and/or KEGG-Pathways. There are several open access tools such as BiNGO [51] or DAVID [52] available. In this study the BiNGO app for Cytoscape was used for GO-Enrichment tests. Table 1 and 2 show the 10 most significant enriched biological functions for the PC3 and DU145 samples. Using this information, it is possible to map the PC3 samples to a more neuronal-like phenotype. There are several similarities between neuron and cancer cells [53] and thus this pathway analysis supports the metastatic nature for PC3 cells.

For an even more in-depth look into the significance of the differentially regulated proteins, manual literature research or extensive prior knowledge of the researcher need to be applied. Because this is typically the most time-consuming part of the analysis a literature mining script was developed to help this process using an informed literature search. Because the first iteration of the script relied on the data found in the iHOP database [54] which was shut down in the second quarter of 2018 a second iteration was developed using the UniProt database which is less likely to be discontinued as it is a partly government funded project. Most of the literature found for the 32 proteins showing higher concentration in the PC3 samples directly suggests that upregulation of these proteins is indeed associated with an increase in metastatic potential (i.e. ATAD3A [55], EPHA2 [56], FABP5 [57], HMGA1 [58], S100A2 [59], WASF2 [60]). This kind of literature search is rather stringent as only articles with search terms found in the title and only gene name synonyms are considered in this version of the script. However, using this approach results in high confidant hits which in case of planning further experiments (i.e. planning of knockdown/out experiments) are generally more valuable than providing the highest number of hits possible. Another advantage of this analysis approach over a network analysis is that the search can be specific for a certain condition or background with which the experiment was planned (in this case the aim was to identify marker connected to metastasis). Although this tool eases the time-consuming literature research it is of utmost importance that each result is checked and reviewed because as can be seen in table 4, false-positive hits are still a concern. Using this script in conjunction with other tools such as network analysis or GO-Enrichment the data can be explained much more comprehensively.

Before continuing with the human prostate cancer samples, a multi-laboratory experiment was designed to assess the influence of non-cooled transport of lyophilized peptides derived from a complex biological sample (here a lysate of SW480 cells was used). This study was conducted as part of the DAAD funded trilateral partnership between the University of Hamburg, the Macquarie University and the Fudan University. During the time of writing of this thesis there was no literature found on the topic of stability of lyophilized peptides for mass spectrometric purposes. Although it is common practice to ship dried peptides, there are apparently only empirically derived conditions and directions available [61, 62] which assume that the sample was cooled along the way during transport. In 1997, Bell investigated the stability of several peptides in solids and solution but this study shows no relevance concerning the integrity of peptides for mass spectrometry and only assesses the stability of simple peptide mixtures [63]. Therefore, the data presented in this thesis and published in the journal Analytical Biochemistry titled “Multi-laboratory analysis of the variability of shipped samples for proteomics following non-cooled international transport” in May 2018 [64], represents the first documented study about sample integrity without cooling during transport. Basic knowledge of chemistry would dictate that a cleaned-up mixture of peptides in a dried state does not undergo drastic changes. Issues could arise from small amounts of liquid still present in the sample which would increase hydrolysis reactions like cyclization or deamidation. Oxidation might also be favored at elevated temperatures. The data presented here shows however, that the non-cooled transport had no significant effect on the samples when using them for shotgun proteomics. The main source of variation between the laboratories was the sample handling from different researchers using different pipettes and probably slightly different calibrations of the MS instruments.

The in the UKE prepared, MS ready human prostate cancer samples were transported without cooling to Sydney where they were analyzed in the Macquarie University using a SWATH-MS (DIA) approach. There are several advantages of using DIA over DDA for label-free quantitation. When working with a large number of samples DIA is much more time efficient as shorter gradients can be chosen for the DIA analysis because of the previously generated spectral library. Because the post-processing of the DIA data relies on the spectral library it is important to use a high quality (meaning extensive) library. Here, a pooled sample of all human samples was generated and fractionated before measurement to achieve the most depth. Using this approach, a library containing 2559 proteins could be generated. Table 6 shows that several peptides show an peptide N-terminal (in this case inter protein) formylation. The modification was present in over 8% of the peptides identified which is an uncharacteristically high amount. This suggests that at least some of the samples were not fresh frozen but rather formalin fixed tissue samples [65]. This would also explain the relatively low number of proteins in the spectral library because the chosen sample preparation was not suitable for formalin fixed tissues. This would also explain the low amount of quantified proteins using PeakView (467 proteins across all samples). PeakView uses q-Values for the quantitation which considers the reproducibility (number of missing values) across the samples. Typically, in a DIA experiment the number of missing values is very small [33]. If the samples used in this study were indeed formalin fixed instead of fresh frozen the reproducibility across the samples would be very low because of the chosen lysis method resulting in higher q-values and therefore reducing the number of confidently quantified proteins. Nevertheless, 15 proteins could be identified that showed significant regulation between the GS 3+4 and 4+3. Using the text mining script Agrin was found to have been previously correlated with prostate cancer in the literature [42]. More specifically the downregulation of Agrin seems to play an important role for the reduction of proliferation of prostate cancer cells. This correlates with the findings in both the mouse and human experiments conducted in this study making Agrin a very interesting marker candidate for the differentiation between GS 3+4 and 4+3. A possible future study could include immunohistological staining against Agrin of several FFPE tissue arrays.

In the last part of this thesis a new semi-online 3D-LC method was developed. Multidimensional protein identification technology (MudPIT) developed by Washburn *et al.* [66] in 2001 is the commonly used term for an online 2D-fractionation (which is the second part of the 3D-LC method presented in this thesis). The original MudPIT featured an analytical column packed with two complementary chromatographic materials (SCX and RP). These bi-phasic columns have the major drawback that they generally cannot be regenerated and thus a new column must be prepared for each experiment. More recent approaches use a bi-phasic trapping column in front of the RP analytical column eliminating the need to prepare new columns before each experiment [67]. Although this is a step in the right direction for easier handling it is still necessary to change the trapping column before being able to conduct standard experiments on the LC-System resulting in measurement down-time. Another drawback of this kind of MudPIT are the salt pulses used for fractionation which put elevated strain on the LC instrument and may result in clogging the system when not handled correctly. Using the LC-setup presented in Figure 36 it is possible to switch the SCX column in front of the trapping column using a second valve, enabling the use of the instrument in either MudPIT or classic configuration, saving measurement time. Additionally, the fractionation on the SCX column is done in displacement mode instead of gradient mode eliminating the need to use salt in the LC system. Another advantage of this setup is the possibility to include iRT peptides in the measurement for a better spectral library generation. Figure 38 shows the power of using displacement instead of gradient mode. It is possible to separate peptides based on their charge in solution. This separation leads to an increase in identified peptides carrying a protein N-terminal acetylation compared to a HpH offline fractionation (see table 9). This method demonstrates a possible advancement of the online 2D-LC separation technique shown by Kwiatkowski M. *et al.* [37] with the addition of an orthogonal HpH reversed-phase fractionation and the possibility of integrating iRT calibration peptides in the measurements which allows this technique to be used in spectral library generation for DIA label free quantification. In this thesis for the first time the application of displacement mode chromatography in a 3-dimensional setup was demonstrated.

7 Material and Methods

7.1 Instruments and Chemicals

Table 10 shows the chemicals and instruments utilized in this study.

Table 10: Chemicals and Instruments as well as their distributor which were utilized in this study.

Chemical/Instrument	**Distributor**
Acclaim PepMap	ThermoFisher
Acetonitrile	Sigma-Aldrich
BSA Kit	ThermoFisher
C18 SepPak Cartridges	Waters
Criterion XT precast Gel	BioRad
DTT	Sigma-Aldrich
Formic acid	Sigma-Aldrich
Fusion	ThermoFisher
IAA	Sigma-Aldrich
nano cHiPLC columns ChromXPTM C18-CL	Eksigent
NanoLCTM ultra and cHiPLC® system	Eksigent
PBS	Sigma-Aldrich
Probe Sonicator	Athena Technology
Qexactive	ThermoFisher
Reducing agent	BioRad
TripleTOF 5600	Sciex
TripleTOF 6600	Sciex
Trypsin	Promega
Ultimate 3000 RSLCnano	ThermoFisher

7.2 Structure and experiment location

Because this thesis was conducted as a Joint PhD between the university of Hamburg and the Macquarie University (Sydney) as well as supported by the DAAD funded Trilateral partnership MQ-FU-HAM (Macquarie, Fudan, Hamburg) the data shown in this work was collected in different laboratories. Figure 40 shows a flowchart depicting the different parts of the experiments as well as their main executing location depicted using colors.

