Diving into cellular signalling

Functional proteome analysis by mass spectrometry based approaches

Niels Leijten

The research presented in this thesis was performed in the Biomolecular Mass Spectrometry & Proteomics group at the Bijvoet Center of Biomolecular Research, Utrecht University, The Netherlands.

Copyright © 2021 Niels Leijten, Dordrecht, The Netherlands.
All rights reserved. No part of this publication may be produced or transmitted in any form or by any means, electronical or mechanical, including photocopy, recording, or any information storage or retrieval system, without prior permission from the author.

ISBN/EAN: 978-90-831713-0-2
Author: Niels Leijten

Cover design, layout and printed by: Print Service Ede | proefschriftenprinten.nl

Diving into cellular signalling

Functional proteome analysis by mass spectrometry based approaches

Een duik in cellulaire signalering

Functionele proteome analyse door
massa spectrometrie gebaseerde methodes

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 13 september 2021 des middags te 2.15 uur

door

Niels Marinus Leijten

geboren op 22 augustus 1991
te Dordrecht

Promotoren:
Prof. dr. A.J.R. Heck
Prof. dr. J. den Hertog

Copromotor:
Dr. S.M. Lemeer

Not all those who wander are lost

J.R.R. Tolkien

Table of contents

Chapter 1: Introduction 9
Scope of this thesis 57

Chapter 2: Thermal proteome profiling in zebrafish reveals effects of napabucasin on retinoic acid metabolism 69

Chapter 3: Histidine phosphorylation in human cells; a needle or phantom in the haystack? 97

Chapter 4: The activated variant of the tyrosine phosphatase SHP2 is more susceptible to oxidation compared to the wildtype 133

Chapter 5: Future perspectives and outlook 160
Lay summary of this thesis 168
Lekensamenvatting van dit proefschrift 172
Curriculum vitae 176
List of publications 178
Acknowledgements 180

Chapter 1

Introduction

In biology, several important findings are found by comparing one functional biological state *versus* another[1]. By comparing a perturbed state to the unperturbed so-called wildtype, important molecular players in the underlying processes can be found. Traditionally, these studies were performed using genomic approaches. The term genomics can be described as the analysis of genes and the genome[2]. In this field, individual genes and gene expression patterns are identified which contribute to malignancies such as cancer. Even though a lot of knowledge has been gained by identifying genetic alterations in malignancies, multiple limitations have also surfaced[2,3]. For instance, most cellular functions are carried out by proteins, instead of DNA or RNA, limiting the link between phenotype and gene expression. Additionally, nucleic acid sequences do not contain information about post-translational modifications (PTMs), such as phosphorylation or ubiquitination, which have a major impact on protein function. The genomic sequence also does not detail information about which proteins interact under certain conditions or where this occurs. Lastly, and most importantly, transcript abundance does not necessarily correlate with protein abundance. By performing some experiments on the genomic scale, important insights can be missed, which can be investigated on the protein level. Moreover, some important sample types, such as blood or body fluids, only have proteins making RNA measurements impossible[1]. Additionally, protein abundance in different organelles can be determined to gain more information about protein localization. Taking this all together, it shows that biological differences should be preferably measured on the protein level compared to the gene level to gain more functional insights.

In 1995, Marc Wilkins already dubbed the term proteome, which stands for "the PROTEin complement expressed by a genOME"[4]. The proteome of a cell is a highly structured entity, in which the constituent proteins carry out their function in a temporal and spatial specific manner[5]. In proteomics, the aim is to identify and quantify all proteins, including their expression, localization, PTMs and interactions on a time, space and cell type dependent basis[6]. This makes the study of proteomics way more complex than genomics: there are possibly more than a million proteins encoded by the circa 25000 genes of the human genome[7,8], showcasing the difficulty of finding the explicit function of all of these (Fig. 1).

There are three different types of proteomics approaches which can be used to investigate the proteome, namely "Top-down", "Middle-down" and "Bottom-up" (Fig. 2). In protein centric top-down proteomics, the intact protein is first fractionated using liquid chromatography before being sprayed into the mass spectrometer, where both the intact mass and their fragments can be measured[9]. By measuring the fragments of the protein by performing MS/MS analysis, the exact location of PTMs can be determined[10]. This technique allows 100% sequence coverage and the full characterization of proteoforms, with proteoforms being the specific isoform resulting from combinations of alternative splicing, PTMs and genetic variations. However, downsides include difficulties with protein fractionation, ionization and fragmentation[6]. Additionally, the high abundance of charges and solvent adducts in electrospray ionization (ESI) hinder the efficient detection of high mass proteins. It was shown that top-down proteomics can be used in the field of biomarkers. For example, Zhang *et al*[11] used top-down proteomics for the identification of the phosphorylation rate of cardiac troponin I as a biomarker for chronic heart failure.

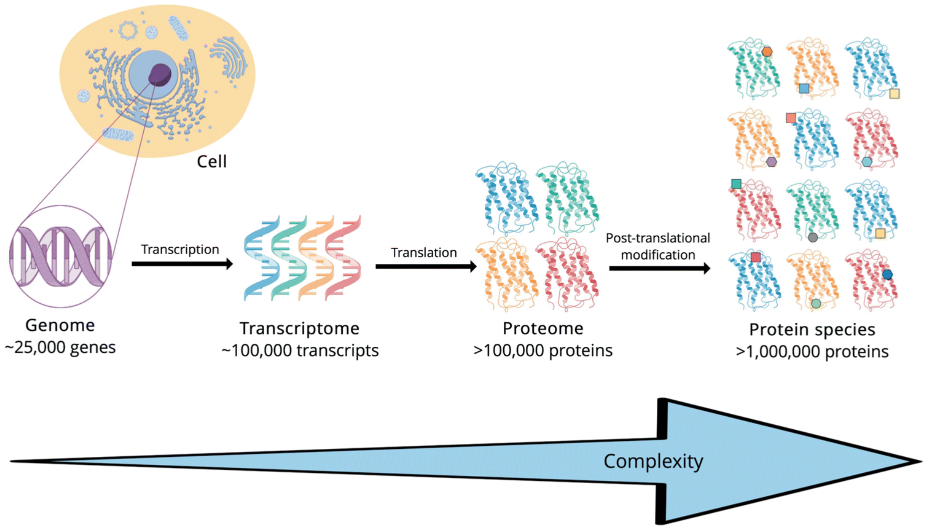


Figure 1: The increasing complexity from genome to proteome. The cellular genome contains circa 25,000 genes. These, when transcribed, will lead to circa 100,000 transcripts, which will lead to more than 1,000,000 proteins after translation. When taking PTMs into account, the complexity is even increased further by tenfold. This increase in complexity between the genome and corresponding proteome makes proteomics a challenging field of studies. Adapted from Virág *et al*[8]

In contrast to top-down proteomics, where the protein complexes are denatured and the mass of their constituents is measured, native mass spectrometry sprays protein complexes directly into the mass spectrometer in their native folded state and measures their mass[12,13]. This allows for the conservation of non-covalent interactions between proteins, with which questions can be answered concerning structure stoichiometry and the arrangement of subcomplexes. Additionally, native

MS can be used to measure multiple other macromolecular assemblies, such as ribonucleoprotein complexes (ribosome) or nucleic acid structures[13]. Lastly, native MS shows a lower distribution of charge states, which aids in the correct identification of mass spectra[9]. Using native MS, Snijder *et al*[14] managed to measure an 18 MDa virus capsid using a QTOF, while Rosati *et al*[15] showed that combining native MS with the orbitrap mass analyzer allowed for the measurement of different glycoforms of monoclonal antibodies. Additionally, they also showed the possibility of measuring non-covalent antibody-antigen interactions using native MS. These examples illustrate the great versatility of native MS.

A step down from top-down and native MS is middle-down proteomics, where proteins are digested into large peptide fragments with (ideally) a mass above, arbitrarily, ~3 kDa[17]. By studying these larger peptides, more information can be gained about isoforms and co-occurring neighboring PTMs. A result from the generation of only larger peptides is a decrease in the number of peptides formed, decreasing complexity and allowing better sequence coverage[18]. Proteases with a single residue target, such as Asp-N and Glu-C, are commonly used in the sample preparation for middle-down proteomics. This technique has already seen some use in literature, for example on the determination of PTMs on histone H4[19]. In this study of Jiang *et al* they studied the role of combinatorial histone PTMs in proteoforms during cell cycle progression of breast carcinoma cells. They showed distinct epigenetic phosphorylation and acetylating patterning dependent on the cell cycle, providing possible diagnostic and prognostic biomarkers. In other research of Valkevich *et al*[20] they investigated the branching of ubiquitin chains, which is important to measure since different lengths employ different biological effects. They showed that the middle-down approach allows the identification of the branching of ubiquitin chains, which would be lost in standard proteomics workflows.

Lastly, and the so far most widespread used, bottom-up proteomics can be utilized to measure the presence of proteins in a complex mixture[6]. When bottom-up analysis is performed on a complex protein mixture, it is also known as shotgun proteomics. Shotgun proteomics is an indirect way of measuring proteins through the peptides which derive from them after proteolytic digestion. In a standard workflow (Fig. 3), these peptides are fractionated after which they are subjected to mass spectrometric analysis. The identity of these peptides is then found by matching their spectra to theoretical spectra from peptides that were generated *in silico*. The identified peptides are then mapped back to protein sequences by computational means to identify the proteins present in the sample. Since some peptides are not unique for certain proteins but can be assigned to multiple, the identified proteins are scored[6].

Shotgun proteomics is the most commonly used method to identify proteins and their PTMs in a sample and all steps of this workflow will be discussed in more detail. Also, the applications which can be studied using shotgun proteomics are discussed.

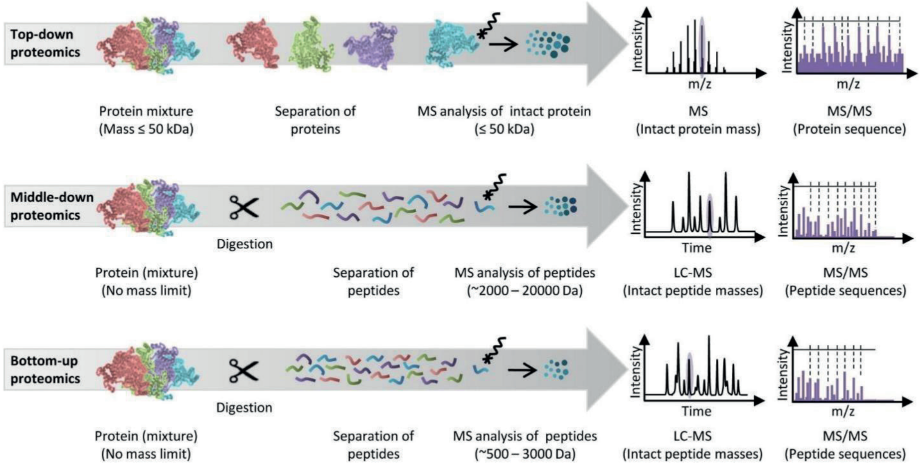
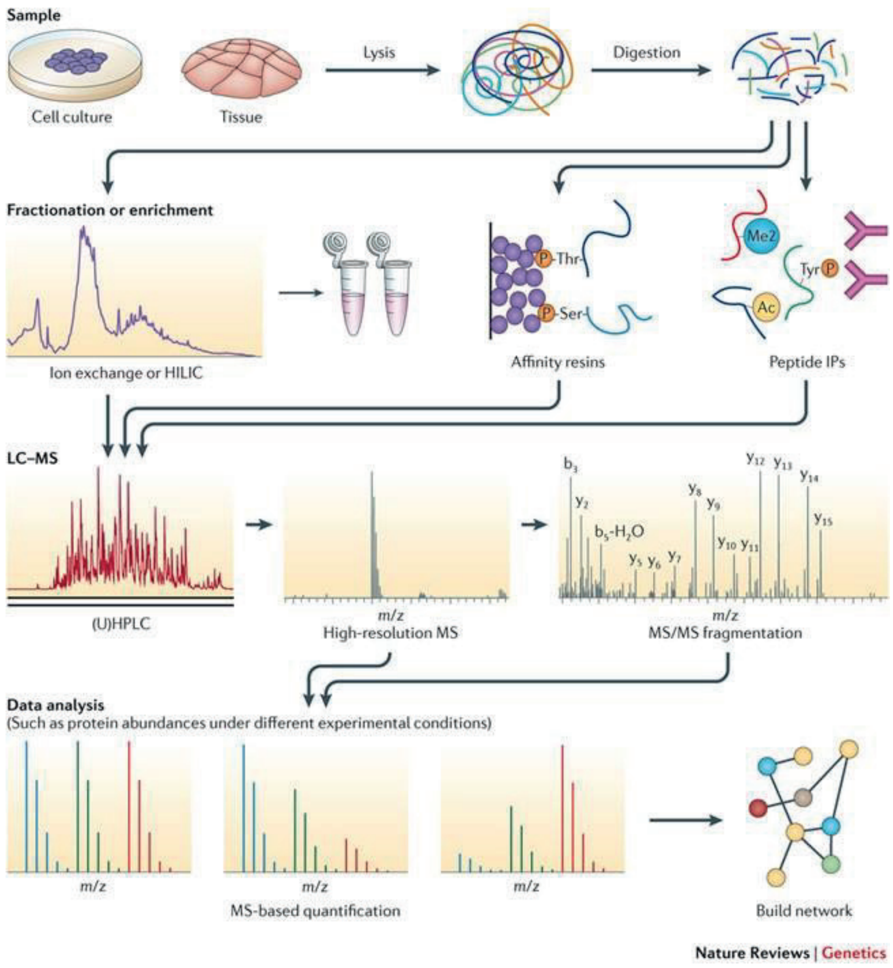

Top-down proteomics
Intensity
m/z
Intensity
m/z
Protein mixture (Mass ≤ 50 kDa)
Separation of proteins
MS analysis of intact protein (≤ 50 kDa)
MS (Intact protein mass)
MS/MS (Protein sequence)
Middle-down proteomics
Digestion
Intensity
Time
Intensity
m/z
Protein (mixture) (No mass limit)
Separation of peptides
MS analysis of peptides (~2000 – 20000 Da)
LC-MS (Intact peptide masses)
MS/MS (Peptide sequences)
Bottom-up proteomics
Digestion
Intensity
Time
Intensity
m/z
Protein (mixture) (No mass limit)
Separation of peptides
MS analysis of peptides (~500 – 3000 Da)
LC-MS (Intact peptide masses)
MS/MS (Peptide sequences)

Figure 2: The different types of mass spectrometry based proteomics. In top-down proteomics, the intact protein is measured, which allows for the determination of the intact mass of the protein but also to achieve 100% sequence coverage. In middle-down proteomics, proteins are digested in peptides with a mass larger than 3 kDa, allowing for the study of proteoforms and co-occurring PTMs. In bottom-up proteomics, the proteins are digested in peptides, which are used to identify the proteins present in the complex cellular lysate and their PTMs. Adapted from Switzar *et al*[16].



Nature Reviews | Genetics

Figure 3: A standard shotgun proteomics workflow. In a standard workflow, either cultured cells or tissue samples are lysed to release their protein content. These proteins are then digested into peptides, which can be fractionated or enriched to study certain aspects of the proteome. The peptide mixture is separated using liquid chromatography, after which the parent mass of the peptide is determined using mass spectrometry. By performing fragmentation studies and comparing the formed fragments to an *in silico* database, the peptide sequence and corresponding protein can be determined. Adapted from Altelaar *et al*[21].

Shotgun proteomics, the basics and applications

Picking an appropriate sample

The first step in a shotgun proteomics experiment is the choice of an appropriate sample. In theory, every organism, tissue or cell containing proteins is amendable for proteomic analysis. In recent research, almost every cell and tissue type present in the human body has been profiled using shotgun proteomics. This effort has been catalogued in the Human Protein Atlas[22], where they collected all the proteomics data on all cell and tissue types and categorized them. At the time of publication, they had protein evidence for 18097 out of the 19628 annotated human genes (Swiss-Prot), showing the magnitude of this resource. They showed a total proteome coverage of circa 16000-17000 proteins, of which 10000-12000 form the so-called core proteome. These core proteins appear in all cell types and take care of the general control and maintenance of cells. Interestingly, they found that protein expression in cell types belonging to a certain tissue are broadly similar, while there are more abundant differences between organs. Additionally, they found that proteins in a certain class, such as kinases, are differentially distributed. Most of them are broadly expressed, but some of them only appear in specialized tissues[22]. Altogether, this shows the importance of taking the appropriate cell type for your experiment. A big field which commonly employs proteomic analysis on cell lines is cancer research. Using proteomics, one has the potential to identify protein markers for early detection, classification and prognosis of cancer[23]. By globally profiling the abundance of proteins between distinct conditions, such as healthy versus cancer cells or before and after drug treatment, one can find proteins that potentially play a role in this cancer type.

In addition to measuring the proteome of human cell lines, one can also perform shotgun analysis on whole organisms, including the model systems *Escherichia coli* (*E. coli*), yeast (*Saccharomyces cerevisiae*), zebrafish (*Danio rerio*) and mouse (*Mus musculus*). The genome of model organisms has often been completely annotated which makes them amendable for experimental studies in the field of heredity, development, physiology and the underlying cellular processes[24,25]. There is a model organism for every biological question: simple organisms (bacteria and yeast) allow for the investigation of conserved biological processes, while more complex organisms (vertebrate zebrafish) allows for the study of more complex traits due to mutant strains. This way, one can study disease models of human physiology in healthy and diseased states, but also physiological processes such as aging[25]. In the past, model organisms were mainly studied using genomics approaches, but nowadays advances in proteomics workflows allow for the measurement of all proteins in the model organism under differential conditions. This has already been widely done, as shown by Gouw *et al*[25]. Here, they gave an overview of model organisms, ranging from simple bacteria to mammals such as mouse and rat, being metabolically labeled to elucidate biological processes. Hebert *et al*[26] showed the possibility of comprehensively measuring the whole yeast proteome in an hour, while Huttlin *et al*[27] measured the proteome and phosphoproteome of multiple mouse tissues, revealing a tissue specific phosphorylation pattern which tunes protein activity. This is only the tip of the iceberg, with more studies being performed on model organisms every year.

Next, I will look into more detail into the zebrafish as model organism, since this is used in the research performed in this thesis. The zebrafish is well suited as a model organism, since its genome is very similar to the human genome and, more importantly, orthologues of 84% of the human disease associated genes have been found[27]. The embryonic development of zebrafish is fast, with the body axis being established 24 hours post fertilization (hpf) and most organs being fully developed 96 hpf[28] (Fig. 4). An added benefit is that zebrafish embryos develop *ex utero*, allowing them to be easily manipulated by microinjection. Since the embryos of zebrafish are transparent, it allows for the monitoring of phenotypic changes after perturbation. Additionally, the transparent character allows for the visualization of genetic expression using fluorescent reporters[29]. Lastly, numerous mutant strains of zebrafish exist, with knock-outs of specific genes, which allows for a tailored model organism for disease states. Taken together, this shows the very big potential of using zebrafish as a model organism in multiple research fields. For example, zebrafish has already seen use in the fields of immunology[30], cancer[31] and toxicology[32].

Even though the zebrafish is widely used as a model organism in the field of genomics, it is also studied in the field of proteomics. Studies range from performing proteomics on zebrafish embryos until analysis on adult zebrafish tissues. For example, Lößner *et al*[35] performed multiple fractionation methods to extend the proteomic coverage of zebrafish embryos. They identified a total of 3464 proteins (with at least 2 peptides) and showed that SDS-PAGE can be used for fractionation. More recently, Purushothaman *et al*[36] also screened the proteome of zebrafish embryos. They optimized the removal of the egg yolk protein vitellogenin, which can greatly hamper identification of co-occurring less abundant proteins. This new deyolking procedure allowed them to screen the embryo starting from the 1-cell stage and increased the number of identified proteins by 3-4 fold to 2575. Lastly, Shrader *et al*[37] investigated the influence of estrogen mimics on protein expression in zebrafish, which they showed to be distinct. In addition to performing shotgun proteomics experiments on zebrafish embryos, adult fish can also be used. For instance, Saxena *et al*[38] studied caudal fin regeneration in adult zebrafish to gain more knowledge about the biomolecular environment in wound repair. Additionally, De Souza *et al*[39] generated a map of the gill proteome of the zebrafish, which might be used in further toxicological studies. As can be seen, zebrafish is already being used widely in proteomics studies and this interest will probably be expanding further in the future.

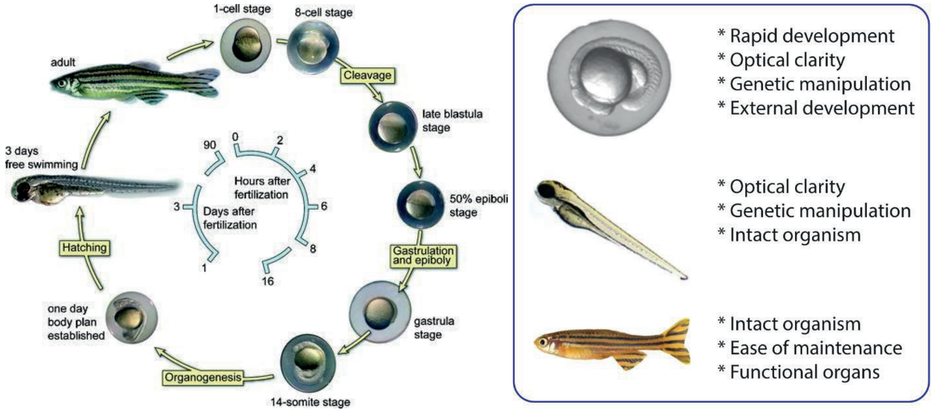


Figure 4: The zebrafish is very suited for proteomics studies. Zebrafish embryos develop fast, are transparent and already form organs after 3 days[33]. These characteristics make it very suited for experimental studies. Every stage of zebrafish embryonal development has it perks. For example, in the egg stage the genes are easy to manipulate due to external development. In the larval state, the zebrafish is an intact organism, allowing for the study of many tissue types. Lastly, the fish stage allows for the study of functioning organs and tissue types. Figures adapted from Willemsen *et al*[33] (left) and Saleem *et al*[34] (right).

Lysis and sample preparation

After the appropriate sample type has been chosen, the organism or cell needs to be lysed to release the proteins. Lysis is the disruption of the outer or cellular membrane of an organism or cells to release the intra-cellular materials such as proteins and DNA[40]. The choice of lysis methods depends greatly on the organism under study, for example eukaryotic versus prokaryotic systems, and the analytes which you want to release. Methods can be categorized as mechanical or non-mechanical lysis methods[40]. In mechanical lysis, the cellular membrane is broken up by using shearing force. Examples of these methods include bead beating and dounce homogenizing. In the bead beating approach, cells are disrupted through collisions with small glass or ceramic beads, which causes the cell membrane to break open through shear forces and release the intra-cellular proteins. This technique is mainly suited for the lysis of more robust organisms, such as yeast and bacteria[40]. In this thesis, we show that bead beating is also a requisite for the efficient lysis of zebrafish embryos. This technique has a very high efficiency, but also heats up the sample which may cause protein degradation. A different technique which can be used to mechanically lyse cells is the dounce homogenizer[41]. In the dounce homogenizer, the sample is pushed between an outer glass wall and a pestle, causing lysis by shearing forces. The homogenizer can be used to lyse tissue and cells. When a certain size pestle is chosen, it even allows for the isolation of organelles such as the nucleus. The homogenizer is ideal for mild lysis.

Besides using mechanical forces to lyse cells, also non-mechanical techniques can be used. These fall into three categories, namely physical, chemical and biological[40]. An example of a physical lysis technique is freeze-thaw cycling. In this lysis method, the cells are flash frozen, causing the formation of ice crystals on the cell membrane, which will subsequently tear down the membrane. By repeating this process numerous times, the cells are lysed in a mild fashion keeping most proteins intact. Another commonly used method to lyse cells is to utilize chemical disruption. In this methodology, lysis buffers are used which disrupt the cell membrane by using a detergent. The detergent will disrupt the hydrophilic and hydrophobic interactions of the cellular membrane, disintegrating it and causing the release of the internal proteins. Most commonly, mammalian cells are disrupted by using lysis buffers. Some examples of commonly used detergents include Triton X, NP-40, Tween and sodiumdeoxycholate (SDC).

In contrast to performing proteomics on the whole cell, one can also perform analysis on specific organelles of the cell. This process of subcellular fractionation can be used to generate proteomics data on one specific organelle, which allows for a better understanding of processes occurring within[42]. Additionally, it is very suited to detecting low abundant proteins that are specific for a certain organelle, such as kinases or phosphatases. The process of subcellular fractionation entails two steps: first the cellular membrane is disrupted, after which the released organelles are fractionated based on their physical properties. Examples of fractionation techniques are differential and density-gradient centrifugation, but a big disadvantage of these technique is the presence of co-purifying contaminants which can complicate analysis[43]. Subcellular fractionation has already seen widespread use to study the organellar proteome of, for example, nucleolus[44] and mitochondria[45].

Zooming in even further, it is also possible to extract individual proteins after lysis using affinity purification to better study single proteins or complexes. In affinity purification, the desired biological material is captured via a specific enrichment with a ligand which is coupled to a solid support[46]. Multiple ligands can be used, ranging from oligonucleotides to peptides or proteins, but most commonly protein-specific antibodies are used. When antibodies are used, the method is also referred to as immunoprecipitation. The purification using antibodies has many advantages: it captures the protein in its native state and has the ability to capture all isoforms of the protein of interest. However, for this technique to work an antibody specific for the protein target needs to be found and developed, which is an expensive and challenging aspect[46]. Also, the antibody needs to have enough affinity to the target and have minimal interaction with non-specific interactors. Once the desired antibody is found, it is bound to the solid support, which is used to pull down the protein (complex) of interest in solution. The support is then washed to remove non-specific binders and the protein of interest is eluted and measured using mass spectrometry. Multiple elution methods can be used, which need to be optimized for each experiment. Additionally, small epitope tags can also be used for immunoprecipitation approaches. Here, a protein tag is introduced into the target protein using genetic approaches, which allows it to be selectively enriched. An example of this methodology is the widespread His-Tag. As can be seen, affinity purification has a large potential in mass spectrometry, but needs many optimization steps.

The next step in the bottom-up proteomics workflow is to proteolytically digest the proteins into peptides. As stated before, the proteins are digested into peptides to circumvent some issues associated with intact protein mass spectrometry[6]. By digesting the proteins into peptides their biochemical heterogeneity is normalized to create a less heterogeneous mixture. Additionally, by generating multiple peptides per protein, there is a bigger chance that low abundant proteins are detected in the mass spectrometer. Proteins are digested using proteases, which perform cleavage through hydrolysis of the amide bond before or after a specific amino acid residue or residues. There are many different proteases which can be used for the digestion of proteins, but most commonly trypsin is used. Trypsin is a serine protease which selectively cleaves after arginine and lysine residues[6]. It has multiple characteristics which have made it the golden standard for digestion in shotgun proteomics: it is a very efficient and specific protease which is widely available at a low cost[47]. Additionally, since it generates short peptides with an arginine or lysine residue at the C-terminus, it is very amendable to current chromatographic separation, fragmentation and search algorithm techniques. The presence of arginine or lysine also causes the presence of two protonatable sites, the amino group at the N terminus and the C terminal basic amino acid, which enhances the generation of ions in the gas phase and their fragmentation[48]. However, trypsin also has some apparent disadvantages. Trypsin generates relatively short peptides, with 56% under 6 amino acids long, which are generally too small for mass spectrometric analysis, causing them to not be identified[49]. Additionally, when a negatively charged amino acid, such as aspartic acid or a phosphorylated serine, is in close proximity to the lysine or arginine, it will lead to missed cleavages and longer peptides[50]. Therefore, multiple other proteases have been introduced in proteomics, with different cleavage specificity, characteristics and (dis)advantages[50]. An overview of these proteases is given in table 1. Besides using only one protease to digest proteins, it has also been shown that the use of multiple proteases in parallel can increase the number of protein identifications in proteomics[49] and phosphoproteomics[51] experiments. However, in the scope of this thesis, trypsin and Lys-C are the only used proteases.

Table 1: Commonly used proteases in proteomics and their mode of action[47].

Protein name	Commercial name	Specificity	Optimal pH
Trypsin	Trypsin	C: K, R	7.5
Chymotrypsinogen A	Chymotrypsin	C: Y, F, W	7.8 – 8
Lysyl endopeptidase	Lys-C	C: K	7 – 9
Glutamyl peptidase I	Glu-C	C: E, D	4 or 8
Peptidyl-Asp metallopeptidase	Asp-N	N: D	4 – 9
Peptidyl-Lys metalloendopeptidase	Lys-N	N: K	9.5
Pepsin A	Pepsin	C: Y, F, W, L	1 – 4

After the correct protease has been picked, different methods of digestion can be chosen: in-solution, in-gel or filter based. Traditionally, the proteins were digested using in-gel digestion methods. In this methodology, which was already introduced in 1996, the complex protein mixture is first separated by SDS-PAGE electrophoresis before being digested in the gel[6,52]. This prefractionation step allows the high capacity of yielding single protein species per band[53]. The bands corresponding to certain proteins in the gel are then selectively digested, measured and identified using mass spectrometric methods. This method has multiple advantages: the fractionation of the proteome over multiple bands decreases the complexity of the sample. Additionally, small molecule interferents that can hamper mass spectrometric analysis, such as urea and detergents, are readily removed. However, this technique also has some apparent downsides. Since the protease needs to diffuse through the gel to reach the protein, a large amount of enzyme is needed for efficient digestion[52]. Also, the auto-digestion of proteases can lead to a large background signal. Lastly, the handling of the gel will commonly lead to the contamination of samples with human proteins such as keratin. Therefore, in recent times, in-solution digestion methods have seen more use.

During in-solution digestion, the proteins are kept in solution, where they are sequentially denatured, reduced, alkylated and digested. The first step is to denature the protein to destabilize higher order protein structures and to improve the access of the proteolytic enzyme to the protein backbone[48]. Most commonly chaotropes or surfactants are employed to denature the protein. A chaotropic agent, of which urea is the most commonly used, destabilizes the protein structure by disrupting hydrogen bonding in protein complexes. Urea is most commonly used, since it is compatible with proteases and is easily cleaned up afterwards. A surfactant which is commonly used to denature proteins is SDC, which can afterwards be easily cleaned up using acidification of the sample. It was shown that SDC produced the highest average digestion efficiency, which was 80%[54]. An interaction which needs to be broken to efficiently denature proteins are disulfide bridges. Disulfide bridges are covalent interactions between cysteine residues, which help maintain the folded state of the protein[54]. It is paramount to break these bonds to allow for the most optimal digestion efficiency. Most commonly reducing reagents, such as dithiothreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP), are used to break the disulfide bonds. However, the free sulfhydryl groups are very reactive, allowing them to spontaneously oxidize with other sulfhydryl groups to reform disulfide bridges. Therefore, the free sulfhydryl group needs to be blocked rapidly by alkylation to prevent this interaction. Most commonly chloroacetamide is used to irreversibly alkylate cysteines. The combination of denaturation, reduction and alkylation makes sure that the subsequent digestion step is performed as efficiently as possible, causing the generation of as many peptides as possible. In-solution digestion is more commonly picked over in-gel digestion, since it allows for optimization of numerous conditions (buffers, additives and pH) and the recovery of the digest is more reliable[54]. The use of salts as buffers and numerous additives warrants the use of a clean-up step, since these can interfere with ionization efficiency and sensitivity in the subsequent mass-spectrometric analysis. Therefore, a desalting step is needed to remove these after in-solution digestion. Solid phase extraction is the most commonly used technique to desalt samples. In solid phase extraction, the peptides of interest are bound to a hydrophobic resin while small molecules and salts do not bind to the resin or have such weak interactions that they can be washed off[54]. By using organic solvents, the peptides can be washed off, free of salts and ready for measurements.

A final method which can be used to digest proteins is by using filter assisted methods. A big advantage of these methods is that harsh detergents can be used, for example sodium dodecyl sulfate (SDS), to help detect hydrophobic proteins such as membrane proteins, which is then removed before MS. Examples of this technique include Filter-aided sample preparation (FASP) and the S-trap[54]. In FASP, SDS containing samples are trapped on the filter after which the SDS micelles are disrupted through the use of urea. The filters are then washed to remove the urea and the proteins are digested in the filter. On the other hand, in the S-trap methodology the SDS containing samples are transformed in a fine protein particulate suspension, which is then trapped on a filter. The SDS is washed away and the proteins are digested on the filter. It was shown that these filter based methods have a higher trypsin efficiency, since the local concentration of protease is higher compared to in-solution methods[54]. However, these methods can be quite time consuming.

Decreasing the proteome complexity through fractionation

An inherent downside to proteomics is the dynamic range of the proteome; the protein abundance inside of the cell spans at least seven orders of magnitude, from one copy to ten million copies per cell[54]. When comparing this number to the dynamic range of the mass spectrometer, which can perform analysis up to four orders of magnitude, the mismatch between instrument and biology is apparent. Even more striking is that after digestion, the resulting peptides have a dynamic range which is even an order of magnitude larger than the corresponding proteins. This limits the detection of lower abundant proteins, which will be overshadowed by more abundant species[54]. A solution was found by decreasing the dynamic range of the peptides before mass-spectrometric analysis by peptide fractionation techniques such as strong cation exchange (SCX) and high pH fractionation. In these fractionation techniques, a physical principle orthogonal to hydrophobicity, which is employed in reversed phase liquid chromatography, is used to better resolve peptides. The principles of size, charge and biological interaction can be used. It is important that the two techniques used are orthogonal to allow for the most optimal separation. Most commonly these fractionation techniques are employed in an offline way, which means that the technique is not directly coupled to the mass spectrometer. This allows for better optimization, the use of MS incompatible buffers (for example salts), the choice of the amount of input material and the combination of separation techniques which cannot be directly interfaced[6].

SCX, which is an ion-exchange chromatography method, separates peptides according to their charge. The mechanism of separation employs electrostatic interactions between the peptide and the functional groups of the stationary phase, which have opposite charges. In SCX, the negative groups of the stationary phase will attract the positively charged peptides at acidic pH. The peptides will subsequently be released from the stationary phase by the use of a mobile phase with increasing salt concentration: with the peptides with the least amount of positive charge eluting first. The peptides are collected into fractions, which can then be concatenated as seen fit. However, SCX faces a bottleneck in shotgun proteomics applications. After tryptic digestion peptides will have two positive charges in acidic conditions, on the N-terminal amine and the C-terminal lysine or arginine. Some peptides may have more charges due to histidines or internal lysines or arginines due to miscleavages. This large presence of 2+ and 3+ charged peptides might elute at the same time in SCX fractionation, limiting the resolution of this technique. However, it has been shown that most ion exchangers also exert some hydrophobic influence on the peptides, which partially explains the resolution of peptides with the same charge in SCX[55]. It was shown that SCX is very suited for the enrichment of cross-linked peptides, due to their higher charge state compared to unmodified peptides[56].

A different fractionation method which is commonly used in proteomics is high pH fractionation. In high pH fractionation, reversed phase liquid chromatography (RPLC) is used to fractionate the sample under high pH buffer conditions[57]. RPLC has traditionally been the method of choice of online separation before mass spectrometric analysis, but it has been shown that when RPLC is operated at different pH (e.g. 3 and 10) it provides the same separation orthogonality as SCX-RPLC[57]. This orthogonality stems from the fact that the charge distribution of peptides changes under high or low pH conditions. Therefore, by using a high pH offline RPLC fractionation step before LC-MS analysis, an efficient fractionation can occur. There are numerous advantages in the use of high pH fractionation compared to SCX; the technique achieves better resolution and higher peak capacities due to faster chromatographic partitioning. Additionally, the use of salt-free buffers negates the need of a clean-up step after fractionation. The separation can even be further improved by using different materials for the stationary phases in both RPLC steps. It has been shown that the process of concatenation further increases the performance of fractionation[57]. Concatenation entails the combination of multiple early, middle and late fractions, which elute over equal time intervals and with little overlap, into one new combined fraction. This process can be seen in figure 5. Concatenation of fractions has three advantages: I) it decreases the amount of fractions needed to be run in the LC-MS analysis, II) it compensates for the imperfect orthogonality of the two techniques and III) makes more use of the whole gradient during LC-MS analysis compared to running 1 fraction. Concatenation is very valuable, since it was shown that concatenation after high pH fractionation increased the peptide identification by 80% compared to SCX-RPLC[57]. In addition, research of Batth *et al*[58] showed that high pH fractionation is complementary with phosphoproteomics strategies, even more so than traditional SCX approaches. These results clearly show the added benefit of performing offline fractionation to increase the coverage of the proteome.

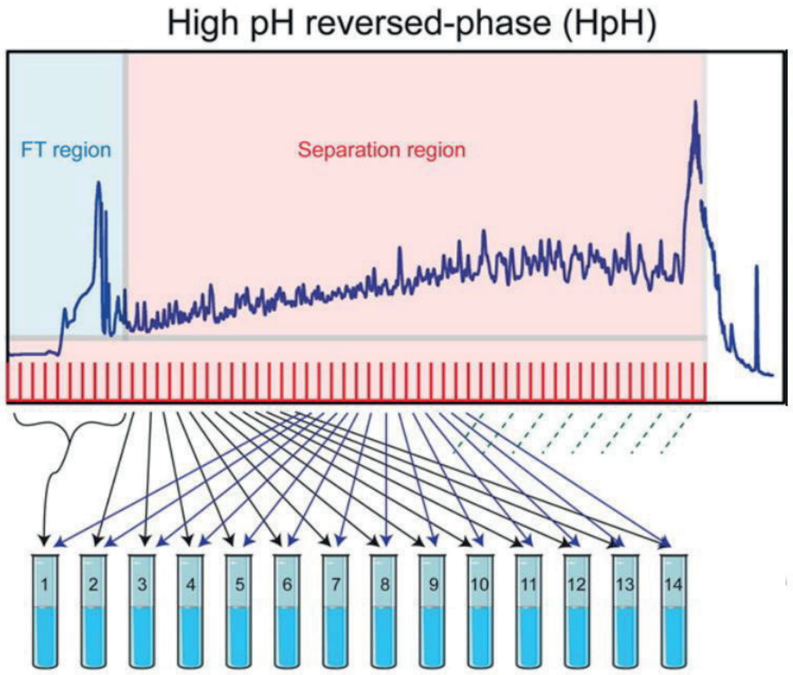


Figure 5: The principle of concatenation in high pH fractionation. By combining fractions ranging from different parts of the high pH chromatographic run, peptide coverage and protein identifications can be increased. Figure adapted from Batth *et al*[58].

Liquid chromatography mass spectrometry for the investigation of the proteome

After sample preparation, the generated peptides will be further fractionated by reversed phase liquid chromatography (RPLC) to decrease complexity and to allow the maximal number of peptides to be identified per MS run. RPLC employs the difference in hydrophobicity of peptides to separate them. Peptides are first bound to the hydrophobic stationary phase under a polar mobile phase[59]. By then increasing the hydrophobicity of the mobile phase, peptides will be sequentially eluted, with the more polar peptides eluting first and the most hydrophobic peptides eluting last. For the stationary phase, most commonly octadecyl carbon groups (C18) on a silica support in a column format are chosen. For the mobile phase, a mixture of water and acetonitrile is chosen, which can be altered in its hydrophobicity through increasing the percentage of acetonitrile. It was shown that a gradient of 13 – 32% acetonitrile is ideal for peptide separation[60]. RPLC offers multiple advantages, such as high-resolution separation capacity and the use of mobile phases compatible with subsequent electrospray ionization.

Multiple parameters can be optimized to improve the resolving power of RPLC[61]. First, it has been shown that longer columns have a better separation efficiency due to an increase in theoretical plates. However, a downside to this is the longer analysis time and higher backpressure generated. Secondly, a smaller inner diameter of the column also increases efficiency. Lastly, by decreasing the size of the particles of the stationary phase resolution will be further increased, again at the cost of a higher back pressure. When combining all these optimized parameters, which greatly increases the backpressure, one speaks of ultra-high-pressure liquid chromatography (UHPLC). Currently, almost all online RPLC fractionation occur at high pressures (maximum 1000 bar), nano-flow (200 nl/min) and small particle sizes (smaller than 2 µm diameter) to achieve optimal separation.

After optimal separation using RPLC, the peptides are measured using mass spectrometry. In mass spectrometry (MS), one measures the mass to charge (m/z) ratio of ions in the gas phase. MS can be used to measure the mass of a peptide or protein, but can also be used to measure the amino acid composition or the location of PTMs[62]. In principle, a mass spectrometer is made up of an ion source that brings ions into the gas phase, a mass analyzer which determines the m/z of the analytes and a detector to measure the abundance of ions at a certain m/z range[63].

The first step in mass spectrometry is to ionize the analytes and put them in the gas phase. In the past, it was only possible to ionize small and thermostable compounds into the gas phase due to fragmentation of the analyte[62]. However, this changed with the introduction of electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI). Using these soft ionization techniques, it is possible to ionize biomolecules without extensive fragmentation.

In ESI, which was introduced in 1989 by Fenn *et al*[64], analytes are ionized directly from the liquid phase, allowing it to be coupled to chromatographic separation techniques[65]. ESI has low chemical specificity, releases very stable ions and has an unlimited mass range. These factors make ESI the most commonly used ionization technique used in proteomics. ESI starts when analytes are pumped through the high voltage emitter, where positive charges are placed on the molecules[66]. These positively charged analytes are repulsed by the positively charged emitter, causing them to be transported to the liquid surface at the end. Under the presence of many charges, the surface will be destabilized and subsequently will form a cone. This is called the Taylor cone, which will start to eject fine charged droplets, which will be pulled towards the counter electrode at the transfer capillary of the MS. This process is schematically shown in figure 6. The protons in the droplet are positioned at the edge to minimize potential energy. However, due to evaporation, the droplet will decrease in size until the surface tension can no longer contain the Coulomb force of repulsion, which is called the Rayleigh limit, causing disintegration into a stream of smaller droplets. This process of disintegration is called Coulomb fission. The process of Coulomb fission continues until the droplet contains only one charged ion[66], which will then be entered into the MS. ESI was shown to generate ions of multiple charge states, which depends on the amino acid sequence and size of the peptide.

A different soft ionization method which can be used to ionize peptides is MALDI, which ionizes analytes from the solid to the gaseous phase[67]. In MALDI, the analyte of interest is mixed with an appropriate matrix, which is then spotted on a MALDI plate and allowed to dry. This process allows the analyte to cocrystallize with the matrix. The analytes are then brought into the gas phase by use of photons generated by a (nitrogen) laser, which are absorbed by the matrix and subsequently cause desorption and ionization of the analyte. The ionized analytes are then transferred towards the mass analyzer. The choice of laser energy and matrix greatly affect the results[68]. A big advantage of MALDI is the robustness related to salts and detergents, which may have a detrimental effect on the analysis when ESI is used. Additionally, MALDI mostly generates singly charged species, which can be useful to determine the molecular ion of proteins, carbohydrates and lipids[67]. Still, as the coupling of ESI with LC is much more straightforward than that of MALDI with LC, ESI-based LC-MS has become the most widespread used method in proteomics, and is also used throughout the work described in this thesis.

After the ions have been ionized and transported to the MS, the m/z value of them is read out using mass analyzers. The first mass analyzer was already developed in 1897 by J.J Thomson[69]. In this early research, he used gas discharge tubes to generate ions, which were subsequently passed though parallel magnetic fields. These ions were then deflected in parabolic trajectories and detected using photographic plates. Using this technique, he was able to measure charged atoms. Since then, numerous other mass analyzers have been developed, but the quadrupole, linear ion trap, time of flight (TOF) and orbitrap mass analyzers are most commonly used. An overview of characteristics of all mass analyzers can be seen in table 2.

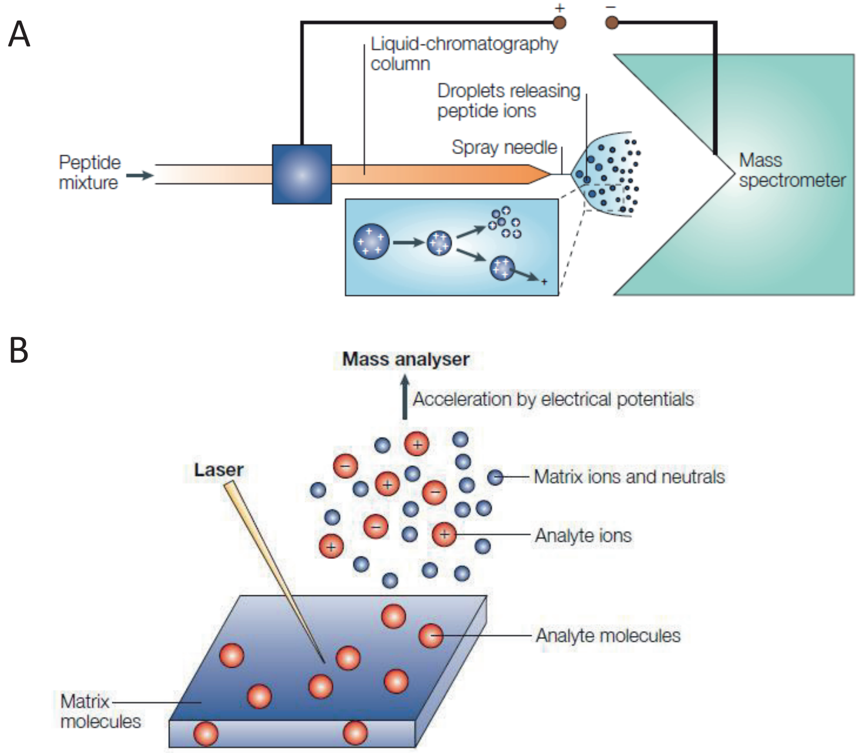


Figure 6: The mechanisms of electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI). In ESI, ions are directly ionized from the liquid phase and transferred to the gas phase, which allows them to be measured using mass spectrometry (A). In MALDI, the ions are directly transferred from the solid phase to the gaseous phase by using photons, generated by a (nitrogen) laser (B). Adapted from Steen *et al*[70].

Table 2: The characteristics of commonly used mass analyzers[6]. ‰ stands for parts per thousands, dynamic range is given in orders of magnitude.

Analyzer	Type	Resolution	Mass accuracy	Dynamic range
Quadrupole	Beam	1 -2 K	~1 ‰	5 - 6
Ion trap	Trapping	1 -2 K	~1 ‰	3 – 4
Time of flight	Beam	10 – 50 K	5 – 10 ppm	4
Orbitrap	Trapping	7.5 – 240 K	10 – 500 ppm	4

The quadrupole mass analyzer consists of four parallel rods: on two of these rods a direct current potential is applied while on the other two rods an alternating rf potential is placed[67]. When the positively charged ions are transferred into the quadrupole, they will drift towards the negatively charged rod. Once the rods switch their polarity, the positively charged ion will switch its trajectory and drift towards the now negatively charged rod. By optimizing the switching of polarity, some ions can be successfully transferred through the quadrupole while others will crash into the rods. This allows quadrupoles to be used as selection tool but also as mass analyzer. Advantages of quadrupoles include low cost, small size and robustness. However, they do have limits in mass range, resolving power and no MS/MS can be performed. To combat the last disadvantage, quadrupoles are commonly used in so called hybrid instruments[67].

The linear ion trap (LIT) consists of four rods ending in lenses which maintain the ions between the rods[71]. The ions will be confined radially by a 2D radio frequency field and at the same time axially by stopping potentials applied to the end electrodes[72]. The ions can either be axially or radially ejected, after which they will be detected using the detector. These LITs have multiple advantages over the earlier developed 3D ion traps, namely reduced space charging effects and higher sensitivities due to higher trapping efficiencies[73]. Even though LITs can be used as stand-alone mass spectrometers, they are more routinely used in hybrid instruments such as the LIT-orbitrap mass spectrometer[74].

A different mass analyzer is the time of flight (TOF) mass analyzer, which uses a different mechanism compared to the trapping of the LIT. TOF is especially useful in combination with pulsed ion sources such as MALDI[75]. In TOF, ions are separated according to their velocity in a free field region called the flight tube[71]. Ions enter the mass analyzer, either through ESI or MALDI, and go through the acceleration region. In this acceleration region, they are accelerated through a difference in potential between an electrode and the extraction grid. All ions will have the same kinetic energy when leaving the acceleration region, after which they will be separated according to their mass in the field free flight tube. Lighter ions will move faster through the flight tube compared to heavier ions. The m/z value is determined by the time it takes for the ion to reach the detector at the end of the flight tube. Big advantages of TOF-based analyzers are the high transmission efficiency and the lack of an upper mass limit[71]. Additionally, a reflectron can be employed to increase the mass resolution[75].

Even though these mass analyzers have seen a lot of use in the past, the most commonly used mass analyzer in the current era of mass spectrometry is the orbitrap. The groundwork for the orbitrap mass analyzer was already put down in 2000 by Alexander Makarov[76]. The analyzer consists of cup-shaped outer electrodes, which are electrically isolated by a thin gap secured by a dielectric central ring. A central spindle-like electrode holds the trap together, as shown in figure 7[77]. When a voltage is applied between the electrodes, an electric field will be formed which is linear along the axis and therefore oscillations along this axis will be pure harmonic. Simultaneously, the radial field will strongly attract ions to the central electrode. The ions are injected into the orbitrap along a tangent through a small slot in the outer electrodes containing a compensation electrode. Through the voltage applied between the outer and central electrode, the ions trajectory will be bend by the radial electric field towards the central electrode while tangential velocity will create an opposing centrifugal force[77]. When the correct parameters are chosen, the ion will remain in a nearly circular spiral along the central axis, like a planet in a solar system. While this is happening, the axial electric field formed by the conical shape of the electrodes will push the ions towards the widest part of the trap initiating harmonic axial oscillations. These outer electrodes will capture the image current of these axial oscillations, which will be Fourier transformed to the frequency domain and converted into a mass spectrum[76]. The orbitrap has numerous advantages when compared to other analyzers: it has a higher sensitivity and dynamic range compared to TOF instruments and has a higher mass accuracy than ion traps[78]. In order to make a pulsed mass analyzer (orbitrap) amendable to a continuous ion source (ESI), the C trap was developed (Fig. 7). The C trap is a rf-only bent quadrupole, in which ions are stored and subsequently transferred to the orbitrap through a high voltage pulse[74,77]. The C trap allows for fast and uniform extraction for large ion populations, making it amendable for proteomics experiments. In all the research performed in this paper, an orbitrap mass analyzer is employed.

A downside to the use of an orbitrap mass analyzer is the strong acceleration used for the injections of ions, leading to a large kinetic energy spread when ions will be fragmented within the analyzer. This greatly compromises the possibility of performing MS^n analysis in the orbitrap[77]. For this reason, the orbitrap mass analyzer is rarely used in isolation but more so coupled to a source of fragmented ions in a hybrid instrument setup. Examples of these hybrid instruments, used also in the work described here in this thesis, are the Q-Exactive and the orbitrap-Fusion line of instruments. For the Q-Exactive line of instruments, the orbitrap is combined with a mass filtering quadrupole, which can be used to transfer all masses or just select one precursor mass for MS/MS analysis[78,79]. Meanwhile, the orbitrap is used for detection of full scans and MS/MS spectra, which are generated in a HCD cell. The schematic can be seen in figure 8.

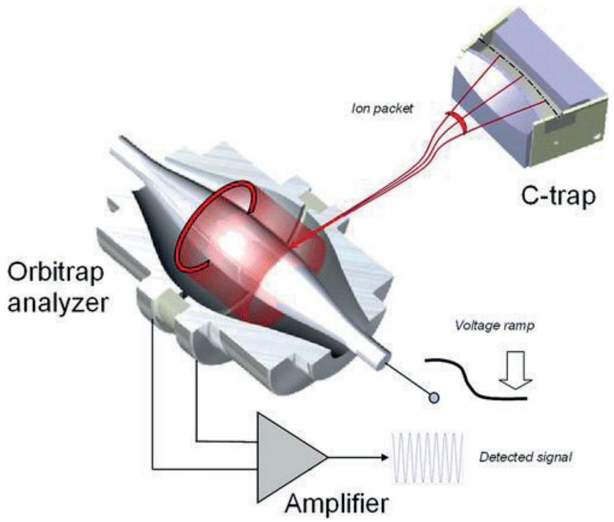


Figure 7: The global schematic of the orbitrap mass analyzer. Ions are first stored in the C-trap, after which they are transferred to the orbitrap. In the orbitrap, ions will move in harmonic axial oscillations around the central spindle. The outer electrodes of the mass analyzer will capture the image current of these axial oscillations. This frequency domain signal is Fourier transformed and converted to mass spectra. Adapted from Zubarev *et al*[77].

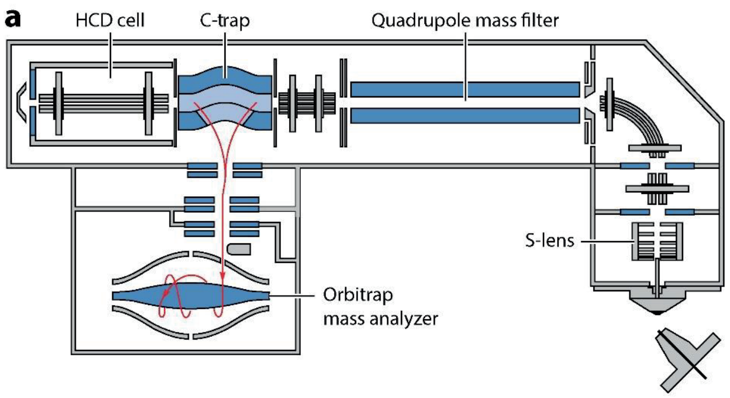


Figure 8: Schematic of the Q-Exactive mass spectrometer. In this hybrid instrument, the quadrupole can filter for specific masses, while the HCD cell is used to fragment ions. The orbitrap mass analyzer is used to measure the corresponding full scan and MS/MS scans. Adapted from Eliuk *et al*[78].

Using fragmentation studies to identify peptides

The previously mentioned mass analyzers are well suited to measure the nominal mass and abundance of peptides using full MS scans. However, in order to get sequence information tandem MS, or MS/MS, needs to be performed. In MS/MS, the peptide ion of interest is mass-selected and fragmented, after which the formed fragments are measured in the mass analyzer[70]. The mass spectrometer will continuously cycle through performing MS scans to determine the precursor mass, followed by MS/MS scans to fragment and determine the sequence of this precursor.

During fragmentation, different methods will cause the formation of different fragments. Traditionally, collision induced dissociation (CID) was used to generate fragment ions. In CID, the peptide ions are subjected to collisions with an inert gas (such as nitrogen or argon), after which the corresponding kinetic energy is converted into internal vibrational energy. Once this vibrational energy reaches a certain threshold, covalent bonds will break[80]. Most commonly the weakest bond will break, which is the amide bond between amino acids. These formed fragments are called b-ions if the charge is maintained on the amino terminal part of the peptide and are called y-ions if the charge remains on the carboxy-terminal part (Fig. 9)[70]. These ions form a series, with each fragment differing in one amino acid from its neighbor. Therefore, by measuring the mass difference between peaks in the MS/MS spectrum, one can determine the peptide sequence. Two variants of CID are commonly used in proteomics workflows: resonant excitation CID and beam-type CID. In resonant excitation CID, or commonly referred to as ion trap CID, peptide ions are trapped in an ion trap and resonance excited by applying a supplemental voltage matching the precursor secular frequency, which fragments the ions[81]. However, to prevent the ions from being ejected from the ion trap the precursor can only be excited by a few electron volt of kinetic energy. Therefore, ion trap CID needs to have many low energy collisions over a long activation time to be able to fragment the peptide ions. Ion trap CID is most commonly used in linear ion trap mass spectrometers and generates both b- and y-ions[6]. On the other hand, in beam-type CID peptide ions are not excited by resonance but are accelerated into the neutral gas bath of a collision cell, causing collisions and subsequent fragmentation. The radiofrequency field of the collision cell allows for higher energy collisions with the neutral gas without ion losses, which greatly decreases the activation time and generates mainly y-ions. In addition, in beam-type CID all ions are activated and fragments can be even further fragmented, resulting in very rich fragmentation spectra and better sequence annotation[81]. An example of a beam-type CID method is higher energy collisional dissociation (HCD). HCD is performed in the collision cell of hybrid orbitrap mass spectrometers[82]. The peptide ions are fragmented in the octopole collision cell, after which they are transferred to the c-trap and measured at high resolution in the orbitrap. HCD has multiple advantages compared to resonance excitation CID[80,83]. It has no low mass cut-off, which allows it to detect chemical reporter ions like iTRAQ and TMT. Additionally, it has a higher resolution and higher MS/MS quality spectra. A drawback is the longer spectral acquisition times due to the Fourier transform detection in the orbitrap.

A different method to fragment peptides is by using electrons. In the technique electron capture dissociation (ECD), which was already developed in 1998 by McLafferty *et al*[84], they reacted low energy electrons with multiply charged peptides. This reaction caused attachment of the electrons to the peptides, which then undergo rearrangements and subsequent fragmentation. ECD is not affected by peptide length and retains labile modifications on peptides. A method employing the same principle is electron transfer dissociation (ETD)[84]. In ETD, the electron is transferred from a radical anion to the peptide ion, causing peptide backbone fragmentation in the same way as ECD does. As an anion, fluoranthene is commonly chosen. In contrast to CID methods, ETD generates complementary c- and z-ions (Fig. 9). Likewise, ETD also maintains labile PTMs and allows for their assignment to certain amino acids in the peptide. Therefore, ETD has seen use in the study of PTMs such as phosphorylation, sulfonation and glycosylation[84]. A downside of ETD is the charge dependent effect; it is mostly suited for longer and more basic peptides with more than 3 charges[85]. When looking at peptides with two charges, ETD has a lower efficiency compared to HCD[81]. Additionally, ETD needs a longer reaction time compared to HCD activation times. This longer reaction time has an important effect on the number of identifiable peptides and limits the dynamic range and depth of ETD experiments.

It has been shown that after ETD fragmentation the charge reduced and unreacted precursor is one of the most abundant peaks in the MS/MS spectra[85]. Therefore, it would be beneficial to improve the fragmentation of the peptide during ETD to increase the product ion yield, especially for lower charged peptides. One way of achieving better dissociation is by first performing an ETD fragmentation step followed by an HCD fragmentation step. This means that all ions, unreacted precursor and fragments following from ETD fragmentation, are fragmented using HCD. This technique was coined EThcD[85] and allows for the generation of b-, y-, c- and z-ions in parallel. It was shown that EThcD substantially increases backbone fragmentation and therefore sequence coverage of the peptide ions. Also, it was shown that EThcD increased the confidence of phosphosite localization[86]. However, since ETD and HCD are sequentially performed it will limit the dynamic range and the depth of experiments.

Database searches and protein identification

After the generation of MS/MS spectra in the mass spectrometer, one needs to identify the peptides these belong to. This identification can occur by different methods: by using *de novo* sequencing or by using database matching (Fig. 10). As stated before, every fragment in a series differs in mass only by one amino acid from the previous. This allows, by checking the mass difference between two peaks, to sequence the peptide from which the spectra originate[70]. This process of determining the peptide sequence from subsequent fragmentation peaks is called *de novo* peptide sequencing. A big advantage of *de novo* sequencing is that it does not need a reference database to identify peptides. This way, unknown peptides or proteins can still be identified[6]. However, in experimental mass spectra some peaks of series are missing or overlapping with fragments of other series, complicating identification. Therefore, in order to successfully sequence peptides one needs high quality data, in the sense of mass accuracy and resolution, in combination with an information rich MS/MS spectrum. Since this is often not the case, most peptide identification strategies now employ database matching strategies.

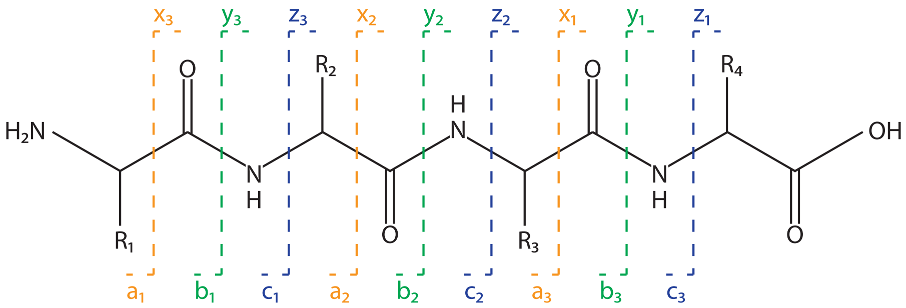


Figure 9: The different fragments which can be formed during MS/MS studies of peptides. If the peptide is subjected to fragmentation by CID, the amide bond between amino acids will break and mostly b- and y ions are formed. On the other hand, if the peptide is subjected to fragmentation by ETD, mostly c- and z-ions are formed.

In the database matching strategy, the spectra generated from the mass spectrometer are compared to spectra which were generated by *in silico* digestion of the proteome using the intended protease and the genome of the organism studied. When the precursor masses match, the experimental fragments are compared to the theoretical masses to determine whether they correspond to the same peptide. This approach is simpler than *de novo* peptide identification: it was found that only a certain amount of peptide compositions occur in nature[70]. Therefore, even if the experimental spectra do not contain enough information to unambiguously determine the whole peptide sequence *de novo*, it might still contain enough information to match it to one of the *in silico* generated spectra. In database searching one settles for the best match in a pool of sequences: if there is a good match, it is accepted as the correct one[87]. A disadvantage of using database matching is that the genome of the organism under study needs to be sequenced in order to generate the *in silico* database. This hinders the use of this strategy in unknown organisms. Numerous algorithms can be used to perform database matching, whereby SEQUEST, Mascot and Andromeda represent some of the most used ones. In the SEQUEST algorithm[88], all experimental fragment ions are compared to the theoretical ions on the spectrum level to determine their correlation[6]. It uses two parameters to determine the goodness of fit. The XCorr parameter is a calculation of the correlation of the theoretical and experimental spectra. The ΔCN parameter is the difference between the best peptide spectrum match (PSM) and the second best. By evaluating these parameters for all matches, the best fit can be found. A different algorithm which can be used is Mascot[89], where a probability score is calculated which determines how likely a theoretical spectra matches an experimental one by chance. The higher the Mascot score, the better the fit between the theoretical and experimental spectra. Lastly, the Andromeda search engine[90] also employs a probability based approach, somewhat similar to Mascot.

After the database search, the algorithm returns a score for the PSMs[87]. When the score is very high, it will show that the correct match was found, a true positive. On the contrary, if the score is below the significance threshold or the peptide sequence is not found in the database it will show a true negative. Problems lay in the other two cases: a false positive occurs when a significant match is found for the wrong sequence while a false negative occurs when no match is found even while the sequence is present in the database. To account for these findings, the false discovery rate (FDR) metric has been introduced. The FDR is a measure for how many matches are allowed by chance: a higher FDR allows for more PSMs (higher sensitivity) while a lower FDR allows for more certainty in the correct matches. The FDR can be estimated by performing a target-decoy search[87]. By repeating the search with the exact same settings, but with the sequences of the database reversed, one can easily determine the amount of false positives expected using these search settings. This FDR is then a marker for the trustworthiness of the search, while the score given to the reversed PSMs can be used as a cut-off value for false positives.

After the peptides have been successfully identified, they are mapped back to the proteins from which they originated. This process is greatly complicated by the fact that some peptides are not unique to certain proteins but can be mapped to numerous different ones[87]. When a protein is inferred by only unique peptides, it can be certain that it is present in the sample. However, more commonly proteins are inferred by shared, or razor, peptides, leading to a probability that this peptide really belongs to the protein. Therefore, more commonly peptides are mapped to protein groups. By determining a protein FDR this can be remedied, but it remains a problem to be able to unambiguously determine which proteins are present in the sample. There are multiple software suites which contain the whole bio-informatics pipeline needed to determine which proteins are present in a sample. They perform database searching, filtering and protein inference in one program, greatly simplifying this whole process. Examples of these packages include MaxQuant[91] and Proteome-Discoverer.

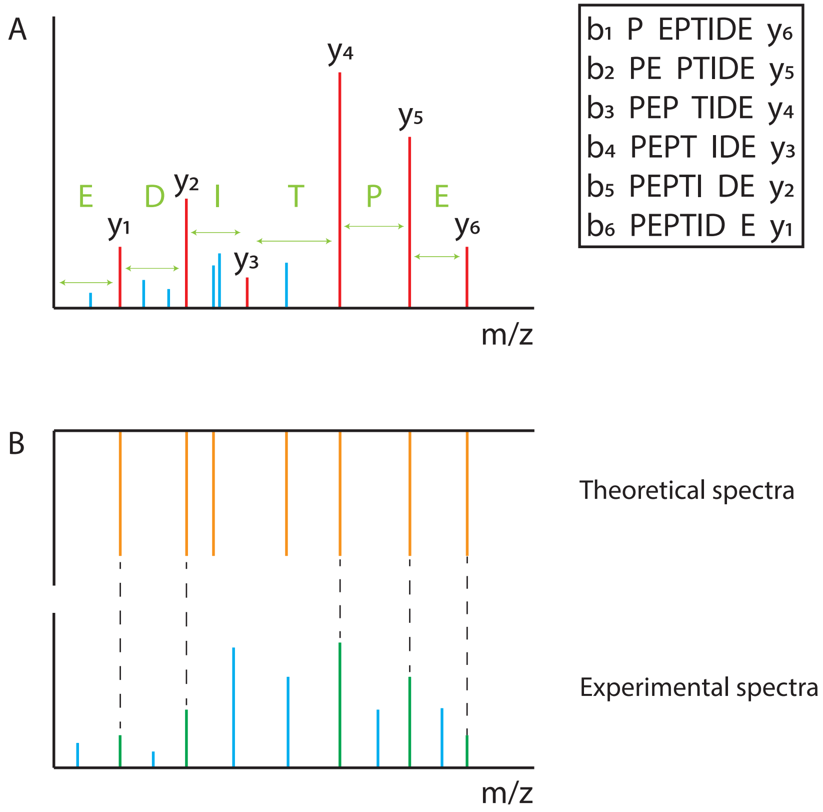


Figure 10: Different methods to determine peptide sequence from experimental fragmentation spectra. In *de novo* peptide sequencing, the peptide composition is determined by finding the mass difference between consecutive fragmentation peaks belonging to a series (A). A different approach is the database matching strategy, where experimental spectra are compared to theoretical *in silico* digested ones to find the correct peptide match (B).

Quantification strategies for bottom-up proteomics

In the early days, proteomics was mainly used as a qualitative tool; investigating the presence or absence of proteins under biological conditions. However, it quickly became apparent that the list of identified proteins does not give all information needed[6]. A change in protein abundance is a better reflection of biological processes or disease states, which can be used to find drivers of disease and druggable protein targets. However, proteomics is not inherently quantitative due to the fact that proteolytic peptides exhibit a wide range of physiochemical properties, such as size, charge and hydrophobicity, which will lead to a difference in mass spectrometric response[92]. Therefore, for accurate quantification it is necessary to compare individual peptides between experiments. Numerous strategies can be employed to achieve this, which are summarized in figure 11. Two major approaches can be taken to quantify the proteome: isotopic labelling and label-free approaches. The isotope labelling strategy is based on the stable isotope dilution strategy, which states that an isotopically labeled peptide is chemical identical to its natural counterpart, leading to it showing the same behavior during chromatographic separation and mass spectrometric analysis. The masses of these labeled and unlabeled peptides can be distinguished from each other in the MS, allowing for accurate quantification. A different strategy is label-free quantification, where peptides are quantified using their mass spectrometric signal intensity or by using the number of acquired spectra. All these strategies will be discussed in more detail below.

Metabolic labelling strategies

The first step where isotopic labels can be incorporated in proteins is during cell growth and division[92]. This technique is called "stable isotope labeling by amino acids in cell culture" (SILAC) and was developed in 2002 by Mann *et al*[93]. In this methodology, cells are cultured in medium where lysine and arginine are replaced by isotopically labeled lysine and arginine (^{13}C, ^{15}N), which ensures that all tryptic peptides will have at least one isotopic label. All the proteins will already be isotopically labeled after 5 cell divisions[94]. These labelled peptides, called the "heavy" peptides, can then be distinguished from the unlabeled ones, the "light" peptides, by a mass difference in the MS/MS spectra. The ratio of the intensities of the light and heavy peptides in the mass spectrometer reflects their relative abundance. SILAC can, for example, be used to compare light labeled wildtype cells to heavy labelled drug treated cells. A big advantage of SILAC labelling is that the mixing of samples already occurs on the cell culture level, meaning that all quantitative errors due to subsequent steps, such as digestion and chromatography, will be shared by both populations. This will greatly increase the accuracy and precision of this quantitative technique. There are some factors that need to be considered: SILAC labelling is not amendable to all sample types. For example, cell types which are sensitive to media composition or are difficult to grow *in vitro* may not be amendable for SILAC. Additionally, SILAC can only be used on cells, meaning it does not work on samples such as serum or body fluids[94]. Lastly, higher organisms, such as mouse[95] or fruit fly[96], are very difficult to fully isotopically label, which is a prerequisite for a successful SILAC experiment. Also, since there is a limit to the repertoire of useful heavily labeled amino acids, only a maximum of three samples are combined in practice[97].

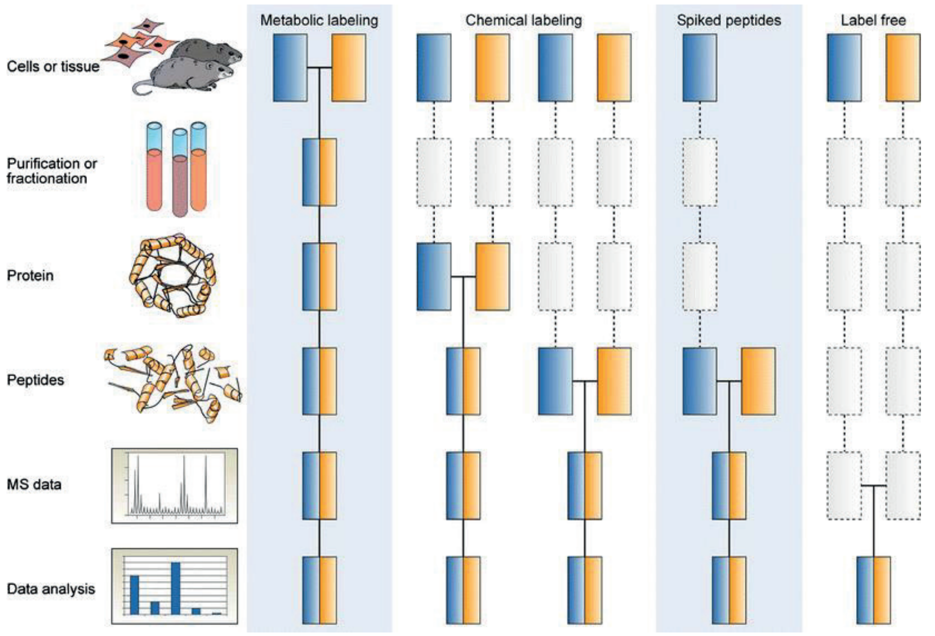


Figure 11: Different strategies employed in quantitative mass spectrometry. Two approaches can be used: isotope labelling strategies, which encompass metabolic labelling, chemical labelling and spiked peptides strategies or label free approaches. All different strategies can be employed in different steps of the sample preparation. Adapted from Bantscheff *et al*[92].

A variant on SILAC is the pulsed SILAC method[98], which can be used to make a quantitative comparison of translation rates on a proteome wide scale. In this approach, cells are first cultured in regular light medium followed by a differential treatment. After this treatment, cells are transferred to medium containing heavy or medium-heavy amino acids. The newly formed proteins after treatment will therefore be heavy or medium labelled. After incubation, the samples are mixed and measured on the mass spectrometer. By calculating the ratio of heavy and medium peptides, conclusions can be made regarding the translation of the corresponding protein in the differential treatment. The proteins present before the initial treatment remain light labeled and therefore can be excluded from the analysis. Using this method, one can detect the translation rate of a protein in a time dependent manner.

Even though SILAC is commonly used for cell culture research, ^{15}N labelling is more commonly used for microorganisms such as bacteria and yeast[97]. In ^{15}N labelling, the number of labels per peptide vary, in contrast to the single label incorporated during SILAC, which greatly complicates the subsequent analysis. Therefore, most commonly SILAC is used as metabolic labelling strategy in proteomics.

Chemical labelling of peptides for quantification

A different quantification strategy which can be used to label peptides or proteins is by using chemical labelling strategies. Numerous approaches exist, which provide a solid alternative for SILAC labelling. An example of a technique used to label peptides is ^{18}O labeling[99]. In this strategy, oxygen isotopes are incorporated in the C-terminus of the peptide during proteolytic digestion using trypsin or Glu-C. A total of 2 heavy oxygen atoms will be incorporated using this strategy, leading to a mass difference of 4 Da which can be distinguished on the MS[92]. An advantage of this technique is the use of an enzyme to label the peptides, reducing artifacts which occur during chemical labelling. In contrast, a big disadvantage is that full labelling is rarely achieved and every peptide incorporates the heavy oxygen at different rates, complicating data analysis.

A differing method to chemically incorporate isotope labels into peptides is by using dimethyl labeling[100,101]. In this approach, proteins are digested after which the primary amines of the N-terminus and lysine residues are converted to dimethylamines. This conversion occurs through the reaction of the primary amine with formaldehyde, which generates a Schiff base which is subsequently reduced by the addition of cyanoborohydride to form the dimethylamine. Three isotopic labels can be formed; regular formaldehyde with regular cyanoborohydride leads to a mass increase of 28 Da per amine, deuterated formaldehyde with regular cyanoborohydride leads to a mass increase of 32 Da per amine and lastly heavy labeled formaldehyde and cyanoborohydride leads to a mass increase of 36 Da per amine. By using dimethyl labeling, three conditions can be mixed and quantified relatively using MS. Dimethyl labeling is cheap, quick and the isotopic labelling does not interfere with ionization efficiency. Additionally, labelling of peptides can also be performed during sample cleanup using desalting columns, which avoids losses during sample preparation[97]. However, the use of deuterium labelling can lead to shifts in chromatographic behavior between labeled and unlabeled peptides[6].

In theory, every reactive amino acid can be used to incorporate isotope tags which can be used for quantification. However, in principle only cysteine and lysine are used for this purpose[92]. In the early work of Gygi *et al*[102] they introduced the isotope-coded affinity tag (ICAT), which is a tag which can be coupled to the thiol side chain of cysteine residues. The tag consists of three elements: a biotin tag used for affinity enrichment, a thiol reactive group and a linker which contains either 8 light or heavy hydrogen isotopes[6]. In this approach, the proteins belonging to the differing biological conditions are treated with either the light or heavy tag, subjected to proteolytic digestion and subsequently enriched by affinity purification. By then measuring the peptides on the mass spectrometer, a conclusion can be made regarding the relative abundance of the proteins in the two biological conditions. ICAT has numerous advantages; since cysteine is not present in every peptide, the affinity purification step allows for a great reduction in complexity of the peptide mixture. However, this also means that all non-cysteine containing peptides are removed and therefore ICAT does not allow for total proteome quantification. Also, the deuterium tags will lead to a retention time shift between light and heavy peptides during chromatographic separation, which can complicate data analysis. However, newer generation ICAT labels contain a ^{13}C linker, which does not show this retention time shift[6].

In all previous methods, quantification occurs due to integration of the MS^1 signal of the heavy and light-labeled peptides. In contrast to this, an MS^2 based approach called isobaric mass tagging has also been developed. In isobaric mass tagging, the peptides of different conditions are derivatized with different variants of isobaric mass tags from a set, pooled and subsequently ran on the mass spectrometer[103]. The isobaric character of these tags causes the peptides to maintain the same nominal mass and chromatographic behavior, causing them to co-elute during chromatographic separation with the same m/z value. However, when the peptides are fragmented, in addition to regular peptide fragments, reporter ion fragments will be detected. These reporter ion masses are unique for every tag in a set. Relative quantification of the peptides is then performed by comparing the intensities of the reporter ions. The most common isobaric mass tags are amine reactive; they are coupled to the N-terminus and lysine residues of peptides through NHS chemistry. This amine reactivity makes almost all the peptides present in a sample amendable to this strategy. Additionally, labelling is efficient for all peptide types regardless of peptide sequence[103].

The use of isobaric mass tags has many advantages[103]. The possibility of combining multiple samples due to multiplexing greatly decreases the analysis time needed. Also, since all conditions are run at the same time, experimental variations in ionization efficiency and mass spectrometric analysis are decreased. However, isobaric mass tags also have some disadvantages. The analysis of isobaric mass tags cannot be performed using conventional fragmentation methods in the ion trap, since during these low mass reporter ions are ejected from the trap and lost. To circumvent this, an orbitrap mass analyzer is commonly used in the study of isobaric mass tags[104]. Additionally, isobaric mass tags suffer from a problem called co-isolation[6]. When the mass spectrometer selects a peptide to fragment, it isolates a certain m/z window. It can occur that besides isolating the intended peptide, another peptide with a similar mass is co-isolated. The subsequent reporter ions formed after fragmentation will be a mixed pool of reporter ions from both precursor peptides, leading to inaccurate quantification. Two methods can be used to combat the co-isolation issues. In the MS^3 based method shown by Ting *et al*[105], the most abundant ion in the MS^2 spectra is subjected to another fragmentation step, after which the reporter ions are quantified. It was shown that this eliminates interference of co-isolation, but requires specialized software to be available. Another technique was developed by Wenger *et al*[106], where they charge reduced the precursor ions to increase the mass difference between the interfering peptides. The charge reduced ion is then fragmented and used for quantification. However, both these methods have the downsides of reduced sensitivity and data acquisition speed. Two isobaric mass tags are most commonly used: isobaric tags for relative and absolute quantification (iTRAQ) and tandem mass tags (TMT).

Tandem Mass Tags were first developed by Thompson *et al*[107] in 2003 and allow for the multiplexing of ten samples in one mass spectrometric run. By creatively combining heavy C and N atoms, it is possible to generate a set of ten TMT tags which all differ in approximately 0.0063 Da[108]. This mass difference can be distinguished after fragmentation on high-end mass analyzers such as the orbitrap. The tags consist of three domains: the reporter domain which generates the differential mass after fragmentation, the balance domain which makes sure the tags are isobaric and the reactive NHS domain which reacts with the N-terminus or lysine of the peptide. The schematic can be seen in figure 12.

The use of TMT tags has multiple advantages: the multiplexing of samples greatly reduces the number of missing peptide quantification values in each experiment and furthermore the quantification reproducibility is less dependent on LC performance[109]. It was shown that TMT has a high inter- and intra-laboratory reproducibility, warranting its use during large patient cohort experiments. These characteristics have made sure that TMT is now commonplace in proteomics research.

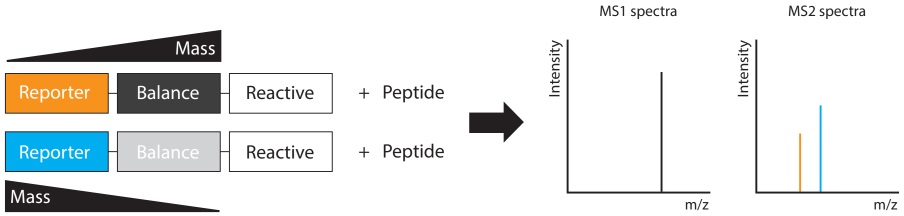


Figure 12: Principle of the tandem mass tag (TMT) isobaric labelling strategy. Due to the complementarity of the masses of the reporter and balance regions of the tag, all TMT mass tags will have the same nominal mass. This causes them to from one peak during MS analysis, however when the tags are fragmented they can be distinguished due to the difference in reporter region mass.

Label free quantification strategies

Instead of relying on the process of chemically labelling peptides in order to quantify them, it is also possible to quantify proteins without the extra step of labelling. These approaches are called label free quantification strategies. Two different strategies exist, which are both widely used but fundamentally quite different[92]. In the first approach, one can quantify the proteins by comparing the mass spectrometric signal intensity of peptide precursors belonging to them. Here, one extracts the ion chromatograms corresponding to the peptide of interest and integrates it over the chromatographic time scale to determine the area under the curve (AUC). By then comparing the intensity value of this AUC for the same peptide in different experiments, one can determine their relative quantity. Three factors need to be considered when performing this approach on very complex peptide mixtures. First, it is very beneficial to perform this type of quantification using a mass analyzer with a high mass accuracy, which allows to distinguish between similar but distinct masses. This way, the effect of interfering signals being co-quantified is minimized. Secondly, robust chromatography is needed to make sure the same peptide elutes at approximately the same retention time between experiments. This is necessary to facilitate the matching of the same peptide between experiments and allow for quantification. Lastly, a balance needs to be struck between MS1 survey scans and MS2 fragmentation scans. It is necessary to perform fragmentation scans to identify the peptide, but a robust quantitative reading requires multiple sampling over the chromatographic peak by survey scans. Therefore, better quantification accuracy will lead to poorer proteome coverage and vice versa[92]. An example of software employing this quantification strategy is MaxLFQ developed by Cox *et al*[110], which is routinely used in the MaxQuant bio-informatics suite.

A different approach to label free quantification is performed by counting and comparing the number of fragmentation spectra belonging to a protein[92]. This approach is called spectral counting and is based on the assumption that the larger the quantity of a certain protein is in the sample, the more fragment spectra of peptides belonging to it will be generated. Relative quantification can be determined by comparing the number of fragment spectra between experiments[92]. In contrast to quantification by peptide intensities, this approach greatly benefits from a larger amount of fragmentation spectra. Here, extensive fragmentation allows for the identification of the peptide and at the same time for better and more accurate quantification. This technique however is still controversial: it does not directly measure physical properties of the peptides. The spectrum count response will be different for every peptide, since they will all have differing chromatographic behaviors, such as retention time and peak width. However, by using alternative spectral counting methods, such as absolute protein expression (APEX)[111], one can circumvent this. In APEX, machine learning is used to determine the detection probabilities of tryptic peptides from experimental data. This probability can then be used to determine the amount of peptides expected to be detected from one copy of a protein, which is then compared to the experimental data to estimate absolute protein quantities.

Among all quantification strategies, label free approaches are the least accurate[92]. This is mainly due to experimental variations occurring during the experimental sample preparation. Samples are not mixed, so every sample will have their own independent experimental variance. Therefore, in label free quantification, the less steps the better. These approaches however can still be a good alternative to labelling strategies. There is no added time needed to incorporate labels into the peptides. Also, there is no limit on the amount of experiments which can be considered. In labelling strategies, there is always a fixed value of conditions which can be compared (e.g. ten for TMT-10plex), but this is not the case for label free approaches. Lastly, the mass spectrometric complexity of the samples is not increased, which may lead to a better analytical depth.

Absolute quantification strategies

All methods discussed so far are relative quantification strategies, where the quantity of a protein is determined relative to different experiments. However, it is also possible to determine the exact amount of a certain protein in the sample by using absolute quantification strategies. The basic principle of absolute quantification is the introduction of stable isotopically labeled standards into the sample[6]. These standards consist of a limited number of peptides, of which the exact amount is known. By comparing the abundance of the isotopically labeled peptide to a chemical identical peptide from the unlabeled digest, the absolute quantity of the corresponding protein can be determined. For absolute quantification, selected reaction monitoring (SRM) strategies are used, which are extremely suited for the quantification and selection of very low amounts of peptides. However, this technique is beyond the scope of this introduction, but a very good review has been written by Picotti *et al*[112]. The labeled standards can be generated in three ways: by chemically synthesized stable isotope labeled peptides (AQUA)[113], biologically synthesized artificial quantification concatemer (QconCAT) peptides[114] and intact isotope labeled protein standard absolute quantification (PSAQ)[115]. These absolute quantification strategies show great promises, but suffer from the drawback that it can only be performed on proteins for which a labeled standard was added[6]. Additionally, it is not known to which degree the measured quantity really matches the true quantity, especially after long sample preparation and chemical treatment steps. Introducing the labels as soon as possible during sample preparation can diminish this issue.

The analysis of post-translational modification using proteomics

The proteome is a very dynamic entity, which regulates all the processes inside of the cell. The activity of a protein is not only governed by its synthesis and degradation, but more so by addition or removal of covalent post-translational modifications (PTMs)[116]. There are more than 200 different PTMs, which regulate essential processes such as protein-protein interactions, cellular localization, turnover and activity[8]. When PTMs are dysregulated it may lead to disease, such as cancer, diabetes and numerous neurological disorders. This warrants the investigation of PTMs to find the underlying processes that can cause disease.

The study of PTMs greatly complicates the mass spectrometric analysis due to a large increase in proteomic complexity. A protein can be modified by numerous different PTMs on many distinct locations, but these do not need to necessarily occur at the same time. The sub-stoichiometric abundance of PTMs have warranted for specialized enrichment techniques to be able to identify them, which are specific for the PTM under investigation. In this introduction, as mostly related to the work described in this thesis, the PTMs phosphorylation and oxidation are discussed in more detail.

Phosphorylation as the driver of signaling cascades

Protein phosphorylation is one of the most important PTMs in biology, due to its versatility and ease of reversibility[117]. The phosphate group was chosen as a key building block in evolution due to multiple suitable characteristics. It is highly soluble in water and due to its chemical versatility can form mono-, di- and tri-esters with alkyl hydroxyl groups. These phosphate esters are stable in aqueous solutions at physiological pH and are formed by using adenosine triphosphate (ATP) as substrate. Through these characteristics, phosphate esters play an important role in cells in the form of nucleic acids and phosphoproteins[117].

The phosphate group is added to amino acids through the enzymatic activity of kinases, using ATP as phosphate donor. In contrast, phosphorylation can also be readily removed through the enzymatic activity of phosphatases. Phosphorylation of proteins can lead to a wide range of effects, ranging from changing its localization to facilitating protein-protein interactions and conformational changes (Fig. 13)[118]. This whole range of effects is due to the fact that when an amino acid gets phosphorylated, it will lead to a chemical characteristic unlike any other amino acid[117]. The phosphate group, with its large hydrated shell and its negative charge greater than one, provides a way for diversifying the chemical nature of the protein surface. By changing the protein surface, phosphorylation has the ability to change the conformation of the protein[118]. Also, the unique characteristics of the phosphate group allow specific and inducible interactions between phosphoproteins and phosphospecific binding domains in other proteins, which are crucial for the propagation of inter and intra-cellular signaling cascades. These factors combined show the major influence protein phosphorylation can have on cellular functioning, warranting further investigation.

It has been shown that phosphorylation can occur on 9 out of 20 amino acids as a regulatory mechanism [117]. These nine include serine, threonine, tyrosine, arginine, lysine, histidine, cysteine, aspartate and glutamine. Serine, threonine and tyrosine are the most prevalently modified residues, at least in eukaryotic systems, and consequently the most studied (Figure 14). In eukaryotes serine phosphorylation occurs the most (86.4%), followed by threonine (11.8%) and tyrosine (1.8%)[119]. These hydroxyl O-linked phospho-residues have seen the most study due to their stability in current acidic phosphoproteomics workflows. The other N-linked phosphoramidate (His, Lys and Arg), carboxy O-linked acyl phosphate (Asp, Glu) and S-linked phosphorothiolate (Cys) are acid labile and therefore more difficult to study using conventional methods. However, the study of histidine phosphorylation recently gained traction. This is discussed in more detail in chapter 3 of my thesis.

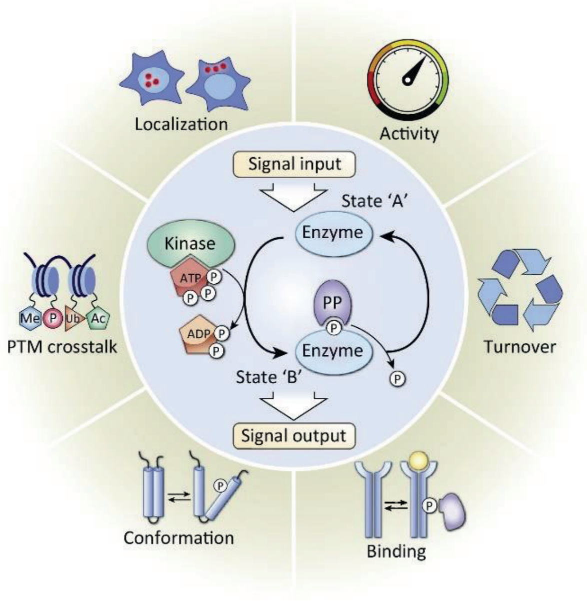


Figure 13: Effects of phosphorylation on the proteome. When a protein is phosphorylated by a kinase, it can lead to a whole range of effects. It can, amongst others, change the localization of the protein, change it activity or turnover. This process is reversible, since the phosphate group can be removed by phosphatases. Adapted from Humphrey et al[118].

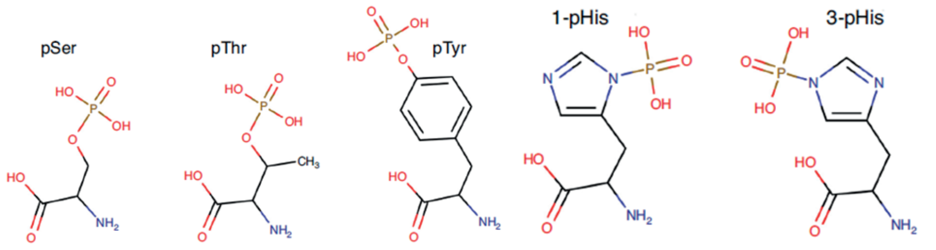


Figure 14: Amino acids that can be phosphorylated. The O-linked phosphoresidues serine, threonine and tyrosine have received most attention. However, phosphohistidine was also shown to play an important role, at least in bacteria. Adapted from Fuhs *et al*[120].

It can be seen that phosphorylation plays an important role in the functioning of an organism through specific signaling cascades. Therefore, instead of just measuring the protein abundance in differential biological conditions, more information can be gained by also measuring the difference in phosphorylation of the proteome. This field is called phosphoproteomics, where multiple factors need to be taken into account, which will be discussed further below.

Phosphoproteomics, the study of signaling inside of the cell

Initially, the phosphorylation of proteins was measured by radioactive ^{32}P labelling or western blotting[121]. In ^{32}P labelling, cells are incubated with radioactive phosphate, which will then be incorporated in proteins by kinases and subsequently be detected using autoradiography. Additionally, purified proteins can also be mixed with a kinase and radioactive ATP to determine whether the protein is a substrate of the kinase. These techniques suffer from disadvantages. The radioactive phosphate is toxic to cells and often labelling will be inefficient due to endogenous ATP present in the cell. Alternatively, western blotting can be performed. In western blotting, antibodies can be employed to measure the presence of proteins in a complex mixture. By using antibodies against phosphoserine, -threonine or -tyrosine the global phosphorylation state of the sample can be determined, but no knowledge is gained on individual proteins. In contrast, when samples are incubated with antibodies against specific phosphosites in proteins, the phosphorylation state of a protein can be determined. However, this way only a certain amount of proteins can be screened at once, which need to be known beforehand. Moreover, in theory for each phosphoprotein and phosphosite another antibody needs to be generated, which is practically impossible.

These methods are thus not ideal, as they are laborious and time consuming while it has been shown that more than 100,000 distinct phosphorylation events may happen in human cells[122]. Additionally, the temporal aspect of the phosphoproteome needs to be considered, which for example can change dramatically during the different stages of the cell cycle[123]. Lastly, it is important to measure the relative abundance of phosphorylation events between biological states to completely understand the underlying signaling pathways[124]. By using mass spectrometric approaches, all these factors can be routinely measured. Mass spectrometry allows for the identification and quantification of phosphopeptides in samples, while simultaneously localizing the amino acid that has been phosphorylated. This makes mass spectrometry the method of choice when measuring the global phosphorylation state under differential biological conditions. However, multiple factors need to be considered in order to successfully perform a phosphoproteomics experiment.