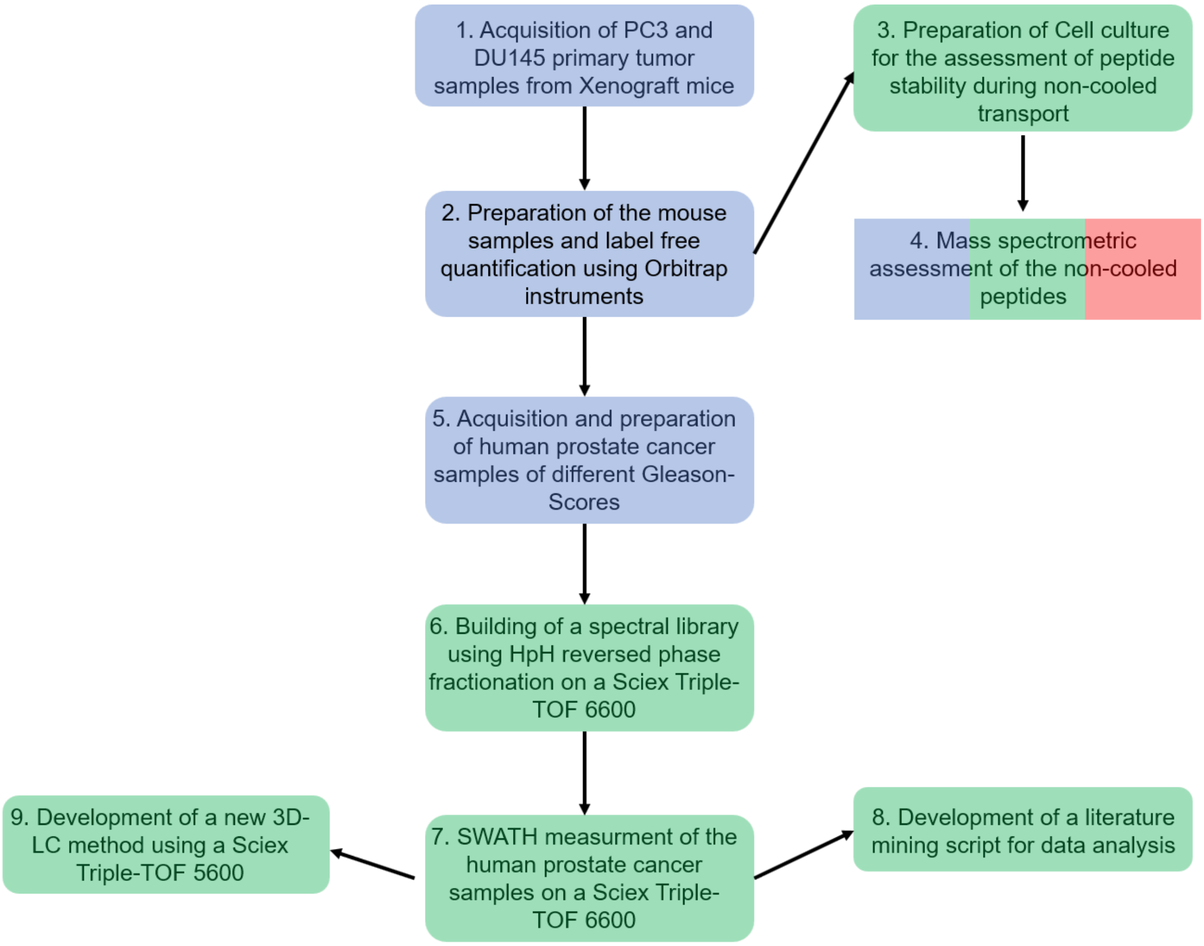

1. Acquisition of PC3 and DU145 primary tumor samples from Xenograft mice
2. Preparation of the mouse samples and label free quantification using Orbitrap instruments
3. Preparation of Cell culture for the assessment of peptide stability during non-cooled transport
4. Mass spectrometric assessment of the non-cooled peptides
5. Acquisition and preparation of human prostate cancer samples of different Gleason-Scores
6. Building of a spectral library using HpH reversed phase fractionation on a Sciex Triple-TOF 6600
7. SWATH measurment of the human prostate cancer samples on a Sciex Triple-TOF 6600
8. Development of a literature mining script for data analysis
9. Development of a new 3D-LC method using a Sciex Triple-TOF 5600

Figure 40: Flowchart of the different parts of the experiments as well as their execution location. Depicted in blue shows experiments conducted in Hamburg, green in Macquarie and red in Fudan.

7.3 PC3 and DU145 primary tumors of xenograft mice

4 biological replicates of each primary tumor derived from either the PC3 or DU145 cells were provided by Prof. Tobias Lange from the anatomy department in the UKE.

Tumors were grown in pfp/rag2$^{-/-}$ xenograft mice as described in [43]. The tumors were excised after sacrificing the mouse and the tumor immediately frozen in liquid nitrogen. The tumors were cut into 20 x 10 µm thick slices using a cryotome. The tissue slices were collected directly into Eppendorf tubes and kept at -80 °C until use.

7.4 Human prostate cancer samples

In total 25 frozen human samples were provided by Dr. Ronald Simon from the pathology department in the UKE. The samples were split as follows: 5 biological replicates of normal prostate tissue, 3 samples of GS 3+3, 5 samples of GS 3+4, 5 samples of GS 4+3, 3 samples of GS 4+5 and 4 samples of GS 5+4. The samples were cut using a cryotome into 10 x 10 µm pieces each and collected directly into Eppendorf tubes and kept at -80 °C until further use.

7.5 Cell culture

SW480 for the assessment of peptide stability, PC3 and DU145 cells for the development of a 3D-LC method were cultured in 10% (v/v) bovine serum supplemented RPMI 1640 medium (Invitrogen) at 37 °C in a 5% CO_2 atmosphere and grown to 80% confluence. Cells were pelleted at 500xg and washed three time using ice cold phosphate buffered saline (PBS) before storing at -80 °C until further use.

7.6 Lysis and protein extraction

Cells as well as tissue samples were lysed by adding 500 µL of a 1% SDC Buffer (1% w/v sodium deoxycholate in 0.1 M triethylammonium bicarbonate) to the Eppendorf tube and sonicated using a probe sonicator (5 sonication cycles at 25% for 15 seconds). Immediately after sonication, the samples were incubated at 100 °C for 5 min. Protein concentration was determined using a BCA test (ThermoFisher).

7.7 In-solution proteolysis of proteins

All lysates were treated the same except for the human tissue samples (see 7.9). Cysteins were reduced using 20 mM dithithreitol (DTT) and incubated at 56 °C for 30 min. After cooling cysteines were blocked using 60 mM 2-iodoacetamide (IAA) and incubated at 37 °C in the dark. Trypsin was added to the solution in a ratio of 1:100

(Trypsin:Protein) and incubated at 37 °C overnight. Peptides were desalted using a reversed phase column (SepPak C18 cartridge, Waters). Peptides were desalted using a reversed phase column (SepPak C18 cartridge, Waters) and lyophilized using a vacuum centrifuge.

7.8 Sample split for the shipping Test (see 5.3)

12 aliquots of the peptides from the SW480 cells were used.

Three aliquots were kept in the laboratory in Macquarie university and kept at -20 °C until measurement (Control). Three aliquots were transported by air without cooling to the University Medical Center Hamburg-Eppendorf (UKE) (HH-Shipped) and the Fudan University (FU-Shipped) in Shanghai. Another three aliquots were transported to the UKE, then returned to the laboratory in Macquarie university (MQ-Shipped).

7.9 SDS-PAGE clean up

Because the human samples contained tissue TEK which is a glycopolymer used to fixate the tissue for better handling during the microtome cutting a clean-up of the samples prior to digestion had to be done.

40 µg of each tissue lysate (30 µL) was mixed with 7.5 µL 4x XT sample buffer (BioRad) and 1.5 µL 20x reducing agent (BioRad) and incubated at 99 °C for 5 min. The samples were loaded onto a 10% Bis-Tris Criterion precast gel (BioRad) and a constant voltage of 200 Volt was applied until the sample was migrated to about 1 cm into the gel. The upper part of the gel (1 cm) for each sample was cut out and transferred into an Eppendorf tube.

7.10 In-Gel Digestion

The 1 cm gel pieces were cut into 1x1 mm^2 pieces before starting the sample preparation. The gel pieces were incubated for 10 min with 100% ACN solution. The ACN was removed and a 10 mM DTT solution (DTT in 0.1 M ammonium bicarbonate) was added and the gel pieces incubated for 30 min at 56 °C. The supernatant was removed and 100% ACN was added and incubated for 10 min and the ACN discarded. A 100 mM IAA solution (IAA in 0.1 M ammonium bicarbonate) was added and incubated for 20 min in the dark after which the supernatant was removed. The gel pieces were incubated another 10 min with 100% ACN which was removed afterwards. A Wash-solution consisting of 50 mM ammonium bicarbonate in 50% ACN was added and the gel pieces incubated for 45 min and the supernatant removed. The gel pieces were incubated a last time with 100% ACN for 10 min and the supernatant removed. A digest solution consisting of 13 ng/µL Trypsin in 50 mM ammonium bicarbonate in 10% ACN was added to the gel pieces and incubated at 37 °C overnight. The supernatant was transferred to a new Eppendorf tube and the gel pieces were incubated with 100% ACN for 15 min after which the supernatant was also transferred to the new collection tube. The gel pieces were incubated with Water for 15 min and the supernatant added to the collection tube. Am extraction solution consisting of 5% formic acid in 65% ACN was added to the gel pieces and incubated for 30 min at 37 °C and the supernatant added to the collection tube. The peptides were lyophilized using a vacuum centrifuge.