Fishing for phosphorylation in a sea of peptides

As stated before, phosphorylation events have a sub-stoichiometric character: it is estimated that only 1-2% of the entire protein population is phosphorylated at a specific moment[125]. This means that of the enormous amount of peptides formed after digestion in a bottom-up proteomics experiment, only a fraction will be phosphorylated. To efficiently measure these phosphopeptides and accurately identify the site of phosphorylation, they need to be enriched from the large pool of non-phosphorylated peptides to decrease the complexity of the sample. Multiple strategies exist for this enrichment, ranging from affinity chromatography approaches to immunoprecipitation and ion exchange chromatography (Fig. 15).

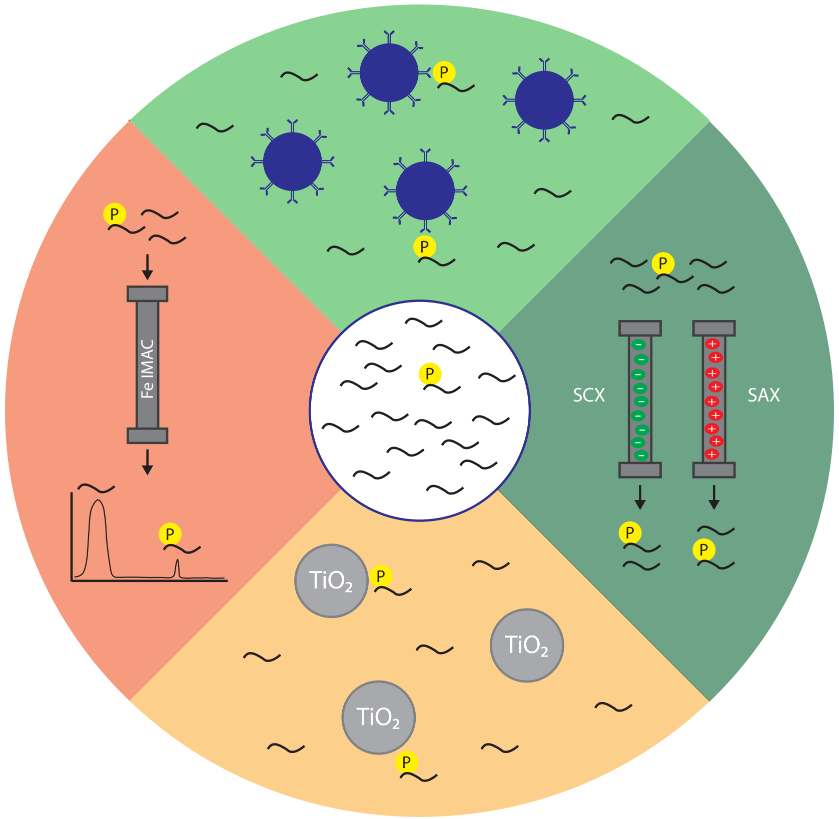


Figure 15: Different approaches which can be taken to enrich phosphorylated peptides. Techniques can be based on the interaction between positively charged metal ions and the negatively charged phosphate group, such as IMAC (left) or MOAC (bottom). Also, ion exchange chromatography approaches, such as SCX or SAX, can be used to enrich phosphopeptides (right). Lastly, immunoprecipitation methods can be used (top).

Many techniques used in the enrichment of phosphopeptides use the interaction between positively charged metal ions and the negatively charged phosphate group. These techniques are called inorganic ion affinity chromatography[125]. A widespread example of these techniques is immobilized metal affinity chromatography (IMAC), which is based on the high affinity of phosphate to specific trivalent metal ions[124]. The metal ions are immobilized on solid supports through chelators, such as iminoacetic acid and nitrolotriacetic acid, after which they can be used to specifically bind phosphopeptides[125]. The phosphopeptides are allowed to interact with the metal, while the non-phosphorylated peptides are washed through. Afterwards, the phosphopeptides are sequentially eluted. Multiple metal ions can be used, such as Fe^{3+}, Al^{3+}, Ga^{3+} or Co^{2+}, which can be bound to different types of solid support formats such as beads or columns[121]. IMAC was originally used for the enrichment of phosphoproteins[126], but is now the most used technique for the enrichment of phosphopeptides. A very nice application of Fe^{3+}-IMAC was shown by Ruprecht *et al*[127], where they describe a workflow of offline chromatographic enrichment of phosphopeptides using a Fe^{3+}-IMAC column format. In this approach, a peptide digest is injected through the Fe^{3+}-IMAC column, where the phosphopeptides will bind to the iron ions while the unmodified peptides will not be retained. The phosphopeptides are then eluted by using a basic buffer, which neutralizes the charge on the phosphate group and diminishes the affinity for the iron ions (Fig. 15). Using this strategy, they could measure more than 10000 unique phosphopeptides from 1 mg of cell digest in 2 hours of measurement time[128]. Recently, Potel *et al*[129] further improved this workflow. In their work they identified negatively charged nucleic acid containing biomolecules as interfering components during the enrichment. These negatively charged molecules will decrease the sensitivity of the enrichment. After cleanup of the samples using DNase and benzonase, they identified more than 17000 unique phosphopeptides from one LC-MS run, which corresponds to an increase of more than 50%. The improvement was even more striking when measuring the phosphoproteome of *E. coli*, where the number of identified phosphopeptides was increased by a factor ten. These results make the enrichment of phosphopeptides using this technique very efficient and is therefore used for all phosphoproteomics experiments discussed in this thesis. However, there is a disadvantage connected to IMAC: it can have a high level of non-specific binding when complex peptide samples are measured. When a peptide contains multiple acidic amino acid residues, it will co-purify during IMAC enrichment, decreasing the selectivity[121]. This can be remedied by acidifying the sample to below pH 1.9, causing the acidic peptides to be neutral while the phosphopeptides retain their charges. This way, the acidic peptides will not bind the column and interfere with analysis.

An alternative approach to enrich phosphopeptides is metal oxide affinity chromatography (MOAC) using TiO_2 beads. This approach is similar to IMAC, but where in IMAC the metal ion needs to be chelated to a solid support, MOAC enrichment is performed using solid beads consisting of the metal oxide. Phosphopeptides are loaded to the beads in an acidic buffer, causing phosphopeptides to bind to the beads while non-modified peptides will not[130]. The phosphopeptides can then be eluted using basic buffers. Similar to IMAC, multiple different metal oxides can be used for the enrichment: TiO_2, ZrO_2, and Al_2O_3, amongst others[125]. These metal oxides all differ slightly in their selectivity towards phosphopeptides. This approach has multiple advantages: it is simple, fast and the TiO_2 material is tolerant to most buffers, detergents and salts used[121]. Additionally, a very high selectivity can be achieved by using a functional acid such as 2,5-dihydroxybenzoic acid as a competitive binder[131]. Even though TiO_2 beads have nice characteristics, they are routinely outperformed by the Fe^{3+}-IMAC column format[127].

A different approach which can be taken to enrich phosphopeptides is by using ion exchange chromatography. Phosphopeptides have a different charge state compared to non-phosphorylated peptides: at pH 2.7 tryptic peptides commonly have 2 charges due to protonation of the N-terminal amino group and the C-terminal arginine or lysine. In contrast, in phosphopeptides the negatively charged phosphate group will decrease the overall charge by 1[124]. This charge difference can be used to separate these populations of peptides using strong cation exchange (SCX) chromatography. As mentioned before, SCX separates peptides according to their charge: the negative groups of the stationary phase will attract the positively charged peptides at acidic pH. This will cause the phosphorylated peptides to elute earlier from the column compared to non-phosphorylated peptides. This approach was successfully used by Beausoleil *et al*[132], where they identified 2000 phosphosites from 8 mg of nuclear extract from HeLa cells. A downside of this approach concerns multiply phosphorylated peptides, which can cause the net-charge of the peptide to go below 1. When this occurs, the peptide will not bind the SCX column and will end up in the flow through. This can be remedied by also measuring the flow through fraction on the MS to identify multiple phosphorylated peptides. Most commonly SCX is not used as a standalone enrichment technique but in combination with either IMAC or TiO_2 enrichment[121]. It was shown that when SCX is used in combination with IMAC, it will lead to at least a threefold increase in identified phosphopeptides compared to the methods used alone[133]. This combination was successfully performed in the research of Olsen *et al*[134], where they combined SCX pre-fractionation with TiO_2 enrichment to identify 6600 phosphorylation sites in 2244 proteins in HeLa cells stimulated with EGF. These facts clearly show the potential of SCX as prefractionation technique before phosphopeptides enrichment.

Likewise, strong anion exchange chromatography (SAX) can also be used to enrich phosphopeptides. This method functions inversely from SCX: here the positively charged stationary phase attracts negatively charged molecules. Under weak acidic conditions, the negatively charged phosphopeptides are more strongly retained to the column compared to their non-phosphorylated counterpart, allowing them to be separated[135]. The validity of this technique was shown by Han *et al*[136], where they showed that phosphopeptides of α and β casein were successfully retained by the column. Additionally, they showed that the technique performs on the same level as Fe^{3+}-IMAC. However, not much literature is known of studies employing SAX for the enrichment of phosphopeptides, so the true potential is not known.

The last chromatographic method which can be used to enrich phosphopeptides is using hydrophilic interaction chromatography (HILIC). HILIC is a separation technique which is mainly suited for the separation of polar analytes: peptides in an organic solvent (mobile phase) bind to the neutral hydrophilic stationary phase through hydrogen bonding[121]. These hydrogen bonds can then be disrupted by decreasing the organic environment through a gradient of increasing percentage of aqueous solvent, causing the peptides to elute according to their hydrophilicities. Since phosphopeptides are more polar compared to non-phosphorylated peptides, they will be retained more strongly on the column, allowing them to be separated[135]. The utility of HILIC in phosphopeptide enrichment was shown by Annan *et al*[137], where they showed that prefractionation using HILIC, followed by enrichment using IMAC led to the identification of 1000 phosphosites in only 300 µg of HeLa lysate. They showed that HILIC is very suited for the prefractionation of complex samples, since it is orthogonal to the reverse phase chromatography used in LC-MS analysis. However, the standalone enrichment efficiency of HILIC is relatively low and phosphopeptides co-elute with non-phosphopeptides[135], warranting the use of a sequential IMAC step.

It was found that when these previously mentioned enrichment techniques are used, that there is a low level of enrichment of phosphotyrosine containing peptides[135]. It is however very important to study tyrosine phosphorylation, as aberrant tyrosine phosphorylation is a hallmark in multiple cancers[138]. Therefore, an optimized enrichment protocol for phosphotyrosine was developed, which employs immunoprecipitation for selective enrichment. There are numerous antibodies commercially available that are specific for phosphotyrosine and show a high efficiency for immunoprecipitation. The success of this approach was initially shown by Rush *et al*[139], where they identified 688 phosphotyrosine peptides and 628 phosphotyrosine sites from a complex cell lysate. The phosphotyrosine containing peptides were enriched using treatment with a pTyr specific antibody immobilized on agarose beads and eluted using dilute acid. By then measuring the peptides eluted from the beads, quantitative conclusions can be made concerning pTyr levels in the samples. It was shown that this technique is simple, sensitive and reproducible. More recently, Boersema *et al*[138] identified 1112 unique phosphopeptides from 4 mg of starting material, of which 80% was shown to be tyrosine phosphorylated. Additionally, they combined this with dimethyl labelling to quantify differences in phosphorylation level under different treatments. These studies show the big potential of measuring pTyr levels using immunoprecipitation. However, this approach is only amendable to pTyr containing peptides, since it was shown that antibodies against phosphoserine and threonine are mostly not specific enough and only a few perform well[140].

After the phosphopeptides have been successfully enriched, they need to be subjected to LC-MS analysis to quantify the differences between biological states and identify the exact location of the phosphorylation event. There are some considerations which need to be taken when measuring phosphopeptides compared to non-modified peptides; the low ionization efficiency of phosphopeptides in combination with the loss of the phosphate group results in a low identification rate for phosphorylated peptides[121]. This partial loss of the phosphate group, also known as a neutral loss, competes with backbone cleavage and can result in less fragment information. This is especially the case in CID fragmentation of phosphoserine and threonine, where a large degree of non-sequence informative neutral losses occur[81]. However, this issue can be averted by using higher energy fragmentation techniques, such as beam-type CID or HCD, in combination with a high-resolution mass spectrometer to produce richer fragment spectra that can facilitate correct sequence annotation. Phosphotyrosine has some interesting characteristics: its phosphate group is much more stable in the gas-phase, leading to less neutral loss effects[121]. Additionally, fragmentation of phosphotyrosine can lead to a characteristic immonium ion in the fragmentation spectra with m/z 216, which can be used as a diagnostic marker in product ion scanning strategies[141]. Recently, it has been shown that pHis can also exhibit an immonium ion at m/z 190, which can be used for diagnostic purposes[142]. The use of ETD as fragmentation technique, in combination with supplemental activation using HCD, gives rise to even better phosphopeptide identification due to the generation of two different ion series in parallel[81]. This combination of techniques was employed by Frese *et al*[85], who dubbed it EThcD, where they denoted an increase of circa 28% in average scoring compared to HCD. However, HCD remains the method of choice due to its higher speed and efficiency[143].

After the peptide has been efficiently fragmented and the peptide sequence has been determined, there is still the difficult task of assigning the phosphate group to the correct amino acid. It is of utmost importance to connect the phosphorylation event to the correct residue to understand its biological function. Since there can be multiple phosphorylation sites (Ser, Thr, Tyr, His etc.) in one peptide, the peptide sequence information alone is not enough to accurately assign the phosphate group to the correct amino acid. Therefore, additional dedicated software is needed to assess the localization confidence[81]. Two approaches can be taken to score the localization of the phosphosite: I) calculating the probability of an incorrect match for each phosphopeptide isoform or II) by using the difference in score between the different phosphopeptide isoforms. The first approach is taken in algorithms such as Ascore[144], PhosphoRS[145] or the PTM-score of Andromeda[90] while the second approach is utilized in algorithms such as the Mascot delta score[146] or Luciphor[147]. Even though these algorithms share basic principles, the results can vary due to small variations in localization strategies. Nevertheless, all these algorithms deploy a database search approach where they match the MS/MS spectra with theoretical spectra where the phosphate group is placed on every possible location. By comparing all these matches, a localization probability is generated that the phosphate group is present on a certain amino acid, which can be used to determine the goodness of fit. Mostly cut-off scores, such as a score larger than 0.75 in the Andromeda PTM-score, are used to determine true hits. This list of phosphosites matching the filtering criteria can then be used to compare biological situations, similar to regular database search approaches.

Similar to normal bottom-up proteomics experiments, more knowledge can be gained when phosphosites are not only detected but also quantified. This way, not only on-off phosphorylation events can be compared but also the presence of subtle changes in phosphorylation, which can have enormous effects, can be determined. Similar to regular bottom-up proteomics experiments, phosphopeptides can be quantified in an absolute or relative manner with the same techniques as discussed earlier. In the study of Hogrebe *et al*[148], they benchmarked multiple quantification methods (LFQ, SILAC and TMT) for their use in phosphoproteomics experiments. They showed that MS^2 based TMT strategies gave the best results, with the highest precision and identification numbers. Additionally, they showed that the high accuracy in MS^3 based TMT gave the best information on phosphosite stoichiometry. Even though quantification strategies in phosphoproteomics work nicely, they have some points of attention. A caveat of quantification in phosphoproteomics is the absence of multiple peptides which can be used for quantification, while this is not the case in protein quantification[148]. Additionally, it is necessary to normalize the phosphopeptides according to their protein abundance to make sure that the change is phosphorylation dependent and not due to a change in expression or degradation of the protein[121].

Redox proteomics, the balance between regulation and disease

Although a significant part of my work for this thesis focused on protein phosphorylation, I also investigated another important protein modification, namely protein oxidation. Every organism living in an aerobic environment will be exposed to reactive oxygen species (ROS), such as superoxide anions, hydrogen peroxides and hydroxyl radicals[149,150]. They can be generated endogenously through mitochondrial oxidative phosphorylation or be introduced from exogenous sources such as xenobiotic compounds. These ROS can be used to perform physiological functions, for example they are indispensable for the function of immune cells against pathogens. Additionally, ROS are very suitable for use in signaling pathways; their generation is very controlled, there is no scarcity of substrate and they can be readily removed due to their chemical instability[151].

Most commonly, ROS interact with the thiol group of cysteines where they can cause a variety of modifications. When the thiol group is oxidized, it will lead to a sulphenic acid modification, which can change the three-dimensional configuration of the protein by forming disulfide bonds (Fig. 16). The sulphenic acid form of cysteine is readily reducible, for example by the thioredoxin and peroxiredoxin systems, making it useful in signaling cascades. However, when the oxidative impulse is very harsh or prolonged, it can lead to further oxidation of the cysteine. When two oxygen atoms attach to the cysteine, it will achieve a sulphinic acid form, while if three oxygen atoms are attached to the cysteine a sulphonic acid form is maintained. Sulphinic and sulphonic acid are non-reducible and therefore non-reversible. This makes them bad candidates for cellular signaling. So similar to phosphorylation, the addition of oxygen can change the three dimensional structure of the protein, attenuating its function or interaction partners[151]. One important example of proteins being attenuated by redox signaling are tyrosine phosphatases, such as SHP2, which can be reversibly inhibited by oxidation of the catalytic cysteine[152,153]. Additionally, processes such as remodeling of the actin cytoskeleton and transcription can also be regulated by cysteine dependent oxidation[154].

However, there is a tight balance that needs to be dealt with. When the ROS overwhelm the antioxidant systems of the cell, it can lead to a range of diseases caused by oxidative stress. It can cause neurodegenerative diseases, such as Alzheimer and Parkinson disease, but can also lead to cancer, atherosclerosis, diabetes and aging effects[149,150]. When the oxidative stress gets too high, proteins will be irreversibly oxidized, which can cause their fragmentation, aggregation and a higher susceptibility to proteolysis[149]. This oxidation induced damage to proteins will eventually promote cell death. Therefore, the irreversible oxidation of cysteine is a pathological marker of oxidative stress, which can be measured to gain knowledge of the influence of redox-induced damage in different pathological states.

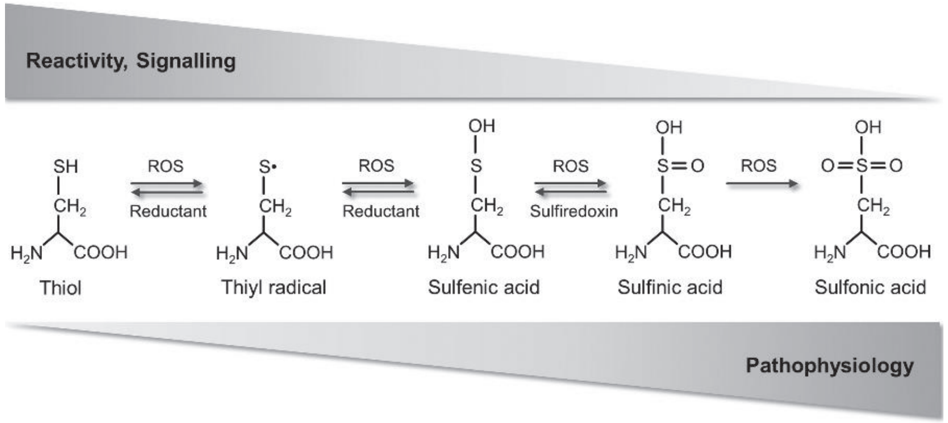


Figure 16: Different oxidative states of cysteine. The longer or harsher the oxidative impulse is, the further the cysteine will be oxidized. In the sulphenic acid state, it can be readily reduced but if the cysteine is converted to the sulphinic or sulphonic acid state, it will be irreversible and lead to inactivation of the protein. Adapted from Jortzik et al[155].

Similar to phosphoproteomics, mass spectrometry is extremely well suited for the site-specific identification and quantification of oxidative PTMs on the proteome level[156]. By comparing the redox proteome between different pathological states, main players that drive disease states can be identified. Sulphinic and sulphonic acid modifications are stable under standard MS sample preparation, making it possible to measure them in standard workflows in a complex sample. In contrast, the reversible sulphenic acid has some complications regarding the measurement of this PTM in a MS setting[156]. It is very labile in the gas-phase, meaning it is often cleaved off during ionization and subsequent fragmentation. In addition, it can be artificially oxidized during sample preparation to a sulphinic or even sulphonic acid state, complicating analysis. Therefore, specific strategies are needed to measure the whole range of reversible and irreversible modifications within the redox proteome.

In living cells, 80 – 90% of the cysteines are in the reduced state, but these can be readily oxidized during cell lysis and additional sample preparation steps. During cell lysis, the compartmentalization of cells is disrupted, mixing the highly oxidizing environment of mitochondria with steady state compartments such as the nucleus, causing artificial changes in the redox state of the cell[157]. To preserve the endogenous oxidation state of the cell, multiple strategies can be used[156]. The most commonly used technique involves the alkylation of the free cysteines during lysis under denaturing conditions, for example using trichloroacetic acid, N-ethylmaleimide (NEM) or iodoacetic acid (IAA). By quickly alkylating the free thiols, this reactive population will be frozen in the alkylated state and be amendable for further sample processing and downstream measurements. When using this strategy, it is vital that complete blocking of free thiols occurs, otherwise quantification will not be accurate. As mentioned before, sulphenic acid modifications are also labile and can easily be converted to alternate forms or be lost completely, complicating analysis. To remedy this a second alkylation step is used to also freeze this population: the sulphenic acid PTMs are first reduced using DTT or TCEP to remove the modification, after which the freed thiol groups are reacted with a different alkylating reagent than used for the alkylation of the reduced thiols. This freezes the population of sulphenic acid modified cysteines, which can then be distinguished from the reduced thiol-alkylated cysteines through a mass difference between the alkylating groups. Since both sulphinic and sulphonic acid cannot be reduced using chemical approaches, they will be stable and differentiated from the other PTMs through mass spectrometry. This strategy, shown in figure 17, is commonly used and referred to as the differential alkylation strategy[157]. Instead of the differential alkylation strategy, other approaches can also be used to selectively enrich these peptides. For example, free thiols can be treated with pyridyldithiol-biotin, which allows their selective purification using streptavidin columns and the subsequent elution using DTT[156]. Alternatively, the alkylating reagents (NEM and IAA) can also be conjugated with fluorophores, epitope labels and stable isotope-code tags which allows for their selective enrichment, detection and quantification. As long as the tag can be covalently attached to cysteine residues, the possibilities are limitless.

Similar to the techniques discussed before, quantitative analysis of the redox proteome gives more information compared to qualitative analysis. It is crucial for the distinction between susceptible and sensitive redox sites: susceptible sites are those who are highly oxidized under physiological conditions while sensitive sites are not oxidized in physiological conditions but can be following oxidative stress[157]. Relative quantification can occur in multiple ways: I) using isotope labeling, II) isobaric labeling or III) label free approaches. Approaches used in isotope labeling include SILAC and the use of selective *in vivo* probes for specific cysteine modifications[156]. In the research of Zhou *et al*[158], they coupled SILAC to the biotin switch assay to determine the degree of endogenous S-nitrosylation, another cysteine redox modification, in RAW264.7 cells. Here, they found over 15 more S-nitrosylated proteins compared to earlier studies and that the level of S-nitrosylation differs for different cysteines in the same protein.

An example of *in vivo* probes is the use of isotope-coded dimedone probes, which specifically label sulphenic acid moieties[159]. Another isotope labeling approach is performed by using differential alkylation strategies with the same alkylating reagent, which is either heavy or light labeled. By comparing the relative abundance of light and heavy pairs, quantitative conclusions can be made. A different approach which can be taken is by using isobaric labeling strategies, which has the added benefit of allowing multiplexing. An example of these is the iodoTMT strategy[160], where the thiol group of cysteines are irreversibly labeled by differential isobaric tags. By then comparing the intensities of the reported ions for the different conditions, quantitave conclusions can be made. Lastly, label-free quantification can be used. However, the sub-stochastic nature of these modification have hampered the use of LFQ as quantitation method in redox proteomics[157], warranting an enrichment step.

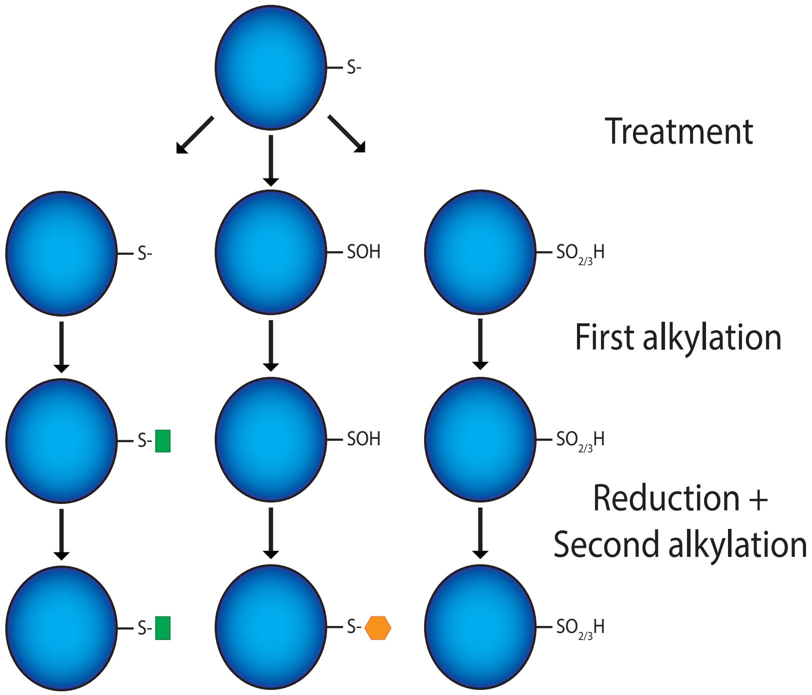


Figure 17: Schematic of the differential alkylation strategy used in redox proteomics. After the stimulus, unreacted cysteines are alkylated by use of an alkylating reagent. The sulphenic acid modification is removed by reduction, after which the now freed cysteines are labeled by a (different) alkylating reagent of different mass. All modifications have different masses, which can be distinguished from each other in the mass spectrometer.

Scope of this thesis

The main aim of my work as described in this thesis has been to improve on proteomics technologies to study protein phosphorylation, protein oxidation and interactions between proteins and drug molecules. These improvements have led to the results discussed in the chapters of this thesis.

In **chapter two** of this thesis, we were the first to perform thermal proteome profiling on zebrafish lysate. Thermal proteome profiling is a valuable technique in which mass spectrometry is used to find the on- and off-targets of small ligands, such as drugs. By finding the toxic off-targets of drugs, their development may be improved. Traditionally, thermal proteome profiling experiments were performed on single cell types, causing information on tissue specific proteins to be lost. Here, we improved on this methodology by performing thermal proteome profiling on zebrafish embryo lysate, which harbors all tissue specific proteins. We first showed, as a proof of principle, that we could detect ligand induced stability changes in pervanadate treated lysate, after which we extended this to the selective STAT3 inhibitor napabucasin. Using our approach, we validated the mode of action of napabucasin, while simultaneously finding aldehyde dehydrogenases as off-targets.

In **chapter three** of this thesis, we investigated the labile post-translational modification phosphohistidine. Phosphohistidine has been very difficult to study in the past, due to not being compatible with standard enrichment strategies. However, recently a novel approach has been developed which allows the identification of this PTM on a proteome wide scale. It was shown that phosphohistidine plays an important role in bacteria such as *E.coli*, however the importance and scope of this PTM in mammalian systems is not known. Here, we investigated the extent of phosphohistidine in mammalian cells using this optimized workflow. Many novel sites were found, but the validity of these was questioned. Therefore, acidification of the samples was used as a negative control. In *E.coli,* this drastically decreased the presence of phosphohistidine, while in mammalian samples this behavior was not replicated. Therefore, we concluded that the sites found in our experiments are false positives, and that the contribution of phosphohistidine in mammalian systems is extremely limited.

In **chapter four** of this thesis, we investigated the oxidative behavior of the catalytic cysteine of the tyrosine phosphatase SHP2 and its mutants. The oxidation of the catalytic cysteine of SHP2 is a known mechanism to (ir)reversibly inactivate it, but we were curious how the rates of oxidation differ between the wildtype phosphatase and the catalytically more active Noonan mutant. This mutant is in a more open conformation compared to the wildtype, which might cause it to be more readily oxidized. Indeed, through a differential alkylation approach we showed that the Noonan mutant is more readily oxidized compared to the wildtype. Additionally, in this chapter we showed that the addition of catalase to SHP2 in a fusion protein can efficiently protect the catalytic cysteine against hydrogen peroxide. In the future, these fusion proteins may be used to determine the oxidation status of SHP2 *in vivo*.

Lastly, in **chapter five** of this thesis I share my view on the future of proteomics. In addition, a lay summary of this thesis and my acknowledgements can be found here.

References

1. Cox, J. & Mann, M. Is Proteomics the New Genomics? *Cell* **130**, 395–398 (2007).
2. Martin, D. B. & Nelson, P. S. From genomics to proteomics: Techniques and applications in cancer research. *Trends Cell Biol.* **11**, 60–65 (2001).
3. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
4. Wilkins, M. R. *et al.* From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology* **14**, 61–65 (1996).
5. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
6. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).
7. Gstaiger, M. & Aebersold, R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* **10**, 617–627 (2009).
8. Virág, D. *et al.* Current Trends in the Analysis of Post-translational Modifications. *Chromatographia* **83**, 1–10 (2020).
9. Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* **445**, 683–693 (2014).
10. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499–519 (2016).
11. Zhang, J. *et al.* Top-down quantitative proteomics identified phosphorylation of cardiac troponin i as a candidate biomarker for chronic heart failure. *J. Proteome Res.* **10**, 4054–4065 (2011).
12. Heck, A. J. R. Native mass spectrometry: A bridge between interactomics and structural biology. *Nat. Methods* **5**, 927–933 (2008).
13. Leney, A. C. & Heck, A. J. R. Native Mass Spectrometry: What is in the Name? *J. Am. Soc. Mass Spectrom.* **28**, 5–13 (2017).
14. Snijder, J., Rose, R. J., Veesler, D., Johnson, J. E. & Heck, A. J. R. Studying 18 MDa virus assemblies with native mass spectrometry. *Angew. Chemie - Int. Ed.* **52**, 4020–4023 (2013).
15. Rosati, S. *et al.* Exploring an orbitrap analyzer for the characterization of intact antibodies by native mass spectrometry. *Angew. Chemie - Int. Ed.* **51**, 12992–12996 (2012).
16. Switzar, L., Giera, M. & Niessen, W. M. A. Protein digestion: An overview of the available techniques and recent developments. *J. Proteome Res.* **12**, 1067–1077 (2013).
17. Cristobal, A. *et al.* Toward an Optimized Workflow for Middle-Down Proteomics. *Anal. Chem.* **89**, 3318–3325 (2017).
18. Pandeswari, P. B. & Sabareesh, V. Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Adv.* **9**, 313–344 (2019).
19. Jiang, T. *et al.* Middle-Down Characterization of the Cell Cycle Dependence of Histone H4 Posttranslational Modifications and Proteoforms. *Proteomics* **18**, 1–11 (2018).

20. Valkevich, E. M., Sanchez, N. A., Ge, Y. & Strieter, E. R. Middle-Down mass spectrometry enables characterization of branched ubiquitin chains. *Biochemistry* **53**, 4979–4989 (2014).
21. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).
22. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
23. Simpson, R. J. & Dorow, D. S. Cancer proteomics: from signaling networks to tumor markers. *Trends Biotechnol.* **19**, 40–48 (2001).
24. Müller, B. & Grossniklaus, U. Model organisms - A historical perspective. *J. Proteomics* **73**, 2054–2063 (2010).
25. Gouw, J. W., Krijgsveld, J. & Heck, A. J. R. Quantitative proteomics by metabolic labeling of model organisms. *Mol. Cell. Proteomics* **9**, 11–24 (2010).
26. Hebert, A. S. *et al.* The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–347 (2014).
27. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).
28. Segner, H. Zebrafish (Danio rerio) as a model organism for investigating endocrine disruption. *Comp. Biochem. Physiol. - C Toxicol. Pharmacol.* **149**, 187–195 (2009).
29. Holtzman, N. G., Kathryn Iovine, M., Liang, J. O. & Morris, J. Learning to fish with genetics: A primer on the vertebrate model Danio rerio. *Genetics* **203**, 1069–1089 (2016).
30. Langenau, D. M. & Zon, L. I. The zebrafish: A new model of T-cell and thymic development. *Nat. Rev. Immunol.* **5**, 307–317 (2005).
31. Feitsma, H. & Cuppen, E. Zebrafish as a cancer model. *Mol. Cancer Res.* **6**, 685–694 (2008).
32. McGrath, P. & Li, C. Q. Zebrafish: a predictive model for assessing drug-induced toxicity. *Drug Discov. Today* **13**, 394–401 (2008).
33. Willemsen, R., Padje, S. V. T., Van Swieten, J. C. & Oostra, B. A. Zebrafish (Danio rerio) as a model organism for dementia. *Neuromethods* **48**, 255–269 (2011).
34. Saleem, S. & Kannan, R. R. Zebrafish: an emerging real-time model system to study Alzheimer's disease and neurospecific drug discovery. *Cell Death Discovery* **4**, 1–13 (2018).
35. Lößner, C. *et al.* Expanding the zebrafish embryo proteome using multiple fractionation approaches and tandem mass spectrometry. *Proteomics* **12**, 1879–1882 (2012).
36. Purushothaman, K. *et al.* Proteomics analysis of early developmental stages of zebrafish embryos. *Int. J. Mol. Sci.* **20**, (2019).
37. Shrader, E. A., Henry, T. R., Greeley, M. S. & Bradley, B. P. Proteomics in Zebrafish Exposed to Endocrine Disrupting Chemicals. *Ecotoxicology* **12**, 485–488 (2003).
38. Saxena, S. *et al.* Proteomic analysis of zebrafish caudal fin regeneration. *Mol. Cell. Proteomics* **11**, (2012).
39. De Souza, A. G., MacCormack, T. J., Wang, N., Li, L. & Goss, G. G. Large-scale proteome profile of the zebrafish (Danio rerio) gill for physiological and biomarker discovery studies. *Zebrafish* **6**, 229–238 (2009).
40. Islam, M. S., Aryasomayajula, A. & Selvaganapathy, P. R. A review on macroscale and microscale cell lysis methods. *Micromachines* **8**, (2017).
41. Burden, D. W. Guide to the Homogenization of Biological Samples. *Random Prim.* 1–14 (2008).
42. Huber, L. A., Pfaller, K. & Vietor, I. Organelle proteomics: Implications for subcellular fractionation in proteomics. *Circ. Res.* **92**, 962–968 (2003).
43. Lee, Y. H., Tan, H. T. & Chung, M. C. M. Subcellular fractionation methods and strategies for proteomics. *Proteomics* **10**, 3935–3956 (2010).
44. Andersen, J. S. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
45. Calvo, S. E. & Mootha, V. K. The Mitochondrial Proteome and Human Disease. *Annu. Rev. Genomics Hum. Genet.* **11**, 25–44 (2010).
46. Dunham, W. H., Mullin, M. & Gingras, A. C. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics* **12**, 1576–1590 (2012).
47. Tsiatsiani, L. & Heck, A. J. R. Proteomics beyond trypsin. *FEBS J.* **282**, 2612–2626 (2015).
48. Brownridge, P. & Beynon, R. J. The importance of the digest: Proteolysis and absolute quantification in proteomics. *Methods* **54**, 351–360 (2011).
49. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).
50. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* **11**, 993–1006 (2016).
51. Giansanti, P. *et al.* An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. *Cell Rep.* **11**, 1834–1843 (2015).
52. Shevchenko, A., Tomas, H., Havliš, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2007).
53. Granvogl, B., Plöscher, M. & Eichacker, L. A. Sample preparation by in-gel digestion for mass spectrometry-based proteomics. *Anal. Bioanal. Chem.* **389**, 991–1002 (2007).
54. Proc, J. L. *et al.* A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin. *J. Proteome Res.* **9**, 5422–5437 (2010).
55. Link, A. J. Multidimensional peptide separations in proteomics. *Trends Biotechnol.* **20**, 8–13 (2002).
56. Fritzsche, R., Ihling, C. H., Götze, M. & Sinz, A. Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. *Rapid Commun. Mass Spectrom.* **26**, 653–658 (2012).
57. Yang, F., Shen, Y., Camp, D. G. & Smith, R. D. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev. Proteomics* **9**, 129–134 (2012).
58. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *J. Proteome Res.* **13**, 6176–6186 (2014).
59. Manadas, B., Mendes, V. M., English, J. & Dunn, M. J. Peptide fractionation in proteomics approaches. *Expert Rev. Proteomics* **7**, 655–663 (2010).
60. Xu, P., Duong, D. M. & Peng, J. Systematical optimization of reverse-phase chromatography for shotgun proteomics. *J. Proteome Res.* **8**, 3944–3950 (2009).

61. Cristobal, A. *et al.* In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. *Analyst* **137**, 3541–3548 (2012).
62. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science (80-.).* **312**, 212–217 (2006).
63. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12**, 483–490 (2008).
64. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).
65. Wilm, M. Principles of electrospray ionization. *Mol. Cell. Proteomics* **10**, 1–8 (2011).
66. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int. J. Anal. Chem.* **2012**, 1–40 (2012).
67. El-Aneed, A., Cohen, A. & Banoub, J. Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers. *Appl. Spectrosc. Rev.* **44**, 210–230 (2009).
68. Dreisewerd, K. *The desorption process in MALDI*. *Chemical Reviews* **103**, (2003).
69. Griffiths, J. A brief history of mass spectrometry. *Anal. Chem.* **80**, 5678–5683 (2008).
70. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
71. Hoffmann Edmond de, S. V. Mass Spectrometry: principles and applications, 3rd Edition - Edmond de Hoffmann, Vincent Stroobant. 1–4 (2007).
72. Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. *Mass Spectrom. Rev.* **24**, 1–29 (2005).
73. Schwartz, J. C., Senko, M. W. & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *Proc. 50th ASMS Conf. Mass Spectrom. Allied Top.* **0305**, 117–118 (2002).
74. Makarov, A. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78**, 2113–2120 (2006).
75. Boesl, U. Time-of-flight mass spectrometry: Introduction to the basics. *Mass Spectrom. Rev.* **36**, 86–109 (2017).
76. Makarov, A. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Anal. Chem.* **72**, 1156–1162 (2000).
77. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **85**, 5288–5296 (2013).
78. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8**, 61–80 (2015).
79. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, 1–12 (2011).
80. Frese, C. K. *et al.* Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. *J. Proteome Res.* **10**, 2377–2388 (2011).
81. Potel, C. M., Lemeer, S. & Heck, A. J. R. Phosphopeptide Fragmentation and Site Localization by Mass Spectrometry: An Update. *Anal. Chem.* **91**, 126–141 (2019).
82. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).
83. Jedrychowski, M. P. *et al.* Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol. Cell. Proteomics* **10**, 1–9 (2011).
84. Mikesh, L. M. *et al.* The utility of ETD mass spectrometry in proteomic analysis. *Biochim. Biophys. Acta - Proteins Proteomics* **1764**, 1811–1822 (2006).
85. Frese, C. K. *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.* **84**, 9668–9673 (2012).
86. Frese, C. K. *et al.* Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *J. Proteome Res.* **12**, 1520–1525 (2013).
87. Cottrell, J. S. Protein identification using MS/MS data. *J. Proteomics* **74**, 1842–1851 (2011).
88. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
89. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–67 (1999).
90. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
91. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
92. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: A critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).
93. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. proteomics* **1**, 376–386 (2002).
94. Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7**, 952–958 (2006).
95. Krüger, M. *et al.* SILAC Mouse for Quantitative Proteomics Uncovers Kindlin-3 as an Essential Factor for Red Blood Cell Function. *Cell* **134**, 353–364 (2008).
96. Sury, M. D., Chen, J. X. & Selbach, M. The SILAC fly allows for accurate protein quantification in vivo. *Mol. Cell. Proteomics* **9**, 2173–2183 (2010).
97. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).
98. Schwanhäusser, B., Gossen, M., Dittmar, G. & Selbach, M. Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* **9**, 205–209 (2009).
99. Stewart, I. I., Thomson, T. & Figeys, D. O labeling: A tool for proteomics. *Rapid Commun. Mass Spectrom.* **15**, 2456–2465 (2001).
100. Hsu, J. L., Huang, S. Y., Chow, N. H. & Chen, S. H. Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Anal. Chem.* **75**, 6843–6852 (2003).

101. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
102. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
103. Rauniyar, N. & Yates, J. R. Isobaric labeling-based relative quantification in shotgun proteomics. *Journal of Proteome Research* **13**, 5293–5309 (2014).
104. Louris, J. N. *et al.* Instrumentation, Applications, and Energy Deposition in Quadrupole Ion-Trap Tandem Mass Spectrometry. *Anal. Chem.* **59**, 1677–1685 (1987).
105. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).
106. Wenger, C. D. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* **8**, 933–935 (2011).
107. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).
108. McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84**, 7469–7478 (2012).
109. Zecha, J. *et al.* TMT labeling for the masses: A robust and cost-efficient, in-solution labeling approach. *Mol. Cell. Proteomics* **18**, 1468–1478 (2019).
110. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
111. Braisted, J. C. *et al.* The APEX quantitative proteomics tool: Generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* **9**, 1–11 (2008).
112. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nature Methods* **9**, 555–566 (2012).
113. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6940–6945 (2003).
114. Beynon, R. J., Doherty, M. K., Pratt, J. M. & Gaskell, S. J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2**, 587–589 (2005).
115. Brun, V. *et al.* Isotope-labeled protein standards: Toward absolute quantitative proteomics. *Mol. Cell. Proteomics* **6**, 2139–2149 (2007).
116. Jensen, O. N. Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **8**, 33–41 (2004).
117. Hunter, T. Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2513–2516 (2012).
118. Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **26**, 676–687 (2015).
119. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Muzio, L. Lo. The crucial role of protein phosphorylation in cell signalingand its use as targeted therapy (Review). *Int. J. Mol. Med.* **40**, 271–280 (2017).
120. Fuhs, S. R. & Hunter, T. pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. *Curr. Opin. Cell Biol.* **45**, 8–16 (2017).
121. Thingholm, T. E., Jensen, O. N. & Larsen, M. R. Analytical strategies for phosphoproteomics. *Proteomics* **9**, 1451–1468 (2009).
122. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, 1–19 (2019).
123. Press, A., Press, E., Massoulie, J. & Massoulie, J. Identification of Phosphohistidine Digests from a Probable Intermediate of Qxidative Phosphorylation *. **237**,
124. Macek, B., Mann, M. & Olsen, J. V. Global and Site-Specific Quantitative Phosphoproteomics: Principles and Applications. *Annu. Rev. Pharmacol. Toxicol.* **49**, 199–221 (2009).
125. Li, X. S., Yuan, B. F. & Feng, Y. Q. Recent advances in phosphopeptide enrichment: Strategies and techniques. *TrAC - Trends Anal. Chem.* **78**, 70–83 (2016).
126. Porath, J., Carlsson, J., Olsson, I. & Belfrage, G. Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature* **258**, 598–599 (1975).
127. Ruprecht, B. *et al.* Comprehensive and reproducible phosphopeptide enrichment using iron immobilized metal ion affinity chromatography (Fe-IMAC) columns. *Mol. Cell. Proteomics* **14**, 205–215 (2015).
128. Ruprecht, B. *et al.* Optimized enrichment of phosphoproteomes by Fe-IMAC column chromatography. in *Methods in Molecular Biology* **1550**, 47–60 (Humana Press Inc., 2017).
129. Potel, C. M., Lin, M. H., Heck, A. J. R. & Lemeer, S. Defeating major contaminants in Fe 3- immobilized metal ion affinity chromatography (IMAC) phosphopeptide enrichment. *Mol. Cell. Proteomics* **17**, 1028–1034 (2018).
130. Pinkse, M. W. H., Uitto, P. M., Hilhorst, M. J., Ooms, B. & Heck, A. J. R. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal. Chem.* **76**, 3935–3943 (2004).
131. Zhou, H. *et al.* Robust phosphoproteome enrichment using monodisperse microsphere-based immobilized titanium (IV) ion affinity chromatography. *Nat. Protoc.* **8**, 461–480 (2013).
132. Beausoleil, S. A. *et al.* Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12130–12135 (2004).
133. Trinidad, J. C., Specht, C. G., Thalhammer, A., Schoepfer, R. & Burlingame, A. L. Comprehensive identification of phosphorylation sites in postsynaptic density preparations. *Mol. Cell. Proteomics* **5**, 914–922 (2006).
134. Olsen, J. V. *et al.* Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127**, 635–648 (2006).
135. Beltran, L. & Cutillas, P. R. Advances in phosphopeptide enrichment techniques for phosphoproteomics. *Amino Acids* **43**, 1009–1024 (2012).