7.11 LC and MS parameter

If not otherwise specified, buffer A consisted of 0.1% FA in H_2O.

7.11.1 LC and MS parameters for the Shipping test

The measurement parameters were used as described in [64]. Measurement of the samples was carried out in all three institutes on a tandem mass spectrometer (QExactive, Thermo Fisher Scientific) using the same analytical column (from the same production batch) and the same LC configuration with minor changes in Macquarie, here peptides were directly injected onto the analytical column without prior trapping. The same LC and MS method was used in each laboratory.

Samples were analyzed on a nano-ultra-pressure-liquid chromatography system (Ultimate 3000 RSLCnano, Thermo Fisher Scientific) coupled to a tandem mass spectrometer (QExactive, Thermo Fisher Scientific) with a nano-spray source. Peptides were trapped on a reversed phase trap column (2 cm x 75 µm ID; Acclaim PepMap trap column packed with 3 µm beads, Thermo Fisher Scientific) and separated on a reversed phase column (25 cm x 75 µm ID, Acclaim PepMap, 3 µm beads, Thermo Fisher Scientific). Column temperature was kept at 40 ºC. Peptides were separated using a 120 min stepped gradient starting at 5% buffer B (100%

acetonitrile (ACN) and 0.1% formic acid (FA)) to 22% in 100 min, increasing to 32% in 10 min and ramping to 90% in 10 min at a flow rate of 300 nL/min. Data were acquired in data dependent mode. Spray voltage was set to 2600 V and the transfer capillary temperature set to 275 ºC. All data were acquired in positive mode using a dynamic exclusion for precursor ions of 30 sec. Fullscan spectra were acquired using a resolution of 70000 with a scan range of 400 to 1220 m/z. AGC target was set to 1x106 with a maximum injection time of 120 ms. All Fullscan spectra were acquired in profile mode. The top 12 precursor ions were selected for fragmentation with a minimum intensity of 1x105. Signals with unassigned, singly charged or with 8 or higher charges were excluded from fragmentation. Peptide match option was turned off. Ions were isolated using a 2.0 m/z window and fragmented using higher energy collisional dissociation (HCD) with stepped normalized collision energy (22.5, 25 and 27.5). Fragment spectra were acquired using a resolution of 17500 with a scan range from 200 to 2000 m/z. AGC target was set to 5x10^5 with a maximum injection time of 60 ms. All fragment spectra were acquired in profile mode.

7.11.2 LC and MS parameters for the mouse experiments

Samples were analyzed on a nano-ultra-pressure-liquid chromatography system (Ultimate 3000 RSLCnano, Thermo Fisher Scientific) coupled to a tandem mass spectrometer (Fusion, Thermo Fisher Scientific) with a nano-spray source. Peptides were trapped on a reversed phase trap column (2 cm x 75 µm ID; Acclaim PepMap trap column packed with 3 µm beads, Thermo Fisher Scientific) and separated on a reversed phase column (25 cm x 75 µm ID, Acclaim PepMap, 3 µm beads, Thermo Fisher Scientific). Column temperature was kept at 45 ºC. Peptides were separated using a 112 min stepped gradient starting at 2% buffer B (100% acetonitrile (ACN) and 0.1% formic acid (FA)) to 20% in 85 min, increasing to 32% in 15 min and ramping to 90% in 2 min at a flow rate of 250 nL/min. Data were acquired in data dependent mode. Spray voltage was set to 1700 V and the transfer capillary temperature set to 300 ºC. All data were acquired in positive mode using a dynamic exclusion for precursor ions of 30 sec. Fullscan spectra were acquired in the Orbitrap using a resolution of 120000 with a scan range of 400 to 1300 m/z. AGC target was set to 2x10^5 with a maximum injection time of 100 ms. All Fullscan spectra were acquired in profile mode. The top speed method for precursor ion selection was used for fragmentation with a minimum intensity of 2x10^5. Signals with unassigned, singly charged or with 6 or higher charges were excluded from fragmentation. Ions were isolated using a 1.6 m/z window and fragmented using higher energy collisional dissociation (HCD) with stepped normalized collision energy of 10% and a normalized collision energy of 30. Fragment spectra were acquired in the Iontrap using the rapid scan rate setting with a fixed first mass of 120 m/z. AGC target was set to 5x10^4 with a maximum injection time of 120 ms. Ions were injected for all available parallelizable time. All fragment spectra were acquired in profile mode.

7.11.3 HpH-reversed phase for the fractionation of pooled samples

Dried peptides were dissolved in buffer A (5 mM ammonia solution pH 10.5) to yield a 100 µg peptide solution. The sample was loaded on a 300 Extend-C18 column (2.1 mm x 150 mm, 3.5 µm, 300Å column, Agilent) at a flowrate of 300 µL/min at room temperature with 97% Buffer A and 3% buffer B (5 mM ammonia in 90% ACN) for 10 min. Peptides were separated using a 60 min gradient starting at 3% buffer B to 30% in 55 min and an increase to 70% buffer B in 10 min with a subsequent ramping to 90% buffer B for 5 min at 300 µL/min. For the first 16 min the eluent was collected every two minutes and for the remainder of the run every min (see supplement Figure 1). In case of the pooled PC3 and DU145 samples used for the 3D-LC (see section 5.5) this process was done two times to yield enough peptides for a normal and the 3D-LC analysis.

7.11.4 LC and MS parameter for peptide ion library generation for the human experiments

Samples were analyzed on a nano-ultra-pressure-liquid chromatography system (NanoLCTM ultra and cHiPLC® system, Eksigent) coupled to a tandem mass spectrometer (6600 TripleTOF, Sciex) with a nano-spray source. Peptides were trapped on a reversed phase trap column (1 cm x 75 µm ID; C18 self-packed 3 µm beads, Dr. Maisch) and separated on a reversed phase chip (15 cm × 200 µm nano cHiPLC columns ChromXPTM C18-CL 3 µm 120 Å; Eksigent). Peptides were separated using a 120 min increasing ACN gradient starting at 5% buffer B (90% acetonitrile (ACN) and 0.1% formic acid (FA)) to 35% in 120 min and ramping to 95% in 2 min at a flow rate of 600 nL/min. Data were acquired in data dependent mode. Spray voltage was set to 2500 V. All data were acquired in positive mode using a dynamic exclusion for precursor ions of 30 sec. Fullscan spectra were acquired in a scan range of 350 to 1500 m/z. The top 20 most intense peaks were selected for fragmentation with a minimum threshold of 200 counts per second (CPS) with 100 msec accumulation time. Signals with with charge states between 2+ and 4+ were selected for fragmentation. Ions were fragmented using collision induced dissociation (CID) with a rolling collision energy setting for fragment ion scans of 0.05 x m/z + 4 for z = 2, 0.05 x m/z + 3 for z = 3 and 0.05 x m/z + 2 for z = 4.

7.11.5 LC and MS parameter for SWATH acquisition for the human experiments

Samples were analyzed on a nano-ultra-pressure-liquid chromatography system (NanoLCTM ultra and cHiPLC® system, Eksigent) coupled to a tandem mass spectrometer (6600 TripleTOF, Sciex) with a nano-spray source. Peptides were trapped on a reversed phase trap column (1 cm x 75 µm ID; C18 self-packed 3 µm beads, Dr. Maisch) and separated on a reversed phase chip (15 cm × 200 µm nano cHiPLC columns ChromXPTM C18-CL 3 µm 120 Å; Eksigent). Peptides were separated using a 60 min increasing ACN gradient starting at 5% buffer B (90% acetonitrile (ACN) and 0.1% formic acid (FA)) to 35% in 60 min and ramping to 95% in 2 min at a flow rate of 600 nL/min. Data were acquired in SWATH mode. Spray voltage was set to 2500 V. All data were acquired in positive mode. Fullscan spectra were acquired in a scan range of 350 to 1500 m/z followed by 100 fragment ion scans with predefined consecutive variable Q1 windows from 400 to 1250 m/z. Fragment ion spectra were accumulated for 30 msec with a rolling collision energy for lowest m/z in Q1 windows (assuming z= 2) + 10%. Supplement Table 4 shows the SWATH windows.

7.12 3D-LC

7.12.1 Determination of the binding capacity of the SCX column for the 3D-LC

To determine the binding capacity of the SCX column sequential injections of 1 µg of a pool of tryptic peptides from PC3 and DU145 cell lysates (dissolved in 0.1% FA in H_2O). Peptides were loaded on the SCX column with a flow-rate of 3 µL/min with 5% buffer B (90% ACN and 0.1% FA). Separation and measurement was performed as described in 7.11.4 with the change that the gradient length was adjusted to 40 min and a 5600 TripleTOF instrument was used with the top 10 precursor ion selected for fragmentation. For the regeneration of the SCX column 2 sequential injections of 1 M

NH_4Ac (dissolved in 0.1% FA in H_2O) was used. After analysis of the data using MaxQuant as described in 7.13.1 the loading capacity of the SCX column was determined to be 4 x 1 µg = 4 µg (see Figure 34 in the result part).