136. Han, G. *et al.* Large-scale phosphoproteome analysis of human liver tissue by enrichment and fractionation of phosphopeptides with strong anion exchange chromatography. *Proteomics* **8**, 1346–1361 (2008).
137. McNulty, D. E. & Annan, R. S. Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. *Mol. Cell. Proteomics* **7**, 971–980 (2008).
138. Boersema, P. J. *et al.* In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. *Mol. Cell. Proteomics* **9**, 84–99 (2010).
139. Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23**, 94–101 (2005).
140. Grønborg, M. *et al.* A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. *Mol. Cell. Proteomics* **1**, 517–527 (2002).
141. Steen, H., Kuster, B., Fernandez, M., Pandey, A. & Mann, M. Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway. *J. Biol. Chem.* **277**, 1031–1039 (2002).
142. Potel, C. M. *et al.* Gaining Confidence in the Elusive Histidine Phosphoproteome. *Anal. Chem.* **91**, 5542–5547 (2019).
143. Ferries, S. *et al.* Evaluation of Parameters for Confident Phosphorylation Site Localization Using an Orbitrap Fusion Tribrid Mass Spectrometer. *J. Proteome Res.* **16**, 3448–3459 (2017).
144. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
145. Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **10**, 5354–5362 (2011).
146. Savitski, M. M. *et al.* Confident phosphorylation site localization using the mascot delta score. *Mol. Cell. Proteomics* **10**, (2011).
147. Fermin, D., Walmsley, S. J., Gingras, A. C., Choi, H. & Nesvizhskii, A. I. LuciPHOr: Algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol. Cell. Proteomics* **12**, 3409–3419 (2013).
148. Hogrebe, A. *et al.* Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* **9**, (2018).
149. Butterfield, D. A., Gu, L., Domenico, F. Di & Robinson, R. A. S. Mass spectrometry and redox proteomics: Applications in disease. *Mass Spectrom. Rev.* **33**, 277–301 (2014).
150. Ray, P. D., Huang, B. W. & Tsuji, Y. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell. Signal.* **24**, 981–990 (2012).
151. Hancock, J. T. The role of redox mechanisms in cell signalling. *Mol. Biotechnol.* **43**, 162–166 (2009).
152. Blanchetot, C., Tertoolen, L. G. J. & Den Hertog, J. Regulation of receptor protein-tyrosine phosphatase α by oxidative stress. *EMBO J.* **21**, 493–503 (2002).
153. Böhmer, F., Szedlacsek, S., Tabernero, L., Östman, A. & den Hertog, J. Protein tyrosine phosphatase structure-function relationships in regulation and pathogenesis. *FEBS J.* **280**, 413–431 (2013).
154. Go, Y. M., Chandler, J. D. & Jones, D. P. The cysteine proteome. *Free Radic. Biol. Med.* **84**, 227–245 (2015).
155. Jortzik, E., Wang, L. & Becker, K. Thiol-based posttranslational modifications in parasites. *Antioxidants and Redox Signaling* **17**, 657–673 (2012).
156. Duan, J., Gaffrey, M. J. & Qian, W. J. Quantitative proteomic characterization of redox-dependent post-translational modifications on protein cysteines. *Mol. Biosyst.* **13**, 816–829 (2017).
157. Wojdyla, K. & Rogowska-Wrzesinska, A. Differential alkylation-based redox proteomics - Lessons learnt. *Redox Biol.* **6**, 240–252 (2015).
158. Zhou, X. *et al.* ESNOQ, Proteomic Quantification of Endogenous S-Nitrosation. *PLoS One* **5**, e10015 (2010).
159. Yang, J. *et al.* Global, in situ, site-specific analysis of protein S-sulfenylation. *Nat. Protoc.* **10**, 1022–1037 (2015).
160. Pan, K.-T. *et al.* Mass Spectrometry-Based Quantitative Proteomics for Dissecting Multiplexed Redox Cysteine Modifications in Nitric Oxide-Protected Cardiomyocyte Under Hypoxia. *Antioxid. Redox Signal.* **20**, 1365–1381 (2014).

Chapter 2

Thermal proteome profiling in zebrafish reveals effects of napabucasin on retinoic acid metabolism

Niels M. Leijten[1*], Petra Bakker[2,3*], Herman P. Spaink[3], Jeroen den Hertog[2,3#] and Simone Lemeer[1#]

1. Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

2. Hubrecht Institute – KNAW and University Medical Center Utrecht, Utrecht, the Netherlands

3. Institute Biology Leiden, Leiden University, Leiden, the Netherlands

* These authors contributed equally

Published in Molecular & Cellular Proteomics (2021), mcp.RA120.002273

Abstract

Thermal proteome profiling (TPP) allows for the unbiased detection of drug – target protein engagements *in vivo*. Traditionally, one cell type is used for TPP studies, with the risk of missing important differentially expressed target proteins. The use of whole organisms would circumvent this problem. Zebrafish embryos are amenable to such an approach. Here, we used TPP on whole zebrafish embryo lysate to identify protein targets of napabucasin, a compound that may affect Signal transducer and activator of transcription 3 (Stat3) signaling through an ill-understood mechanism. In zebrafish embryos, napabucasin induced developmental defects consistent with inhibition of Stat3 signaling. TPP profiling showed no distinct shift in Stat3 upon napabucasin treatment, but effects were detected on the oxidoreductase Pora, which might explain effects on Stat3 signaling. Interestingly, thermal stability of several aldehyde dehydrogenases (Aldhs) was affected. Moreover, napabucasin activated ALDH enzymatic activity *in vitro*. Aldhs have crucial roles in retinoic acid metabolism and functionally we validated napabucasin-mediated activation of the retinoic acid pathway in zebrafish *in vivo*. We conclude that TPP profiling in whole zebrafish embryo lysate is feasible and facilitates direct correlation of *in vivo* effects of small molecule drugs with their protein targets.

Introduction

The thermal stability of proteins is influenced by their interaction with other (small) molecules[1]. In recent years, large scale studies using thermal proteome profiling (TPP) facilitated analysis of protein-protein, protein-metabolite and protein-drug interactions, based on the changing thermostability of proteins[2]. Especially in the field of drug discovery, the use of TPP facilitated the identification of new off-targets of clinically relevant drugs[3,4,5,6].

Typically, a single cell type or - more recently - a single tissue[6] is being used as source of proteins. By using this approach, one may miss important drug targets, because proteins are differentially expressed in cells and tissues[7] and the target proteins may not be expressed in the cells or tissues analyzed. Therefore, it would be interesting to perform TPP at an organism-wide scale, to take all possible target proteins and their interactions in different tissues and cell types into account and obtain global insight into the mode of action of the drug-target interaction.

The zebrafish (*D. rerio*) is an ideal model system, since its genome is very similar to the human genome[8] and orthologues of 84% of human disease associated genes have been identified. Furthermore, external embryonic development is rapid with practically all organs developed at 2 days post fertilization (dpf). The embryos are transparent, allowing imaging of development and genetic manipulation of the zebrafish genome is relatively easy. Finally, the zebrafish model system is amenable to drug screens, in that treatment of developing zebrafish embryos with small molecule compounds may induce specific phenotypic changes[9]. Because of these assets of the zebrafish, we hypothesized that zebrafish is an ideal model system for TPP on whole organisms.

To test the hypothesis that zebrafish embryos may be used to identify drug targets by TPP, we set out to identify protein targets of the small molecule napabucasin (BBI608). Napabucasin, a naphthoquinone, blocks stem cell activity in cancer cells and is being tested in multiple clinical trials as a mono or combination therapy[10]. Napabucasin is thought to inhibit STAT3 signaling through a yet ill understood mechanism[11]. In response to stimuli, cytosolic STAT3 is phosphorylated on Tyr705 by upstream tyrosine kinases[12], after which the phosphorylated protein homo-dimerizes due to an interaction between the phosphorylated tyrosine of one protein and the SH2 domain of another. The dimerized complex translocates to the nucleus, binds specific DNA fragments and induces gene transcription. STAT3 activation has a central role in many biological processes. Transient STAT3 activation mediates wound healing and tissue integrity[13], and persistent STAT3 activation promotes tumor cell proliferation, survival, invasion and immunosuppression[14]. Treatment of cells with napabucasin leads to a decrease in the amount of phosphorylated STAT3 and hence, suppression of STAT3 signaling[15,16,17].

Here, we report TPP at an organism-wide scale for the first time (Fig. 1). Using 5dpf zebrafish embryo lysates, targets of napabucasin were identified. Whereas napabucasin induced similar developmental defects as Stat3 knockdown, Stat3 did not show a stabilization shift, suggesting that Stat3 is not a direct target. However, several members of the aldehyde dehydrogenase (Aldh) family showed a thermal shift. Napabucasin enhanced aldehyde dehydrogenase activity and we demonstrate that at least part of the effect of napabucasin on zebrafish embryonic development may be explained by its effect on retinoic acid metabolism via activation of Aldhs.

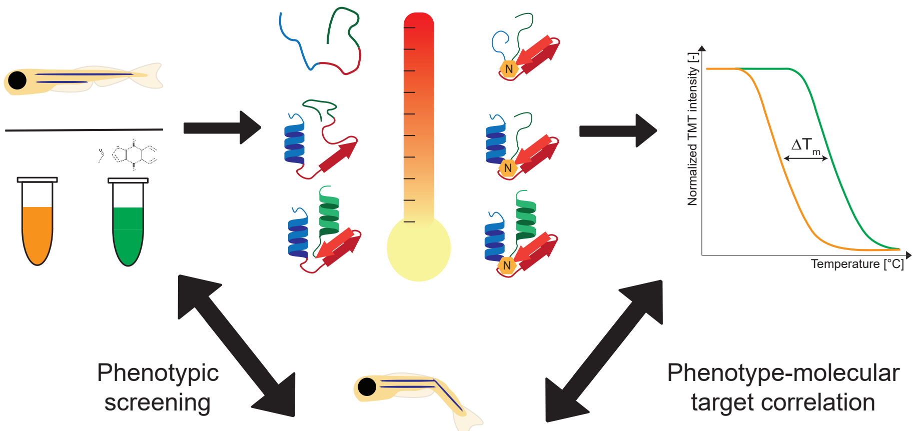

N
Normalized TMT intensity [-]
ΔT$_m$
Temperature [°C]
Phenotypic screening
Phenotype-molecular target correlation

Figure 1: Graphical representation of the workflow for zebrafish thermal proteome profiling. Zebrafish embryo lysates were treated with either DMSO (control) or napabucasin before being subjected to heat treatment. If the drug binds to a protein, the complex can be stabilized leading to an increase in melting point. By quantification of the remaining soluble protein amount, melting curves can be generated. Phenotypic screening of zebrafish embryos using the same drug was done in parallel, which allows for direct correlation of the biological effects to the molecular drug targets.

Materials and methods

Animal husbandry

Adult zebrafish were maintained and embryos were collected following natural mating as previously described[18,19]. All procedures involving experimental animals were approved by the local animal experiments committee (Koninklijke Nederlandse Akademie van Weterschappen-Dierexperimenten commissie) and performed according to local guidelines and policies in compliance with national and European law.

Zebrafish embryo assays

Zebrafish embryos were injected at the one-cell stage with antisense Stat3 morpholino as described before[20]. Alternatively, embryos were treated with napabucasin (5 -10 μM), RA (10^{-10} - 10^{-8} M), Cyp26 inhibitor R115866 (1 -10 μM) or control (0.2% DMSO). The embryos were imaged at 28 hpf to determine the distance from the otic vesicle to the tip of the nose using Image J. Alternatively, the embryos were fixed at 10.5 hpf or 18 hpf for *in situ* hybridization using *ntl*-specific[21] or *krox20/myoD*-specific probes[22,23], respectively. The embryos were imaged and the length of the *ntl*-stained notochord was determined using Image J.

Experiment design and Statistical Rationale

A total of 560 D. rerio embryos were pooled and lysed. From this pooled lysate 4 replicates were taken, which were either treated with vehicle (2 samples) or drug of interest (2 samples). Samples were divided in ten aliquots until TMT labelling, subsequently pooled and divided in 10 fractions by high pH fractionation. All these fractions were injected separately in the LC-MS/MS system and the following raw files were processed by MaxQuant.

Thermal profiling, digestion, TMT-labelling and fractionation

Zebrafish embryos (5 days post fertilization) were pooled and snap frozen. Embryos were reconstituted in lysis buffer (0.4% NP-40 in PBS + complete mini EDTA free protease inhibitors (Roche)) and lysed using bead-beating (Digital Disruptor Genie, Scientific industries) with zirconium oxide beads (1 mm). Subsequently, lysates were subjected to sonication (10 cycles, 30 secs on/off) (Diagenode). Lysates were centrifuged for 30 min at 14,000 rcf and 4 °C to remove insoluble debris, after which the supernatant was transferred to a new tube and used for thermal proteome profiling.

Thermal proteome profiling was performed as previously described[2]. Briefly, protein concentration was determined by a BCA protein assay kit (Thermo Fisher Scientific). The concentration of the lysate was adjusted to 2 mg/ml and divided in control and treatment. For the pervanadate experiment, lysates were treated with either PBS (control) or 100 µM pervanadate[24] for 20 minutes at room temperature. For the napabucasin (Bio-connect) experiment, lysates were incubated for 20 minutes with either DMSO (control) or 50 µM napabucasin (treatment). After treatment, lysates were divided in ten 100 µl aliquots. Heat treatment was performed in a PCR cycler (T100 thermal cycler, Bio Rad) for three minutes, cooled down to 25 °C for three minutes and afterwards placed on ice. The temperature range spanned from 34 to 64 degrees, with increments of 3.3 °C. Precipitates were removed by ultracentrifugation (Beckman Coulter) at 125,000 rcf for 1 hour at 4 °C. A volume corresponding to 100 µg of protein at the lowest temperature point was taken for further processing. Samples were reduced and alkylated by incubation with respectively 10 and 40 mM of tris(2-carboxyethyl)phosphine (TCEP) and chloroacetamide (CAA) for 15 minutes. Afterwards, samples were subjected to methanol/chloroform precipitation. The supernatant was removed and the protein pellet was air dried. The protein pellet was dissolved in 6M urea in PBS, after which a second reduction/alkylation step was performed using 10 mM TCEP and 40 mM CAA for 15 minutes on the shaker. Samples were predigested with 1:100 LysC (protein: protease ratio) (Wako) for 2 hours at 37 °C on a shaker. After predigestion, samples were further diluted to 1.5 M urea using 50 mM TEAB buffer (pH 8.5). Trypsin (Sigma) was added to a 1:100 ratio and samples were digested overnight at 37 °C on a shaker. Digestion was stopped by acidifying to pH 2 by adding formic acid. Samples were centrifuged at 20,000 rcf for 10 minutes at 4 °C before they were desalted using 1cc SEPPAK SPE cartridges (Waters). Briefly, Cartridges were washed 3 times with 1 ml acetonitrile followed by washing three times with 1 ml of 0.1 M acetic acid. Samples were loaded after which the flow through was passed through the cartridges again. Cartridges were washed three times with 1 ml 0.1 M acetic acid after which the peptides were eluted using three times 250 µl 0.1 M acetic acid/80% acetonitrile. Samples were dried in a Thermo Savant SPD SpeedVac (Thermo Fisher Scientific).

TMT labelling was performed as previously described[25]. Briefly, protein digests were dissolved in 40 µl of 50 mM HEPES (pH 8.5) and mixed for 10 minutes at 20 °C. TMT reagents were dissolved in 42 µl 100% anhydrous acetonitrile, after which 10 µl of this solution was added to the peptides. Labelling was performed for 1 hour at 20 °C while shaking at 400 rpm. The reaction was stopped by adding hydroxylamine to a final concentration of 0.4% and incubating for 15 minutes at 20 °C and 400 rpm. Subsequently, samples were pooled and acidified to pH 2 using formic acid. Samples were desalted using 1cc SEPPAK SPE cartridges as described before. After desalting samples were dried using speedvac. Sample pools were fractionated using high pH reverse-phase HPLC fractionation using a Kinetex 5u EVO C18 100A column (Phenomenex) on a HPLC 1200 system (Agilent) operating at a flow rate of 200 µl/min. Briefly, Dried pellet was reconstituted in 20 µl of buffer A (10 mM NH4OH, pH 10) and injected. Samples were first loaded on the column at a flow rate of 20 µl/min for 2 minutes. Peptides were eluted stepwise using the following gradient: 2 – 12 % buffer B (10 mM NH4OH/ 90% acetonitrile, pH 10) in 6 minutes, 12 – 35% buffer B in 47 minutes, 35 - 55% buffer B in 7 minutes, 55 - 100% buffer B in 3 minutes, 0 - 100% buffer A in 9 minutes, 100% buffer A for 31 minutes. A total gradient time of 105 minutes was used. Fractions corresponding to 1 minute of gradient time were collected on a 1260 infinity fraction collector (Agilent). Only fractions eluting after 8 minutes were collected. These fractions were concatenated in 10 fractions. All the fractions were dried down using speedvac and stored at -80 °C until further use.

LC-MS/MS analysis

For the TPP analysis, peptides were dissolved in 10% formic acid and a volume corresponding to 2 µg of peptides was injected on a UHPLC 1290 system (Agilent) coupled to a Q Exactive HF-X mass spectrometer (Thermo Fisher scientific). Peptides were trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) before being separated using an analytical column (Agilent Poroshell EC-C18, 2.7 µm, 50 cm x 75 µm). Trapping was performed for 5 minutes in buffer A (0.1% formic acid) at a flow rate of 0.005 ml/min. The following gradient was used for separation: 12 - 42% buffer B (80% acetonitrile + 0.1% formic acid) in 95 minutes, 100% buffer B for 2 minutes followed by 100% buffer A for 11 minutes. The flow was split to generate a final flow of 300 nl/min. The Q Exactive HF-X was operated in a data dependent acquisition mode with positive ionization. Full MS spectra were acquired from 375-1500 m/z at 60000 resolution, using an automatic gain control (AGC) target value of 3 x 10^6 charges and a maximum injection time of 20 ms. A maximum of 12 precursors were allowed to be fragmented. A dynamic exclusion of 18 seconds was used. MS2 fragmentation spectra were obtained with a fixed first mass of 120 m/z at 45000 resolution, using an AGC target of 1 x 10^5 and a maximum injection time of 85 ms. Fragmentation was performed using HCD at a NCE of 32.

Protein identification and quantification

In order to identify peptides and proteins, raw files were processed using MaxQuant (version 1.6.5.0) and the Andromeda search engine, using the full Trembl database for zebrafish (55769 entries, downloaded 22-07-2019). The following parameters were used: digestion by trypsin/P with a maximum of 2 missed cleavages, carbamidomethylation of cysteine as a fixed modification, oxidation of methionine and N-terminal acetylation were selected as variable modifications. TMT 10plex labelling was used for quantification. The mass tolerance of precursor ions was chosen as ±5 ppm and the mass tolerance of MS/MS was chosen as ±20 ppm. Results were adjusted to 1% PSM and 1% FDR using a target-decoy approach using reverted protein sequences.

Melting points were determined as previously described[2]. In the first step the relative abundances of the TMT reporter ions compared to the lowest temperature point were calculated. The lowest temperature point was set to 1. The experiments were normalized using the TPP R script[2] and melting curves were fitted according to the chemical denaturation theory:

$$\boldsymbol{f(T) = \frac{1 - plateau}{1 + e^{-\left(\frac{a}{T} - b\right)}} + plateau}$$

In this equation T is the temperature and a, b and plateau are constants. The melting point of a proteins is determined as the temperature where half of the protein has denatured: f (T) = 0.5. The generated melting curves were inspected for a change in melting behavior. All melting curves shown were generated in GraphPad prism (8.3.0).

The generated melting curves were checked for significant difference by use of the NPARC method developed by Childs *et al*[26], where we determined curves to be different if $P \leq 0.01$. Additionally, we applied two more filters: I) the melt point differences between vehicles and controls must have the same sign and II) the difference between the melting points of vehicle and control must be bigger than the difference between melting points of both vehicles. The remaining hits were manually screened.

Sequence alignment

All pair wise sequence alignments were performed with the EMBOSS Needle algorithm[27].

ALDH activity assay

Aldehyde dehydrogenase activity colorimetric assay kit (Sigma-Aldrich) was performed according to manufacturer's protocol. Briefly, HepG2 cells (30 million) were resuspended in ALDH assay buffer before being lysed by freeze/thaw cycling. The lysate was treated with 1% DMSO (control), 50/100 µM Napabucasin or 20/50/100 µM ALDA-1 for 30 minutes before the colorimetric assay was started. Absorbance was measured at 450 nm on a Multiskan GO plate reader (Thermo Scientific), which was subsequently converted to the ALDH activity.

Data availability

The datasets reported in this paper have been deposited in the ProteomeXchange Consortium via the PRIDE[28] partner repository PXD017418 (pervanadate) and PDX017419 (napabucasin). Mass labelled MS/MS spectra were supplied to MS viewer[27] with search keys **lea6ryrs0n** (pervanadate) and **bbxsm2el6a** (napabucasin).

Results

Optimizing thermal proteome profiling for zebrafish embryos

Zebrafish embryo development is well underway at 5 days post fertilization, and most organs and tissues have formed, including cartilage. Accordingly, the original TPP protocol[2,3] was adjusted to make it amenable to zebrafish. Lysis was performed by bead beating followed by sonication, which was necessary to disrupt the embryo. The lysis buffer contained the mild detergent NP-40 to allow the generation of melting curves for membrane proteins[29]. In this first attempt at organism-wide TPP, the workflow was performed on lysate to keep the complexity of analysis as low as possible. By way of a positive control, we anticipated that pervanadate, a strong oxidative agent, would have a global effect on the thermal stability of the proteome. The lysates were treated with either 100 µM pervanadate or PBS as negative control and incubated for 20 minutes at RT before they were subjected to heat treatment. Zebrafish normally live at 28 °C. Human cell lines that have been used to date for TPP grow at 37 °C. To avoid potential problems resulting from this temperature difference, a broad range of temperatures was tested for TPP on zebrafish lysates. A temperature range of 34 to 64 °C is optimal for the generation of melting curves of zebrafish embryo lysates (Supplementary Fig. 1).

Our pilot study resulted in the identification of a total of 6592 proteins, of which 5159 had at least 2 unique peptides (Supplementary Table 1, 2). Differential melting behavior was determined through the non-parametric analysis of response curves (NPARC) method[26]. The global analysis of the melting behavior of proteins shows a small global increase in protein stability in the pervanadate treated samples (Supplementary Fig. 2) and indicates an infliction point at around 50 °C where the majority of proteins start to precipitate. Distinct classes of proteins showed a shift, including multiple ATPases and proteins involved in the citric acid cycle and glycolysis (Supplementary Fig. 2). All melting characteristics of the identified proteins can be found in Supplementary Table 3. Taken together, our results showcase the validity of thermal proteome profiling on total zebrafish embryo lysates.

Probing the targets of Napabucasin by Thermal Proteome Profiling in zebrafish embryos

Next, we sought to probe the feasibility of using zebrafish TPP with a more selective inhibitor. Napabucasin has an effect on stemness of cancer cells and may act through inhibition of STAT3 signaling[11]. Liu *et al*[30] have shown that zebrafish embryos lacking functional Stat3 display a shortened notochord compared to wildtype embryos, which is caused by reduced cell proliferation and increased apoptosis. We tested napabucasin (5 µM) treatment on zebrafish embryos for phenotypic changes, using *ntl in situ* hybridization staining as read-out for a shortened notochord. As a control, morpholino-based antisense oligonucleotide knockdown of Stat3 was used, which induced a shortened notochord, comparable to the genetic knockout. Treatment with DMSO (solvent) was used as a negative control. Treatment with napabucasin induced significant shortening of the notochord, comparable to the effect of morpholino-mediated Stat3 knockdown (Fig. 2A, B). These results suggest that napabucasin may have a direct effect on Stat3 signaling and thus affect zebrafish embryonic development.

A TPP experiment was performed to identify direct protein targets of napabucasin. Lysates from 5 dpf zebrafish embryos were treated with 50 µM napabucasin or DMSO as a control for 20 min, prior to heat treatment. 7646 proteins were identified, of which 6114 proteins with at least 2 unique peptides, making this the largest proteomics dataset for zebrafish till date (Supplementary Table 4, 5). The experiment captured a large dynamic range of proteins: high abundant proteins such as the egg yolk protein vitellogenin (vtg1) and muscle components titin (ttnb) and myosin (myhz1.2) were detected (Fig. 2F). At the same time, low abundant proteins such as the protein deacetylase sirtuin 1 (sirt1), heterochromatin component cbx5, but also the ciliar protein bbs7 are detected. This showcases the potential of detecting proteins with intensities spanning multiple orders of magnitude using this protocol. Differential melting behavior was determined through the non-parametric analysis of response curves (NPARC) method[26]. Heat maps displaying the global profile show that overall thermal behavior between DMSO and napabucasin treated samples was highly similar (Fig. 2C, D). Comparison of the distribution of melting points indicates that only a small population of proteins showed a different melting behavior after drug treatment (Fig. 2E). All melting characteristics of the identified proteins can be found in Supplementary Table 6. We selected multiple proteins as examples and found that the change in thermal stability of these proteins is distinct between napabucasin and pervanadate treatment (Supplementary Fig. 3).

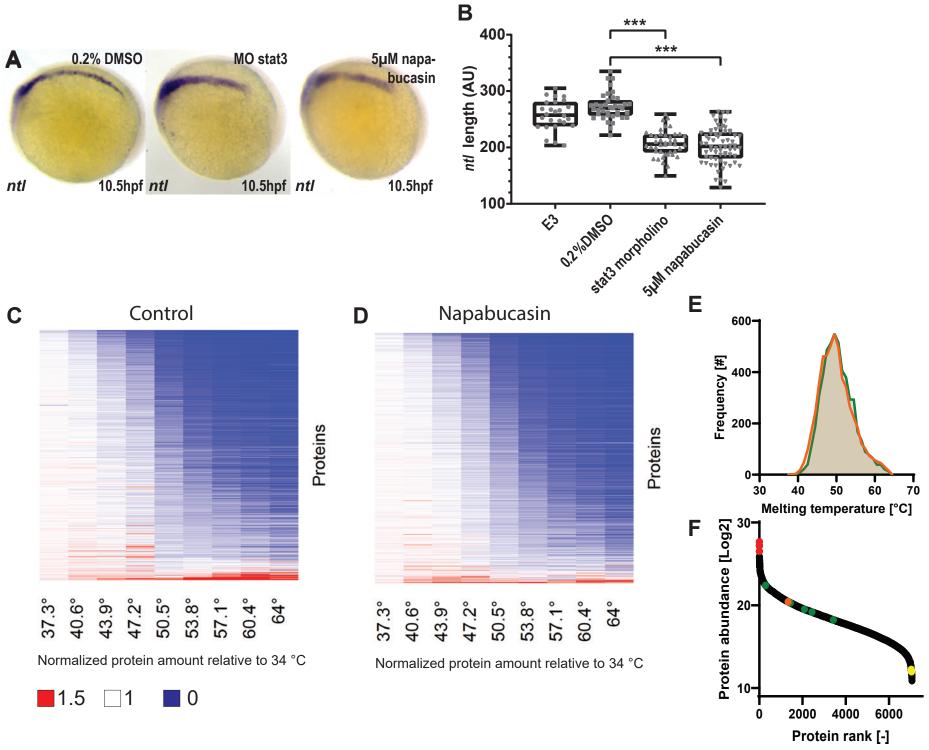


Figure 2: Thermal proteome profiling of Napabucasin in zebrafish embryos.
Embryos were treated with napabucasin (5 µM) or solvent control (0.2 % DMSO). Alternatively, antisense STAT3-specific morpholino (MO stat3) was microinjected at the one cell stage. At 10.5 hpf, the embryos were fixed and in situ hybridization was performed using a *ntl*-specific probe, which stains the notochord (A). The length of the notochord was determined in napabucasin treated and STAT3-morpholino knockdown embryos, compared to DMSO and non-treated E3 medium control. Significance was determined using a one-way ANOVA with Dunnetts multiple comparisons test, n = 26 - 66 embryos per condition; *** =P < 0.001 (B). The global precipitation behavior of proteins in DMSO (C) or napabucasin treated (D) lysate is similar, in agreement with overall comparable melting points for proteins under these two conditions (E). The protocol captured a large dynamic range of proteins, ranging from high abundant proteins such as vtg1, ttnb and myhz1.2 (red) to low abundant proteins such as sirt1, cbx5 and bbs7 (yellow). Additionally, STAT3 (orange) and all shifting ALDH proteins (green) are shown. (F).

No shift in Stat3 and only a minimal non-significant shift in Stat5 were observed (Fig. 3A, B). Stat3 was readily detected with high abundance (Fig. 2F) and good sequence coverage (37.3%), indicating that the absence of shift is likely due to the absence of interaction between protein and drug. However, an interesting class of proteins showed a stabilizing effect due to napabucasin, the aldehyde dehydrogenases

(Aldhs) (Fig. 3C, D). These stabilized aldehyde dehydrogenases are detected with high abundances, indicating high data quality (Fig. 2F). All Aldhs that were detected in this assay can be found in Supplementary Fig. 4A. Interestingly, the stabilized Aldh proteins show a high degree of conservation with each other, whereas conservation with non-shifting Aldhs is limited (Supplementary Fig. 4B). Aldh proteins play a role in converting endo- and exogenous aldehydes to the corresponding carboxylic acids[31], and one of the most important *in vivo* substrates of Aldh enzymes is retinaldehyde, which is converted into retinoic acid (RA), an essential morphogen in vertebrate development. Pora, an oxidoreductase, was also destabilized by napabucasin, whereas another oxidoreductase, NQO1, did not shift (Fig. 3E, F). It is noteworthy that napabucasin is a substrate of oxidoreductases[17]. Altered thermal stability of Pora, but not NQO1, may reflect direct interaction of napabucasin with Pora, but not NQO1.

We investigated the interaction between napabucasin and ALDH proteins by analysis of human ALDH enzymatic activity. ALDH expression is especially high in liver cells and therefore we assessed the effect of napabucasin on ALDH in lysate of HepG2 human liver cells in a colorimetric assay. Surprisingly, napabucasin had a significant activating effect on ALDH enzymatic activity. As a control, we used the known ALDH2 activator ALDA-1[32], which also showed elevated ALDH activity in HepG2 cell lysates (Fig. 4A, B). Taken together, these data show that napabucasin interacts with human ALDHs, thereby activating their activity.

Sequence alignment of zebrafish and human ALDH proteins reveals high sequence conservation between zebrafish retinaldehyde converting Aldhs (aldh1a2 and aldh1a3) and the human retinaldehyde converting ALDHs, providing a proper explanation for an interaction of napabucasin with human ALDHs (Fig. 4C). Additionally, the retinaldehyde converting Aldhs show a large conservation with human acetaldehyde converting ALDH2, explaining the results of the colorimetric assay on HepG2 cells, which used acetaldehyde as substrate. It is noteworthy that the zebrafish Aldh2.1 also shows a stabilizing effect due to napabucasin, indicating an interaction, however due to large conservation only a single unique peptide for this protein could be detected. These results combined prove the use of our zebrafish model as a platform to study drug function in humans.

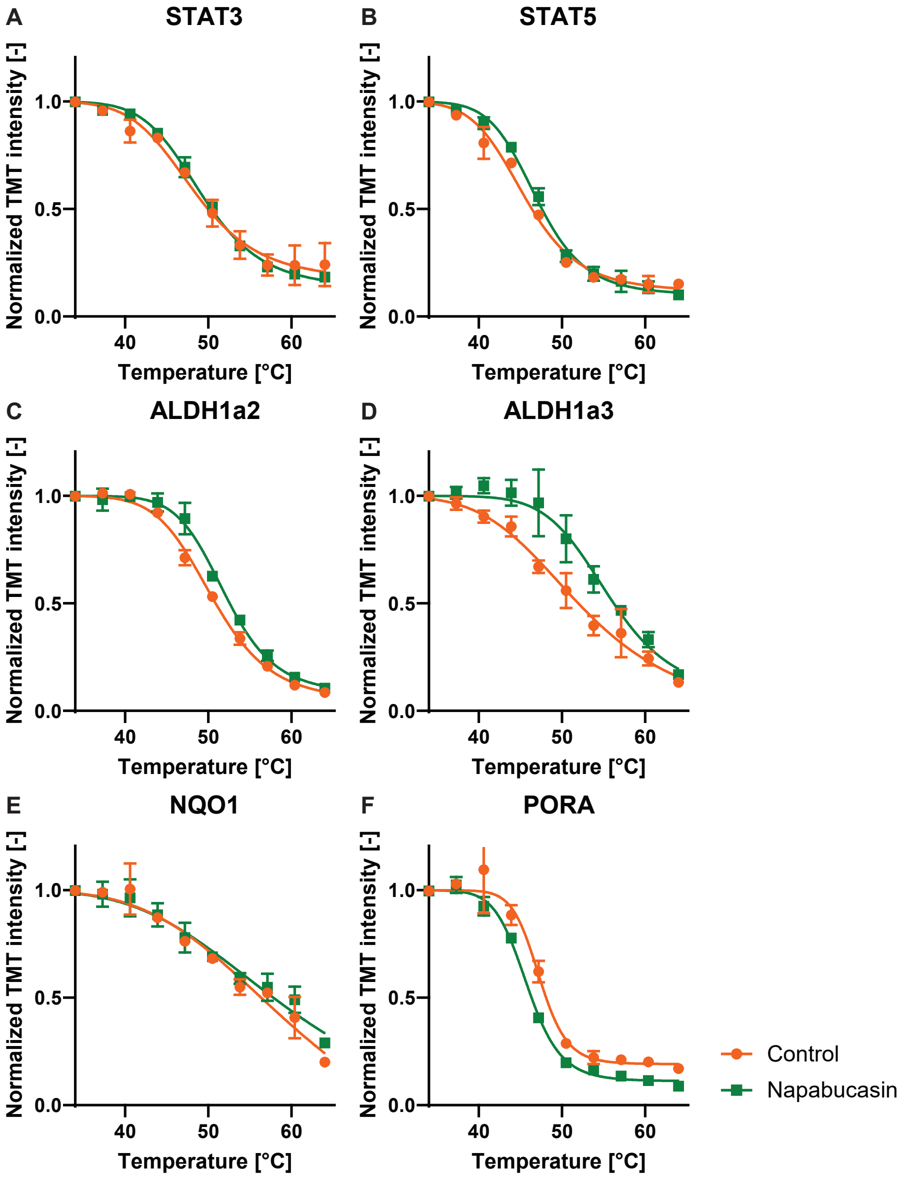


Figure 3: TPP indicates differential effects of napabucasin treatment on selected proteins. STAT3 (A) does not show any shift, while STAT5 (B) shows a small shift. However, a shift was detected in ALDH1a2 (C) and ALDH1a3 (D). The known interactor of napabucasin NQO1 (E) shows no shift. Interestingly, another known interactor, PORA, (F) shows a shift indicating destabilization. Data points (n = 2 independent experiments) are shown as mean ± SEM, melting curve fitting was performed according to chemical denaturation theory.

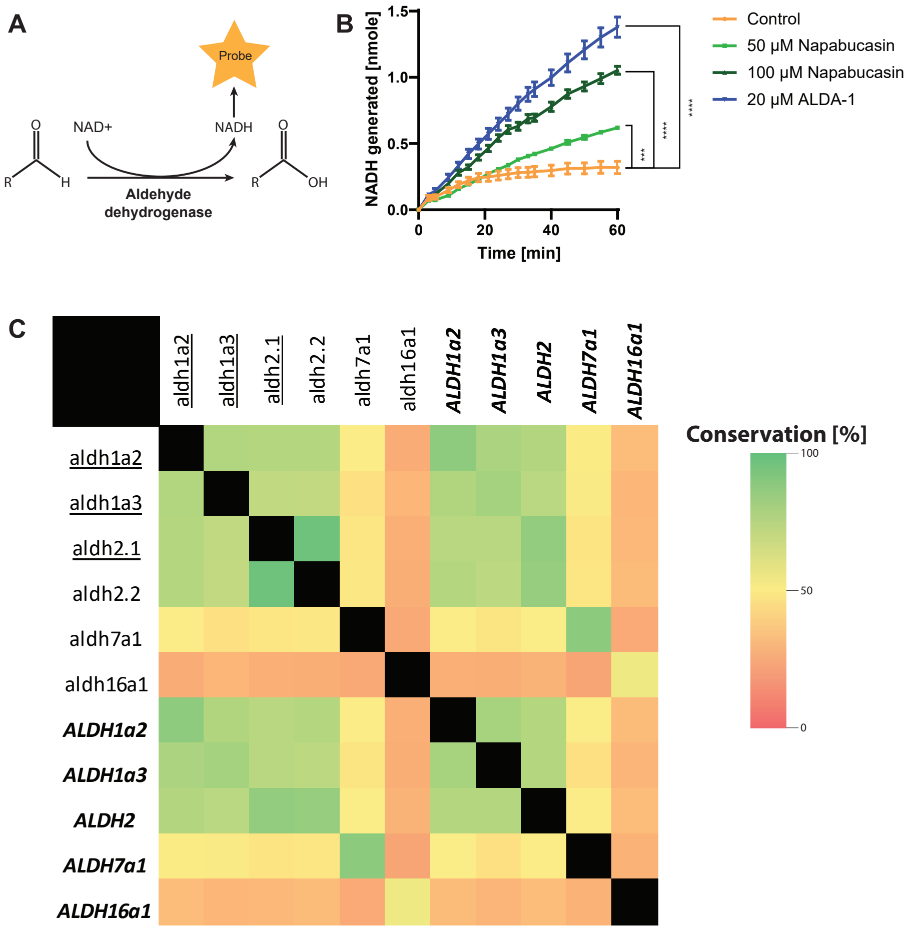


Figure 4: Activation and conservation of human ALDHs. The ALDH assay measures the conversion of NAD+ to NADH using a colorimetric probe (A). ALDH enzymatic activity was determined in lysates of HepG2 cells. Two concentrations of napabucasin were used (50 µM and 100 µM) as well as a known ALDH activator, ALDA-1 (20 µM). 1% DMSO was used as a control and experiments were performed in triplicate. P*** < 0.001, P**** < 0.0001. Significance was determined by one-way ANOVA (B). The shifting (underlined) zebrafish aldh1a2, aldh1a3 and aldh2.1 show a large degree of conservation with their human counterparts (shown in italic and bold), but show no large conservation with other zebrafish or human Aldh enzymes (C).

Validation of the interaction of napabucasin with ALDHs by using a zebrafish phenotypic model

The Aldh family of enzymes has a crucial role in retinoic acid metabolism *in vivo*[33]. Retinoic acid is an essential morphogen in vertebrate development and regulates cellular processes, embryo patterning and organogenesis. Enhancing retinoic acid levels by exogenous addition of retinoic acid or by inhibition of retinoic acid catabolizing enzymes, such as Cyp26, results in severe developmental defects, most prominently in development and patterning of anterior neural structures. In zebrafish embryos, treatment with exogenous retinoic acid induces defects in hindbrain development, resulting in a significant reduction of the distance from the otic vesicle to the tip of the nose[34]. Treatment with exogenous RA and inhibition of Cyp26 was reported to interfere with normal development of rhombomeres 3 and 5[35,36,37]. Since napabucasin enhanced Aldh enzymatic activity, we hypothesized that napabucasin treatment would induce similar developmental defects as treatment with exogenous retinoic acid and/or inhibition of Cyp26. Treatment with 10^{-8} M RA induced a significant reduction of the distance between the otic vesicle and the tip of the nose, whereas treatment with 10^{-9} M or 10^{-10} M did not (Fig. 5A, B, Supplementary Fig. 5). Napabucasin treatment (10 µM) by itself did not induce significant defects. Interestingly, combined treatment of embryos with 10 µM napabucasin and 10^{-9} M RA induced a significant reduction in the distance between the otic vesicle and the tip of the nose, suggesting that napabucasin and RA cooperate to induce these defects. Note that at these concentrations, napabucasin and RA by themselves did not induce developmental defects. Higher concentrations of napabucasin induced severe developmental defects, which precluded assessment of the distance between the otic vesicle and the tip of the nose.

Inhibition of Cyp26 with R115866 induced defects in the development of rhombomeres 3 and 5 in a dose-dependent manner, assessed by *krox20 in situ* hybridization (Fig. 5C-E). The myoD-specific marker was included to mark the somites. Napabucasin (10 µM) by itself did not affect rhombomere development, but co-treatment of embryos with 1 µM R115866, which did not affect rhombomere development by itself, did induce defects (Fig. 5F, G), indicating that napabucasin and R115866 cooperate. Napabucasin enhanced Aldh enzymatic activity, resulting in enhanced RA production and R115866 reduced Cyp26 activity, which also results in enhanced RA levels (Fig. 5H). Our results suggest that at least part of the *in vivo* function of napabucasin is mediated by elevation of RA levels in zebrafish embryos.

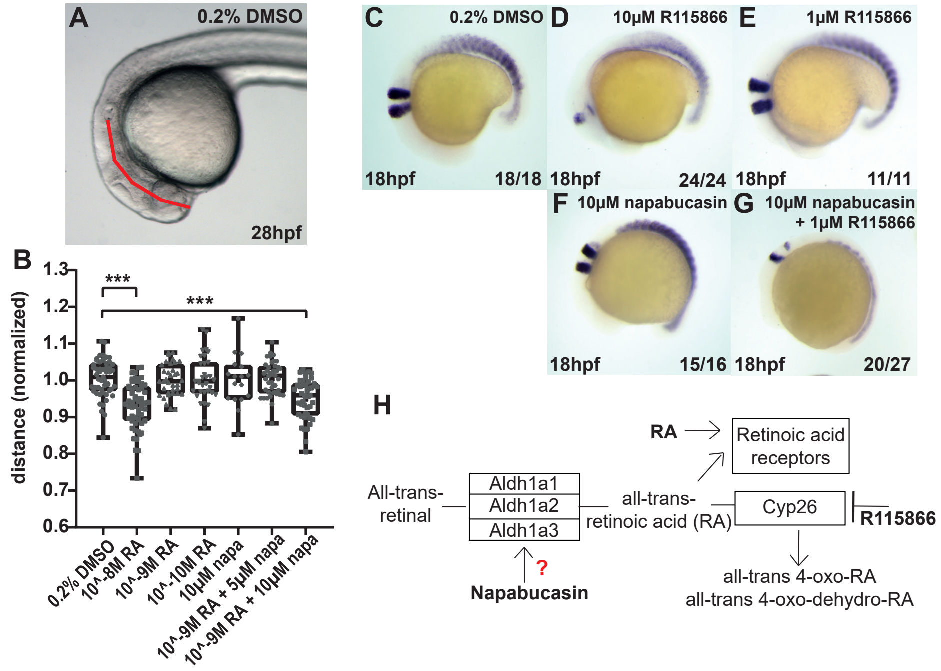


Figure 5: Validation of the off-targets of napabucasin using a zebrafish phenotypic model. The distance between the otolith and the tip of the nose was determined (indicated with a red line) in control (0.2% DMSO), RA treated (10^{-8}, 10^{-9} or 10^{-10} M), napabucasin (10 µM) or combinations, as indicated (A, B). Significance was determined using a one-way ANOVA with Dunnetts multiple comparisons test. Individual dots represent individual embryos; minimally 26 and maximally 66 embryos were used per condition; *** =P < 0.001. Zebrafish embryos were treated with the Cyp26 inhibitor R115866, with napabucasin or combinations. The embryos were fixed at 18 hpf and in situ hybridization was performed using krox20 and myod-specific probes, marking rhombomeres (3 and 5) and the somites, respectively (C-G). Schematic representation of RA metabolism and the role of Aldhs and Cyp26 in the process (H).

Discussion

This study shows the applicability of zebrafish as model system for thermal proteome profiling on whole organisms. First, a proof of principle experiment was performed with the broad-range oxidative agent pervanadate. A part of the proteome shows a stabilizing effect due to pervanadate treatment, clearly indicating the possibility of measuring treatment induced stability changes. Focusing on specific proteins showed a shift in ATPases (Supplementary Fig. 2). It has been reported that vanadate binds to ATPases in the catalytic site, which causes thermal stabilization[38]. These results show that ligand induced stabilization can be detected in zebrafish lysates on a proteome-wide scale.

The biggest advantage of using whole organisms compared to a single cell type is the increased diversity of proteins that can be investigated. In our proteomics data, multiple proteins from specific tissue types were found[39]. Amylase (pancreas), glial fibrillary acidic protein (cerebral cortex), myosin binding protein C (heart muscle) and apolipoprotein A-II (liver) were found, amongst others. This indicates the possibility of screening all proteins of a whole organism using TPP, including tissue specific proteins that would be missed if single cell types or tissues would have been used. It is evident that highly and broadly expressed proteins are more highly represented in whole organism lysates than lowly expressed proteins or proteins that are exclusively expressed in specific, low abundant cell types. Current mass spectrometry technology is instrumental in identifying low abundant proteins in a background of highly expressed proteins. In this respect, it is noteworthy that we identified the largest zebrafish proteomics dataset reported to date, containing 7646 unique proteins. We managed to identify proteins at different scales of magnitude, from the highly abundant vtg1 to low abundant proteins such as sirt1.

We used zebrafish embryo lysates to identify targets of napabucasin, a drug that reportedly affects STAT3 signaling. The melting curves of Stat3 did not indicate a thermal shift in response to napabucasin. This may suggest that there is no binding between the drug and Stat3. Recent literature suggests that the mode of action of napabucasin involves the oxidoreductases POR and NQO1, which generate reactive oxygen species (ROS)[17]. Interestingly, we observed a destabilizing effect of napabucasin on the more highly conserved (71%) homolog of POR, Pora, than on the less conserved (49%) homolog of NQO1, (Fig. 3E, F). Hence, napabucasin may modulate Pora in zebrafish, resulting in increased ROS levels, leading to a decrease in phospho-STAT3 and STAT3 levels[40], which might explain the phenotypic similarity between napabucasin treatment and morpholino-mediated knockdown of STAT3 (Fig. 2A, B).

Multiple members of the Aldh family of proteins were stabilized upon napabucasin treatment. ALDHs play an important role in aldehyde metabolism, by catalyzing the oxidation of reactive aldehydes[41], thereby reducing the level of reactive oxygen species. ALDH activity is also necessary for the generation of vital biomolecules, including retinoic acid (RA) and folate[42]. Our data convincingly showed that napabucasin increased ALDH enzymatic activity *in vitro* (Fig. 4B). It is noteworthy that zebrafish Aldh1a2 and -1a3 show a large degree of conservation with human ALDH2. This observation explains why the colorimetric assay, which uses the ALDH2 preferred acetaldehyde as substrate, shows an increase in human ALDH activity after napabucasin treatment. The effect of napabucasin on zebrafish development *in vivo* was consistent with Aldh activation, in that napabucasin treatment cooperated with suboptimal RA treatment and inhibition of RA catabolism (Fig. 5).

Napabucasin blocks stem cell activity in cancer cells and is being tested in multiple clinical trials as anti-cancer drug. It will be interesting to investigate in future experiments whether napabucasin-induced activation of ALDH activity and subsequent elevation of RA levels also has a role in the effect of napabucasin on stemness of cancer stem cells.

In this project, we improved the traditional TPP workflow and applied it to whole zebrafish embryo lysates instead of only a single cell type or tissue. This allowed us to screen the proteome for targets of napabucasin. Multiple Aldh family proteins were identified as targets of napabucasin. Our data support the conclusion that developmental defects in napabucasin treated zebrafish embryos may result from activated Aldh-mediated elevation of RA levels.