7.12.2 Determination of displacer pulse

To determine the amount of displacer (Spermine) needed for a good separation of doubly charged ions, 5 µg of tryptic peptides from the pooled PC3 and DU145 cell lysates were loaded on the SCX column with a flow-rate of 3 µL/min with 5% buffer B (90% ACN and 0.1% FA). Separation and measurement was performed as described in 7.12.1. Peptides were eluted using repeated sequential injections of spermine (25 ng dissolved in 0.1% FA in H_2O) and separated in the second dimension as described in 7.11.4 with the change that the gradient length was adjusted to 40 min and a 5600 TripleTOF instrument was used with the top 10 precursor ion selected for fragmentation. For the regeneration of the SCX column 2 sequential injections of 1 M NH_4Ac (dissolved in 0.1% FA in H_2O) was used. After analysis of the data using MaxQuant as described in section 7.13.1 the optimal displacer amount to elute mostly doubly charged peptides from the SCX column was determined to be 7 x 25 ng = 175 ng (see Figure 35 in the result part).

7.12.3 LC and MS parameter for the 3D-LC

Approximately 4 µg of each of the 13 fractions obtained from the HpH fractionation of the peptide pool of the PC3 and DU145 cell lysate was loaded onto the SCX column with a flow-rate of 3 µL/min with 5% buffer B (90% ACN and 0.1% FA). The flowthrough fraction was recorded using the LC and MS parameter described in section 7.11.4 with the change that a 5600 TripleTOF instrument was used with the top 10 precursor ion selected for fragmentation. Next, 175 ng of spermine was injected on the SCX column and the data for the spermine fraction recorded as described above. Lastly, 1 M NH_4Ac was injected on the SCX column and the regeneration fraction recorded as described above. This process was repeated for all 13 HpH fractions of the polled sample.

7.13 Data analysis

7.13.1 Peptide and protein identification using MaxQuant and ProteinPilot

Obtained data files were processed using either the MaxQuant software [24] version 1.5.3.30 or the ProteinPilot software version 5.0 (Sciex) (for building the spectral library for the human experiments and the 3D-LC analysis). Spectra were searched against a reviewed human FASTA database, obtained in October 2014 (for MaxQuant) or in Januar 2016 (for ProteinPilot) from UniProtKB. MaxQuant provided an additional contamination database. Cystein carbamidomethylation was set as fixed modification, Oxidation on methionine and acetylation on protein N-Terminus was set as variable modification. In case of ProteinPilot the option of biological modifications was enabled. Trypsin was set as specific enzyme and up to two missed cleavages were allowed. Peptides and protein identification were filtered to a false-discovery rate (FDR) of 1%. For all other parameters the standard values were used. In MaxQuant the LFQ option was activated using the fast LFQ algorithm.

7.13.2 Generation of a spectral library and post-processing of SWATH data

To generate the spectral library for the human experiment the group file obtained by ProteinPilot of the DDA runs was loaded into PeakView version 2.1 with SWATH Quantitation plug-in (Sciex) and exported as spectral library by filtering for 1% FDR in CSV format.

The spectral library as well as the generated raw SWATH data files were loaded together in PeakView. SWATH MS peak areas were extracted using in filter criteria of PeakView. These include maximum number of peptides per protein which was set to 6, number fragment ions per peptide which was set to 6, peptide confidence (1% FDR of ProteinPilot), an FDR cutoff for quantification which was set to 1% (this value cannot be changed by the user), XIC retention time window of 5 min and a XIC mass tolerance window which was set to 75 ppm. After peak area extraction the calculated protein peak areas were exported in Excel format for statistical testing.

7.13.3 Statistical testing of label free quantitation data

The data was processed using Perseus version 1.6.0.7 [68], Excel 2016 and the Differential Enrichment Package (DEP) [46] in the R computation language.

CVs and LFQ comparisons shown in section 5.3 were computed by loading the ProteinGroups text file in to Perseus and filtering out contamination, reverse and identified by site hits. The data was exported as a text file and processed in Excel for presentation. The PCA was computed using Perseus after filtering out proteins with missing values. All other data shown in section 5.3 was directly processed in Excel.

Label free quantification data as presented in sections 5.1 and 5.4 was processed using the DEP package in R. Thresholds for the t-Tests were set as minimum fold-change of 2 and a maximum adjusted p-Value of 5%. For the computation of the PCA the result files of the DEP analysis were exported as text files and loaded in to Perseus. The volcano-plots were computed using a self-written R Script using the ggplot2 package and the result files obtained with the DEP package. All Venn diagrams shown in this thesis were computed using the Venny 2.1 algorithm [69].

7.13.4 Network and enrichment analysis

For the network analysis shown in section 5.2.1, the Cytoscape environment [70] was used with the ReactomeFI application. The network was created using the significant proteins as determined by t-Test. The resulting network was filtered to only contain interactions between these proteins and for better visualization the data was appended with the calculated fold-changes of the proteins. The network was clustered using the clustermaker2 app [50] using the GLay algorithm. Cluster with less than 2 interacting proteins were removed.

GO-enrichment analysis was done using BiNGO [51] inputs were the significant proteins as determined by t-Test which showed differential protein concentration in either of the two datasets.

8 Literature

1. WHO Fact Sheet on Cancer. Available from: http://www.who.int/mediacentre/factsheets/fs297/en/.

2. NIH Cancer Stat Facts: Common Cancer Sites. Available from: https://seer.cancer.gov/statfacts/html/common.html.

3. PDQ® Prostate Cancer Treatment [21.07.2018]. Available from: http://www.cancer.gov/cancertopics/pdq/treatment/prostate/Patient/page1/AllPages.

4. Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, et al. Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin. 2016;66(4):271-89. Epub 2016/06/03. doi: 10.3322/caac.21349. PubMed PMID: 27253694.

5. Jahn JL, Giovannucci EL, Stampfer MJ. The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era. Int J Cancer. 2015;137(12):2795-802. Epub 2015/01/06. doi: 10.1002/ijc.29408. PubMed PMID: 25557753; PubMed Central PMCID: PMCPMC4485977.

6. Livi L, Isidori AM, Sherris D, Gravina GL. Advances in prostate cancer research and treatment. Biomed Res Int. 2014;2014:708383. Epub 2014/09/13. doi: 10.1155/2014/708383. PubMed PMID: 25215290; PubMed Central PMCID: PMCPMC4151599.

7. Krahn MD. Screening for Prostate Cancer. Jama. 1994;272(10). doi: 10.1001/jama.1994.03520100035030.

8. Prostate-Specific Antigen (PSA) Test [21.07.2018]. Available from: https://www.cancer.gov/types/prostate/psa-fact-sheet.

9. Prostate-specific antigen (UniProt.org). Available from: https://www.uniprot.org/uniprot/P07288.

10. US Preventive Services Task Force, Final Recommendation Statement, Prostate Cancer: Screening. Available from: https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/prostate-cancer-screening1.

11. Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89-95. Epub 2001/03/10. doi: 10.1067/mcp.2001.113989. PubMed PMID: 11240971.

12. Drabovich AP, Martinez-Morillo E, Diamandis EP. Toward an integrated pipeline for protein biomarker development. Biochim Biophys Acta. 2015;1854(6):677-86. Epub 2014/09/15. doi: 10.1016/j.bbapap.2014.09.006. PubMed PMID: 25218201.

13. Jungblut PR, Thiede B, Schluter H. Towards deciphering proteomes via the proteoform, protein speciation, moonlighting and protein code concepts. J Proteomics. 2016;134:1-4. Epub 2016/03/15. doi: 10.1016/j.jprot.2016.01.012. PubMed PMID: 26972666.

14. Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, et al. Progress with gene-product mapping of the Mollicutes:Mycoplasma genitalium. Electrophoresis. 1995;16(1):1090-4. doi: 10.1002/elps.11501601185.

15. James P. Protein identification in the post-genome era: the rapid rise of proteomics. Quarterly Reviews of Biophysics. 1997;30(4):279-331. doi: 10.1017/s0033583597003399.

16. Anderson NL, Anderson NG. Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis. 1998;19(11):1853-61. Epub 1998/09/18. doi: 10.1002/elps.1150191103. PubMed PMID: 9740045.

17. Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000;405(6788):837-46. Epub 2000/06/24. doi: 10.1038/35015709. PubMed PMID: 10866210.

18. The Human Genome Project. Available from: https://www.genome.gov/11006929/2003-release-international-consortium-completes-hgp/.

19. Eliuk S, Makarov A. Evolution of Orbitrap Mass Spectrometry Instrumentation. Annu Rev Anal Chem (Palo Alto Calif). 2015;8:61-80. Epub 2015/07/15. doi: 10.1146/annurev-anchem-071114-040325. PubMed PMID: 26161972.