Acknowledgements

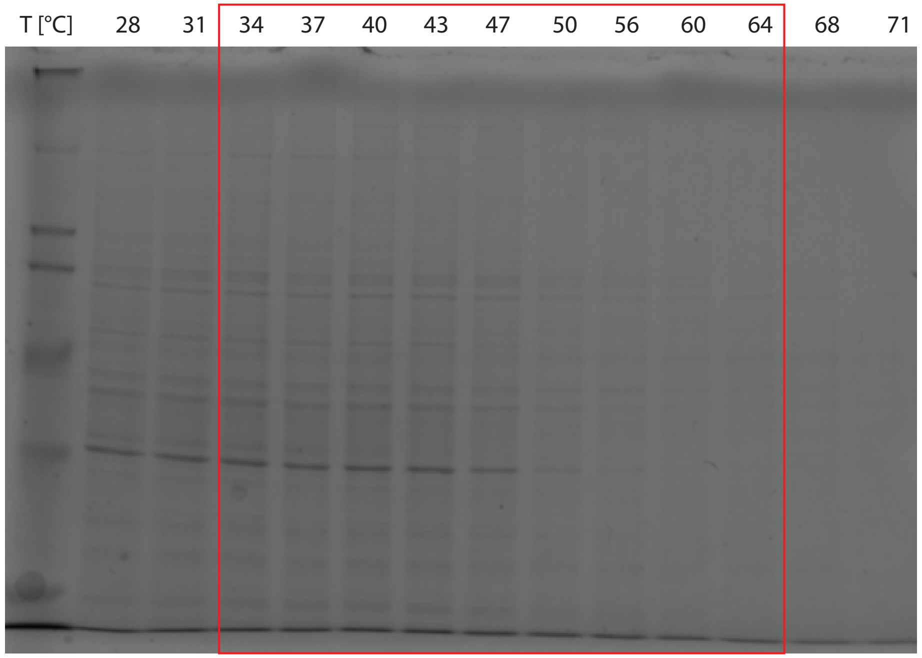
S.L. acknowledges support from the Netherlands Organization for Scientific Research (NWO) through a VIDI grant (project 723.013.008).

References

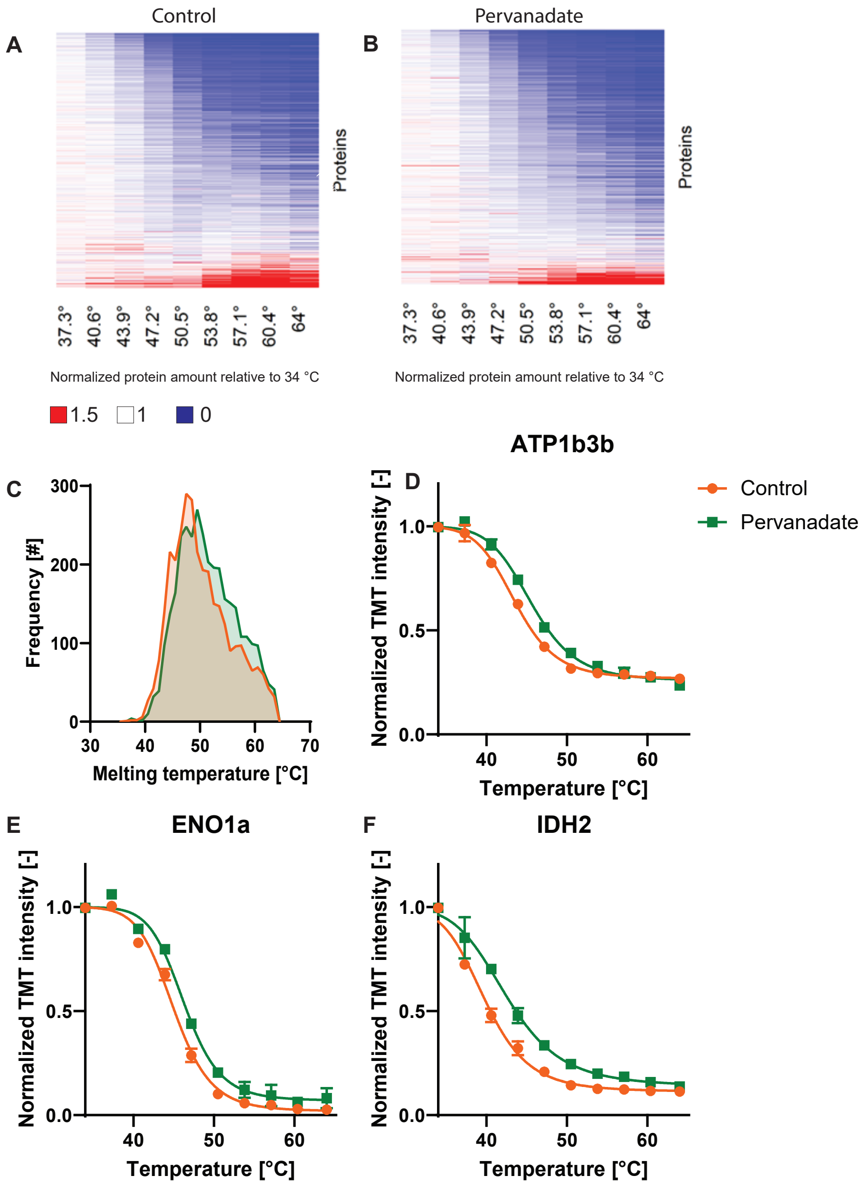
1. Molina, D. M., Jafari, R., Ignatushchenko, M. & Seki, T. Monitoring Drug Target Engagement in Cells and Tissues Using the Cellular Thermal Shift Assay. *Science (80-.).* **341**, 84–88 (2013).
2. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science (80-.).* **346**, (2014).
3. Franken, H. *et al.* Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* **10**, 1567–1593 (2015).
4. Becher, I. *et al.* Thermal profiling reveals phenylalanine hydroxylase as an off-target of panobinostat. *Nat. Chem. Biol.* **12**, 908–910 (2016).
5. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state in vivo . *Mol. Syst. Biol.* **14**, 1–15 (2018).
6. Perrin, J. *et al.* Identifying drug targets in tissues and whole blood with thermal-shift profiling. *Nat. Biotechnol.* doi:10.1038/s41587-019-0388-4
7. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
8. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
9. Horzmann, K. A. & Freeman, J. L. Making waves: New developments in toxicology with the zebrafish. *Toxicol. Sci.* **163**, 5–12 (2018).
10. Hubbard, J. M. & Grothey, A. Napabucasin: An Update on the First-in-Class Cancer Stemness Inhibitor. *Drugs* **77**, 1091–1103 (2017).
11. Li, Y. *et al.* Suppression of cancer relapse and metastasis by inhibiting cancer stemness. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1839–1844 (2015).
12. Wong, A. L. A. *et al.* Do STAT3 inhibitors have potential in the future for cancer therapy? *Expert Opin. Investig. Drugs* **26**, 883–887 (2017).
13. Huynh, J., Chand, A., Gough, D. & Ernst, M. Therapeutically exploiting STAT3 activity in cancer — using tissue repair as a road map. *Nat. Rev. Cancer* **19**, 82–96 (2019).
14. Zhong, J. H., Huang, D. H. & Chen, Z. Y. Prognostic role of systemic immune-inflammation index in solid tumors: A systematic review and meta-analysis. *Oncotarget* **8**, 75381–75388 (2017).
15. Löcken, H., Clamor, C. & Müller, K. Napabucasin and Related Heterocycle-Fused Naphthoquinones as STAT3 Inhibitors with Antiproliferative Activity against Cancer Cells. *J. Nat. Prod.* **81**, 1636–1644 (2018).
16. Zuo, D. *et al.* Inhibition of STAT3 blocks protein synthesis and tumor metastasis in osteosarcoma cells. *J. Exp. Clin. Cancer Res.* **37**, 1–11 (2018).
17. Froeling, F. E. M. *et al.* Bioactivation of Napabucasin Triggers Reactive Oxygen Species–Mediated Cancer Cell Death. *Clin. Cancer Res.* **1**, 7162–7175 (2019).
18. Westerfield, M. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish Danio (‘ Brachydanio Rerio’).* (University of Oregon, 2007).

19. Aleström, P. *et al.* Zebrafish: Housing and husbandry recommendations. *Lab. Anim.* **0**, 1–12 (2019).
20. Yamashita, S. *et al.* Stat3 controls cell movements during zebrafish gastrulation. *Dev. Cell* **2**, 363–375 (2002).
21. Schulte-Merker, S., Ho, R. K., Herrmann, B. G. & Nusslein-Volhard, C. The protein product of the zebrafish homologue of the mouse T gene is expressed in nuclei of the germ ring and the notochord of the early embryo. *Development* **116**, 1021–1032 (1992).
22. Oxtoby, E. & Jowett, T. Cloning of the zebrafish krox-20 gene (krx-20) and its expression during hindbrain development. *Nucleic Acids Res.* **21**, 1087–1095 (1993).
23. Weinberg, E. S. *et al.* Developmental regulation of zebrafish MyoD in wild-type, no tail and spadetail embryos. *Development* **122**, 271–280 (1996).
24. Huyer, G. *et al.* Mechanism of inhibition of protein-tyrosine phosphatases by vanadate and pervanadate. *J. Biol. Chem.* **272**, 843–851 (1997).
25. Zecha, J. *et al.* TMT labeling for the masses: A robust and cost-efficient, in-solution labeling approach. *Mol. Cell. Proteomics* **18**, 1468–1478 (2019).
26. Childs, D. *et al.* Nonparametric analysis of thermal proteome profiles reveals novel drug-binding proteins. *Mol. Cell. Proteomics* **18**, 2506–2515 (2019).
27. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
28. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
29. Reinhard, F. B. M. *et al.* Thermal proteome profiling monitors ligand interactions with cellular membrane proteins. *Nat. Methods* **12**, 1129–1131 (2015).
30. Liu, Y., Sepich, D. S. & Solnica-krezel, L. Stat3 / Cdc25a-dependent cell proliferation promotes embryonic axis extension during zebrafish gastrulation. *PLoS Genet.* **13**, 1–32 (2017).
31. Rodríguez-Zavala, J. S., Calleja, L. F., Moreno-Sánchez, R. & Yoval-Sánchez, B. Role of Aldehyde Dehydrogenases in Physiopathological Processes. *Chem. Res. Toxicol.* **32**, 405–420 (2019).
32. Perez-Miller, S. *et al.* Alda-1 is an agonist and chemical chaperone for the common human aldehyde dehydrogenase 2 variant. *Nat. Struct. Mol. Biol.* **17**, 159–164 (2010).
33. Duester, G. Alcohol Dehydrogenase as a Critical Mediator of Retinoic Acid Synthesis from Vitamin A in the Mouse Embryo. *J. Nutr.* **128**, 459S-462S (1998).
34. Holder, N. & Hill, J. Retinoic acid modifies development of the midbrain-hindbrain border and affects cranial ganglion formation in zebrafish embryos. *Development* **113**, 1159–1170 (1991).
35. Marshall, H. *et al.* Retinoic acid alters hindbrain Hox code and induces transformation of rhombomeres 2/3 into a 4/5 identity. *Nature* **360**, 737–741 (1992).
36. Hernandez, R. E., Putzke, A. P., Myers, J. P., Margaretha, L. & Moens, C. B. Cyp26 enzymes generate the retinoic acid response pattern necessary for hindbrain development. *Development* **134**, 177–187 (2007).
37. Cai, A. Q. *et al.* Cellular retinoic acid-binding proteins are essential for hindbrain patterning and signal robustness in zebrafish. *Development* **139**, 2150–2155 (2012).
38. Clausen, J. D. *et al.* Crystal Structure of the Vanadate-Inhibited Ca2+-ATPase. *Structure* **24**, 617–623 (2016).
39. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
40. Han, D. *et al.* Napabucasin, a novel STAT3 inhibitor suppresses proliferation, invasion and stemness of glioblastoma cells. *J. Exp. Clin. Cancer Res.* **38**, 1–12 (2019).
41. Vassalli, G. Aldehyde dehydrogenases: Not just markers, but functional regulators of stem cells. *Stem Cells Int.* **2019**, (2019).
42. Moreb, J. S., Ucar-Bilyeu, D. A. & Khan, A. Use of retinoic acid/aldehyde dehydrogenase pathway as potential targeted therapy against cancer stem cells. *Cancer Chemother. Pharmacol.* **79**, 295–301 (2017).

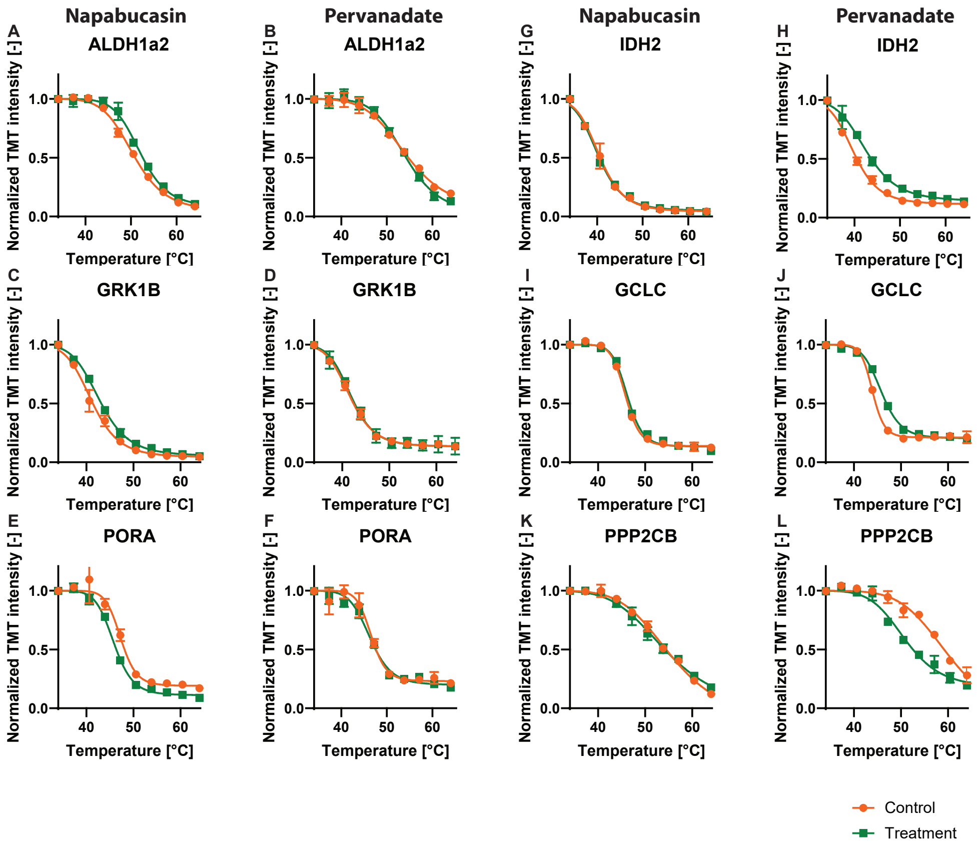
Supplementary figures



Supplementary figure S1: SDS-page of zebrafish lysates heated to different temperatures. The temperature range of 34 - 64 °C is optimal for the generation of melting curves.



Supplementary figure S2: TPP experiment using pervanadate. Heat maps of the precipitation behavior of control (A) and pervanadate treated (B) lysate shows that there is a global stabilization in the pervanadate treated samples, which is reflected in a global increase in melting points in pervanadate treated lysates (C). Some proteins have a shifted melting curve, such as ATPases (D), proteins involved in glycolysis such as enolase (E) and proteins involved in the citric acid cycle such as isocitrate dehydrogenase (F). Data points (n = 2 independent experiments) are shown as mean ± SEM, melting curve fitting was performed according to chemical denaturation theory.

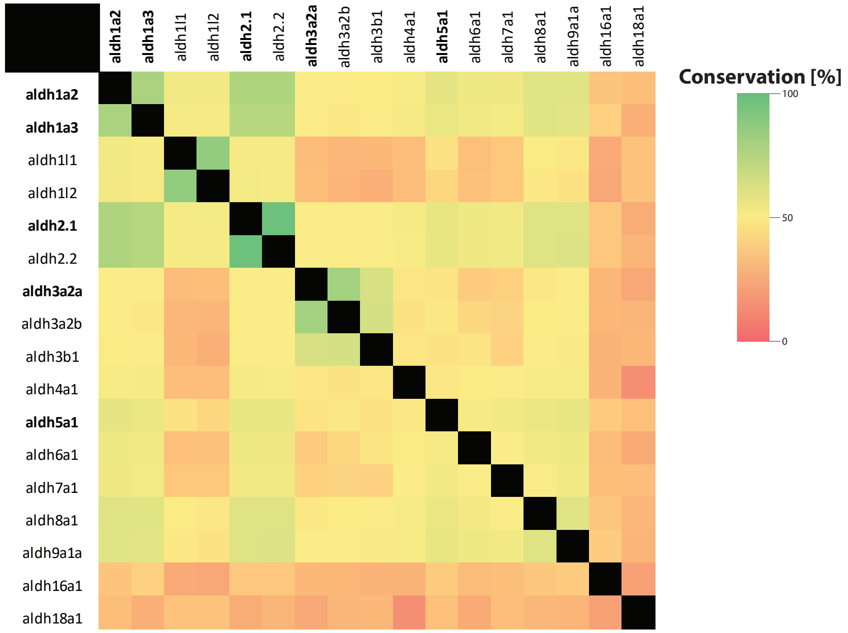


Supplementary figure S3: Comparison between the same proteins across experiments. Melting curves for the proteins ALDH1a2, GRK1B, PORA, IDH2, GCLC and PPP2CB after Napabucasin or pervanadate treatment show that the induced shift is selective and unique for the treatment. Data points (n = 2 independent experiments) are shown as mean ± SEM, melting curve fitting was performed according to chemical denaturation theory.

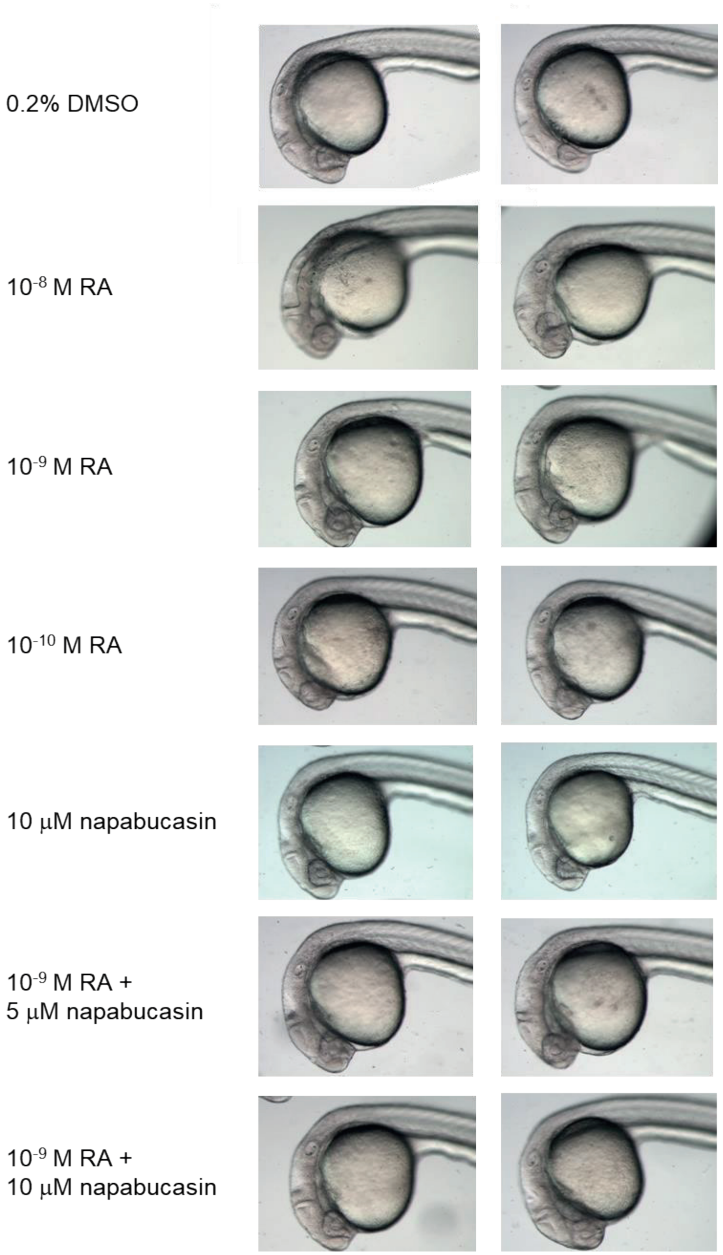
A

Gene name	All peptides	Unique peptides	Unique sequence coverage [%]	Thermal shift?
aldh1a2	37	33	61.4	Yes
aldh1a3	15	12	30.4	Yes
aldh1l1	62	54	59	No
aldh1l2	62	54	63.1	No
aldh2.1	20	1	2.7	Yes
aldh2.2	28	9	17.6	No
aldh3a2a	9	8	18	Yes
aldh3a2b	24	23	45.9	No
aldh3b1	12	1	1.7	No
aldh4a1	20	20	33.3	No
aldh5a1	21	20	53.5	Yes
aldh6a1	23	23	57.1	No
aldh7a1	37	3	5.3	No
aldh8a1	19	19	45.6	No
aldh9a1a	44	42	77	No
aldh16a1	20	2	2.9	No
aldh18a1	24	24	37.7	No

B



Supplementary figure S4: Details of all Aldh proteins found in the thermal proteome profiling experiment using napabucasin and whether they show a thermal stabilization effect (A). The heat map shows the conservation (in percentage) between all Aldh proteins found in the experiment (B). The pair wise conservation was determined using the EMBOSS Needle algorithm[27].


0.2% DMSO
10^{-8} M RA
10^{-9} M RA
10^{-10} M RA
10 μM napabucasin
10^{-9} M RA + 5 μM napabucasin
10^{-9} M RA + 10 μM napabucasin

Supplementary figure S5. Examples of 28 hpf embryos that were used for measurement of the distance from the tip of the nose to the otolith (Fig. 5A, B). Embryos were treated with the agents as indicated. Two representative embryos of each treatment are depicted, illustrating that the treatments did not induce gross developmental defects, but rather subtle defects which were quantified in Fig. 5B.

Chapter 3

Histidine phosphorylation in human cells; a needle or phantom in the haystack?

Niels M. Leijten[1], Albert J. R. Heck[1] and Simone Lemeer[1#]

1. Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

Manuscript in preparation

Abstract

Many researchers have attempted, with variable success, to identify protein histidine phosphorylation in mammalian cells. Despite these efforts, the extent and biological implication of eukaryotic histidine phosphorylation has still not been resolved. Here, we described our attempts to observe histidine phosphorylation in 4 different human cell lines, combining efficient phosphoproteomics technologies, such as column-based Fe^{3+}-IMAC enrichment and immonium ion triggering, which proved to be successful for the identification of widespread histidine phosphorylation in the bacterium *E.coli*.

Initially, using these approaches we seemingly picked up hundreds of specific protein histidine phosphorylations, of which multiple were co-occurring in the different cell lines observed. However, by extending our experiments, using extensive fractionation, immonium ion triggering and *in vitro* kinase assays, and after careful evaluation of our data, we had to conclude that the vast majority of these protein histidine phosphorylations were due to mislocalization, largely caused by frequently present neighboring Ser and Thr residues. Therefore, we conclude that there is likely an over interpretation in the literature of the extent of histidine phosphorylation in mammalian systems.

Our conclusions were backed by parallel analysis of protein histidine phosphorylation in *E.coli* using an identical workflow, where site localization scores were way higher and histidine phosphorylation reasonably widespread. Moreover, in *E.coli*, acid treatment readily and selectively removed the histidine phosphorylated complement of the phosphoproteome, while leaving other phosphorylations intact. Acid hydrolysis in this case was supported by the total disappearance of the true indicative pHis immonium ions in those samples. In human cells however, no specific acid hydrolysis of phosphohistidine was observed and the seemingly high extent of histidine phosphorylation was equal before and after acidification, indicating that these identifications indeed resulted from mislocalization.

Introduction

The proteome is a very dynamic entity that governs all processes in a cell in a spatial and temporal manner. By regulating the expression and degradation of proteins, the cell can tune many physiological processes such as cellular growth and survival. However, the activity of proteins is more tightly regulated by post-translational modifications (PTMs). These PTMs can alter the activity, localization and turnover of proteins, but can also change protein-protein interactions[1]. When these PTMs are dysregulated, it can lead to a range of malignancies such as cancer, diabetes and neurological disorders.

Due to its reversibility and versatility, protein phosphorylation is one of the most important and most studied PTMs. Phosphorylation events are mediated by kinases, which add a phosphate group to an acceptor protein[2]. Reversely, phosphatases can readily remove the phosphate moiety. A phosphorylation event can mediate protein function in two distinct ways[3]. First, it can act as an anchor for new intra- or intermolecular protein-protein interactions. Secondly, it can change the conformation of the protein and through this way change its activity and function. Through these mechanisms, phosphorylation acts as a rapid molecular switch, which has a large impact on cellular function.

A total of nine out of twenty amino acids can be phosphorylated[4,5]. These include serine, threonine, tyrosine, arginine, lysine, histidine, cysteine, aspartate and glutamine. In mammalian cells, serine phosphorylation (pSer) is the most abundant in frequency (~86%), followed by phosphothreonine (pThr) (~12%) and phosphotyrosine (pTyr) (~2%)[6]. N-linked phosphorylation on lysine, arginine and histidine residues has been proposed to also occur in mammalian cells. However, evidence on their existence, role and biological implications is limited or lacking, primarily due to lack of proper technologies to study these modifications. One of the most intriguing N-linked phosphorylations to study is histidine phosphorylation (pHis), as this modification has unique chemical properties. First, there are two isomers possible for pHis; 1- and 3-pHis[7]. These isomers have differences in structure, reactivity and stability. It has been shown that 1-pHis is slightly less stable than 3-pHis, while it is possible for the phosphate group to transfer from the 1- to 3- position[4]. Therefore, it would be beneficial to be able to distinguish between the isomers to determine their exact redundant or distinctive functions. Additionally, pHis serves as a high-energy intermediate in the transfer of the phosphoryl group to other amino acids. No other amino acid which can be phosphorylated shows this behavior. Therefore, in addition to changing the molecular surface of proteins, pHis can also relay information inside signaling cascades. These characteristics, not shared by any other amino acid, makes pHis a particular interesting PTM to study.

The study of histidine phosphorylation has trailed decades behind the study of O-linked phosphorylation. The phosphoramidate bond formed after phosphorylation of histidine has a very high free energy in comparison to the stable phosphoester bond formed in pSer, pThr and pTyr residues. This high free energy of hydrolysis makes pHis quite unstable, causing it to rapidly hydrolyze at acidic pH or high temperatures[8,9]. Because of this, standard methods to enrich for sub-stoichiometric phosphopeptides, such as immobilized metal affinity chromatography (IMAC) or metal oxide affinity chromatography (MOAC), performed at strong acidic pH, are not very suitable[10]. The lability of the phosphoramidate bond in combination with its presumed low abundance has thus complicated research of this modification.

Despite the challenges in detecting site specific histidine phosphorylation in proteins, its existence has been well described in bacterial systems, where it plays an important role especially in the two-component system[4,9]. Even though phosphohistidine plays an established signaling role in bacteria, both gram positive and negative[11] and simple eukaryotes, such as yeast (*Saccharomyces cerevisiae*) as well as the slime mold (*Dictyostelium discoideum*)[11], similar phosphohistidine-based signaling has not been detected in higher eukaryotes such as mammals, birds and fish[7]. It has been hypothesized that this is due to the fact that the signals following from the His-Asp relay system are too unstable to be efficiently transferred to the nucleus of mammalian cells[4]. An alternative hypothesis is that it is due to the development of receptor tyrosine kinases or G-protein coupled receptors, which might have replaced the two-component system as cell surface sensors[4]. Notwithstanding, some specific cases of histidine phosphorylation have been described in mammalian systems.

Histidine phosphorylation in eukaryotes was first reported in 1962 by Boyer *et al*[12], where they found it as an enzyme intermediate of oxidative phosphorylation in the mitochondria of bovine liver. Later, this enzyme was identified to be succinyl-CoA synthetase (SUCLG1)[8]. In addition, the nucleoside diphosphate kinases NME1 and NME2 have been described to also potentially have histidine kinase activity. In this case, the activity depends upon catalytic transfer of the activated phosphate from the autophosphorylated histidine 118 residue (H118) onto a histidine in a target protein, instead of onto a nucleoside diphosphate. There are a total of 10 NME family genes reported in humans, with orthologues present in all eukaryotes[4], but only NME1 and NME2 are reported to have histidine kinase activity[8]. No clear sequence motif has been found for these kinases, but their high conservation supports the idea that pHis plays an important role in basic cellular processes. Some putative targets of these kinases have been thoroughly investigated, which include the ion channels KCa3.1 and TRPV5, and the G-protein beta subunit GNB1. Likewise, several putative histidine phosphatases have been reported. These include PHPT1, LHPP and PGAM5, but in most cases their phosphatase activity is not restricted to pHis[4]. So far, the presence of pHis has mainly been validated in metabolic enzymes[8]. It has been established that pHis plays a role as reactive phospho-intermediate in proteins such as SUCLG1, PGAM1 and ACLY. In order to get a more comprehensive view on pHis signaling in mammalian systems, better and more robust analytical tools are needed.

A big step towards proper investigation of pHis occurred in 2010, when Kee *et al*[13] developed the first antibody which can recognize pHis containing proteins and peptides. Before this time, no antibodies against pHis could be generated, since immunogens against pHis are readily dephosphorylated in serum. Kee *et al* circumvented this by using nonhydrolyzable and nonisomerizable phosphoryl-triazolylalanine as a mimetic for pHis, which they incorporated in a peptide from histone H4 and used to immunize rabbits[13]. Using this method, they generated the first sequence specific pHis-3 antibody against histone H4. Later, they further improved[14] this process to generate a pan-specific 3-pHis antibody, but this had limited use due to significant cross-reactivity with pTyr. In 2016, Fuhs *et al*[15] used the same principle and further improved on these antibodies. Here, they incorporated the phosphoryl-triazolylalanine pHis analogues in peptide libraries, which they then injected into rabbits. They were able to generate selective monoclonal anti-1-pHis and anti-3-pHis antibodies, which were subsequently used for immunoprecipitation experiments in human cell lines. Based on the proteins identified from these immunoprecipitation experiments, the authors claimed a role for 1-pHis in phagocytosis and 3-pHis in mitosis[15]. The most important drawback of the use of pHis antibodies, is the current inability to acquire site specific information on the site of histidine phosphorylation. Therefore, for most of the immunopurified proteins no direct unambiguous evidence was provided that these proteins were indeed pHis phosphorylated.

Although antibodies are frequently used in phosphoproteomics, most comprehensive studies on phosphoproteomes in the last decade rely on LC-MS based analysis, especially when studying pSer and pThr. However, acidic conditions in both liquid chromatography and current phosphopeptide enrichment strategies complicate the analysis of pHis, as pHis is known to readily hydrolyze under these conditions. Very recently, Potel *et al*[10] developed a phosphoproteomics workflow which was shown to be suitable for the analysis of histidine phosphorylated peptides. Multiple optimization steps were taken, which resulted in an overall increased phosphoproteome coverage, while at the same time maintaining the labile pHis modification[10,16]. These optimizations include: I) performing lysis with a stronger lysis buffer, II) the removal of phosphate-moiety containing molecular interferents by protein precipitation, III) enzymatic depletion of nucleic acids and IV) keeping the temperature during digestion and sample handling under room temperature. However, the most important improvement was made in the Fe^{3+}-IMAC enrichment: hydrolysis of pHis during IMAC enrichment was significantly reduced by performing it at mild acidic conditions (pH 2.3), which largely retained the labile modification while at the same time allowed for selective and efficient enrichment. Using this improved workflow, the number of phosphorylation events identified in *Escherichia coli (E.coli)* was increased to a cumulative number of 2129 phosphosites, a ten-fold increase to previous reports. More importantly, 246 pHis sites on 173 phosphoproteins were identified, indicating the successful detection of the labile modification using this technique. Therefore, we hypothesized that this exact same workflow might be suitable to shed light on the existence and extent of histidine phosphorylation in mammalian cells. Evidently, the increased complexity of the human phosphoproteome (possibly 100,000 sites) compared to the bacterial one (likely a few thousand sites) needs to be taken into account, since this may further complicate the accurate identification of pHis in human cells.

Here, we adopted the optimized protocol developed by Potel *et al*[10,16] to detect histidine phosphorylation in human cells. By using four different cell lines, initially a large population of novel pHis sites were detected. However, careful evaluation of the obtained results indicated that the validity of these sites was ambiguous. The localization unambiguity of histidine phosphorylation was significantly worse than for Ser, Thr and Tyr phosphoresidues. In addition, pHis sites were often detected on phosphopeptides harboring multiple other pSer and pThr phosphosites. Validation by using immonium ion triggering, high pH fractionation and *in vitro* kinase assays using NME1/NME2 did further minimize our confidence of pHis identification and site-localization (Fig. 1). Finally, acid lability of phosphohistidine provided compelling evidence for the widespread and abundant existence of pHis in *E.coli*, but this observation could not be substantiated in human cells. In summary, pHis seems to play a minor signaling role in human cells, although we were able to observe the auto-phosphorylation of NME1/NME2 and a few reactive tele-phosphohistidine intermediates in a selected number of proteins.

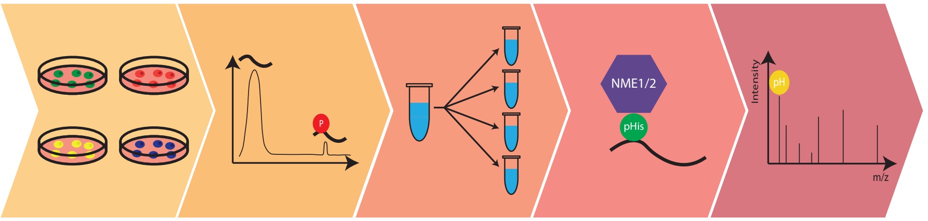


Figure 1: **Approaches taken to identify phosphohistidine in mammalian cells**. Four different cell types (A431, HEK293T, HeLa and PC9) were analyzed separately for the presence of pHis. These were lysed, digested and the phosphopeptides enriched using Fe^{3+}-IMAC and measured by LC-MS. To potentially gain more data on low abundant pHis, additionally high pH fractionation and kinase assay approaches were used to make the lysate less complex. Lastly, histidine immonium ion triggering was used to confidently confirm pHis phosphoresidues.

Results

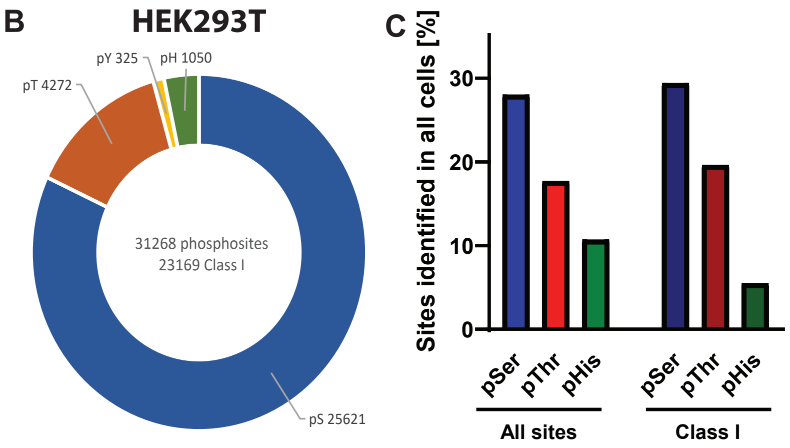
Chasing histidine phosphorylation in 4 human cell lines

With the aim to extend our knowledge on histidine phosphorylation in mammalian systems, the phosphoproteome of four different human cell lines: A431, HEK293T, Hela and PC9 was investigated, whereby the phosphopeptides were enriched using previously described conditions[10,16]. Five replicates were measured per cell type to achieve good analytical depth. The resulting data are summarized in Figure 2 and show that a total of 782, 1050, 967 and 779 phosphohistidine sites were identified in respectively A431, HEK293T, HeLa and PC9 cells (Fig. 2A). However, less than a third of these pHis sites have a localization probability higher than 0.75 (Class I sites), showing that the confidence of localization of more than 66% of detected pHis sites is low. Filtering Class I phosphosites detected on the other identified phosphoresidues does not show such a dramatic decrease: approximately 55 - 60% of pThr and 79% of pSer sites are Class I (Supplementary figure 1). The total number of phosphosites detected in HEK293T cells is also shown in figure 2B, all other cell types have a similar distribution in occurrence of the different phosphoresidues (Supplementary figure 2A).

In a further attempt to gain more confidence in the identified phosphohistidine sites, a comparison was made between all datasets. Only ~10% of all pHis sites were identified in all 4 cell lines, while only ~5% overlap remained for class I pHis sites (Fig. 2C). This behavior is not shared by pSer and pThr sites, where respectively 28 and 18% of sites were detected in all 4 cell types. These percentages even increased when

A

Cell type	Class I pHis sites	All pHis sites
A431	207	782
HEK293T	262	1050
HeLa	273	967
PC9	218	779



D

Protein	Phosphosite	Best Loc. Prob.	A431	HEK293T	HeLa	PC9
CEP135	1133	0.61	X			
DENND4C	736	0.35			X	
PFKFB3	254	0.99		X	X	X
PGAM1	11	0.99	X	X	X	X
SUCLG1	299	1	X		X	

E

Protein	Phosphosite	Localization prob.
DCLK1	35	0.48
PGAM1	11	0.96
SUCLG1	299	1
SRRM2	433	0.52

Figure 2: Initial results of the study of phosphohistidine in cellular lysates of human A431, HEK293T, HeLa and PC9 cells. All different cell types revealed around thousand pHis sites of which a quarter were class I sites (A). The distribution of all detected phosphosites in the HEK293T lysate showed an occurrence of circa 3% for pHis, seemingly 3-4 fold more frequent than pTyr, and only 4-fold less frequent than pThr (B). Of all pHis sites ~10% are identified in all 4 cell types. For class I sites this overlap diminishes to ~5%. For the identified pSer and pThr sites the overlap is substantially larger, respectively 28 and 18%. The overlap is even larger for class I sites, 29% for pSer and 20% for pThr (C). Only 5 pHis peptides in the different cell types exhibit the characteristic histidine immonium ion in their fragmentation spectra (D), but when using the immonium ion triggering method only 4 histidine sites were detected (E).

looking at class I sites, where the overlap increased to 29 and 20% for respectively pSer and pThr. These results showed that the variation of identified pHis sites is larger compared to pSer and pThr. Additionally, the decrease in overlap for class I pHis sites indicates possible issues with correct localization.

When looking at the sites in more detail, a total of 217 phosphohistidine sites were detected in all experiments (Supplementary Fig. 2B), of which 35 are class I sites (Supplementary Fig. 2C). Peptides carrying pHis 118 of the supposed histidine kinases NME1 and NME2 were detected in all cell types, indicating that indeed the experimental conditions used are suitable to detect histidine phosphorylation. Several phosphohistidine sites were identified in enzymes that have previously been reported to carry pHis, most are known to play a role as reactive tele-phosphohistidine intermediate. These proteins include PGAM1 (pHis 11), PFKFB3 (pHis 254) and SUCLG1 (pHis 299). Besides these known phosphohistidine sites, a large part of the identified sites have not been studied or reported before. When looking for motifs in the histidine phosphorylated peptides, no motif could be found. This is in agreement with what has been reported in literature for human pHis[17]. Also in earlier studies in *E.coli*[10] no specific motif could be found.

Immonium triggering to improve confidence in pHis localization in human cell lines

Confident site-specific localization of histidine phosphorylation remains very challenging[18,19]. To tackle this issue, Potel *et al*[20] developed a mass spectrometry based method to improve the unambiguous identification of phosphohistidine. In this method, the phosphohistidine immonium ion (m/z 190.0376), which can be generated by using high-energy HCD fragmentation, is used to trigger a second, high quality EThcD based fragmentation step, which allows for comprehensive peptide backbone information and phosphosite localization. This method allowed the identification of an equal amount of pHis sites as previously reported for *E.coli*[10], however the localization confidence of these sites had significantly increased. It is important to note that in *E.coli*, about 35% of all identified pHis phosphorylated bacterial peptides displayed the pHis immonium ion upon fragmentation by using high-energy HCD.

Surprisingly, in our study, the pHis immonium ion could only be detected 11 times on 5 different proteins in the 4 different cell lines used (Fig. 2D). This is a marginal fraction of the total number of identified phosphohistidine sites, i.e. less than 1%, compared to the 30% of immonium ions detected for pHis peptides reported in *E.coli*. Based on this finding, we started to worry that histidine phosphorylation in mammalian samples may just be far less abundant when compared to the bacterial *E.coli* sample. Despite the low number of initial immonium ions detected, 3 replicates of HEK293T lysate were subjected to the triggering method. In all 3 replicates, only 4 pHis sites could be identified (Fig. 2E). Two of these sites were already detected with high confidence using the standard HCD method. Two new sites were detected in the pHis triggering assay; on the kinase DCLK1 and the matrix protein SRRM2. It is to note that both phosphohistidine sites on these proteins only have a localization probability of circa 0.5, indicating an uncertainty to which residue in the peptide is really phosphorylated. Taken together, these findings revealed that the pHis immonium ion assay could not be used to boost confidence in the pHis assignments in human cells, probably due to an extreme low abundance and occurrence, or even non-occurrence of phosphohistidine in human cell samples.

Decreasing the complexity of the phosphoproteome using high pH fractionation

To deal with the potential low abundance of pHis in mammalian cells and to obtain more depth in the analysis of the phosphoproteome, we decided to further fractionate the samples. There are numerous fractionation techniques that can be employed for the fractionation of peptides, but most routinely high pH and strong cation exchange (SCX) fractionation are used. Batth *et al*[21] showed that by performing high pH fractionation before IMAC enrichment, the total sample complexity can be reduced, which increased the proteomic depth and the number of phosphosites analyzed. They identified circa 16000 phosphosites compared to the 7000 using SCX fractionation. Here, we attempted to increase the number of identified phosphohistidine sites and the confidence in their localization probability by performing high pH fractionation on peptides from HEK293T cell lysate before enrichment using Fe^{3+}-IMAC.

A total of 10 aliquots of HEK293T lysate were fractionated using high pH fractionation, after which all fractions were concatenated to form 10 samples consisting of 2 mg of material (Fig. 3A). These 10 samples were enriched using Fe^{3+}-IMAC and subsequently analyzed by LC-MS. The number and distribution of (class I) phosphosites can be seen in figure 3B. When comparing the fractionated HEK293T lysate to its unfractionated counterpart (Supplementary Fig. 1), 10000 more phosphosites could be identified (42052 versus 31268), of which 5000 class I sites (28205 versus 23169). In the fractionated samples an additional 200 pHis sites were identified (Supplementary Fig. 1), increasing the number to 1271. However, when only considering class I pHis sites, an almost negligible increase of 23 pHis sites was observed. Only a limited overlap of ~29% between the identified phosphohistidine sites was detected (Fig. 3C), while when looking at the class I identified sites only ~20% overlap can be seen.

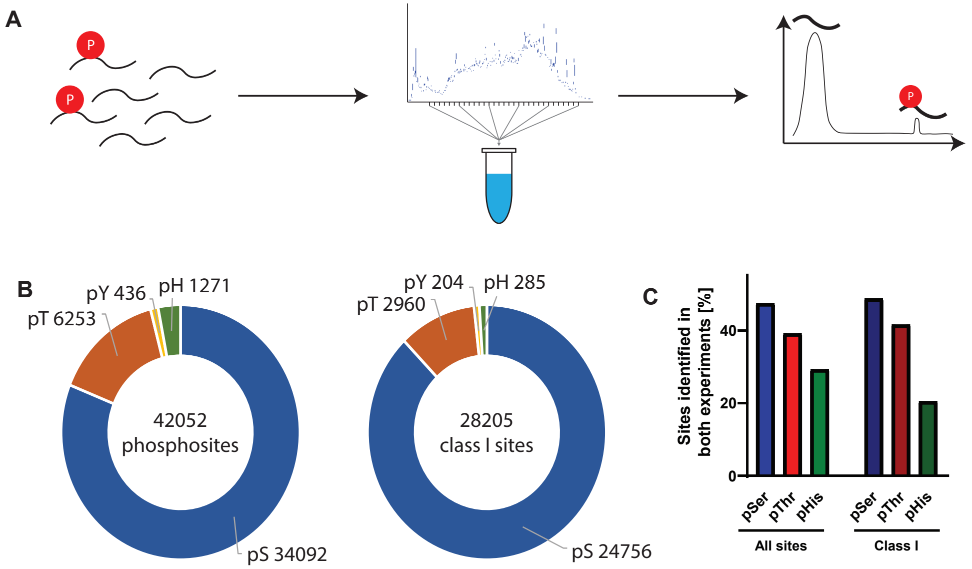


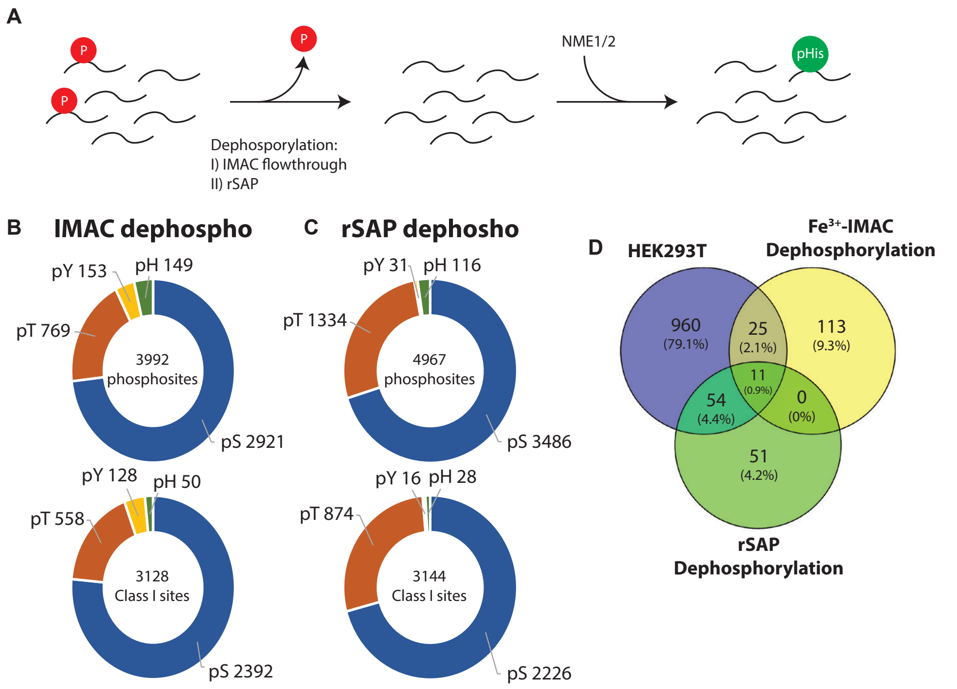
Figure 3: High pH fractionation of the HEK293T cellular lysate normally provides a more comprehensive view of the phosphoproteome, but here did not lead to a substantial increase of detected pHis peptides, and did also not lead to a boost in confidence in their identification and correct site-localization. The digested HEK293 cellular lysate was fractionated using high pH fractionation, after which the concatenated fractions were enriched for phosphopeptides using Fe^{3+}-IMAC (A). A large population of phosphosites was identified, more than identified in the non-fractionated experiments, with a large population of phosphohistidine sites. When looking at the class I sites, many histidine sites are lost (B). Only ~29% of pHis sites detected were shared between the analyses of the fractionated and unfractionated HEK293T lysate. For class I sites just ~20% overlap was seen. This overlap was substantially lower than for pSer (47%, 49% for class I) and pThr (39%, 42% for class I) containing phosphopeptides (C).