20. Trypsin (Sigma) [21.07.2018]. Available from: https://www.sigmaaldrich.com/technical-documents/articles/biology/trypsin.html.

21. Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem (Palo Alto Calif). 2016;9(1):499-519. Epub 2016/06/17. doi: 10.1146/annurev-anchem-071015-041550. PubMed PMID: 27306313; PubMed Central PMCID: PMCPMC5373801.

22. Fragmentation methods (ThermoFisher). Available from: https://www.thermofisher.com/de/de/home/industrial/mass-spectrometry/mass-spectrometry-learning-center/mass-spectrometry-technology-overview/dissociation-technique-technology-overview.html.

23. Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol. 2004;5(9):699-711. doi: 10.1038/nrm1468. PubMed PMID: 15340378.

24. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 2016;11(12):2301-19. Epub 2016/11/04. doi: 10.1038/nprot.2016.136. PubMed PMID: 27809316.

25. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry. 1994;5(11):976-89. doi: 10.1016/1044-0305(94)80016-2.

26. Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol. 2007;367:87-119. doi: 10.1385/1-59745-275-0:87. PubMed PMID: 17185772.

27. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010;73(11):2092-123. Epub 2010/09/08. doi: 10.1016/j.jprot.2010.08.009. PubMed PMID: 20816881; PubMed Central PMCID: PMCPMC2956504.

28. Kruger M, Moser M, Ussar S, Thievessen I, Luber CA, Forner F, et al. SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. Cell. 2008;134(2):353-64. Epub 2008/07/30. doi: 10.1016/j.cell.2008.05.033. PubMed PMID: 18662549.

29. Li Z, Adams RM, Chourey K, Hurst GB, Hettich RL, Pan C. Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. J Proteome Res. 2012;11(3):1582-90. Epub 2011/12/23. doi: 10.1021/pr200748h. PubMed PMID: 22188275.

30. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014;13(9):2513-26. Epub 2014/06/20. doi: 10.1074/mcp.M113.031591. PubMed PMID: 24942700; PubMed Central PMCID: PMCPMC4159666.

31. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007;389(4):1017-31. Epub 2007/08/02. doi: 10.1007/s00216-007-1486-6. PubMed PMID: 17668192.

32. Liu H, Sadygov RG, Yates JR, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem. 2004;76(14):4193-201. doi: 10.1021/ac0498563. PubMed PMID: 15253663.

33. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;11(6):O111 016717. doi: 10.1074/mcp.O111.016717. PubMed PMID: 22261725; PubMed Central PMCID: PMCPMC3433915.

34. Hu A, Noble WS, Wolf-Yadlin A. Technical advances in proteomics: new developments in data-independent acquisition. F1000Res. 2016;5. Epub 2016/04/20. doi: 10.12688/f1000research.7042.1. PubMed PMID: 27092249; PubMed Central PMCID: PMCPMC4821292.

35. Kocova Vlckova H, Pilarova V, Svobodova P, Plisek J, Svec F, Novakova L. Current state of bioanalytical chromatography in clinical analysis. Analyst. 2018;143(6):1305-25. Epub 2018/02/21. doi: 10.1039/c7an01807j. PubMed PMID: 29461553.

36. Frenz JH, C. High performance displacement chromatography: Calculation and experimental verification of zone development. AIChE. 1985;31(3):400-9. doi: 10.1002/aic.690310307

37. Kwiatkowski M, Krosser D, Wurlitzer M, Steffen P, Barcaru A, Krisp C, et al. Application of Displacement Chromatography to Online Two-Dimensional Liquid Chromatography Coupled to Tandem Mass Spectrometry Improves Peptide Separation Efficiency and Detectability for the Analysis of Complex Proteomes. Anal Chem. 2018. Epub 2018/07/18. doi: 10.1021/acs.analchem.8b02189. PubMed PMID: 30014690.

38. Kamel MH, Khalil MI, Alobuia WM, Su J, Davis R. Incidence of metastasis and prostate-specific antigen levels at diagnosis in Gleason 3+4 versus 4+3 prostate cancer. Urol Ann. 2018;10(2):203-8. Epub 2018/05/03. doi: 10.4103/UA.UA_124_17. PubMed PMID: 29719335; PubMed Central PMCID: PMCPMC5907332.

39. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol. 2016;40(2):244-52. Epub 2015/10/23. doi: 10.1097/PAS.0000000000000530. PubMed PMID: 26492179.

40. Valko M, Izakovic M, Mazur M, Rhodes CJ, Telser J. Role of oxygen radicals in DNA damage and cancer incidence. Molecular and Cellular Biochemistry. 2004;266(1/2):37-56. doi: 10.1023/b:Mcbi.0000049134.69131.89.

41. Agrin (Uniprot). Available from: https://www.uniprot.org/uniprot/O00468.

42. Li X, Wang X, Song W, Xu H, Huang R, Wang Y, et al. Oncogenic properties of NEAT1 in prostate cancer cells depend on the CDC5L-AGRN transcriptional regulation circuit. Cancer Res. 2018. Epub 2018/06/07. doi: 10.1158/0008-5472.CAN-18-0688. PubMed PMID: 29871935.

43. Lange T, Ullrich S, Muller I, Nentwich MF, Stubke K, Feldhaus S, et al. Human prostate cancer in a clinically relevant xenograft mouse model: identification of beta(1,6)-branched oligosaccharides as a marker of tumor progression. Clin Cancer Res. 2012;18(5):1364-73. Epub 2012/01/21. doi: 10.1158/1078-0432.CCR-11-2900. PubMed PMID: 22261809.

44. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res. 2010;9(2):761-76. doi:

10.1021/pr9006365. PubMed PMID: 19921851; PubMed Central PMCID: PMCPMC2818771.

45. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. Epub 2015/01/22. doi: 10.1093/nar/gkv007. PubMed PMID: 25605792; PubMed Central PMCID: PMCPMC4402510.

46. Zhang X, Smits AH, van Tilburg GB, Ovaa H, Huber W, Vermeulen M. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc. 2018;13(3):530-50. Epub 2018/02/16. doi: 10.1038/nprot.2017.147. PubMed PMID: 29446774.

47. Bensimon A, Heck AJ, Aebersold R. Mass spectrometry-based proteomics and network biology. Annu Rev Biochem. 2012;81:379-405. Epub 2012/03/24. doi: 10.1146/annurev-biochem-072909-100424. PubMed PMID: 22439968.

48. Reactome [10.07.2018]. Available from: https://reactome.org/.

49. STRING-DB. Available from: https://string-db.org.

50. ClusterMaker2. Available from: http://www.rbvi.ucsf.edu/cytoscape/clusterMaker2/.

51. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448-9. Epub 2005/06/24. doi: 10.1093/bioinformatics/bti551. PubMed PMID: 15972284.

52. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57. Epub 2009/01/10. doi: 10.1038/nprot.2008.211. PubMed PMID: 19131956.

53. Heine P, Ehrlicher A, Kas J. Neuronal and metastatic cancer cells: Unlike brothers. Biochim Biophys Acta. 2015;1853(11 Pt B):3126-31. Epub 2015/06/30. doi: 10.1016/j.bbamcr.2015.06.011. PubMed PMID: 26119327.

54. Fernandez JM, Hoffmann R, Valencia A. iHOP web services. Nucleic Acids Res. 2007;35(Web Server issue):W21-6. Epub 2007/05/09. doi: 10.1093/nar/gkm298. PubMed PMID: 17485473; PubMed Central PMCID: PMCPMC1933131.

55. Teng Y, Ren X, Li H, Shull A, Kim J, Cowell JK. Mitochondrial ATAD3A combines with GRP78 to regulate the WASF3 metastasis-promoting protein. Oncogene. 2016;35(3):333-43. Epub 2015/03/31. doi: 10.1038/onc.2015.86. PubMed PMID: 25823022; PubMed Central PMCID: PMCPMC4828935.

56. Taddei ML, Parri M, Angelucci A, Onnis B, Bianchini F, Giannoni E, et al. Kinase-dependent and -independent roles of EphA2 in the regulation of prostate cancer invasion and metastasis. Am J Pathol. 2009;174(4):1492-503. Epub 2009/03/07. doi: 10.2353/ajpath.2009.080473. PubMed PMID: 19264906; PubMed Central PMCID: PMCPMC2671379.

57. Wang W, Chu HJ, Liang YC, Huang JM, Shang CL, Tan H, et al. FABP5 correlates with poor prognosis and promotes tumor cell growth and metastasis in cervical cancer. Tumour Biol. 2016;37(11):14873-83. Epub 2016/09/21. doi: 10.1007/s13277-016-5350-1. PubMed PMID: 27644245.

58. Di Cello F, Shin J, Harbom K, Brayton C. Knockdown of HMGA1 inhibits human breast cancer cell growth and metastasis in immunodeficient mice. Biochem Biophys Res Commun. 2013;434(1):70-4. Epub 2013/04/03. doi: 10.1016/j.bbrc.2013.03.064. PubMed PMID: 23545254; PubMed Central PMCID: PMCPMC3662800.

59. Bulk E, Sargin B, Krug U, Hascher A, Jun Y, Knop M, et al. S100A2 induces metastasis in non-small cell lung cancer. Clin Cancer Res. 2009;15(1):22-9. Epub 2009/01/02. doi: 10.1158/1078-0432.CCR-08-0953. PubMed PMID: 19118029.

60. Yao Q, Cao Z, Tu C, Zhao Y, Liu H, Zhang S. MicroRNA-146a acts as a metastasis suppressor in gastric cancer by targeting WASF2. Cancer Lett. 2013;335(1):219-24. Epub 2013/02/26. doi: 10.1016/j.canlet.2013.02.031. PubMed PMID: 23435376.

61. Thermo Scientific, HeLa Data sheet 2018 [cited 2018 15.01.18]. HeLa Protein Digest Standard]. Available from: https://assets.thermofisher.com/TFS-Assets/LSG/certificate/Certificates%20of%20Analysis/PC199008_88329.pdf.

62. Stability of peptides, Sigma-Aldrich.

63. Bell LN. Peptide Stability in Solids and Solutions. Biotechnology Progress. 1997;13(4):342-6. doi: 10.1021/bp970057y.

64. Steffen P, Krisp C, Yi W, Yang P, Molloy MP, Schluter H. Multi-laboratory analysis of the variability of shipped samples for proteomics following non-cooled international transport. Anal Biochem. 2018;548:60-5. Epub 2018/02/28. doi: 10.1016/j.ab.2018.02.026. PubMed PMID: 29486204.

65. Klockenbusch C, O'Hara JE, Kast J. Advancing formaldehyde cross-linking towards quantitative proteomic applications. Anal Bioanal Chem. 2012;404(4):1057-67. doi: 10.1007/s00216-012-6065-9. PubMed PMID: 22610548.

66. Washburn MP, Wolters D, Yates JR, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001;19(3):242-7. doi: 10.1038/85686. PubMed PMID: 11231557.

67. Webb KJ, Xu T, Park SK, Yates JR, 3rd. Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. J Proteome Res. 2013;12(5):2177-84. Epub 2013/04/02. doi: 10.1021/pr400027m. PubMed PMID: 23540446; PubMed Central PMCID: PMCPMC3815592.

68. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731-40. doi: 10.1038/nmeth.3901. PubMed PMID: 27348712.

69. Oliveros JC. Venny. An interactive tool for comparing lists with Venn's diagrams. 2007-2015 [cited 2017 02.05.2017]. Available from: http://bioinfogp.cnb.csic.es/tools/venny/index.html.

70. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504. Epub 2003/11/05. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed Central PMCID: PMCPMC403769.

9 Supplement

9.1 GHS classification of used chemicals

Hazard symbols and H- and P-Statements for the chemicals used:

Chemical	Hazard Symbol	Hazard statements	Precautionary Statements
Acetonitrile	GHS02, GHS07	H225, H302, H312, H319, H332	P210, P280, P305+351+338
Formic Acid	GHS02, GHS05	H226-H314	P280-P305 + P351 + P338-P310
Dithiothreitol	GHS07	H302-H315-H319-H335	P261-P305 + P351 + P338
Iodoacetamide	GHS06, GHS08	H301-H317-H334-H413	P261-P280-P301 + P310-P342 + P311
Trypsin	GHS07	H315-H317-H319-H335	P261-P280-P305 + P351 + P338
PBS, 1x in H_2O	-	-	-

9.2 Tables

S. Table 1: Proteins responsible for the left/right shift of the PCA, first marker candidates. Additionally, the log2 ratios and the adjusted p-Value as calculated by the t-Test are shown. The last column depicts if a Protein from this List was also found as significant by the t-Test (x).

name	Cell	DU145_vs_PC3_diff	DU145_vs_PC3_p.adj	Is Significant in T-Test
ACOT7	PC3	-3.173374329	0.094883559	
ACTBL2	PC3	-4.093399075	0.866041238	
ADIRF	PC3	-6.41798583	2.21E-13	x
AGRN	DU145	4.03841134	0.243139788	
AK4	PC3	-3.617912073	0.44593147	
AKR1B1	DU145	3.690513532	0.257900499	
AKR1B10	DU145	5.416170462	6.88E-06	x
AKR1C1	DU145	8.991905489	1.95E-10	x
AKR1C2	DU145	8.21518361	4.43E-13	x

AKR1C3	DU145	7.900415381	2.52E-06	x
ALDH1A1	DU145	3.218117922	0.021656571	x
ASRGL1	DU145	3.552361618	8.69E-05	x
ASS1	DU145	4.291585287	1.61E-05	x
BASP1	DU145	3.910622438	8.24E-08	x
C4A	DU145	3.30807973	0.748499611	
CDH1	DU145	4.7302548	5.53E-13	x
CEACAM5	DU145	5.382395326	0.027100083	x
CEACAM6	DU145	7.787948717	2.71E-10	x
CKB	DU145	8.395353261	3.37E-08	x
CNN2	DU145	4.433525648	5.43E-05	x
COTL1	PC3	-4.27253889	3.61E-07	x
CPD	DU145	3.401903061	6.37E-09	x
CPM	DU145	4.140995941	0.011897277	x
CTNNA1	DU145	6.31050056	0.000140517	x
DHRS7	DU145	3.018544952	0.738552694	
DMBT1	DU145	3.220998973	0.77037297	
DSP	DU145	5.141130093	0.000407014	x
EDF1	PC3	-3.25530011	0.582168544	
EFR3A	PC3	-3.344930879	0.814512304	
EPCAM	DU145	4.147988919	0.171257114	
EPHX1	DU145	4.582617996	0.607425011	
FBP1	DU145	4.35607161	0.020507968	x
FGA	DU145	7.10196186	0.011058915	x
FGB	DU145	3.619651286	0.624940317	
FGG	DU145	4.26734983	0.220836493	
FTL	DU145	3.409931253	0.472171059	
GCLC	DU145	4.643142949	0.004554606	x
GPX2	DU145	5.191226365	0.000110279	x
HIST1H1D	PC3	-3.987571135	0.013900286	x
HK1	PC3	-4.252733087	0.301810846	
HPCAL1	PC3	-3.487826935	0.077487802	
HSD17B11	DU145	3.486360971	0.54310936	
ICAM1	DU145	3.179647146	0.429240292	
JUP	DU145	3.773943095	0.011162031	x
KRT18	DU145	5.218408829	1.60E-09	x
KYNU	DU145	4.322520178	7.18E-06	x
LPGAT1	DU145	5.390812857	0.008355873	x
MAP1B	PC3	-4.52503219	0.000578714	x
ME1	DU145	3.387695219	0.058206786	
MFF	DU145	3.232735287	0.481259172	
MGST1	DU145	3.709840153	0.328534285	
MSMP	PC3	-6.168762266	1.41E-08	x
MUC1	DU145	4.675384216	0.00035686	x
MYH10	DU145	5.185323626	8.85E-13	x

NAPRT	DU145	4.117270495	0.469453255	
NAPSA	DU145	5.778012968	0.133996246	
NES	PC3	-6.918637776	2.21E-13	x
NNMT	DU145	4.42339451	6.78E-05	x
NQO1	DU145	3.885301082	8.29E-09	x
PC	PC3	-3.318577391	0.43815931	
PRODH	DU145	4.290046054	0.002414432	x
PTGR1	DU145	6.274164955	0.000778498	x
PXN	PC3	-3.593485497	0.003714138	x
S100A2	PC3	-4.618018963	1.09E-06	x
S100A4	PC3	-3.621069738	0.077799965	
SERPINA1	DU145	5.244379888	0.000484656	x
SFTPB	DU145	6.370219317	0.057004956	
SHMT1	DU145	3.505741328	0.057024519	
SIGLEC16	DU145	3.168629577	0.758815337	
SPAG9	PC3	-3.844544821	0.005819288	x
SQSTM1	DU145	3.549462347	1.37E-05	x
STAT3	DU145	3.453427354	0.000413713	x
SULT1A4	DU145	6.188764722	4.24E-09	x
TGFBI	DU145	4.988906154	0.221060125	
TGM2	DU145	5.905392372	0.000325078	x
TOMM34	PC3	-4.043571336	0.147842742	
TYMP	DU145	3.554597676	0.01264802	
UGDH	DU145	4.003778279	9.75E-08	x
ZYX	PC3	-3.741275698	0.116011841	

S. Table 2: Significantly regulated proteins as determined by the t-Test.

name	DU145_vs_PC3_diff	DU145_vs_PC3_p.adj	DU145_vs_PC3_significant
ADIRF	-6.41798583	2.21E-13	TRUE
AHNAK	-2.206533862	0.023786472	TRUE
AKR1B10	5.416170462	6.88E-06	TRUE
AKR1C1	8.991905489	1.95E-10	TRUE
AKR1C2	8.21518361	4.43E-13	TRUE
AKR1C3	7.900415381	2.52E-06	TRUE
ALDH16A1	3.218117922	0.021656571	TRUE
ANP32E	1.763448373	0.000426138	TRUE
ANXA4	2.778191271	0.011869045	TRUE
ANXA7	-1.697296752	0.031881844	TRUE
ASRGL1	3.552361618	8.69E-05	TRUE
ASS1	4.291585287	1.61E-05	TRUE