This overlap is consistently much lower when compared to pSer and pThr, which showed overlaps of respectively ~47% (~49% for class I) and ~39% (~42% for class I). Additionally, ~29% (~37% if only looking at class I) of the phosphohistidine sites were uniquely identified in the unfractionated sample. We had anticipated that the extensive fractionation efforts would give a more comprehensive dataset of all phosphohistidine sites present in HEK293T cells, but surprisingly there is a large number of pHis sites uniquely identified in the unfractionated samples. Although the overlap between measurements is never 100% due to the stochastic nature of the mass spectrometer, part of the limited overlap can again be explained by the difficulty in localizing pHis sites, due to presence of interfering phospho amino acids and the lability of the pHis bond. We hypothesized that by decreasing the complexity of the sample, suppression of low abundant pHis peptides might become less and consequently more histidine immonium ions could be observed. However, only immonium ions for PGAM1 (pHis 11) and SRRM2 (pHis 1724) were detected. Overall, our extended fractionation did not lead to a substantial increase of the detected pHis phosphoproteome, and more importantly did also not enhance our confidence in their identification and correct site-localization.

Finding mammalian histidine phosphorylation by an *in vitro* kinase assay using NME1/2

Triggered by our growing lack of confidence in detecting pHis peptides, we next performed *in vitro* kinase assays using the as histidine kinase annotated NME1/2 protein (Fig. 4A). Our methodology was based on the Kinase Assay-Linked phosphoproteomics (KALIP) assay developed by Xue *et al*[22]. This assay allows for the validation of pHis sites from earlier *in vivo* experiments, while it at the same time allows for finding novel residues being phosphorylated *in vitro* by the kinase of interest.

We first purified NME1/2 from HEK293T cells, wherein it is endogenously reasonable abundant, following the protocol from Potel *et al*[23]. HEK293T cells were lysed, after which an affinity purification using cGMP-beads was used to selectively enrich for NME1/2 (see Material and Methods). The enrichment gave a pure fraction containing NME1/2 (Supplementary figure 3A, B). Next, we generated an endogenous peptide substrate pool, carefully making sure it did not contain any phosphopeptides. To achieve this, two methods were employed: I) collection of the flow through peak after Fe^{3+}-IMAC enrichment and II) dephosphorylation of Fe^{3+}-IMAC enriched phosphopeptides using shrimp alkaline phosphatase (rSAP). Here, we hypothesized that the flow through peak of the Fe^{3+}-IMAC enrichment contains only non-phosphorylated peptides, which would be suitable for the *in vitro* kinase assay. Indeed, it was observed that when the peptides present in the flow through peak of the Fe^{3+}-IMAC enrichment are enriched again, only an extremely small number of phosphosites (96 Class I) could be identified (Supplementary figure 3C-E). This confirms the reliability of the column based Fe^{3+}-IMAC protocol as a very efficient dephosphorylation tool. For the second method of dephosphorylation, endogenous phosphopeptides enriched by Fe^{3+}-IMAC were enzymatically dephosphorylated using rSAP. This procedure generated a potentially more specific pool of peptides, being substrates for kinases *in vivo*. Dephosphorylation with rSAP proved to be efficient as well as only 352 class I phosphopeptides were identified after dephosphorylation (Supplementary figure 3F).



E

Protein	Phosphosite	Localization Prob.
ACTB	40	1
HIST1H2AJ	83	1
HSPA1A / HSPA1B	89	1
PGAM1 / PGAM2 / PGAM4	91	0.99
TPI1	138	0.99

Figure 4: Kinase assay to find *in vitro* putative NME1/2 substrates. In this assay, phosphopeptides were first dephosphorylated by one of two different methods: I) by taking the flow through peak of the IMAC column or II) by using alkaline phosphatase rSAP. Subsequently, the dephosphorylated peptides were phosphorylated by NME1/2 (A). Unexpectedly, widespread phosphorylation was observed, however primarily on Ser and Thr. Distribution of the number of (class I) phosphosites per receptor-residue detected in the kinase assay after Fe^{3+}-IMAC dephosphorylation (B) or rSAP dephosphorylation (C). The observed overlap in putative histidine phosphorylation from both the *in vitro* and *in vivo* experiments. A total of 90 sites are shared (D). Of the *in vitro* identified sites, 5 phosphohistidine sites could be detected using the immonium ion triggering method (E).

The non-phosphorylated samples were phosphorylated *in vitro* by using NME1/2, adding ATP to start the reaction. Subsequently, phosphopeptides were enriched using Fe^{3+}-IMAC and analyzed by mass spectrometry. A total of 149 phosphohistidine sites were identified from the flow through Fe^{3+}-IMAC approach, of which 50 were class I (Fig. 4B). For the dephosphorylation using rSAP approach, a total of 116 phosphohistidine sites were detected, of which 28 class I (Fig. 4C). Unexpectedly, for both approaches a large number of phosphorylation events on serine and threonine were detected; 2921 pSer and 769 pThr for the Fe^{3+}-IMAC approach and 3486 pSer and 1334 pThr for the rSAP approach. It has previously been shown that NME1/2 also has the ability to phosphorylate serine and threonine[24]. Still the number of potential substrates is surprisingly high, which likely may be attributed to the *in vitro* nature of the assay.

Likely, the most genuine pHis sites should be present in both *in vivo* and *in vitro* datasets. The overlap in phosphohistidine sites detected in both experiments can be seen in figure 4D. A total of 90 phosphohistidine sites observed in the *in vivo* HEK293T experiment were also detected in either or both *in vitro* kinase assay experiments. However, when only class I sites are considered, only 18 pHis sites were identified in both the *in vivo* and *in vitro* experiments. Similar to before, no clear motif could be detected for pHis in the datasets. We anticipated that the decreased sample complexity due to the preceding dephosphorylation step might allow for the identification of more pHis immonium ions compared to the earlier described *in vivo* experiments, allowing for improved confidence and better localization. In the experiment where the peptides were dephosphorylated by rSAP, only one pHis immonium ion was detected on SRRM2 (pHis 1724). For the experiment in which the peptides were taken from the flow-through of the Fe^{3+}-IMAC column, only two immonium ions were detected for ACTB (pHis 40) and HIST3H2A (pHis 83). Even though these numbers were again very low, we still used the immonium ion triggering method to see if more sites could be confidently detected, extending the number to 5 pHis sites (Fig. 4E), all with high localization probabilities. pHis sites on the proteins ACTB and HIST3H2A were detected again, with the same localization probability as in the standard (non-triggered) mass spectrometric runs. Newly identified sites on proteins HSPA1A and TPI were detected, but these were not identified in any previous experiment. Lastly, in these experiments the well-described PGAM1 pHis phosphorylation could again be detected.

Acid lability as a true indicator of histidine phosphorylation

In line with our earlier work on *E.coli*[10], we hypothesized that genuine phosphohistidine sites would be labile under acidic conditions and thus vanish after acid treatment. Falsely annotated pHis sites, i.e. identified by false localization, would in contrast still likely be detected after such treatment. We first performed a control experiment, again using an *E.coli* lysate[10]. *E.coli* phosphopeptides were enriched using Fe^{3+}-IMAC, dried down and reconstituted in 20 mM citric acid and 1% formic acid. Three replicates were analyzed directly by LC-MS/MS, essentially as described previously[10]. Another 3 enriched samples were left on the bench at room temperature for 24 hours, to allow acid hydrolysis of the phosphoramidate bond in pHis. Subsequently, these samples were also analyzed by LC-MS/MS to monitor the potential loss of pHis after acidification. When measuring 3 replicates of *E.coli*, a total of 146 phosphohistidine sites were detected, of which 107 were class I (Fig. 5A). In contrast, when 3 replicates of the acidified *E.coli* were measured, a total of 16 phosphohistidine sites were identified, of which 6 were class I (Fig. 5B). The percentage of phosphohistidine sites in the total of *E.coli* phosphosites identified decreased from ~13 to 2.5%, clearly indicating that the acidification removed histidine phosphorylation efficiently from these peptides. When looking at the overlap in pHis sites between the two conditions (Supplementary figure 4A), 90% of all sites were only identified in the non-acidified, non-incubated samples. This indicated that these pHis sites are genuine hits, displaying the pHis specific instability under acid conditions. The pHis sites only detected in the acidified sample are either very stable or could still also represent false positives. Additionally, in the non-acidified samples a total of 23 pHis immonium ions could be detected, while all such pHis immonium ions were fully absent in the acidified samples. This was quite distinct from what was observed for the phosphotyrosine immonium ions, of which 27 were detected in the untreated samples versus 23 in the acidified samples. These results show that by acidifying phosphopeptide samples, pHis can be selectively and efficiently removed from the phosphopeptide pool.

Next, we employed the same strategy to the human cell samples. Three aliquots of HEK293T lysate were used to enrich, using Fe^{3+}-IMAC, phosphopeptides from. These samples were alike acidified and incubated for 24 hours as described above and subsequently analyzed by LC-MS. When looking at the untreated HEK293T phosphopeptides (Fig. 5C), a total of 1050 pHis sites were identified. In contrast, for the acidified samples a total of 619 pHis sites were identified (Fig. 5D). A decrease in absolute number of sites can be seen, but when looking at the percentage of phosphohistidine peptides in the total pool of phosphopeptides no difference could be detected (3.4% for untreated *versus* 3% for acidified samples). For HEK293T lysate, ~46% of all identified pHis sites were identified in the acidified lysate, indicating that these might be false positives (Supplementary figure 4B). The intense decrease in pHis occurrence seen in the *E.coli* experiment following acidification, is not reproduced in the human cell line samples, indicating that the majority of detected pHis sites in the human cell lines are likely false positives. When comparing the ratios of pH/pSTY in both samples before and after acid treatment (Fig. 5E), a striking difference can be seen. The ratio stays the same when acidifying HEK293T lysates, while a big decrease in this ratio can be seen in the *E.coli* experiments, again indicating that the sites detected in HEK293T are likely mostly false positives. As acid lability proved to be a good indicator for confident identification of histidine phosphorylation in *E.coli* samples, we hypothesized that class I phosphosites which are repeatedly detected in human cell lines and are dephosphorylated in the acid treatment could be potential 'true' hits. Such analysis results in 13 class I phosphohistidine sites (Supplementary Figure 5A, B)

However, because of the low overlap in identified phosphohistidine sites in general between different experiments, either of the same or different cell lines, the mere absence of phosphohistidine peptides in the acid sample is not providing sufficient evidence for histidine phosphorylation. To provide even more compelling evidence, we searched for the presence of the dephosphorylated counterpart peptides of reported histidine phosphorylated peptides in the acid treated sample. In addition, we checked if the phosphorylated form also persisted in the acidified sample. This analysis resulted in a list of 4 confident pHis sites for which the dephosphorylated peptide was detected following acidification, and no phosphopeptide remained in the acid treatment; the notorious earlier observed PFKFB3, PGAM1, SUCLG1 and NME1 sites. This showed that our approach does pick up proteins that are known to have a telephosphohistidine as active site, proving the reliability of our analysis and selection methods. Following our stringent criteria, no other pHis sites were detected, indicating that the abundance of histidine phosphorylation in mammalian systems is limited, despite recent claims predicting otherwise[25].

Discussion

The aim of our study was to answer whether phosphohistidine plays a substantial role in mammalian cells, building further upon the work and methods developed by Potel *et al*[10]. Using this method and extensive analysis, we could confidently detect just a handful of histidine phosphorylations sites in human cell lines. We are not the first to study possible widespread histidine phosphorylation in human cells. Important work from Fuhs *et al*[15] used a very different approach based on monoclonal antibodies supposedly specific for both isoforms of phosphohistidine. In their study, a total of 630 proteins were identified which were enriched using the anti-1-pHis antibody and 506 proteins were identified using the anti-3-pHis antibody. Notably, although these proteins were pulled-down by these pHis targeting antibodies, no direct evidence was presented for the presence of a pHis on these proteins, and no site-localization was provided either. Still, taken all together, they reported 786 different proteins that were supposedly histidine phosphorylated. When comparing this dataset to the initial 1531 phosphohistidine containing proteins found in our 4 cell lines study, only a small overlap of 75 proteins (3.3%) is observed. When comparing the identified

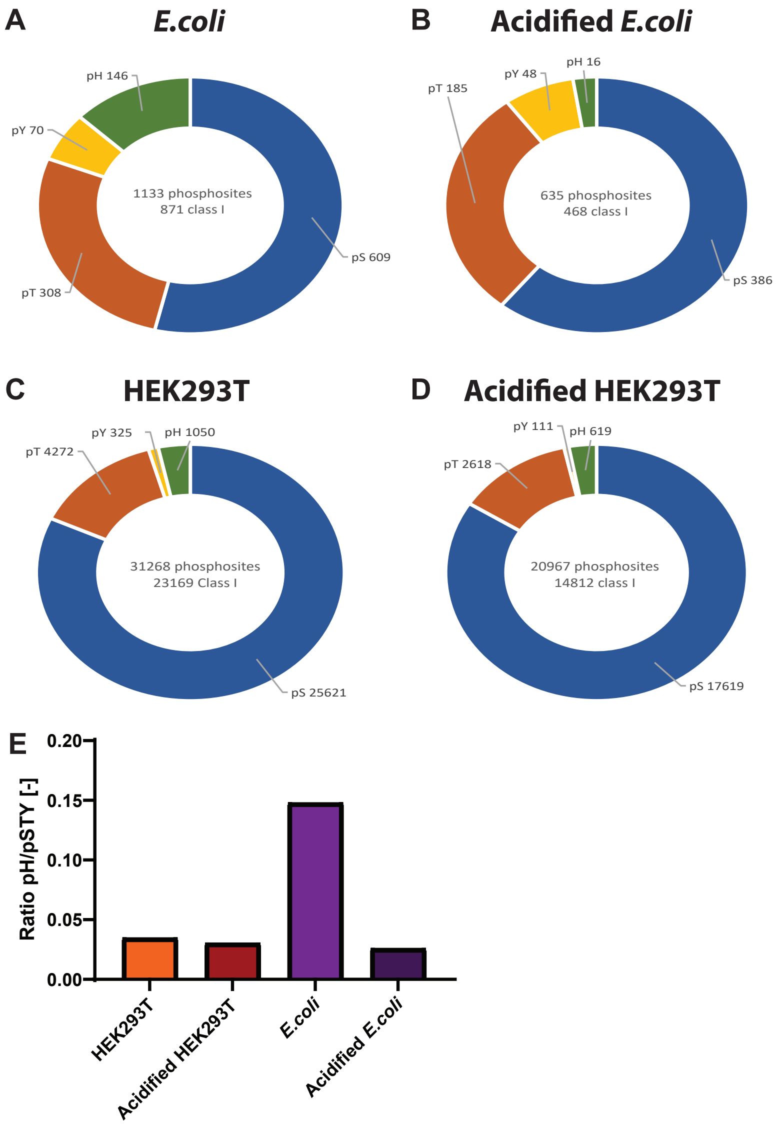


Figure 5: Acidification of pHis containing phosphopeptide samples diminishes their numbers in *E.coli* but not in human cell samples. When comparing the number of observed pHis sites in *E.coli* prior to (A) and following incubation under acidified conditions (B), it can be seen that almost all pHis peptides are hydrolyzed and not detected anymore. This shows that acidification of the lysate can act as a control for the true existence of pHis peptides. However, when the same approach is taken to analyze a HEK293T cell lysate, almost no difference in the percentage of identified pHis is seen between untreated (C) and the acidified sample (D). When looking at the ratio of pH/pSTY between untreated and acidified samples, a clear difference can be seen between the *E.coli* and the HEK293T samples (E), from which we conclude that the majority of identified pHis peptides are false-positives.

proteins with a class I site (localization probability > 0.75) with the proteins identified using the antibody-based approach, a minimal overlap of 31 proteins (2.4%) was observed. This small overlap between these two methodologies is illustrative for the challenges to detect pHis in mammalian cells.

By using a mass spectrometric readout, site specific localization of phosphohistidine can be achieved, making the assignment of pHis protein unambiguous. To obtain a first view at the scope of pHis in mammalian systems, four different cell types (A431, HEK293T, HeLa and PC9) were here investigated for the presence of the pHis modification. Initially, an exciting number of 2021 distinct phosphohistidine sites were identified, although just 217 phosphohistidine sites were found consistently in all of the studied cell lines (Supplementary Fig. 2B). Even though many pHis sites were initially identified, a large part of these were lost when only looking at class I phosphosites, i.e. having a localization probability larger than 0.75. In typical proteomics experiments, only class I sites are taken into account. Applying this criteria to pHis sites, only 635 phosphohistidine sites remain, of which just 35 were detected in all cell lines (Supplementary Fig. 2C).

Phosphohistidine thus turns out to be very difficult to localize confidently. During collision induced dissociation (CID), neutral loss of the phosphate group occurs, which further can complicate accurate localization[20]. This effect is more severe for phosphohistidine peptides, where the labile phosphorus-nitrogen bond is prone to extensive neutral losses as well as gas-phase rearrangements. Additionally, due to its relative lower frequency of occurrence in the human proteome, most of the times pHis containing peptides harbor multiple other residues which could be phosphorylated (notably Ser and Thr), increasing the risk of false-positive identification and incorrect localization. This can be seen when looking at the distribution of localization probabilities for the phosphosites identified in the four different cell types (Fig. 6A-E). From this analysis it is clear that pSer sites are the most confidently localized, followed by pThr and pTyr (Fig. 6E). In contrast, for pHis a substantial decrease in localization confidence is observed, clearly indicating that these are way more difficult to localize. These trends were observed consistently in the analyses of all four different cell types. However, when looking at the localization distribution per residue in *E.coli*, a different behavior is observed (Fig. 6A-D). The pSer sites in *E.coli* are localized as confidently as in mammalian cells, while the pThr sites are localized slightly more confident. The largest difference is seen when comparing pHis and pTyr localization in *E.coli* and mammalian cells, which both are significantly more confidently localized in *E.coli*. The deviation in localization for pTyr sites is probably due to the complexity of mammalian cell lysate in combination with the low abundance of phosphotyrosine

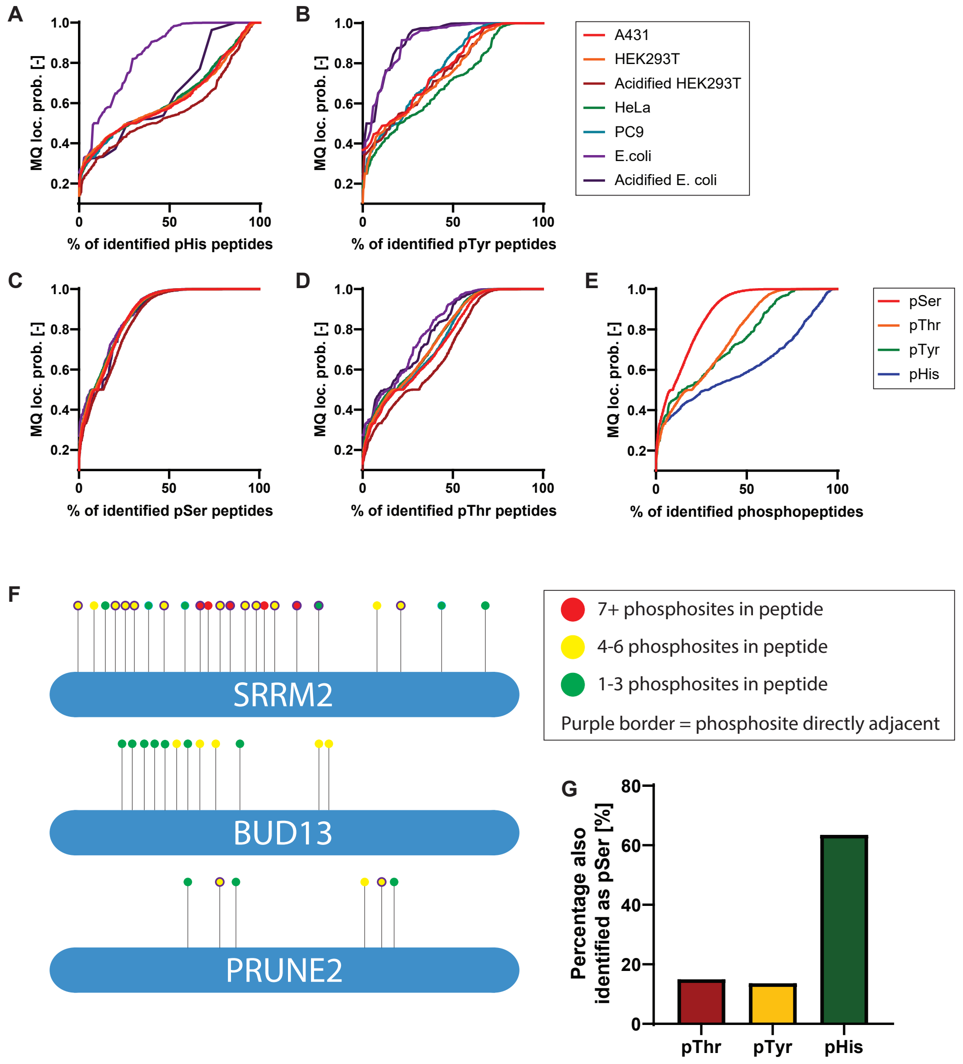


Figure 6: The unique challenge in correctly assigning pHis site localizations. When comparing the localization distributions of pSer, pThr, pTyr and pHis (A-D), it is found that in particular pHis localization is way more ambiguous in mammalian cells as in *E.coli*. This can be negated by acidifying the samples. When looking at the distribution of the phosphoresidues within HEK293T cells (E), it can be seen that site localization in pThr and pTyr is of lower quality than pSer but that site localization for pHis is really not good. Due to the low frequency of occurrence of histidine, the identified peptide almost always also contains (multiple) serine or threonine residues (F). Here, some example proteins are visualized, with the color of the dot indicating how many other phosphoresidues are in the identified peptides: green has 1 to 3, yellow has 4 to 6 and red has more than seven other phosphoresidues. A purple ring indicates that the observed pHis site had an adjacent residue which can be phosphorylated. The chance that a class I, singly phosphorylated pHis peptide was also identified as a class I serine phosphorylated peptide in the same dataset was substantially larger than for pThr and pTyr peptides (G) making again in particular the localization confidence for pHis very low.

(Supplementary Fig. 5C, D). The study of pTyr typically requires an immunopurification step to successfully enrich and measure this modification, which might explain the lowered confidence. For pHis, an even larger difference in localization confidence can be observed. This decrease in localization confidence can be explained by the fact that pHis sites in mammalian cells are almost always in peptides with neighboring residues that can be phosphorylated. Examples of this can be seen in figure 6F, where phosphosites detected on these proteins always have multiple other phosphoresidues present on the identified peptide. This greatly complicates the correct localization of phosphohistidine, as the fragmentation spectra of multiple phosphorylated peptides are harder to assign and often more ambiguous. We analyzed how often a class I, singly phosphorylated Tyr, Thr or His peptide was also identified as a class I serine phosphorylated peptide in the same dataset. This analysis showed that 65% of all pHis peptides are also reported to be serine phosphorylated, whereas for pThr and pTyr this percentage was found to be only around 15% (Fig. 6G). Although it is clear that proteins can sometimes be differentially phosphorylated on neighboring sites, the observed extreme difference between pHis and pThr/pTyr again questions whether the occurrence of pHis in human cells is really as widespread as sometimes being argued[8,15,25].

A method which can be used to localize pHis more accurately and with greater confidence is phosphohistidine immonium ion triggering, as demonstrated by Potel *et al*[20]. To more accurately localize the labile histidine modification, immonium ion triggering was performed on the four different cell types. However, only 5 pHis immonium ions could be detected in our experiments on mammalian cells (Fig. 2D). The immonium ion was observed clearly for well-known histidine phosphorylated proteins, such as PGAM1 and SUCLG1. When using the immonium ion triggering method on the cellular lysate of HEK293T, only 4 histidine sites were detected (Fig. 2E). In contrast, in experiments performed on *E.coli*, a total of 25 immonium ions could be detected for 146 proteins (17%), which is a much larger percentage compared to HEK293T lysate (0.2%). It was shown by Potel *et al*[20] that the immonium ion is present for up to 33% of all pHis peptides in *E.coli*. It has been hypothesized that the low occurrence of the pHis immonium ion especially in mammalian cells might be due to the huge increase in complexity between bacterial and mammalian lysates. Therefore, we also attempted to decrease the complexity by using a high pH pre-fractionation, but also this 10-times more elaborate approach led to the detection of immonium ions only for PGAM1 and SRRM2. Therefore, no further triggering studies were performed on the fractionated lysate.

Besides using an immonium ion triggering method for validation, an *in vitro* kinase assay was also developed. In this assay, two methods of dephosphorylation were used after which the dephosphorylated peptides were phosphorylated again by NME1/2. It can be seen that there is limited overlap in pHis sites identified between these two methods of dephosphorylation (Fig. 4D). Obviously, the source of the peptides is different; in the case of dephosphorylation using Fe^{3+}-IMAC a pool of peptides was produced which were not phosphorylated *in vivo*. In contrast, in the case of dephosphorylation by rSAP the pool of phosphopeptides was taken which was subsequently dephosphorylated. This means that both peptide substrates prior to the kinase assay are potentially orthogonal and could give complementary information, although stoichiometry for phosphorylation is unlikely to be 100%. A total of 254 pHis sites were identified, of which 90 were shared with the HEK293T dataset. Surprisingly, also a large population of pSer and pThr sites were detected. It is known[24] that NME1/2 also has serine/threonine specific kinase activity, but it was not expected that it would be more efficient in phosphorylating these residues compared to histidine. This might be due to the fact that the kinase assay was performed on peptide substrates instead of proteins, which might limit the selectiveness of the kinase due to a loss of substrate structure and therefore lead to a high false positive rate. This might be especially true for NME1/2, where no motif is known[17] and which therefore might use the tertiary structure of the substrate protein for its specificity. Performing the kinase assay on proteins might give results more similar to the *in vivo* situation, but then potential interference of endogenous kinases need to be taken into account[22].

To deal with the potential promiscuity of the kinase in the *in vitro* kinase assay on peptides, we compared the detected phosphosites with the *in vivo* phosphosites identified from the cell lysates. Here, 11 sites were identified to be phosphorylated in cellular lysate and in both kinase assays, while 25 and 54 sites were shared between the cellular lysate and respectively dephosphorylation by Fe^{3+}-IMAC or phosphatase. Similar to the high pH fractionation method, only 2 immonium ions were found in the kinase assay experiments, namely for ACTB and HIST3H2A. Nevertheless, the immonium triggering method was used to see if more sites could be confidently localized. Again, only 5 sites were identified using the triggering method (Fig. 4E).

Since the previous experiments did not give an unambiguous picture of pHis in mammalian systems, we became increasingly concerned that the observed histidine phosphorylation was just an artefact of database searching. Thus, next we attempted to remove pHis by acidification to determine which hits are false positives and which are true. Samples of *E.coli* and HEK293T cells were acidified to determine which sites are stable under acidic conditions and therefore the most likely to be false positives. In particular, histidine phosphorylation is hydrolyzed efficiently under acidic conditions[9], so it was expected that all pHis sites would be gone after 24 hours of incubation under acidic conditions. This was indeed observed when performing these experiments on *E.coli*, where the contribution of pHis decreased from ~13 to 2.5% (Fig. 5A, B), clearly showing that most of the bacterial pHis sites were acid labile. However, when the same treatment was performed on mammalian samples, no decrease in contribution of pHis to the total phosphopeptide pool could be observed (Fig. 5C, D). For both untreated and acidified samples, pHis contributed to circa 3% of the total. This is also shown by comparing the ratios of pH/pSTY for the samples (Fig. 5E), where a large decrease was seen after acidification for bacterial samples while the ratio stays constant for mammalian samples. A decrease in the total number of identified phosphosites was seen between acidified and untreated samples, for both *E.coli* and HEK293T, however this is probably caused by adsorption of peptides to plastics during the 24 hours of acid treatment. Still, the fact that a specific decrease in pHis can be seen in *E.coli* while this is not the case for HEK293T indicates that there is likely a very substantial degree of false positives due to site mislocalization in mammalian cell lysate.

When looking at the distribution of localization probabilities between acidified samples and their untreated counterpart, an interesting trend was observed. When the *E.coli* lysate is acidified, the localization probabilities of remaining pHis sites are worsened compared to the untreated counterpart, and show the same localization behavior as the mammalian pHis sites (Fig. 6A). This again indicated that most of the pHis sites identified in HEK293T cells are not true, since it was shown that pHis is efficiently removed from *E.coli* lysate after the acid treatment and the remaining sites are thus most likely incorrectly localized. In support of this notion, the localization behavior of pTyr sites in both *E.coli* and HEK293T does not change upon acidification of the samples (Fig. 6B).

Summarizing, although we excitingly started with a considerable number of putative pHis sites in each of the studied 4 human cell lines, attempts to increase the confidence of localization by performing high pH fractionation and kinase assays actually dramatically decreased the number of trustworthy sites. Furthermore, after immonium ion triggering and the acidification of samples, even less pHis sites withstood the test, but the few remaining sites should have a larger confidence of being true hits (Fig. 7). To arrive finally at a subset for which, in our opinion, histidine phosphorylation is sufficiently proven we looked at those sites that are i) repeatedly found to be histidine phosphorylated in 3 out of 4 cell lines ii) are not found to be phosphorylated in the acid treated sample and for which iii) the dephosphorylated peptide was found in the acidified sample. This filtering resulted in confident pHis sites on PFKFB3, PGAM1, SUCLG1 and NME1. NME2 is also found to be repeatedly histidine phosphorylated, however the peptide is still found to be phosphorylated on serine in the acidified treatment. Through all our efforts we did identify most tele-phosphohistidine sites known in human cells. This indicates that our method is working, but it also indicates that histidine phosphorylation is most likely restricted to metabolic intermediates and has as such no substantial signaling function in mammalian cells, especially when compared to STY phosphorylation.

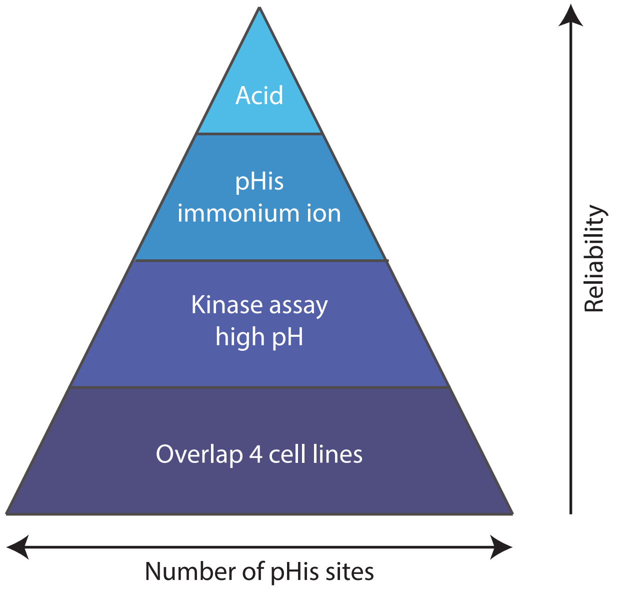


Figure 7: The challenges in studying phosphohistidine. Even though a lot of putative histidine sites were identified in 4 cell types, further experiments significantly decreased the number of confident pHis sites. In the end, only a handful of very reliable sites were detected.

Materials and Methods

Cell culture

Cells were cultured at 37 °C in an atmosphere of 5% CO_2. The cell lines HEK293T, HeLa, A431 and PC9 (ATCC) were grown in their respective growth medium, supplemented with 10% FBS, 1% Pen/Strep and L-glutamine. Cells were grown to 80% confluence, after which they were washed with PBS before being detached using trypsin. Cells were collected in falcon tubes, spun down and washed twice with PBS. Cell pellet was transferred to Eppendorf tubes and stored at -80 °C until further processing.

Bacterial cell growth

The *E.coli* (Subcloning Efficiency DH5α) (Invitrogen) were grown overnight in 500 ml of LB broth (10 g tryptone, 5 g yeast extract and 10 g NaCl per liter of water) at 37 °C. The bacteria were grown to the stationary phase, after which they were spun down and resuspended in 50 ml of medium. Next, the bacteria were spun down again and resuspended in 5 ml of medium. Lastly, the bacteria suspension was transferred to 1 ml Eppendorf tubes, spun down after which the supernatant was discarded. The bacterial pellets were stored at -80 °C until use.

Sample preparation

Cell pellet was dissolved in 5 times the volume of lysis buffer (7M urea, 100 mM Tris (pH 8.5), 1% Triton, 5 mM tris(2-carboxyethyl)phosphine (TCEP), 30 mM chloroacetamide, phosphatase and protease inhibitor, 100 units DNase 1, 1 mM pervanadate, 2.5 mM magnesium and 1% benzonase). Afterwards, the cells were further lysed by sonication (45 cycles, 30 sec on / 30 sec off) in a sonication bath (Diagenode). Cellular debris was pelleted by ultracentrifugation (Beckman Coulter) at 125,000 rcf for 1 hour at 4 °C. The supernatant was then incubated for 2 hours at room temperature, after which the protein concentration was determined by BCA protein assay kit (Thermo Fisher scientific). Subsequently, interfering contaminants were removed by methanol chloroform precipitation. Briefly, samples were diluted with 4 volumes methanol, 1 volume chloroform and 3 volumes water. Samples were vortexed thoroughly, after which the samples were centrifuged (5000 g for 10 minutes at room temperature). A white protein disc should form, after which the upper solvent layer is removed. The samples were diluted with 3 volumes of methanol after which the protein pellet is broken. Precipitate was spun down (5000 g for 10 min at room temperature) and the solvent layer was removed. The protein pellet was air dried.

The protein pellet was dissolved in digestion buffer (1% sodiumdeoxycholate (SDC), 100 mM Tris (pH 8.5), 5 mM TCEP and 30 mM CAA) to a concentration of 3 µg/µl. Proteins were digested overnight at room temperature by 1: 25 trypsin (protease: protein) (Sigma) and 1:100 LysC (Wako). After overnight digestion, the lysate was acidified by adding formic acid (FA) to a final concentration of 0.5%. The SDC precipitate was spun down and the supernatant was desalted by using 3cc SEPPAK SPE cartridges (Waters). Briefly, cartridges were first activated by using 4 x 1 ml of acetonitrile. Then the cartridges were equilibrated by using 1 ml of elution buffer (40% acetonitrile, 0.07% FA). The cartridges were washed four times with 1 ml of 0.1% FA after which the samples were loaded. Subsequently, the cartridges were washed twice with 1 ml of 0.1% FA. Lastly, peptides were eluted by adding 3 times 250 µl of elution buffer. The supernatant was divided in aliquots of 2 mg of peptides and dried using a lyophilizer (Labconco).

Fe^{3+}-IMAC enrichment

The peptide pellet (2 mg) was dissolved in 115 µl of buffer A (30% ACN with 0.07% TFA), pH was adjusted to 2.3 using 10% TFA after which 100 µl was injected unto the Fe^{3+}-IMAC column (Propac IMAC-10 4 x 50 column, Thermofisher scientific). The enrichment starts with 100% buffer A for 7 minutes at a flow of 0.1 ml/min to load the peptides on the column. Then, the flow is increased to 1 ml/min for 5 min at 100% buffer A. Next, the solvent composition is changed to 50% buffer B (0.3% NH4OH) at a flow rate of 1 ml/min for 1.5 minutes. Elution of peptides occurs at a flow rate of 0.5 ml/min, 50% buffer B for 2.5 minutes. The system is equilibrated at 100% buffer A at 1 ml/min for 9 minutes. The phosphopeptides were collected and dried down using a lyophilizer.

Liquid chromatography mass spectrometry

Peptides were reconstituted in 12.5 µl of 20 mM citric acid and 1% formic acid after which 10 µl was injected in a UHPLC 1290 system (Agilent) which was coupled to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). Peptides were first trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) after which they were separated on an analytical column (Agilent Poroshell EC-C18, 2.7 µm, 50 cm x 75 µm). Trapping was performed for 5 minutes in buffer A (0.1% formic acid) at a flow rate of 0.005 ml/min. Separation of peptides was achieved using the following gradient: 6 to 32% buffer B (80% acetonitrile + 0.1% formic acid) in 155 minutes, 100% B for 4 minutes followed by 100% buffer A for 11 minutes. The flow of 300 nl/min was achieved by split flow. The Q Exactive HF-X was operated in a data dependent acquisition mode with positive ionization. Full MS spectra were acquired from 375 – 1600 m/z at 60000 resolution, using an automatic gain control (AGC) target value of 3×10^6 charges and a maximum injection time of 20 ms. A maximum of 12 precursors could be chosen for MS/MS fragmentation, with a dynamic exclusion of 24 seconds. The MS2 fragmentation spectra were obtained at 30000 resolution, an AGC target value of 1×10^5 and a maximum injection time of 50 ms. The fragmentation was performed using HCD at a NCE of 27.

For the kinase assay samples, a shorter gradient was used. Separation of peptides was achieved using the following gradient: 8 to 32% buffer B in 65 minutes, 100% B for 2 minutes followed by 100% buffer A for 11 minutes. Additionally, the MS parameters differ from the previously described parameters as follows: A total of 15 precursors were chosen for fragmentation with a dynamic exclusion of 12 seconds.

Immonium ion triggering

For the immonium ion triggering studies, the method developed by Potel *et al*[20] was used. Mass spectrometric analysis was performed by coupling an UHPLC 1290 system (Agilent) to an Orbitrap Fusion Lumos (Thermo Fisher Scientific) mass spectrometer. The mass spectrometer was operated in a data-dependent acquisition mode with positive ionization. The resulting mass spectra were processed using MaxQuant (version 1.5.3.30) and the Andromeda search engine using the reviewed *Homo sapiens* database (Uniprot, March 2016, 20265 entries). Mass tolerance of precursor ions was set as ± 5 ppm and the mass tolerance of MS2 was chosen as ± 20 ppm. Results were adjusted to 1% PSM and FDR using a target-decoy approach using reverted protein sequences. A score cut-off of 40 was chosen for non-modified peptides and the minimal peptide length was set to 7.

High pH fractionation

To reduce complexity, the phosphopeptides were fractionated by high pH reverse-phase HPLC fractionation using a Kinetex 5u EVO C18 100A column (Phenomenex) on a HPLC 1200 system (Agilent) operating at a flow rate of 200 µl/min. Briefly, the lyophilized peptide pellet was reconstituted in 20 µl of buffer A (10 mM NH4OH, pH 10) and injected. Samples were loaded on the column for 2 minutes at a flow rate of 20 µl/min. Peptides were eluted sequentially using the following gradient: 0 – 14.3% buffer B (10 mM NH4OH / 90 acetonitrile, pH 10) in 20 minutes, 14.3 – 35% buffer B in 15 minutes, 100% buffer B for 2 minutes followed by 40 minutes buffer A for 40 minutes. A total gradient time of 77 minutes was used. Fractions corresponding to 1 minute gradient time were collected using a 1260 infinity fraction collector (Agilent). Only fractions eluting after 8 minutes were collected, which were subsequently concatenated in 10 fractions. The fractions were dried down using lyophilizer and stored at -80 °C.

NME1/2 purification

The NME1/2 kinase needed for the *in vitro* kinase assay was purified according to the protocol of Potel *et al*[23]. Briefly, HEK293T cells were grown Dulbecco's modified Eagle's medium (Lonza), supplemented with 10% FBS (Gibco), 2mM L-glutamine (Lonza) and 1% penicillin- streptomycin (Lonza). After the cells reached 80% confluence, they were detached by scraping, washed twice with PBS and stored in a falcon tube at -80 °C until further use.

The cell pellet corresponding to 1 x 10^8 cells was resuspended in 5x lysis buffer (0.1% Tween, phostop phosphatase inhibitors (Roche) and complete-mini EDTA-free protease inhibitors (Roche) in PBS). The cells in suspension were passed to a dounce homogenizer on ice, where they were lysed by passing them twice through pestle A for 30 seconds and then passing them twice through pestle B for 30 seconds. A minute of rest was maintained between cycles. The lysate was transferred to Eppendorf tubes, after which they were centrifuged at 20,000 g for 10 minutes at 4 °C. Subsequently, the supernatant was filtered through a 0.45 µm filter and protein concentration was determined using BCA protein assay kit. The supernatant was diluted to a concentration of 2mg/ml of protein.

Cyclic nucleotides, cAMP and cGMP (Biolog), were added to the lysate at a concentration of 15 µM to block cyclic nucleotide binding proteins. The lysate was incubated at 4 °C for 30 minutes. After incubation, 8-AET-cGMP agarose beads (Biolog) were added to the lysate at a concentration of 10 µl of beads per mg of protein. The lysate was incubated for 2 hours at 4 °C and afterwards transferred to an empty micro biospin polypropylene column (Biorad). The beads were first washed twice with 1 ml of washing buffer (0.1% Tween and 10 µM cyclic nucleotides in PBS) and next washed twice with 1 ml of PBS. The kinase was then sequentially eluted by adding 2 x 100 µl elution buffer A (10 mM ADP in PBS) followed by 2 x 100 µl elution buffer B (100 mM ADP in PBS). The pH of both elution buffers was set to 7.4 by the addition of sodium hydroxide. The fractions corresponding to the same elution buffer were pooled and the concentration was determined using BCA protein assay kit. The kinase was buffer exchanged to 50 mM Tris (pH 7.6) by using 10 kDa Amicon Ultra 0.5 ml centrifugal filters. Purity of the kinase was determined by SDS-page.

In vitro kinase assay

To allow the determination of substrates of the enriched NME1/2 kinase, peptides first needed to be dephosphorylated. The first method of dephosphorylation was performed by taking the flow through peak of the IMAC analysis, which should only contain non phosphorylated peptides. The second method was dephosphorylation using shrimp alkaline phosphatase (New England Biolabs). Briefly, peptides corresponding to 2 mg of protein were dissolved in 75 µl of 1x CutSmart buffer. Dephosphorylation occurred by adding 2 units of the phosphatase and incubating at 37 °C for 24 hours. Afterwards, the reaction was quenched by heating the sample to 65 °C for 30 minutes.

The kinase reaction was performed based on the protocol of Xue *et al*[23]. Briefly, the peptides were dissolved in 75 µl of kinase buffer (2 mM $MgCl_2$ and 5 mM ATP in 50 mM Tris (pH 7.6)). NME1/2 was added to the peptides in a 1:10 ratio and incubated for 30 minutes at 30 °C. Afterwards, the samples were desalted as described before and lyophilized. The phosphorylated peptides were enriched using Fe^{3+}-IMAC enrichment and lyophilized again.

Data analysis

In order to identify proteins, the raw files were searched using MaxQuant (version 1.6.10.43) and the Andromeda search engine, using the reviewed *Homo sapiens* database (Uniprot, March 2016, 20265 entries). The following parameters were used: digestion using trypsin with a maximum of 2 missed cleavages, carbamidomethylation of cysteine as fixed modification. Phosphorylation on serine, threonine, tyrosine or histidine, oxidation of methionine and N-terminal acetylation were chosen as variable modification. Mass tolerance of precursor ions was set as ± 5 ppm and the mass tolerance of MS2 was chosen as ± 20 ppm. Results were adjusted to 1% PSM and FDR using a target-decoy approach using reverted protein sequences. A score cut-off of 40 was chosen for non-modified peptides and the minimal peptide length was set to 7.

Then entry of phospho[HSTY] was entered into MaxQuant as follows:

- Modification composition: [H O(3) P]
- Position: anywhere
- Type: standard
- New terminus: none
- Specificities: H, S, T and Y residues
- Neutral losses of composition [H(3) O(4) P] for residues S, T and H
- Diagnostic peaks of composition [C(5) H(8) N(3) O(3) P] for H and composition [C(8) H(10) O(4) N P] for Y.

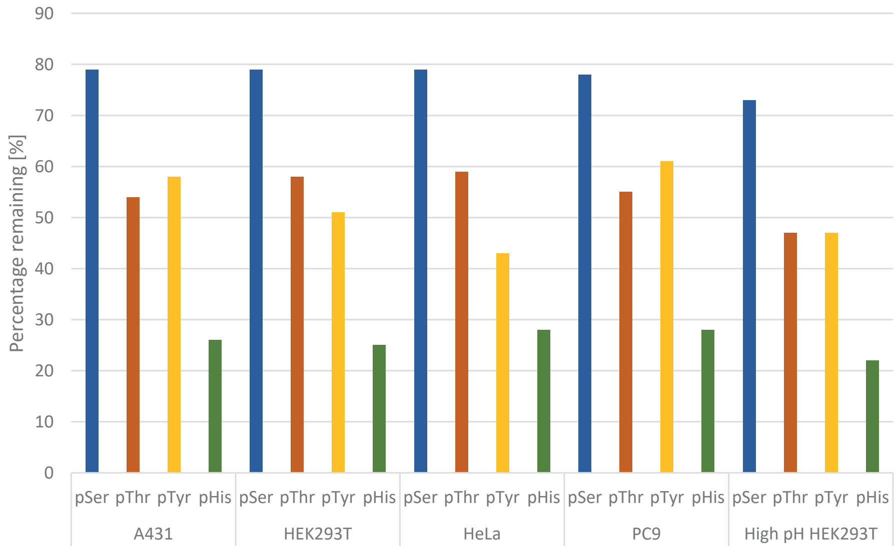
References

1. Jensen, O. N. Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **8**, 33–41 (2004).
2. Li, X. S., Yuan, B. F. & Feng, Y. Q. Recent advances in phosphopeptide enrichment: Strategies and techniques. *TrAC - Trends Anal. Chem.* **78**, 70–83 (2016).
3. Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **26**, 676–687 (2015).
4. Adam, K. & Hunter, T. Histidine kinases and the missing phosphoproteome from prokaryotes to eukaryotes. *Lab. Investig.* **98**, 233–247 (2018).
5. Hunter, T. Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2513–2516 (2012).
6. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Muzio, L. Lo. The crucial role of protein phosphorylation in cell signalingand its use as targeted therapy (Review). *Int. J. Mol. Med.* **40**, 271–280 (2017).
7. Makwana, M. V., Muimo, R. & Jackson, R. F. W. Advances in development of new tools for the study of phosphohistidine. *Lab. Investig.* **98**, 291–303 (2018).
8. Fuhs, S. R. & Hunter, T. pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. *Curr. Opin. Cell Biol.* **45**, 8–16 (2017).
9. Attwood, P. V., Piggott, M. J., Zu, X. L. & Besant, P. G. Focus on phosphohistidine. *Amino Acids* **32**, 145–156 (2007).
10. Potel, C. M., Lin, M. H., Heck, A. J. R. & Lemeer, S. Widespread bacterial protein histidine phosphorylation revealed by mass spectrometrybased proteomics. *Nat. Methods* **15**, 187–190 (2018).
11. Klumpp, S. & Krieglstein, J. Phosphorylation and dephosphorylation of histidine residues in proteins. *Eur. J. Biochem.* **269**, 1067–1071 (2002).
12. Boyer, P. D., Deluca, M., Ebner, K. E., Hultquist, D. E. & Peter, J. B. Identification of phospohistidine in digests from a probable intermediate of oxidative phosphorylation. *J. Biol. Chem.* **237**, PC3306–PC3308 (1962).
13. Kee, J. M., Villani, B., Carpenter, L. R. & Muir, T. W. Development of stable phosphohistidine analogues. *J. Am. Chem. Soc.* **132**, 14327–14329 (2010).
14. Kee, J. M., Oslund, R. C., Perlman, D. H. & Muir, T. W. A pan-specific antibody for direct detection of protein histidine phosphorylation. *Nat. Chem. Biol.* **9**, 416–421 (2013).
15. Fuhs, S. R. *et al.* Monoclonal 1- and 3-Phosphohistidine Antibodies: New Tools to Study Histidine Phosphorylation. *Cell* **162**, 198–210 (2015).
16. Potel, C. M., Lin, M. H., Heck, A. J. R. & Lemeer, S. Defeating major contaminants in Fe 3- immobilized metal ion affinity chromatography (IMAC) phosphopeptide enrichment. *Mol. Cell. Proteomics* **17**, 1028–1034 (2018).

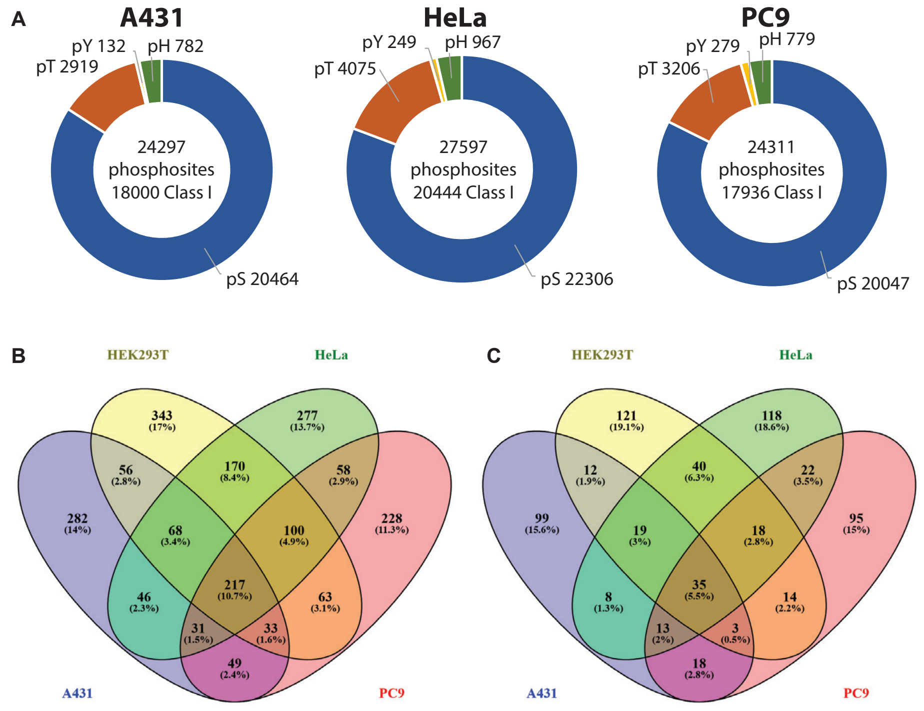
17. Adam, K., Ning, J., Reina, J. & Hunter, T. NME/NM23/NDPK and Histidine Phosphorylation. *Int. J. Mol. Sci.* **21**, 5848 (2020).
18. Gonzalez-Sanchez, M. B., Lanucara, F., Hardman, G. E. & Eyers, C. E. Gas-phase intermolecular phosphate transfer within a phosphohistidine phosphopeptide dimer. *Int. J. Mass Spectrom.* **367**, 28–34 (2014).
19. Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **31**, 557–564 (2013).
20. Potel, C. M. *et al.* Gaining Confidence in the Elusive Histidine Phosphoproteome. *Anal. Chem.* **91**, 5542–5547 (2019).
21. Batth, T. S., Francavilla, C. & Olsen, J. V. Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *J. Proteome Res.* **13**, 6176–6186 (2014).
22. Xue, L. *et al.* Sensitive kinase assay linked with phosphoproteomics for identifying direct kinase substrates. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5615–5620 (2012).
23. Potel, C. M., Fasci, D. & Heck, A. J. R. Mix and match of the tumor metastasis suppressor Nm23 protein isoforms in vitro and in vivo. *FEBS J.* **285**, 2856–2868 (2018).
24. MacDonald, N. J., Freije, J. M. P., Stracke, M. L., Manrow, R. E. & Steeg, P. S. Site-directed mutagenesis of nm23-H1. Mutation of proline 96 or serine 120 abrogates its motility inhibitory activity upon transfection into human breast carcinoma cells. *J. Biol. Chem.* **271**, 25107–25116 (1996).
25. Hardman, G. *et al.* Strong anion exchange-mediated phosphoproteomics reveals extensive human non-canonical phosphorylation. *EMBO J.* **38**, (2019).

Supplementary figures

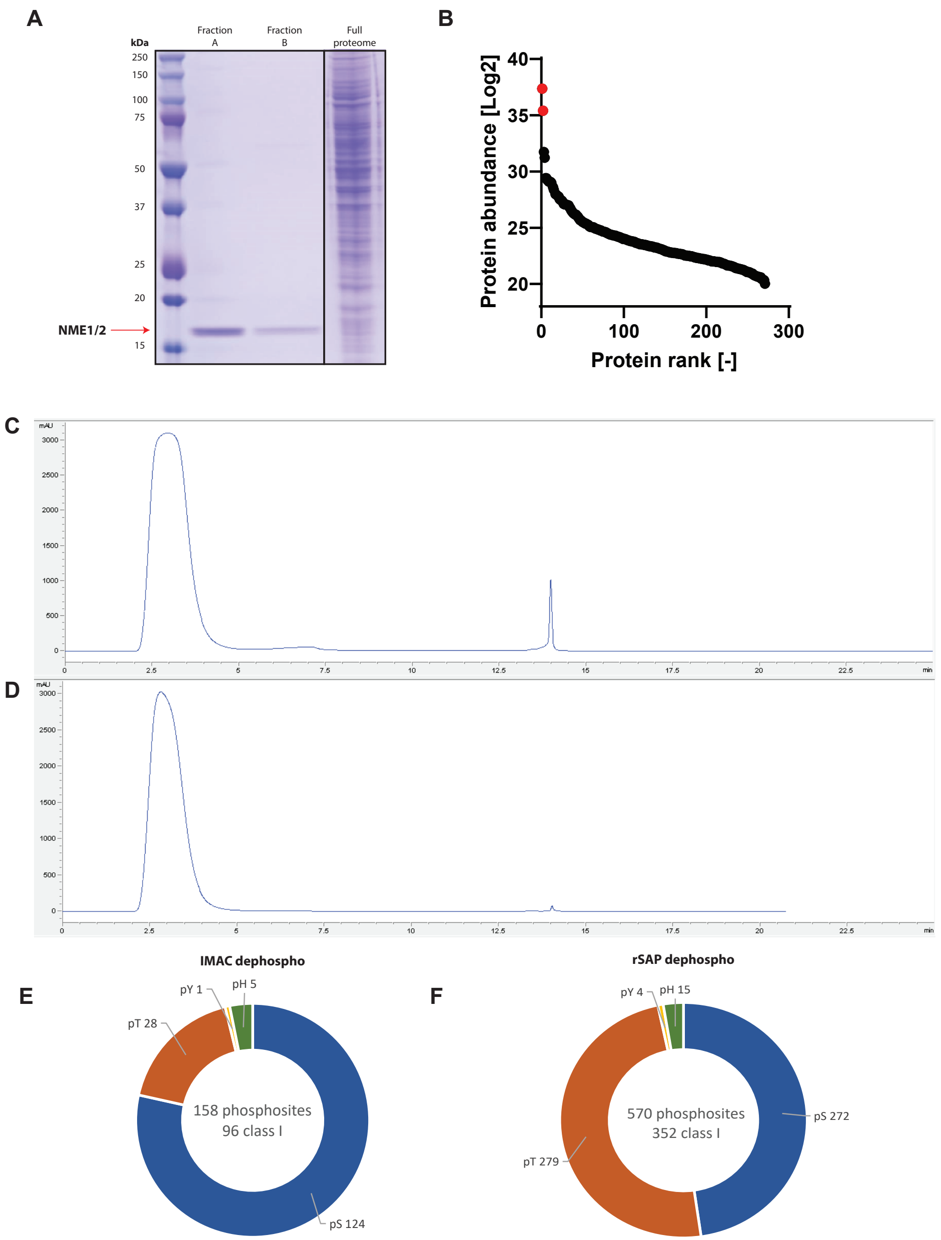
Cell type	Residue	Class I	Total	Remaining [%]
A431	pSer	16143	20464	79
	pThr	1573	2919	54
	pTyr	77	132	58
	pHis	207	782	26
HEK293T	pSer	20244	25621	79
	pThr	2496	4272	58
	pTyr	167	325	51
	pHis	262	1050	25
HeLa	pSer	17668	22306	79
	pThr	2396	4075	59
	pTyr	107	249	43
	pHis	273	967	28
PC9	pSer	15783	20047	78
	pThr	1766	3206	55
	pTyr	169	279	61
	pHis	218	779	28
High pH HEK293T	pSer	24756	34092	73
	pThr	2960	6253	47
	pTyr	204	436	47
	pHis	285	1271	22



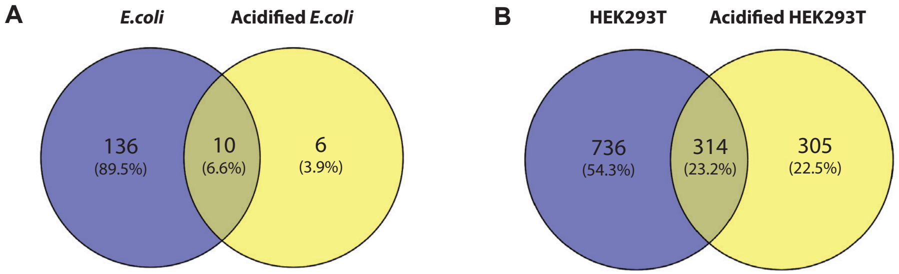
Supplementary figure S1: The distribution of all phosphoresidues in all different cell types. It can be seen that when pHis is considered, bad localization causes a large drop in identified phosphosites when only considering class I sites.



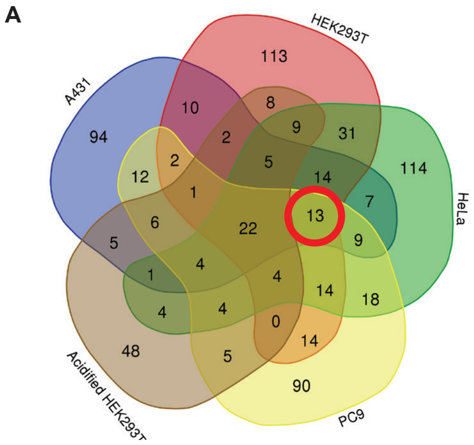
Supplementary figure S2: Distribution of all phosphosites in A431, HeLa and PC. All the different cell types have similar contributions of the different phosphoresidues to the total (A). When comparing all phosphohistidine sites (B) or only the class I sites (C), a subclass of 217 sites (or 35 class I sites) are identified as being histidine phosphorylated in all experiments.



Supplementary figure S3: affinity purification of NME1/2 gives a very pure fraction of kinase. Using SDS-page it was shown that the enrichment of the kinase was successful and pure (A). NME1/2 are shown in red in the abundance plot (B). When re-enriching the phosphopeak from a Fe^{3+}-IMAC enrichment (C), it can be seen that the subsequent enrichment almost has no presence of phosphopeptides (D). Both dephosphorylation methods greatly decrease the amount of phosphopeptides (E, F)

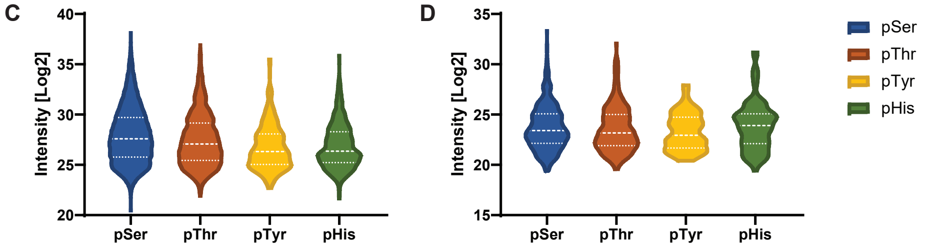


Supplementary figure S4: overlap of pHis sites between acidified and untreated *E.coli* and HEK293T lysate. For *E.coli,* it can be seen that almost all of the pHis was hydrolyzed and not detected anymore after acidification. This shows that acidification of the lysate can act as a control for the true existence of pHis. Only a small overlap was seen between the detected pHis sites, which can be attributed to very stable pHis sites (A). For HEK293T lysate, 45% of pHis sites identified was shared with the acidified lysate, indicating that these are likely false positives (B)



B

Protein	Phosphosite
CCNY	20
CXorf23	13
GCG	92
IWS1	154
IWS1	361
NKAP	160
PPIG	361
SMARCC1	327
SRRM2	1724
SRSF9	213
TOR1AIP1	155
TRA2A	99
WNK1	2031



Supplementary figure S5: pHis sites with the most evidence of being true. When looking at a comparison between all four cell types and the acidified HEK293T, 13 class I sites appear in all 4 cell types, while they are absent from the acidified experiments (A, B). The intensity of pTyr containing peptides is low, and similar to pHis, in HEK293T, which can explain the decrease in localization probability compared to pSer and pThr (C). This is not true for *E.coli*, where pSer, pThr and pTyr have a similar intensity distribution. Interestingly, pHis is the most abundant (D). Bold dashed line is the median value and top and bottom dashed line are respectively the first and third quarter.

Chapter 4

The activated variant of the tyrosine phosphatase SHP2 is more susceptible to oxidation compared to the wildtype

Niels M. Leijten[1], Maaike Allers[2], Jelmer Hoeksma[2], Simone Lemeer[1] and Jeroen den Hertog[2,3]

1. Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
2. Hubrecht Institute – KNAW and University Medical Center Utrecht, Utrecht, the Netherlands
3. Institute Biology Leiden, Leiden University, Leiden, the Netherlands

Manuscript in preparation

Abstract

Signaling through tyrosine phosphorylation is tightly regulated by the coordinated action of tyrosine kinases and phosphatases. Deregulation of this tight balance can result in a variety of diseases and malignancies. Tyrosine phosphatases (PTPs) represent a class of proteins originating from 107 different genes. The group of classical PTPs consists of 38 proteins, for which the phosphatase activity is regulated by an active site cysteine. It has been previously shown that the activity of these PTPs is controlled by redox regulation, through (reversible) oxidation of this catalytic cysteine. Here, we investigated the oxidation behavior of the phosphatase SHP2 and the concomitant change in phosphatase activity. We compared this to the oxidation and activity of the D61G mutant of SHP2, which is known to cause Noonan syndrome. The D61G mutation leads to a more open conformation of the phosphatase, resulting in an increased activity *in vitro*. Here, we showed that the D61G mutant of SHP2 is more readily irreversibly oxidized compared to the wildtype, probably due its more open conformation. This irreversible oxidation decreased the phosphatase activity of D61G-SHP2 under physiological oxidative conditions. In addition, we could show that a SHP2-catalase fusion protein protects the catalytic cysteine against oxidation, resulting in a prolonged activity under oxidative conditions. Finally, preliminary data shows that *in vivo*, the catalase fusion protects the D61G-SHP2 against irreversible oxidation, allowing it to remain active, resulting in severe phenotypic defects *in vivo*.

Introduction

Many cellular processes, such as proliferation, differentiation and cell growth, are mediated by protein phosphorylation[1]. Through phosphorylation, signals from the extra cellular environment can readily be transmitted to the nucleus to influence cellular function[2]. Signaling close to the plasma membrane in eukaryotic systems commonly involves phosphorylation of tyrosine residues, which are phosphorylated by protein tyrosine kinases (PTKs) while they can subsequently be dephosphorylated by protein tyrosine phosphatases (PTPs). When there are aberrations in the regulation of tyrosine signaling, it can lead to a whole range of diseases, including cancer and atherosclerosis[2,3]. Especially mutations in PTPs, which may act as positive or negative regulators of signaling pathways, can play a large role in developing malignancies.

The superfamily of PTPs consists of 107 genes, of which approximately 38 encode for the classical PTPs[1,4]. These include 21 receptor and 17 non-receptor PTPs, which are highly specific for phosphotyrosine residues. All classical PTPs have a conserved catalytic motif, [I/V]H<u>C</u>SXGXGR[S/T]G, where the cysteine acts as nucleophile and is necessary for catalysis[1,5]. This catalytic cysteine has a pKa between 4.5 and 5.5, which causes it to remain in the thiolate state (S^-) at physiological pH and therefore allows it to exert its function[1]. A side effect of this pKa is a high susceptibility to oxidation of the catalytic cysteine by reactive oxygen species (ROS). The ROS, which for example are endogenously generated by NADPH oxidases after growth factor stimulation or exogenously by UV-radiation, induce reversible oxidation of the catalytic cysteine of the PTP, abrogating the nucleophilic function and thereby inhibiting the phosphatase activity[1,5]. The cell may use this reversible oxidation of PTPs to regulate their activity, enhancing or inhibiting the tyrosine phosphorylation dependent signaling[6]. For example, it has been shown that ROS are essential for growth factor signaling[4]. Depending on the extent of the oxidative stimulus, the catalytic cysteine is converted to a sulphenic (SOH), sulphinic (SO_2H) or sulphonic (SO_3H) acid state[5]. When the cysteine is converted to a sulphenic acid state, it is readily reversible and can be activated again by reduction through glutathione or thioredoxin systems[7]. PTPs have developed additional mechanisms to protect against further oxidation[8,9]. For example, when the catalytic cysteine of PTEN is oxidized, it will form an intramolecular disulfide bond, protecting itself from further oxidation. Likewise, PTP1B forms a sulfenylamide bond upon oxidation, which also protects against further oxidation. However, when the cysteine is converted to the sulphinic or sulphonic acid state due to prolonged or harsh oxidation, it will be irreversibly oxidized and the PTP will lose its function[5]. This enhanced level of irreversibly oxidized PTPs has been detected in multiple cancer cell lines, which suggests that the inactivation of PTPs is a mechanism in the transformation process[9]. Therefore, it would be very interesting to investigate at which stimulus PTPs are irreversibly oxidized and the effect which this has on cellular function.

One PTP which is involved in multiple cytokine and growth factor initiated signaling cascades is Src homology region 2 domain-containing phosphatase 2 (SHP2)[10]. SHP2 is a cytosolic PTP that is ubiquitously expressed in multiple tissue and cell types. It plays a role in multiple signal transduction processes, including the Ras-Raf-MAP, Jak-Stat and PI3 kinase pathways[10,11]. Through this signaling, it is involved in multiple cellular processes such as proliferation, cell migration, stem cell self-renewal and differentiation[12]. Interestingly, SHP2 is one of the rare PTPs which appear to promote activation of signaling pathways, instead of downregulating them[11,13]. When mutations occur in the gene coding for SHP2, it can lead to malignancies such as myeloid neoplasms and involvement in cancer[13,14].

SHP2 consists of two N-terminal SH2 domains and a phosphatase domain at the C-terminus. The SH2 domains of SHP2 are used to specifically bind to phosphorylated tyrosine residues on substrate molecules, thereby facilitating the phosphatase function[10]. However, the SH2 domains may also inhibit the function of the phosphatase[15]. Under basal conditions, when phosphatase activity is low, the N-terminal SH2 domain is positioned over the phosphatase domain[16] (Fig. 1A). Because of this, the catalytic cleft of the phosphatase domain as well as the phosphotyrosine binding pocket of the SH2 domain are obstructed. Through this mechanism, the N-SH2 domain efficiently inhibits the function of the phosphatase. However, when SHP2 engages a phosphotyrosine containing interaction partner, the N-SH2 domain will lose its complementarity to the phosphatase domain and will interact with the ligand (Fig. 1A). Subsequently, the phosphatase will adopt an open conformation, freeing the catalytic cysteine and restoring phosphatase activity[6]. Another mechanism how SHP2 activity is regulated is by employing the previously mentioned ROS[7]. For example, it has been shown by Meng *et al*[6] that SHP2 undergoes inactivation by ROS in PDGF-stimulated Rat 1 cells, which was shown to be necessary for PDGFR function. Additionally, collagen induced oxidation of SHP2 in platelets will lead to platelet aggregation and thrombosis, which are necessary for wound healing[17].

When mutations occur in the PTPN11 gene coding for SHP2, it may lead to a range of diseases and syndromes. One syndrome that is caused by mutations in SHP2 is Noonan syndrome (NS). NS is characterized by congenital heart disease, distinctive facial features, short stature and learning difficulties[18]. It has a prevalence of 1 in 1000-2500 births and it was shown that 40-50% of patients suffering from NS had mutations in SHP2[19]. In rare cases, NS patients can progress to juvenile myelomonocytic leukemia, which can be fatal if not treated by bone marrow transplantation. One mutation known to cause NS is the D61G mutation, which is localized in the N-SH2 domain of SHP2[18] (Fig. 1B). The aspartic acid residue at position 61 is essential for hydrogen bonding between the N-SH2 domain and multiple residues close to the catalytic P-loop of the phosphatase domain (residues 458 – 464), thereby playing an important role in the interaction between the N-SH2 and PTP domains[18]. However, when this aspartic acid is changed to a glycine the hydrogen bonds are abolished and the surface charge is changed from very negative to neutral, greatly weakening key interactions between the two domains. Consequently, the N-SH2 domain does not interact effectively anymore with the phosphatase domain, causing the protein to always be in an opened conformation, resulting in a strong increase in catalytic activity of the phosphatase (Fig. 1C).

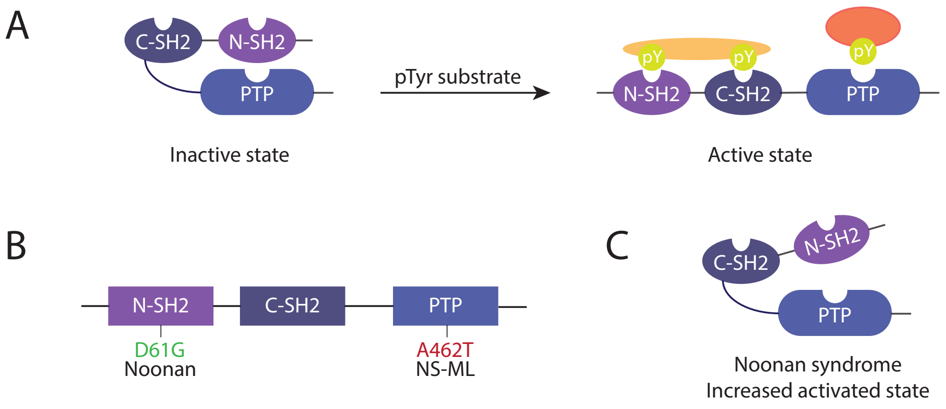


Figure 1: The catalytic mechanism of tyrosine phosphatase SHP2 and its mutations. In the basal state, the N-SH2 domain of SHP2 interacts with the catalytic PTP domain, causing the phosphatase to be inactive. However, when a substrate containing pTyr residues is in proximity, the N-SH2 domain will interact preferentially with this substrate, freeing up the PTP domain and allowing it to exert its phosphatase activity (A). Multiple mutations are known for SHP2, which can either be activating (NS) or inactivating (NS-ML) (B). For NS, there is little to no interaction between the N-SH2 and PTP domains, causing the protein to always be in an open conformation and be catalytically active.

Since it has been shown that ROS play an important role in the regulation of PTPs, among which SHP2[6,20], it would be interesting to investigate whether there is a difference of susceptibility to oxidation between the wildtype SHP2 and the D61G Noonan mutant. Since the Noonan mutant is always in an open conformation, it might be more susceptible to ROS compared to the wildtype phosphatase. This might have an effect on the function of the D61G-SHP2, since oxidation abrogates the catalytic function. Therefore, by studying the oxidative behavior of the catalytic cysteine in wildtype and D61G-SHP2, conclusions might be made regarding the functionality of the enzyme in Noonan syndrome and the influence of the closed N-terminal SH2 domain. Here, we attempted to elucidate this by using a combined approach of mass spectrometry and phosphatase assays.

Since all different oxidative modifications have distinct masses, it is possible to distinguish between them using mass spectrometry. We and others have already shown the potential of using mass spectrometry to probe the oxidative status of PTPs. For example, we have used a differential alkylation approach in combination with MALDI-TOF to measure the level of oxidation in different PTPs, such as both domains of RPTPα, PTP1B and LAR-D1[21]. The pH of the microenvironment of the catalytic cysteine plays an important role in the rate of oxidation. In the research of Karisch *et al*[22], they performed an oxidation step, after which the unreacted active PTPs were alkylated using N-ethylmaleimide to prevent further modifications. Subsequently, the reversibly oxidized species were reduced and hyperoxidized to a sulphonic acid state using pervanadate. By performing immunopurification using antibodies against sulphonic acid containing peptides, the oxidized peptides were identified using mass spectrometric analysis. Using this method, they identified multiple receptor and non-receptor PTPs in multiple tissue types and cell lines. By combining this with selected reaction monitoring, the oxidation levels were quantified.

Here, we applied a differential alkylation approach to detect all the different oxidative states of the catalytic cysteine of SHP2 (Fig. 2). By measuring the oxidative status of both SHP2 and the D61G/Noonan mutant after oxidative stimuli, we hoped to elucidate the protective role that the closed N-SH2 domain may have on the oxidation status of the catalytic cysteine and how this affects the enzymatic activity of the phosphatase.

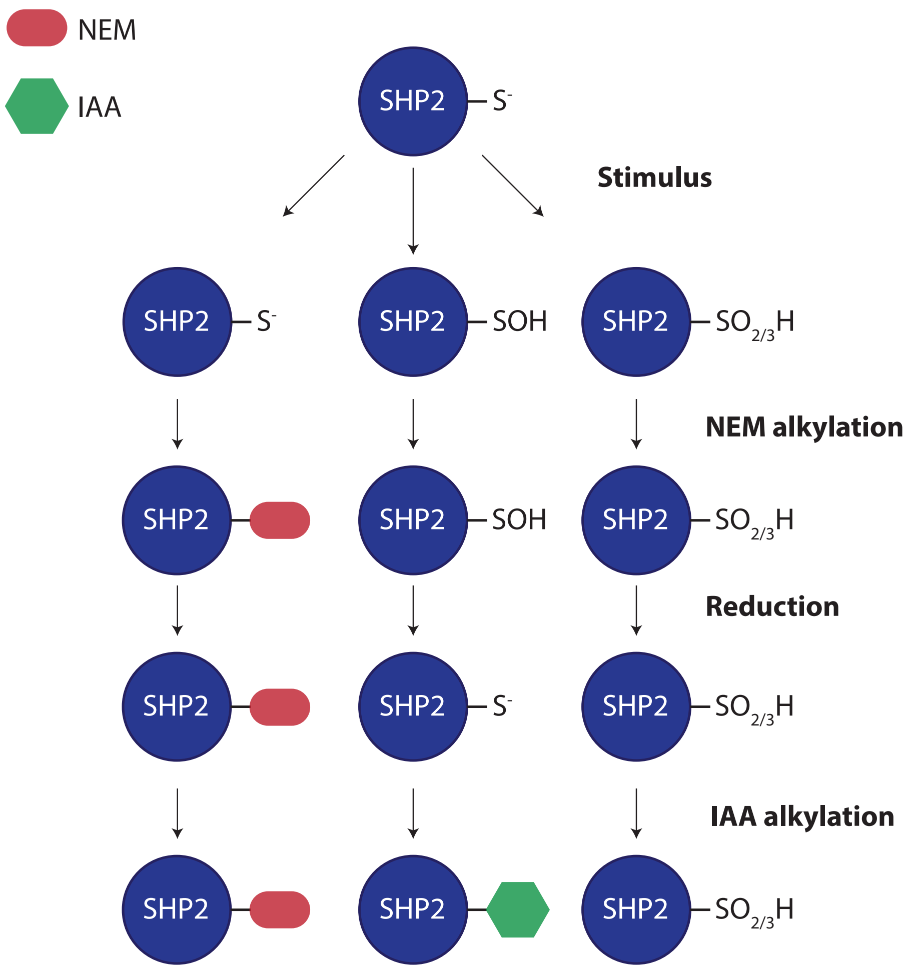


Figure 2: Differential alkylation approach used in this study. Recombinant SHP2 was stimulated with either MilliQ water or hydrogen peroxide at different concentrations to oxidize the catalytic cysteine. This will lead to multiple modifications: the cysteine can stay in the thiolate state, it can be converted to the reversible sulphenic (SOH) acid state or to the irreversible sulphinic (SO_2H) or sulphonic (SO_3H) acid modification. The unreacted cysteines were alkylated by adding N-ethylmaleimide (NEM). Subsequently, the singly oxidized species were reduced by adding DTT, after which the now reduced cysteines were alkylated by adding iodoacetic acid (IAA). After these steps, all populations of peptide were stable and could be measured by mass spectrometry.

Results

Determining the oxidation status of SHP2 by mass spectrometry

First, we wanted to determine the rate of oxidation of the catalytic cysteine of SHP2 and the D61G mutant under oxidizing conditions, *in vitro,* using mass spectrometry. By doing so, the direct influence of the oxidative stimulus on the phosphatase could be measured. To this end, we performed a differential alkylation experiment on recombinantly expressed SHP2 and the D61G mutant. As depicted in figure 2, recombinant proteins were first treated with DTT to bring them in the reduced state. After incubation with a specific stimulus, reactive free thiols of still reduced cysteines were alkylated with N-ethylmaleimide (NEM) to prevent further oxidation at a later stage and to make the peptide amenable for mass spectrometric analysis[22]. Reversible and irreversible oxidized cysteines will not react under these conditions. Reversibly oxidized cysteines (i.e. sulphenic acid modification) were reduced by DTT and subsequently alkylated by iodoacetic acid (IAA). This treatment will not affect the irreversibly oxidized cysteines nor the previously NEM-alkylated cysteines. The irreversible sulphinic and sulphonic acid modifications are readily stable and are already suitable for direct mass spectrometric analysis. In short, in this experimental setup, NEM-alkylated cysteines represent non-oxidized (i.e. reduced) cysteines, IAA-alkylated cysteines represent reversibly oxidized cysteines, and sulphinic and sulphonic cysteines represent the irreversibly oxidized cysteines. Because the different modifications add different masses to the catalytic cysteine, these differences in oxidation status can be readily detected using mass spectrometry at the peptide level.

Recombinant SHP2 and D61G-SHP2 were bound to beads, after which they were subjected to two hours of treatment with water or two different concentrations of hydrogen peroxide (H_2O_2). After following the differential alkylation protocol, the proteins were digested into peptides, desalted and the oxidation status of the catalytic cysteine was determined by LC-MS/MS. The catalytic cysteine containing peptide (QEGITGAGPIVVHCSAGIGR) with the different modifications was readily detected by LC-MS/MS. In addition to a change in mass, the addition of different modifications to the peptide also resulted in a change in retention time (Fig. 3A, B). For the NEM alkylated peptide, 2 peaks were detected, but this was due to the presence of diastereomers[23]. Extracted ion chromatograms were constructed, which were used to determine the area under the curve (Fig. 3A, B). By comparing these areas, the contribution of all different oxidative states was determined (Fig. 3E).

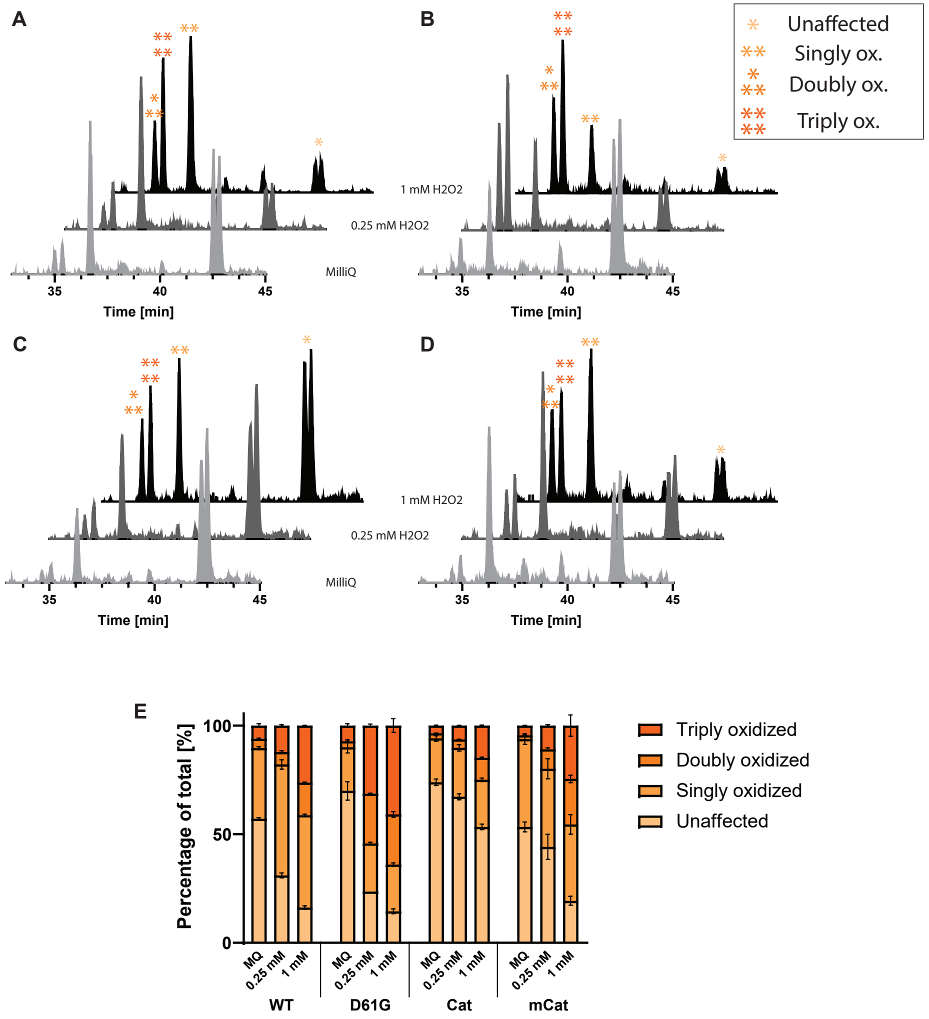


Figure 3: The oxidation patterns of wildtype and mutant SHP2. Extracted ion chromatograms for SHP2 (A), D61G (B), SHP2-Catalase (C) and SHP2-mCatalase (D). All the different oxidation states were detected. Comparison of the normalized different oxidation states of the PTPs at different oxidative stimuli (E).

When SHP2 and the D61G mutant were treated with water as a control, a comparable amount of irreversibly oxidized peptide was detected, while a slight difference was seen for singly oxidized and reduced peptides. However, when a stimulus of 0.25 mM H_2O_2 was used, a large difference in oxidation status was detected between WT and

D61G-SHP2. A significantly larger population of the catalytic cysteine in the D61G mutant was shown to be irreversibly oxidized, whereas WT SHP2 was mainly reversibly oxidized under these conditions. This effect was even more pronounced when the H_2O_2 concentration was increased to 1 mM. When comparing the wildtype SHP2 with the mutant under oxidative stress, a significantly larger population of irreversible doubly and triply oxidized peptide was seen in the D61G mutant. This clearly indicated that the D61G mutant is more prone to irreversible oxidation compared to the wildtype. It was interesting to see that at 1 mM of H_2O_2 stimulation both the WT and D61G mutant had a similar percentage of oxidation, around 85% of the total. However, for the D61G mutant this mainly resulted in irreversible (65%) oxidation, while for the wildtype the oxidation was mainly reversible (42%). This showed that under oxidative stress conditions, the D61G mutant SHP2 was more readily irreversibly oxidized and thereby potentially catalytically inactivated, while the wildtype phosphatase could possibly regain more of its catalytic function through reduction.

Besides the catalytic cysteine, other cysteines might also be oxidized due to the treatment with H_2O_2. Indeed, we identified 3 other peptides containing cysteines (Cys104, Cys 260 and Cys 487). Analysis of the oxidation behavior of these peptides indicated that these cysteines partially but exclusively get singly, and thus reversibly, oxidized due to the treatment with H_2O_2. No big difference in oxidation behavior for these cysteines was detected between WT and the D61G mutant under H_2O_2 treatment (Supplementary Figure 1). This indicates that the specific microenvironment of the catalytic cysteine leads to its unique (irreversible) oxidation behavior in WT and D61G. Unfortunately, the two backdoor cysteines of SHP2 (Cys333 and Cys367), which are reported to act as protectors against irreversible oxidation of the catalytic cysteine of SHP2 by forming an intramolecular disulfide bond, were not detected in our assay[24]. A better understanding of the oxidation status of these cysteines might potentially contribute to a better understanding of the redox regulation of SHP2.

Evaluating the activity of SHP2 by phosphatase assays

To test whether the differences in oxidation status detected in the mass spectrometric measurements indeed affected the activity, phosphatase activity assays were performed. In this assay, p-nitrophenyl phosphate (pNPP) was used as a substrate for the phosphatases. When the phosphate group is removed by the phosphatase, pNPP is converted to para-nitrophenol, which gives an intense yellow color under alkaline conditions, which is quantitatively measured at 405 nm. The phosphatase assay was performed after oxidation of SHP2 and D61G mutant, in the presence of DTT, potentially allowing reversibly oxidized catalytic cysteines to return to the reduced and thus active state. Therefore, when oxidation of the catalytic cysteine would lead to irreversible oxidation and hence inactivation, this would be detected through a loss of signal. Here, we measured the activity of SHP2 and the D61G mutant after 2 hours of differential oxidative treatment (Fig. 4). To compare the global decrease in catalytic activity after oxidation, the activity was normalized to treatment with water for all experiments. When comparing the activity of SHP2 to its activated mutant D61G, it was evident that there was a large decrease in activity in the mutant compared to the wildtype after treatment with 0.25 mM H_2O_2. The enzymatic activity of the WT SHP2 was decreased to 46% of the initial activity, whereas the activity of the D61G mutant was reduced to 9% of the initial activity. For the oxidation with 1 mM H_2O_2, the effect was even more pronounced. For WT SHP2, the activity was reduced to 28%, whereas for D61G the activity was almost entirely absent (1.5% activity remaining). These observations are in line with the mass spectrometric analysis, in which it was shown that D61G was more extensively irreversibly oxidized under the conditions used, compared to WT SHP2. Here, we also convincingly showed that such irreversible oxidation leads to extensive inactivation of the D61G mutant, whereas activity of WT SHP2 is reduced, but not abolished.

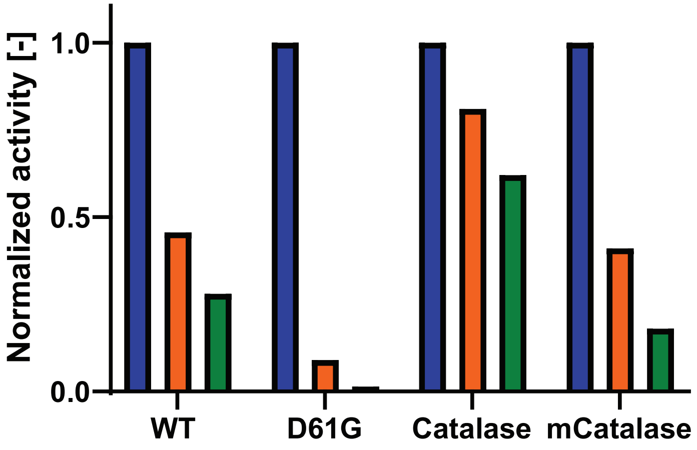


Figure 4: D61G-SHP2 is more sensitive to oxidation and catalase protects against oxidation. SHP2 (mutant) and catalase fusions were treated with water (control), 0.25 mM or 1 mM H_2O_2 for 2 h. Subsequently, samples were treated with DTT and PTP assays were done. Activities relative to the water control are depicted.

Catalase as protector against hydrogen peroxide

Catalases are enzymes which play an important role in the cellular defense against ROS by metabolizing H_2O_2 to form water and oxygen[25]. It has been shown that catalases can efficiently protect cells against hydrogen peroxide. For example, Sundaresan *et al*[26] showed that the increase in H_2O_2 due to PDGF treatment of vascular smooth muscle cells was diminished by increasing the concentration of catalase. Also, Bae *et al*[27] showed that the EGF-induced increase in H_2O_2 in A431 cells was completely abolished by the introduction of catalases. Here, we investigated whether catalase could protect SHP2 against oxidation of the catalytic cysteine. To ensure catalase activity in the proximity of the SHP2 protein, catalase was fused to the C-terminal side of SHP2, to create a SHP2-catalase fusion protein. In addition, as a control, to detect if such fusion would interfere with the functionality of SHP2, a fusion protein of SHP2 with an inactive mutant of catalase (H95F) was also created. Both constructs were treated with H_2O_2 and subjected to the differential alkylation protocol (Fig. 2), as described previously. After protein digestion, the oxidation status of the catalytic cysteine was again determined by LC-MS/MS (Fig. 3C-D).

This analysis showed that the fused catalase indeed protected SHP2 against oxidation (Fig. 3E). Even upon treatment with only water, the SHP2-catalase fusion protein was already less oxidized compared to the wildtype SHP2 without catalase fusion. This indicates that even water treatment caused some minor oxidation. When the SHP2-catalase fusion protein was treated with 0.25 mM H_2O_2, 67% of the catalytic cysteines were present in the reduced form, compared to 31% for WT SHP2. Also, the amount of irreversibly oxidized species was smaller when compared to the wildtype SHP2. At 1 mM H_2O_2, again the effect was more pronounced. For the SHP2-catalase fusion, 53% of catalytic cysteine was still present in the reduced form, whereas for the WT SHP2, only 16% was in the reduced form. Again, also at 1 mM H_2O_2 the percentage of irreversibly oxidized cysteines was higher in the WT SHP2 compared to the catalase fusion. Taken together, these results clearly showed the protective behavior of the catalase on SHP2. To make sure the observed effect was truly caused by catalase activity and not an artefact of the SHP2-catalase fusion, a SHP2 fusion was created with an inactive catalase mutant (H95F). This SHP2-mutant catalase fusion protein was subjected to oxidation with 0.25 mM or 1 mM H_2O_2. As can be seen, this fusion protein with an inactive catalase showed a strikingly similar oxidation behavior compared to wildtype SHP2. This clearly indicated that the protection against H_2O_2 induced oxidation of the catalytic cysteine is indeed caused by the catalase. The extent of oxidation of the other detected cysteines (Cys104, Cys260, Cys487) was also reduced by the catalase fusion, whereas oxidation of the catalase mutant was again similar to the WT (Supplementary Figure 1). This convincingly showed that the addition of catalase to form a fusion protein does by itself not excessively change the oxidative behavior of SHP2.

Also, for the SHP2-catalase fusion proteins phosphatase assays were performed in order to link oxidative status to enzymatic activity (Fig. 4). When the catalase fusion protein was treated with 0.25 mM of H_2O_2, circa 81% of the activity remained. Even at 1 mM H_2O_2, which greatly inactivated the wildtype SHP2, circa 62% of the original activity remained. This clearly showed the potential of catalases to protect against oxidation by H_2O_2. However, when the inactive mutant catalase was fused to SHP2, none of the protective behavior remained. The activity profile of the mutated catalase fusion protein was similar to wildtype SHP2, again showing that the fusion of the catalase to SHP2 did not by itself change the oxidative behavior nor the enzymatic activity of the PTP. Additionally, this showed that the phosphatase is still active after the addition of the catalase.