ATAD3A	-1.924370005	0.012452772	TRUE
BASP1	3.910622438	8.24E-08	TRUE
BCHE	2.593861822	2.58E-06	TRUE
CBR1	2.02661311	0.027450496	TRUE
CBR3	2.103326596	0.005176925	TRUE
CD44	-1.869632567	0.015590735	TRUE
CDH1	4.7302548	5.53E-13	TRUE
CDK6	-1.946522811	0.042498282	TRUE
CEACAM5	5.382395326	0.027100083	TRUE
CEACAM6	7.787948717	2.71E-10	TRUE
CIB1	2.551451109	0.001955449	TRUE
CKB	8.395353261	3.37E-08	TRUE
CNN2	4.433525648	5.43E-05	TRUE
COL12A1	2.321643608	0.012198639	TRUE
COL4A1	3.175476712	0.017074339	TRUE
COTL1	-4.27253889	3.61E-07	TRUE
CPD	3.401903061	6.37E-09	TRUE
CPM	4.140995941	0.011897277	TRUE
CTNNA1	6.31050056	0.000140517	TRUE
CYB5A	2.325455441	0.001840658	TRUE
CYC1	-1.643333424	0.045104584	TRUE
DNAJC9	-1.984312415	0.01329144	TRUE
DSG2	2.96890867	0.040206187	TRUE
DSP	5.141130093	0.000407014	TRUE
EPHA2	-3.069545839	7.13E-06	TRUE
ERMP1	2.056480126	0.043385821	TRUE
ERP29	1.654794017	0.006824011	TRUE
F11R	1.865567938	0.024417282	TRUE
FABP5	-3.079643335	0.000588801	TRUE
FAM177A1	1.748517502	0.030004335	TRUE
FBN2	-3.23674273	0.009576216	TRUE
FBP1	4.35607161	0.020507968	TRUE
FGA	7.10196186	0.011058915	TRUE
GCLC	4.643142949	0.004554606	TRUE
GOLPH3	2.849719552	0.001012381	TRUE
GPX2	5.191226365	0.000110279	TRUE
GRHPR	1.901962604	0.00772445	TRUE
GSR	2.024494484	0.019620216	TRUE
GUSB	3.024788767	0.000443217	TRUE
HEXA	2.676646896	0.039370674	TRUE
HIGD1A	-1.903832673	0.019103671	TRUE
HIST1H1D	-3.987571135	0.013900286	TRUE
HMGA1	-2.440519863	0.006055604	TRUE
HSPD1	-1.861656486	0.000422912	TRUE
HSPE1	-1.82859279	0.002206342	TRUE

JUP	3.773943095	0.011162031	TRUE
KRT18	5.218408829	1.60E-09	TRUE
KYNU	4.322520178	7.18E-06	TRUE
LACTB2	2.093624265	0.007744802	TRUE
LGALS1	-2.736169529	0.001231037	TRUE
LPGAT1	5.390812857	0.008355873	TRUE
LYPLA1	2.780266514	0.007162514	TRUE
MAP1B	-4.52503219	0.000578714	TRUE
MSMP	-6.168762266	1.41E-08	TRUE
MUC1	4.675384216	0.00035686	TRUE
MUT	2.133668844	0.03222491	TRUE
MYH10	5.185323626	8.85E-13	TRUE
NAMPT	1.932470346	0.003963331	TRUE
NEFL	-3.183729834	0.00021785	TRUE
NES	-6.918637776	2.21E-13	TRUE
NNMT	4.42339451	6.78E-05	TRUE
NQO1	3.885301082	8.29E-09	TRUE
NUDT5	1.988353646	0.000418266	TRUE
PGD	2.164814287	0.0002204	TRUE
PHGDH	-1.664554043	0.038474456	TRUE
PLEC	-1.486053347	0.042391323	TRUE
PPM1F	-2.192770862	0.037939044	TRUE
PRODH	4.290046054	0.002414432	TRUE
PSAP	1.794111509	0.045137222	TRUE
PSMD10	-2.511327638	0.001345838	TRUE
PTGR1	6.274164955	0.000778498	TRUE
PXN	-3.593485497	0.003714138	TRUE
RAB2A	1.377672921	0.03163959	TRUE
S100A2	-4.618018963	1.09E-06	TRUE
SCFD1	1.676795902	0.013816433	TRUE
SELENBP1	2.361634417	3.98E-05	TRUE
SERPINA1	5.244379888	0.000484656	TRUE
SMS	-2.72285362	2.37E-06	TRUE
SPAG9	-3.844544821	0.005819288	TRUE
SQRDL	1.969433633	0.004413021	TRUE
SQSTM1	3.549462347	1.37E-05	TRUE
STARD10	2.461262208	0.006722998	TRUE
STAT3	3.453427354	0.000413713	TRUE
SULT1A4	6.188764722	4.24E-09	TRUE
SUSD2	3.302388171	0.002532931	TRUE
TGM2	5.905392372	0.000325078	TRUE
TYMP	3.554597676	0.01264802	TRUE
UBA1	-1.316502737	0.043880589	TRUE
UGDH	4.003778279	9.75E-08	TRUE
VAPA	1.467641403	0.013567086	TRUE

WASF2	-1.873415009	0.018655454	TRUE

S. table 3: Complete summary tab of the output of the text mining script for the 32 significantly up regulated proteins in the PC3 samples.

N	UniProtID	TaxID	Synonyms	Keywords	KeywordInTitleOnly	Results	Category	Other	PubmedQuery
1	P48681	9606	NES,Nbla00170	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("NES"[TI] OR "Nbla00170"[TI])
2	Q15847	9606	ADIRF,AFRO,APM2,C10orf116	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("ADIRF"[TI] OR "AFRO"[TI] OR "APM2"[TI] OR "C10orf116"[TI])
3	Q1L6U9	9606	MSMP,PSMP	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("MSMP"[TI] OR "PSMP"[TI])
4	P29034	9606	S100A2,S100L	Metastasis	Yes	2	2	0	"Metastasis"[TI] AND ("S100A2"[TI] OR "S100L"[TI])
5	P46821	9606	MAP1B	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("MAP1B"[TI])
6	Q14019	9606	COTL1,CLP	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("COTL1"[TI] OR "CLP"[TI])
7	P16402	9606	HIST1H1D,H1F3	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("HIST1H1D"[TI] OR "H1F3"[TI])
8	O60271	9606	SPAG9,HSS,KIAA0516, MAPK8IP4,SYD1,HLC6	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("SPAG9"[TI] OR "HSS"[TI] OR "KIAA0516"[TI] OR "MAPK8IP4"[TI] OR "SYD1"[TI] OR "HLC6"[TI])
9	P49023	9606	PXN	Metastasis	Yes	2	2	0	"Metastasis"[TI] AND ("PXN"[TI])
10	P35556	9606	FBN2	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("FBN2"[TI])
11	P07196	9606	NEFL,NF68,NFL	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("NEFL"[TI] OR "NF68"[TI] OR "NFL"[TI])
12	Q01469	9606	FABP5	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("FABP5"[TI])
13	P29317	9606	EPHA2,ECK	Metastasis	Yes	14	2	0	"Metastasis"[TI] AND ("EPHA2"[TI] OR "ECK"[TI])
14	P09382	9606	LGALS1	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("LGALS1"[TI])
15	P52788	9606	SMS	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("SMS"[TI])
16	O75832	9606	PSMD10	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("PSMD10"[TI])
17	P17096	9606	HMGA1,HMGIY	Metastasis	Yes	3	2	0	"Metastasis"[TI] AND ("HMGA1"[TI] OR "HMGIY"[TI])
18	Q09666	9606	AHNAK,PM227	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("AHNAK"[TI] OR "PM227"[TI])
19	P49593	9606	PPM1F,KIAA0015,POPX2	Metastasis	Yes	2	2	0	"Metastasis"[TI] AND ("PPM1F"[TI] OR "KIAA0015"[TI] OR "POPX2"[TI])
20	Q8WXX5	9606	DNAJC9	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("DNAJC9"[TI])

21	Q00534	9606	CDK6,CDKN6	Metastasis	Yes	3	2	0	"Metastasis"[TI] AND ("CDK6"[TI] OR "CDKN6"[TI])
22	Q9NVI7	9606	ATAD3A	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("ATAD3A"[TI])
23	Q9Y241	9606	HIGD1A,HIG1,HSPC010	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("HIGD1A"[TI] OR "HIG1"[TI] OR "HSPC010"[TI])
24	Q9Y6W5	9606	WASF2,WAVE2	Metastasis	Yes	4	2	0	"Metastasis"[TI] AND ("WASF2"[TI] OR "WAVE2"[TI])
25	P16070	9606	CD44,LHR,MDU2,MDU3,MIC4	Metastasis	Yes	156	1	16	"Metastasis"[TI] AND ("CD44"[TI] OR "LHR"[TI] OR "MDU2"[TI] OR "MDU3"[TI] OR "MIC4"[TI])
26	P10809	9606	HSPD1,HSP60	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("HSPD1"[TI] OR "HSP60"[TI])
27	P61604	9606	HSPE1	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("HSPE1"[TI])
28	P20073	9606	ANXA7,ANX7,SNX,OK/SW-cl.95	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("ANXA7"[TI] OR "ANX7"[TI] OR "SNX"[TI] OR "OK/SW-cl.95"[TI])
29	O43175	9606	PHGDH,PGDH3	Metastasis	Yes	1	2	0	"Metastasis"[TI] AND ("PHGDH"[TI] OR "PGDH3"[TI])
30	P08574	9606	CYC1	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("CYC1"[TI])
31	Q15149	9606	PLEC,PLEC1	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("PLEC"[TI] OR "PLEC1"[TI])
32	P22314	9606	UBA1,A1S9T,UBE1	Metastasis	Yes	0	3	0	"Metastasis"[TI] AND ("UBA1"[TI] OR "A1S9T"[TI] OR "UBE1"[TI])

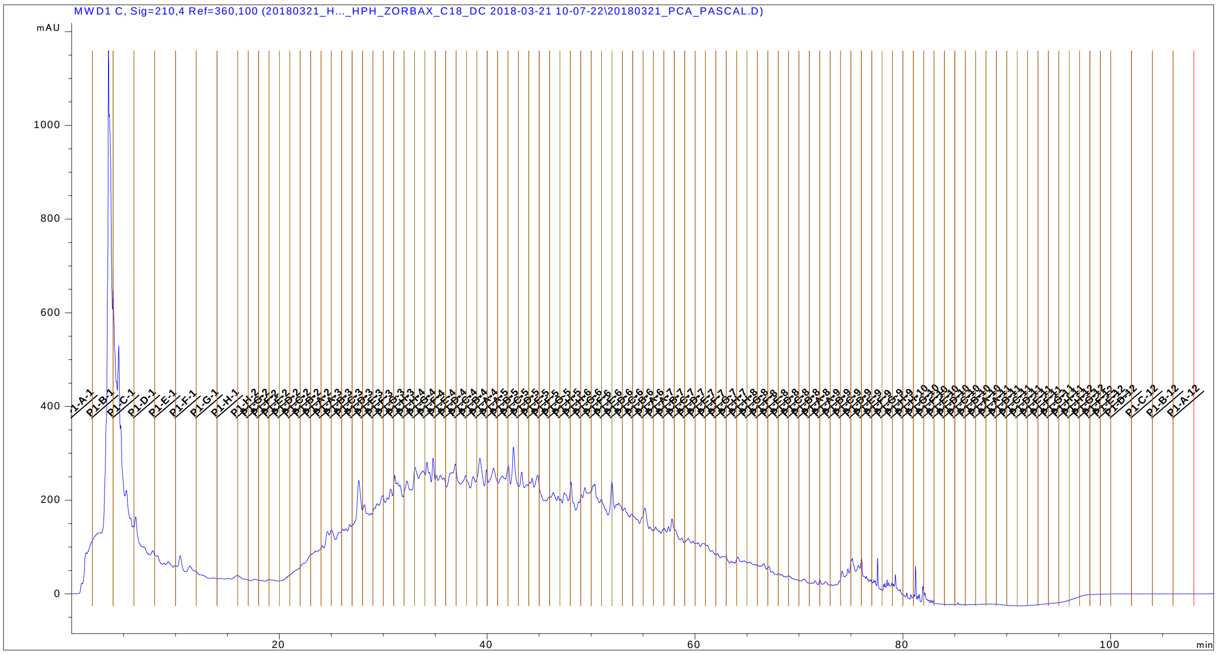
S. table 4: SWATH windows with their start and end m/z values as well as their margins.

Start m/z	End m/z	Margin m/z
399.5	406.5	0.5
405.5	412.5	0.5
411.5	418.5	0.5
417.5	424.5	0.5
423.5	430.5	0.5
429.5	436.5	0.5
435.5	442.5	0.5
441.5	448.5	0.5
447.5	454.5	0.5
453.5	459.5	0.5
458.5	464.5	0.5
463.5	469.5	0.5
468.5	474.5	0.5
473.5	479.5	0.5
478.5	484.5	0.5
483.5	489.5	0.5
488.5	494.5	0.5
493.5	499.5	0.5
498.5	504.5	0.5
503.5	509.5	0.5
508.5	514.5	0.5
513.5	519.5	0.5
518.5	524.5	0.5
523.5	529.5	0.5
528.5	534.5	0.5
533.5	539.5	0.5
538.5	544.5	0.5
543.5	549.5	0.5
548.5	554.5	0.5
553.5	559.5	0.5
558.5	564.5	0.5
563.5	569.5	0.5
568.5	574.5	0.5
573.5	579.5	0.5
578.5	584.5	0.5
583.5	589.5	0.5
588.5	594.5	0.5
593.5	599.5	0.5
598.5	604.5	0.5
603.5	609.5	0.5
608.5	614.5	0.5
613.5	619.5	0.5

618.5	624.5	0.5
623.5	629.5	0.5
628.5	634.5	0.5
633.5	639.5	0.5
638.5	644.5	0.5
643.5	649.5	0.5
648.5	654.5	0.5
653.5	660.5	0.5
659.5	666.5	0.5
665.5	672.5	0.5
671.5	678.5	0.5
677.5	684.5	0.5
683.5	690.5	0.5
689.5	696.5	0.5
695.5	702.5	0.5
701.5	708.5	0.5
707.5	714.5	0.5
713.5	720.5	0.5
719.5	726.5	0.5
725.5	732.5	0.5
731.5	738.5	0.5
737.5	744.5	0.5
743.5	750.5	0.5
749.5	756.5	0.5
755.5	763.5	0.5
762.5	770.5	0.5
769.5	777.5	0.5
776.5	784.5	0.5
783.5	791.5	0.5
790.5	798.5	0.5
797.5	805.5	0.5
804.5	812.5	0.5
811.5	819.5	0.5
818.5	826.5	0.5
825.5	834.5	0.5
833.5	842.5	0.5
841.5	850.5	0.5
849.5	858.5	0.5
857.5	867.5	0.5
866.5	876.5	0.5
875.5	885.5	0.5
884.5	894.5	0.5
893.5	903.5	0.5
902.5	914.5	0.5
913.5	925.5	0.5

924.5	936.5	0.5
935.5	950.5	0.5
949.5	964.5	0.5
963.5	978.5	0.5
977.5	992.5	0.5
991.5	1,011.50	0.5
1,010.50	1,030.50	0.5
1,029.50	1,054.50	0.5
1,053.50	1,078.50	0.5
1,077.50	1,117.50	0.5
1,116.50	1,156.50	0.5
1,155.50	1,200.50	0.5
1,199.50	1,249.50	0.5

9.3 Figures



S.Figure 1: Chromatogram of the pooled human prostate cancer sample with corresponding fractions.

10 Danksagung/Acknowledgement

At this point, I would like to express my gratitude to all the people who supported me during the work on this thesis.

I would like to express my sincere gratitude to my supervisors Prof. Hartmut Schlüter and Prof. Mark Molloy for the continuous support during my study, the interesting and motivating discussions during my time in the UKE and at Macquarie University as well as for their knowledge and guidance during my time of research. I could not have imagined better supervisors who made the work on this thesis so much more enjoyable.

I would also like to thank Prof. Christian Betzel for being my co-supervisor at the University of Hamburg as well as my thesis referees Prof. Sasha Rohn, Prof. Udo Schumacher, Dr. Maria Riedner and Dr. Charlotte Uetrecht.

My sincere thanks also goes to PD Dr. Friedrich Buck for all the intriguing discussions and who imparted so much knowledge to me.

I would also like to thank Antje Katschner, Dr. Wang Yi and Dr. Stine von Förster for their support with all things related to my Joint PhD process.

I would like to thank Prof. Tobias Lange for providing me with the mouse samples used in this thesis as well as the discussions of the data.

I thank my fellow labmates Sönke Harder, Marcel Kwiatkowski, Marcus Wurlitzer, Laura Heikaus, Benjamin Dreyer, Dennis Krösser, Lorena Hendl, Christoph Krisp, Karthik Kamath, Daniel Bucio Nobel and Matthew McKay for the fun times inside and outside the laboratory as well as all the support during downtime of the instruments.

Last but not least, I would like to thank my family: my parents, my grandmother and my brother for their spiritual as well as financial support during all my studying up to this point and my life in general. Without you I would have never made it this far.

11 Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, die vorliegende Dissertation selbst verfasst und keine anderen als die angegebenen Hilfsmittel benutzt zu haben. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Ich versichere, dass diese Dissertation nicht in einem früheren Promotionsverfahren eingereicht wurde.

Hamburg, den 10.08.2018