Taken together, these results showed that the D61G mutant of SHP2 is more readily oxidized compared to the wildtype, thereby reducing its phosphatase activity. In addition, catalase fusion proteins are capable of protecting the catalytic cysteine of SHP2 against oxidation, and by doing so, keeping a high enzymatic activity under oxidative conditions. This effect was truly caused by the catalase activity and not merely an artefact of protein fusion.

Discussion

Reversible oxidation of the catalytic cysteine of PTPs takes place in cells and is assumed to be an important regulator of its function[9]. It has a physiological relevance, for example by locally inactivating PTPs close to activated PTKs to facilitate signal propagation[6]. At the same time, it can also play a role in pathophysiological processes. It was shown that increased levels of reversible and irreversibly oxidized PTPs were detected in cancer cell lines, suggesting a role of oxidative inactivation of PTPs in the transformation process[28]. Therefore, investigating the oxidation status of PTPs can aid the understanding of its function and biological consequences. There is a distinct difference between the susceptibility of different PTPs to ROS[21], which is caused by differing structural features of the catalytic domain or can be caused by additional regulatory domains of the PTP[9]. One of the regulatory domains which might change the susceptibility to oxidation of the catalytic cysteine is the N-SH2 domain of SHP2. When inactive, the N-SH2 domain is folded over the phosphatase domain (Fig. 1A), obscuring the catalytic cleft of the phosphatase domain[16]. Indeed, in the research of

Weibrecht *et al*[9] it was shown that the free catalytic domain of SHP1 and SHP2 have a higher susceptibility to oxidation compared to the full-length phosphatase. This effect was especially distinct for SHP1, which has a similar structure as SHP2. These results showed that the closed N-SH2 domain has a protective influence on oxidation of the catalytic cysteine of SHP1 and SHP2.

A downside to this previous study was that they compared the oxidation of wildtype SHP2 to the isolated catalytic domain, which might miss key aspects of the interaction. Here, we attempted to determine the consequences of an oxidative stimulus on SHP2 and the D61G mutant, which is known to cause Noonan syndrome. The D61G mutant of SHP2 is always in an open conformation, which might allow easier access of ROS to the catalytic cysteine compared to the closed conformation of the wildtype[18]. Through this comparison, a better understanding of the protective behavior of the N-SH2 domain was obtained, while at the same time the ratios of reversible to irreversible oxidized PTP were determined. These experiments were performed by using a differential alkylation approach, which allowed the distinction between all the different oxidation forms of SHP2 using mass spectrometry, shedding more light on the regulation of SHP2 activity by ROS.

First, we compared the influence of hydrogen peroxide on SHP2 and the NS mutant *in vitro* using mass spectrometry. It was shown that after 2 hours of treatment with either 0.25 or 1 mM of H_2O_2, that the NS mutant was more easily irreversibly oxidized compared to the wildtype (Fig. 3E). This difference was especially distinct after treatment with 0.25 mM H_2O_2. The concentrations of H_2O_2 used in these studies (0.25 and 1 mM) were in the same range which occur after cellular stimuli, making them biologically relevant[21]. Interestingly, the contribution of non-oxidized species in WT and the D61G mutant were similar, which showed that the D61G mutant transcends more easily to the irreversible oxidative states compared to the WT, while the ratio of oxidized to non-oxidized PTP stayed the same. This might have an effect on the increased catalytic activity of the D61G mutant *in vivo*, which might be partially inactivated due to the ease of irreversible oxidation of the catalytic cysteine. Therefore, the expected activating effect of the open conformation of the D61G mutant might be (partially) quenched due to irreversible oxidation. This characteristic might explain an interesting phenomenon discussed in literature, where the activating NS mutant has the same disease phenotype as the inactivating mutations observed in NS with multiple lentigines (NS-ML, previously known as LEOPARD) syndrome[29,30]. NS-ML is caused by an inactivating mutation in the phosphatase domain of SHP2 (Fig. 1B), which almost completely abolishes the catalytic function, in contrast to the activating mutations in NS. However, both these syndromes share clinical features such as cardiovascular anomalies, hypertrophic cardiomyopathy, short stature, facial abnormalities and mental retardation[30,31]. The same holds true for zebrafish, where expression of NS and NS-ML mutants caused overlapping phenotypes in embryos[29]. We hypothesized that the open conformation of NS might cause it to be irreversibly oxidized by ROS, causing inactivation of the PTP. This inactivated mutant may then have the same disease phenotype as the genetically inactivated NS-ML mutant. However, this hypothesis needs to be proven by measuring the oxidative behavior and activity of the NS-ML mutant under oxidative stress, while also determining the oxidative behavior and activity of the NS mutant *in vivo.*

It is known that the antioxidant enzyme catalase can protect against H_2O_2[26,27]. We hypothesized that this protective behavior of catalase could be employed to protect SHP2 against H_2O_2 by combining them in one fusion protein. This concept was already shown in the past. In the research of Andre *et al*[32] they found that the alkane producing enzyme aldehyde-deformylating oxygenase (ADO) was reversibly inhibited by H_2O_2. To combat this, they created a catalase-ADO fusion protein, which they hypothesized should be able to detoxify H_2O_2 near the active site of the enzyme. Indeed, the catalase-ADO fusion protein was insensitive to inhibition by oxidation and continued to show activity after treatment with H_2O_2. To see whether catalases also protected SHP2 against the ROS induced oxidation of its catalytic cysteine, SHP2-catalase fusion proteins were made. These fusion proteins were subjected to the same oxidative treatment as the wildtype SHP2 and the level of oxidation was read out using mass spectrometry. It was shown (Fig. 3E) that catalases readily protected SHP2 against H_2O_2. Even after treatment with water, a larger proportion of unoxidized cysteine was detected for the catalase fusion protein compared to the wildtype. This showed that catalases also protected against transient oxidation in water. However, the greatest effect was seen when the fusion protein was treated with 0.25 or 1 mM H_2O_2. More than double of the peptides remained in the unoxidized state in the fusion protein (67%) compared to the wildtype (31%). Additionally, 22% remains in the reversible sulphenic acid state, compared to 51% for the WT. These findings clearly showed that it is possible to protect SHP2 against oxidation through the activity of catalases. To validate this, a fusion protein containing genetically inactivated catalase (H95F mutation) was made and tested for its behavior under oxidative stress. The inactive catalase fusion protein showed oxidative behavior similar to the WT SHP2, proving that the protective behavior of the fusion proteins stemmed from the catalytic function of the catalase. Additionally, this indicated that the addition of catalase to SHP2 did not interfere with oxidation. These results together showed that catalases are very efficient in protecting SHP2 against oxidation. It has been suggested by Yano *et al*[33] that SHP2 can bind to tyrosine phosphorylated catalase via integrin signaling, which subsequently causes it to be protected against ROS. This indicates that the catalase-induced protection of SHP2 might also have a biological role, which might be exploited for possible therapeutic interventions. No fusion proteins consisting of D61G-catalase were made, but it would be very interesting to see whether catalase also protects the more easily oxidized D61G mutant against ROS *in vitro*.

To validate the mass spectrometric findings, phosphatase assays were performed. These assays directly measured the activity of the phosphatase after oxidative treatment, which should correlate with the amount of irreversibly oxidized phosphatase. It should be noted that DTT was added to the incubation step before the read-out of the phosphatase activity. This caused the singly oxidized cysteines to be reduced to the thiolate state and able to exert their function again. Therefore, the phosphatase assay measured the activity of both unaffected and singly oxidized cysteine after treatment, which also is the pool of SHP2 that might still exert their function *in vivo*. The samples which were generated for the mass spectrometry assay were also used for the phosphatase assay to remove any sample preparation influences. It can be seen (Fig. 4) that the activity in the D61G mutant was diminished extensively compared to the wildtype upon oxidative treatment. At both concentrations of H_2O_2, the activity in the mutant was lower compared to the wildtype phosphatase. Notably, at 1 mM of H_2O_2 almost no activity was detected for the D61G mutant, indicating that it was totally inactivated by the ROS. Additionally, phosphatase assays were performed for the fusion proteins. Here, the addition of catalase again provided a protective behavior against ROS, reflected by an increased activity under oxidative stress compared to the wildtype protein. This protective behavior was again abolished by the inclusion of the mutated catalase. These results showed that the phosphatase and mass spectrometric read-outs gave similar results and may be used to give a complementary view of the oxidative status and activity of the phosphatase. Again, no D61G-fusion protein was subjected to the phosphatase activity assays, which would be very interesting to do.

The results of the phosphatase assay showed that the SHP2-catalase fusion protein still exerted phosphatase activity, indicating that the addition of catalase to SHP2 has no effect on its catalytic function or structure. This opens up the path for these kind of fusion proteins in the investigation of other PTPs, such as PTP1B[34], LAR and RPTPα[21]. Besides PTPs, this approach might also be extended to other proteins that are sensitive to H_2O_2 such as thioredoxin, nucleoredoxin and the Src family kinase Lyn[8,35]. Therefore, the proof of principle work on catalase fusion proteins reported here can open up a completely new research field, where the effect of H_2O_2 on redox sensitive proteins can be investigated more thoroughly *in vitro* and *in vivo*.

After establishing that the wildtype and D61G-SHP2 were differentially regulated by H_2O_2 *in vitro*, it would be very interesting to expand these observations to the *in vivo* situation. By using the catalase and mutant-catalase fusion proteins, the direct effect of H_2O_2 on PTP activity *in vivo* may be identified. Preliminary experiments were performed by expressing the different SHP2-catalase fusion proteins in zebrafish embryos and monitoring the phenotypic changes. This time, D61G-catalase fusion proteins were included. It is known that expression of D61G-SHP2 in zebrafish embryos results in shorter embryos compared to non-injected counterparts due to defects in convergence and extension cell movements during epiboly[29]. Additionally, the embryo suffers from craniofacial abnormalities and defects in cardiac development. Therefore, the embryos transfected with the fusion proteins were screened for these abnormalities to detect issues with SHP2 signaling. The preliminary results can be seen in supplementary figure 2. Embryos were classified as normal, small, severe or dead. Embryos classified as small also missed their swimming bladder and lower jaw, whereas embryos classified as severe were truncated and developed heart edema. Expression of wildtype SHP2 did not induce distinct differences in phenotype at 2 and 5 days post fertilization (dpf). Most embryos developed normally, with a small proportion being classified as small. However, a clear phenotype was observed in the D61G-mutant catalase fusion expressing embryos, which resembled the D61G-SHP2 expressing embryos, in that more small and severe classified embryos were seen compared to wildtype SHP2 expressing embryos. This showed that the D61G mutation in SHP2 has an effect on the phenotype of zebrafish as expected. Interestingly, the D61G-catalase fusion protein showed an even bigger increase in deformities in the embryo at 2 and 5 dpf. Likely, this stemmed from the fact that wildtype catalase protected the D61G-SHP2 mutant from inactivation through oxidation, maintaining its increased activity and causing more severe developmental defects. The effect was most pronounced at 5 dpf, where most embryos died. It is noteworthy that the expression of catalase fusions to wildtype SHP2, be it wildtype or mutant catalase, did not affect embryonic development. Taken together, these preliminary results suggest that catalase protects D61G-SHP2 against oxidation *in vivo* and warrant further investigation and validation of these experiments.

Following these promising results, it would be very interesting to assess the oxidative status of these fusion proteins *in vivo*. By performing these experiments, the extent to which SHP2 and the D61G mutant are oxidized *in vivo* may be detected, while simultaneously the protective behavior of catalase is seen. Additionally, the interaction partners of these fusion proteins might be found. The more open conformation of D61G-SHP2 might allow it to interact with a completely different range of protein partners when compared to the wildtype, which might play a role in the NS phenotype.

All these questions can be answered by performing co-immunopurification experiments. First, the transfected SHP2-catalase fusion protein may be purified from cellular lysate by immunoprecipitation using antibodies attached to sepharose beads, which will facilitate the subsequent differential alkylation approach (Fig. 2) to identify the oxidative status of the fusion proteins *in vivo*. Next, with a similar approach, the interaction partners of SHP2 may be identified. These will co-purify in immunoprecipitations, which allows them to be identified using mass spectrometry. Preliminary experiments were performed to detect the oxidative status and interaction partners of the fusion proteins in cells, but these were unsuccessful due to technical problems. Optimization of these experiments will ultimately result in assessment of the oxidative status and interaction partners *in vivo*, which will provide valuable new information regarding the role of oxidation in the function of SHP2 and its variants.

Conclusion

In conclusion, there was a distinct difference in the behavior of wildtype and D61G-SHP2 after oxidative stimulus. The D61G mutant was more readily oxidized *in vitro* compared to the wildtype SHP2, which led to its inactivation. This effect probably stemmed from the open conformation of the D61G mutant, which granted easier access to H_2O_2. In the future, it would be very interesting to determine the behavior of D61G-catalase fusion proteins *in vitro*, to see if catalase also protects this mutant against ROS and how this affects its activity. In addition, it would be very interesting to study the fusion proteins *in vivo* in both cells and zebrafish embryos. By determining the oxidative status and interaction partners of the fusion proteins *in vivo*, a better understanding of the developmental defects may be obtained. Lastly, it would be very interesting to expand the catalase-fusion protein methodology to other redox sensitive proteins to be able to study these in more detail.

Materials and Methods

Oxidative treatment

Agarose beads were dissolved in TBS, after which 40 µl was taken to perform experiments with. The beads were washed twice with PBS0, after which the recombinant SHP2 protein or mutant was added to bind to the beads. During bead binding, dithiothreitol (DTT) was added to a concentration of 1 mM to remove prior oxidation from the protein. An additional 100 µl of PBS0 was added and incubated for 30 minutes on a spinner at 4 °C. After bead binding the oxidative treatment was performed. Either MilliQ or H_2O_2 was added to the desired concentration and incubated for the desired time. Afterwards, the beads were washed twice with PBS0.

After oxidative treatment, non-reacted cysteines were alkylated by adding N-ethylmaleimide (NEM) to a final concentration of 4 mM and incubating for 2 hours at 37 °C. The beads were washed twice with PBS0, after which the singly oxidized cysteine were reduced by adding 10mM DTT in PBS0 with 1% sodium deoxycholate. The beads were incubated for 30 minutes at room temperature. Subsequently, the beads were washed twice with PBS0. Finally, the now reduced cysteines were alkylated by adding 40mM iodoacetic acid (IAA). Beads were incubated at room temperature for 30 minutes. The reaction was quenched by adding 1 mM DTT, after which the beads were stored in the -80 freezer.

Digestion and desalting

The beads were centrifuged at 20,000 rcf for 5 minutes at 4 °C to pellet the beads. The supernatant was removed and 240 µl of digestion buffer was added (1.5M Urea in 50 mM ammonium bicarbonate (AMBIC)). A predigestion step was performed by adding trypsin in a 1:100 ratio (protein:protease). The samples were incubated for four hours at 37 °C. A second digestion step with trypsin (Sigma) (1:100) was performed overnight at 37 °C. After digestion, the samples were diluted with 200 µl of 50 mM AMBIC and acidified by adding 12 µl of formic acid. The samples were centrifuged at 20,000 rcf for 10 minutes and 4°C, after which they were desalted using SEPPAK SPE cartridges (Waters). Briefly, the cartridges were washed three times with 1 ml acetonitrile, followed by washing three times with 1 ml of 0.1 M acetic acid. Afterwards, samples were loaded and the flow through was passed through the cartridge again. The cartridges were then washed three times with 0.1 M acetic acid, after which the peptides were eluted by adding three times 250 µl of 0.1 M acetic acid / 80% acetonitrile. The samples were subsequently dried using a Thermo Savant SPD SpeedVac (Thermofisher scientific).

Mass spectrometric analysis

For the mass spectrometric analysis, peptides were dissolved in 2% formic acid and a volume corresponding to 2 µg was injected on a UHPLC 1290 system (Agilent) coupled to a Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). The peptides were trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) for 5 minutes in buffer A (0.1% formic acid) at a flow rate of 0.005 ml/min. Afterwards they were separated using an analytical column (Agilent Poroshell EC-C18, 2.7 µm, 50 cm x 75 µm). The following gradient was used: 13 – 44% buffer B (80% acetonitrile + 0.1% formic acid) in 65 minutes, 100% buffer B for 2 minutes followed by 100% buffer A for 11 minutes. A split flow was used to generate a flow rate of 300 nl/min. The Q Exactive HF-X was operated in a data dependent acquisition mode with positive ionization. The full MS spectra were acquired from 375 to 1600 m/z at 60000 resolution, using an automatic gain control (AGC) target value of 3 x 10^6 charges and a maximum injection time of 20 ms. A maximum of 15 precursors were allowed to be fragmented. The dynamic exclusion was set to 12 seconds. MS/MS fragmentation spectra were obtained with a fixed first mass of 120 m/z and a resolution of 30000. An AGC target of 1 x 10^5 was chosen and a maximum injection time of 50 ms. The fragmentation was performed using HCD at a NCE of 27.

Quantification of oxidative modifications

The relative presence of the different oxidative modification was determined by creating extracted ion chromatograms (XIC) of the masses corresponding to the modified peptide (QEGITGAGPIVVH**C**SAGIGR). The following masses were chosen: NEM alkylated (572.81, 4+), IAA alkylated (741.06, 3+), sulphinic acid (732.39, 3+) and sulphonic acid (737.72, 3+). The extracted XICs were integrated in the xCalibur Qual browser (Thermo Fisher scientific, version 4.0.27.21) using the Genesis algorithm. The areas of these XICs were used to compare the intensity of the different modifications.

Phosphatase assay

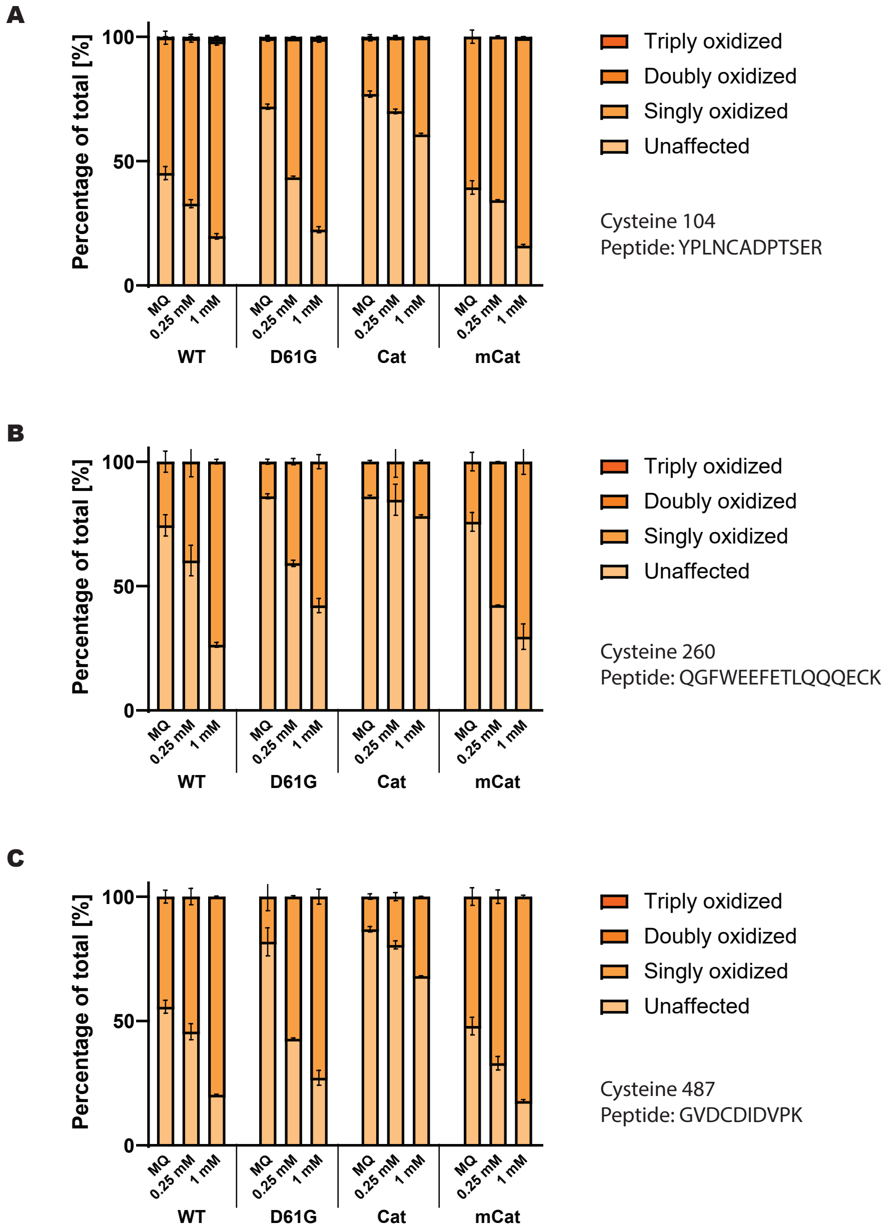
The phosphatase proteins were dissolved in 33.3 mM 2-morpholinoethanesulfonic acid (MES) (pH 6). A total of 8 µl of either water or H_2O_2 (5x concentrated) was added to 32 µl of the dissolved phosphatase to get the correct concentration of oxidative agent in the sample. The samples were incubated for 2 hours at room temperature. A total of 40 µl of pNPP mix was added (40 mM MES (pH 6), 2 mM DTT, 2 mM EDTA, 300 mM NaCl, 20 mM pNPP) and the samples were incubated for 30 minutes at 30 °C. After incubation, 50 µl of 0.5 M NaOH was added to stop the reaction. The activity was determined by measuring the absorption at OD 405 using a plate reader.

References

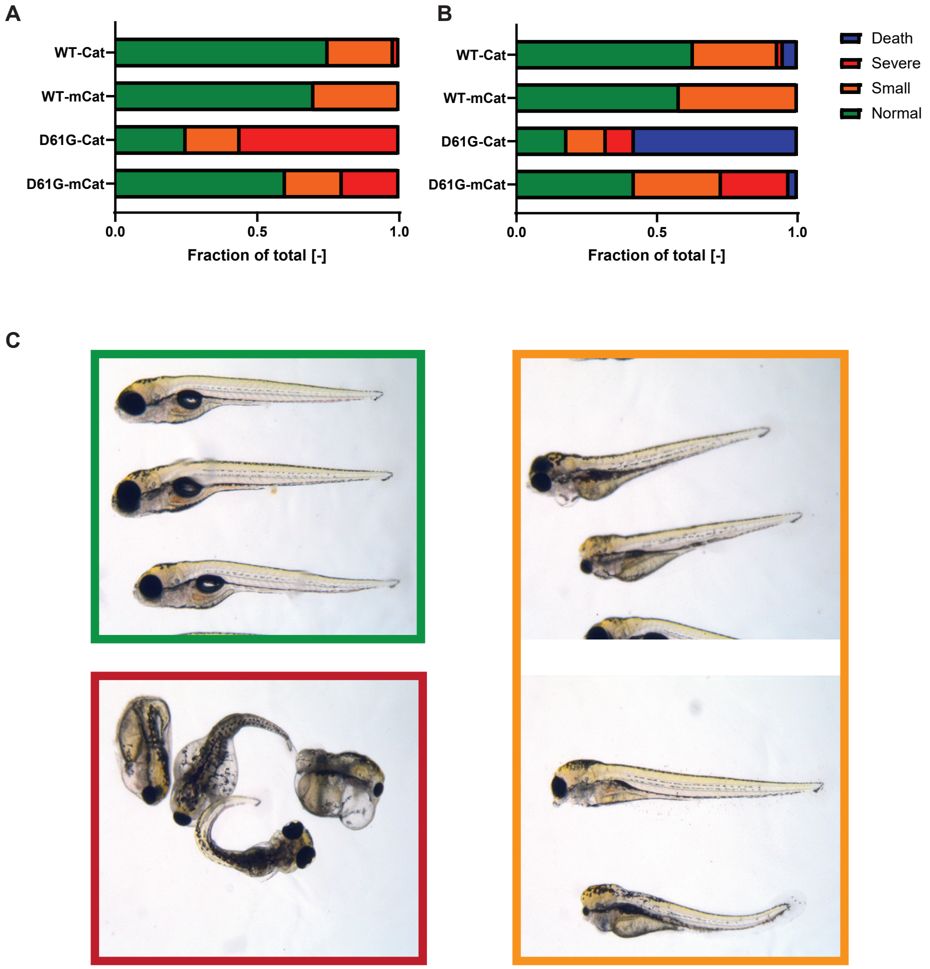
1. Karisch, R. & Neel, B. G. Methods to monitor classical protein-tyrosine phosphatase oxidation. *FEBS J.* **280**, 459–475 (2013).
2. Motiwala, T. & Jacob, S. T. Role of Protein Tyrosine Phosphatases in Cancer. *Progress in Nucleic Acid Research and Molecular Biology* **81**, 297–329 (2006).
3. Östman, A. & Böhmer, F. D. Regulation of receptor tyrosine kinase signaling by protein tyrosine phosphatases. *Trends in Cell Biology* **11**, 258–266 (2001).
4. Den Hertog, J., Groen, A. & Van Der Wijk, T. Redox regulation of protein-tyrosine phosphatases. *Arch. Biochem. Biophys.* **434**, 11–15 (2005).
5. Tonks, N. K. Protein tyrosine phosphatases: from genes, to function, to disease. *Nat. Rev. Mol. Cell Biol.* **7**, 833–846 (2006).
6. Meng, T. C., Fukada, T. & Tonks, N. K. Reversible oxidation and inactivation of protein tyrosine phosphatases in vivo. *Mol. Cell* **9**, 387–399 (2002).
7. Tsutsumi, R. *et al.* Assay to visualize specific protein oxidation reveals spatio-temporal regulation of SHP2. *Nat. Commun.* **8**, 1–14 (2017).
8. Miki, H. & Funato, Y. Regulation of intracellular signalling through cysteine oxidation by reactive oxygen species. *J. Biochem.* **151**, 255–261 (2012).
9. Weibrecht, I. *et al.* Oxidation sensitivity of the catalytic cysteine of the protein-tyrosine phosphatases SHP-1 and SHP-2. *Free Radic. Biol. Med.* **43**, 100–110 (2007).
10. Qu, C. K. The SHP-2 tyrosine phosphatase: Signaling mechanisms and biological functions. *Cell Res.* **10**, 279–288 (2000).
11. Dance, M., Montagner, A., Salles, J. P., Yart, A. & Raynal, P. The molecular functions of Shp2 in the Ras/Mitogen-activated protein kinase (ERK1/2) pathway. *Cell. Signal.* **20**, 453–459 (2008).
12. Hale, A. J. & den Hertog, J. Shp2–Mitogen-Activated Protein Kinase Signaling Drives Proliferation during Zebrafish Embryo Caudal Fin Fold Regeneration. *Mol. Cell. Biol.* **38**, (2017).
13. Tajan, M., de Rocca Serra, A., Valet, P., Edouard, T. & Yart, A. SHP2 sails from physiology to pathology. *Eur. J. Med. Genet.* **58**, 509–525 (2015).
14. Keilhack, H., David, F. S., McGregor, M., Cantley, L. C. & Neel, B. G. Diverse biochemical properties of Shp2 mutants: Implications for disease phenotypes. *J. Biol. Chem.* **280**, 30984–30993 (2005).
15. Neel, B. G., Gu, H. & Pao, L. The 'Shp'ing news: SH2 domain-containing tyrosine phosphatases in cell signaling. *Trends Biochem. Sci.* **28**, 284–293 (2003).
16. Hof, P., Pluskey, S., Dhe-Paganon, S., Eck, M. J. & Shoelson, S. E. Crystal structure of the tyrosine phosphatase SHP-2. *Cell* **92**, 441–450 (1998).
17. Jang, J. Y. *et al.* Reactive oxygen species play a critical role in collagen-induced platelet activation via shp-2 oxidation. *Antioxidants Redox Signal.* **20**, 2528–2540 (2014).
18. Qiu, W. *et al.* Structural insights into Noonan/LEOPARD syndrome-related mutants of protein-tyrosine phosphatase SHP2 (PTPN11). *BMC Struct. Biol.* **14**, 1–11 (2014).

19. Li, H. L. *et al.* Exploring the effect of D61G mutation on SHP2 cause gain of function activity by a molecular dynamics study. *J. Biomol. Struct. Dyn.* **36**, 3856–3868 (2018).
20. Persson, C. *et al.* Preferential oxidation of the second phosphatase domain of receptor-like PTP-α revealed by an antibody against oxidized protein tyrosine phosphatases. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 1886–1891 (2004).
21. Groen, A. *et al.* Differential oxidation of protein-tyrosine phosphatases. *J. Biol. Chem.* **280**, 10298–10304 (2005).
22. Karisch, R. *et al.* Global proteomic assessment of the classical protein-tyrosine phosphatome and 'redoxome'. *Cell* **146**, 826–840 (2011).
23. Smyth, D. G., Blumenfeld, O. O. & Konigsberg, W. Reactions of N-ethylmaleimide with peptides and amino acids. *Biochem. J.* **91**, 589–595 (1964).
24. Chen, C. Y., Willard, D. & Rudolph, J. Redox regulation of SH2-domain-containing protein tyrosine phosphatases by two backdoor cysteines. *Biochemistry* **48**, 1399–1409 (2009).
25. Glorieux, C. & Calderon, P. B. Catalase, a remarkable enzyme: Targeting the oldest antioxidant enzyme to find a new cancer treatment approach. *Biol. Chem.* **398**, 1095–1108 (2017).
26. Sundaresan, M., Yu, Z. X., Ferrans, V. J., Irani, K. & Finkel, T. Requirement for generation of H2O2 for platelet-derived growth factor signal transduction. *Science (80-.).* **270**, 296–299 (1995).
27. Bae, Y. S. *et al.* Epidermal Growth Factor (EGF)-induced Generation of Hydrogen Peroxide. *J. Biol. Chem.* **272**, 217–221 (1997).
28. Östman, A., Hellberg, C. & Böhmer, F. D. Protein-tyrosine phosphatases and cancer. *Nat. Rev. Cancer* **6**, 307–320 (2006).
29. Jopling, C., Van Geemen, D. & Den Hertog, J. Shp2 knockdown and noonan/LEOPARD mutant Shp2-induced gastrulation defects. *PLoS Genet.* **3**, 2468–2476 (2007).
30. Kontaridis, M. I., Swanson, K. D., David, F. S., Barford, D. & Neel, B. G. PTPN11 (Shp2) mutations in LEOPARD syndrome have dominant negative, not activating, effects. *J. Biol. Chem.* **281**, 6785–6792 (2006).
31. Ogata, T. *et al.* Two novel and one recurrent PTPN11 mutations in LEOPARD syndrome. *Am. J. Med. Genet.* **130 A**, 432–434 (2004).
32. Andre, C., Kim, S. W., Yu, X. H. & Shanklin, J. Fusing catalase to an alkane-producing enzyme maintains enzymatic activity by converting the inhibitory byproduct H2O2 to the cosubstrate O2. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3191–3196 (2013).
33. Yano, S., Arroyo, N. & Yano, N. SHP2 binds catalase and acquires a hydrogen peroxide-resistant phosphatase activity via integrin-signaling. *FEBS Lett.* **577**, 327–332 (2004).
34. Mahadev, K., Zilbering, A., Zhu, L. & Goldstein, B. J. Insulin-stimulated Hydrogen Peroxide Reversibly Inhibits Protein-tyrosine Phosphatase 1B in Vivo and Enhances the Early Insulin Action Cascade. *J. Biol. Chem.* **276**, 21938–21942 (2001).
35. Yoo, S. K., Starnes, T. W., Deng, Q. & Huttenlocher, A. Lyn is a redox sensor that mediates leukocyte wound attraction in vivo. *Nature* **480**, 109–112 (2011).

Supplementary figures



Supplementary figure S1: Oxidative behavior of the other cysteines of SHP2. The oxidative behavior of the other cysteines of SHP2, Cys 104 (A), Cys260 (B) and Cys487 (C). The intensities of the different modifications were normalized.



Supplementary figure S2: Fusion of catalase increases the effect of D61G- SHP2, but not wild type SHP2. Zebrafish embryos were microinjected at the one-cell stage with synthetic mRNA encoding fusion proteins of wildtype Shp2a and catalase (WT-Cat), wildtype Shp2a and mutant Catalase (WT-mCat), mutant Shp2a-D61G and wildtype Catalase (D61G-Cat) or mutant Shp2a-D61G and mutant Catalase (D61G-mCat). Developmental defects were assessed at 2 dpf (A) and 5 dpf (B) and classified by eye as normal (no change), small (missing swimming bladder and lower jaw), severe (truncated and heart edema) or dead. Examples of the classification are depicted in (C).

Chapter 5

Synopsis

Future perspectives and outlook

Shotgun proteomics, as used in this thesis, is very well suited for generating novel hypotheses. By investigating the whole proteome under specific conditions, many interesting leads can be found. However, all these interesting leads need to be investigated further and validated to find their true biological meaning. Thorough validation of the data is necessary due to the inherent characteristics of a shotgun proteomics experiment. Since the proteins in a sample are first digested into peptides before they are measured on the mass spectrometer, there is indirect evidence to which proteins were originally present in the sample due to shared peptides. This is amplified by the stochastic nature of mass spectrometry, where replicate experiments can lead to slightly different results. Thorough bio-informatics pipelines (such as Proteome discoverer and MaxQuant) are used to generate a list of proteins which have a high confidence of being in your sample, but orthogonal validation methods always need to be used to verify this before biological conclusions can be made. Another inherent downside to studying the proteome by shotgun proteomics is the extreme complexity of it. The presence of thousands of different proteins, which can have multiple post-translational modifications on different sites, creates an extremely complex system to investigate. Since we now have the possibility to measure all these proteins and their modifications at the same time due to sensitive mass spectrometers, it can be even more difficult to find interesting leads which can be followed-up on in this big quantity of data. In addition to this, only a small part of the identified proteins have been studied in more detail to elucidate their exact mechanism and function in a biological setting, meaning that many other proteins found in the increasingly larger proteomics datasets have not been fully annotated yet. Many times, interesting leads present themselves after mass spectrometric analysis, of which no further information can be found in literature, greatly limiting the potential of these leads. Therefore, even though the datasets continue to increase in size due to better and more sensitive mass spectrometers, the biological relevance of these datasets is lacking behind. This has greatly limited the field of proteomics, where sometimes more questions arise from an experiment than answers are given.

Therefore, a shift in paradigm in the field of proteomics might be necessary, from trying to generate increasingly larger datasets to generating more functional data that holds more biological promise. One way to generate more functional data on interesting proteins is done by using targeted techniques such as selected reaction monitoring (SRM)[1]. In SRM, certain peptides corresponding to the protein of interest are chosen and selectively measured, allowing for extremely accurate protein identification and quantification. By using SRM, low abundant biomarkers can be accurately quantified, allowing for better diagnosis of disease. Additionally, a panel of proteins can be reproducibly identified and quantified across different samples, allowing for a comprehensive view of the behavior of these proteins under certain conditions. As can be seen, SRM is one way to investigate a subset of the whole proteome, allowing for a more functional and comprehensive look at the role of proteins under specific conditions. However, these interesting proteins need to already be known in literature or need to be found using shotgun proteomics approaches, limiting the use of SRM techniques.

On the other hand, by spending more time during sample preparation, more biologically relevant samples can be introduced to the mass spectrometer. This way, more functional data can be generated, which might then lead to more biologically relevant conclusions. I think this is where the field of shotgun proteomics should evolve towards, where the more time spend on sample preparation can save a lot of time in the analysis of the data and can lead to more valuable results. In this thesis, techniques were used which transcended past the classical expression proteomics approach to study the proteome on a deeper and more functional level. This way, we generated more functional and valuable data, which led to answers to fundamental biological questions.

Boiling the proteome to find novel drug interactors

Thermal proteome profiling (TPP) is one method where data generated from the mass spectrometer can be interpreted more directly and functionally. In TPP, lysates or cells, which were treated with either drug or control, are heated up to multiple temperature points to generate melting curves for all the proteins in the proteome. If the melting point of a protein shifts between the treated and control samples, it can be assumed that the drug interacts with this protein. By discarding the non-shifting proteins and by only looking at the proteins that show a (de)stabilizing shift, more focused follow-up and validation can be achieved. A big advantage of TPP is the ability to measure drug-protein interactions *in vivo*, where the crowded cellular environment and the possibility of protein-protein interactions can have a distinct effect on target engagement. Additionally, TPP makes it possible to measure the effect of a drug on the whole proteome, which allows for the identification of toxic off-target effects. Taken together, TPP is a valuable new technique in the drug development pipeline.

Since the first implementation by Savitski *et al*[2] in 2014, the field of TPP has matured substantially. For example, the scope has now extended from cultured cells to bacteria[3], parasites[4] and even tissue[5]. Last year the meltome atlas[6], which consists of the thermal behavior of 48.000 proteins across 13 species, was developed, greatly increasing the knowledge on the thermal behavior of proteins. We added to this important work, as described in chapter 2 of this thesis, where we described the use of thermal proteome profiling to determine drug behavior in whole zebrafish embryo lysates. In this study, we screened for interactions of the drug napabucasin with all proteins that can occur within an organism, including tissue specific proteins that would normally be lost when only investigating one cell type. This methodology might have great implications in the field of drug development in the future. When performing this methodology on intact zebrafish embryos, aspects of pharmacology such as "absorption, distribution, metabolism and excretion" (ADME) and passage of the blood-brain barrier can be investigated. The proof-of-principle experiments that we showed in chapter 2 of this thesis can be the first step towards this assay.

All the recent publications in the field of TPP show the great potential which this technique can have in the field of drug development. For example, by showing the mechanism how anti-microbial drugs combat certain pathogens or by finding the exact proteins that can interact with a novel drug to identify off-target effects, TPP gives many new valuable insights that can streamline drug development. However, there are still some drawbacks, which can withhold the routine use of TPP in the drug development pipeline. For example, some classes of proteins need very high temperatures before they precipitate. Additionally, some proteins do not show any ligand induced stabilization effects, making them unusable as targets for TPP. Lastly, it might be very difficult to investigate low abundant proteins; however, the introduction of novel more sensitive mass spectrometers can alleviate this. Nevertheless, these drawbacks will not hold back the great potential that TPP has in the field of drug discovery, where it can be very valuable in finding the (un)expected off-target effects of the drug under investigation. Methods like TPP are a great example of techniques that can be used to distill more relevant information from mass spectrometric experiments.

The big steps that the field of TPP has taken in the latest years make it a valuable tool in the drug discovery pipeline. However, the scope of TPP is already transcending past drug development. For example, TPP was also used to determine the effect of genetic perturbations on the thermal stability of proteins in *E. coli*[7]. Additionally, it was used to determine the thermal stability of host proteins during SARS-CoV-2 infection, which might then be used for pharmacological inhibition[8]. These studies show that TPP is adaptable to many different types of experiments, making it very interesting to see where the field will evolve towards in the future. In my opinion, TPP will evolve to be a technique that will be routinely used in proteomics labs around the world, to achieve more functional data on diverse research subjects.

Learning more of signaling in the cell through phosphorylation events

A different approach that can be taken to investigate the proteome in a more functional way is by looking at post-translational modifications (PTMs). PTMs are small modifications that are placed on the protein, which can have a big influence on its function[9]. For example, it can change the activity of the protein, its localization and even its degradation. Therefore, if a protein is modified, it can be assumed that it is involved in signaling processes inside of the cell. By looking at PTMs, the dynamic behavior of the cell can be more thoroughly investigated. In chapter 3 of this thesis, we focused on the labile modification phosphohistidine in mammalian systems. We showed that phosphohistidine is barely present in human cells, where it also does not seem to play a role in signaling but instead acts as a reactive intermediate. In contrast, it was shown by Potel *et al*[10] that phosphohistidine plays a large role in bacteria such as *E. coli*. The presence of phosphohistidine in bacterial cells, but not in mammalian, might make it a prospective target of novel antibacterial drugs. By interfering with the transmission of phosphohistidine signaling in bacteria, for example by targeting the two-component system[11], novel new drugs might be developed which selectively target bacteria and not human cells, limiting toxic side effects. Additionally, it is has been proposed that phosphohistidine can play a role in simple eukaryotes such as yeast (*Saccharomyces cerevisiae*) and slime mold (*Dictyostelium discoideum*)[12]. Therefore, it might be interesting to investigate the scope of phosphohistidine in these systems using mass spectrometry, to investigate how the role of phosphohistidine might have evolved in more complex organisms than bacteria and to see when it became obsolete. Since the recently developed mass spectrometry tool[10] allows the investigation of the labile modification phosphohistidine, which in the past could not be routinely measured, it might also be possible to measure other more labile phosphorylation events. Other amino acids, such as arginine, lysine, cysteine, aspartate and glutamine, can also be phosphorylated[13]. By using mass spectrometry, the role and scope of these non-traditional phosphorylation events might be investigated in both pro- and eukaryotic sources. However, an Achilles heel is the fact that it will be more difficult to accurately localize them. The extensive phosphorylation of human proteins by pSer and pThr ensures that the peptides generated after digestion will almost always have multiple phosphorylation sites, which increases the chance of mislocalization, as was seen for pHis. Therefore, it will be very difficult to validate the presence of phosphorylation events on these non-traditional residues and to see how their phosphorylation might affect cellular function. Novel approaches, such as immonium triggering assays[14], might alleviate this, but these need to be developed first.

The future of investigating novel phosphoresidues is very exciting, where, if true, it can lead to valuable new insights in cellular signaling. The methodology to maintain the labile modifications and measure them on the mass spectrometer is there, but the validation of the found sites remains the biggest bottleneck. Therefore, to be certain that the novel findings are true, and not artifacts of the method like phosphohistidine in mammalian cells, robust validation strategies need to be developed. For example, by using SRM to selectively detect possible phosphorylation events on peptides[15], sites can be accurately validated and the scope of novel phosphoresidues can be investigated. Additionally, by developing software tools that can more accurately and confidently localize the phosphate group on peptides, the large abundance of false positives can be diminished and the true scope can be found. Therefore, by focusing more on validation strategies in the future, light can be shed on the true influence of these novel phosphoresidues.

Oxidation as regulatory mechanism of protein function

PTMs can be regulated by enzymatic reactions; however, some of them can be transiently generated while still having a large effect on protein function. One of these PTMs is oxidation, which occurs transiently due to the presence of reactive oxygen species (ROS) in cells. An example of proteins which can be easily oxidized are the tyrosine phosphatases, of which the catalytic cysteine is prone to oxidation due to its micro-environment[16]. The tyrosine phosphatases lose their function when their catalytic cysteine is oxidized, making ROS a direct mechanism how their activity can be tuned. Therefore, by measuring the oxidative status of subclasses of proteins which are sensitive to it, direct biological information regarding their function can be found and the influence which this might have on the cell. However, extra steps need to be taken during sample handling to make sure that the true oxidative status of the protein is determined and not an artifact. For example, the harsh oxidants released during cellular lysis can artificially oxidize proteins, but this can easily be avoided by directly alkylating the cysteines prone to oxidation during the lysis. Also, it can be difficult to measure peptides containing cysteine in the thiolate or sulphenic acid state, since these are known to be labile and be removed during mass spectrometric analysis[17]. This can be accounted for by performing differential alkylation approaches, as performed on the tyrosine phosphatase SHP2 in chapter 4. As can be seen, the study of oxidative modified proteins can be difficult, but it can give valuable information on how proteins function under oxidative stress.

Novel tools are still being developed to gain more biological information on oxidized proteins. The fusion proteins introduced in chapter 4 of this thesis can be a valuable new tool in the study of oxidation sensitive proteins. Through the addition of catalase to the protein, the direct influence of ROS can be determined. It is a useful approach for *in vitro* studies, where it is possible to directly measure the extent to which proteins are oxidized under pre-determined oxidative stresses. However, I think that this methodology has the biggest potential to be used *in vivo*. By transfecting the constructs in cells or zebrafish embryo's, the endogenous levels of oxidation can be determined, which gives the most direct read-out of protein function under physiological conditions. At the same time, the effect of oxidation on cellular function or zebrafish phenotype can be determined. I think this approach of measuring the oxidative status of proteins by mass spectrometry and correlating this to phenotypic changes will be a very valuable tool in the study of redox proteomics.

Closing remarks

As described above, a more focused approach regarding the proteome can lead to more distinct and interpretable results. By only looking at a subclass of proteins, which either show a thermal shift, are oxidized or have a low abundant modification, more direct biological information can be obtained. I think this will be the future of proteomics, where more effort should be put into sample preparation instead of just trying to measure increasingly more complex data sets. By shying away from the generation of these increasingly more complex datasets, maybe more questions can be answered, instead of just generating new ones.

References

1. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nature Methods* **9**, 555–566 (2012).
2. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science (80-.).* **346**, 1255784 (2014).
3. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state *in vivo*. *Mol. Syst. Biol.* **14**, e8242 (2018).
4. Herneisen, A. L. *et al.* Identifying the Target of an Antiparasitic Compound in Toxoplasma Using Thermal Proteome Profiling. *ACS Chem. Biol.* **15**, 1801–1807 (2020).
5. Perrin, J. *et al.* Identifying drug targets in tissues and whole blood with thermal-shift profiling. *Nature Biotechnology* **38**, 303–308 (2020).
6. Jarzab, A. *et al.* Meltome atlas—thermal proteome stability across the tree of life. *Nat. Methods* **17**, 495–503 (2020).
7. Mateus, A. *et al.* The functional proteome landscape of Escherichia coli. *Nature* **588**, 473–478 (2020).
8. Selkrig, J. *et al.* SARS-CoV-2 infection remodels the host protein thermal stability landscape. *Mol. Syst. Biol.* **17**, e10188 (2021).
9. Jensen, O. N. Modification-specific proteomics: Characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology* **8**, 33–41 (2004).
10. Potel, C. M., Lin, M. H., Heck, A. J. R. & Lemeer, S. Widespread bacterial protein histidine phosphorylation revealed by mass spectrometrybased proteomics. *Nat. Methods* **15**, 187–190 (2018).
11. Adam, K. & Hunter, T. Histidine kinases and the missing phosphoproteome from prokaryotes to eukaryotes. *Lab. Investig.* **98**, 233–247 (2018).
12. Klumpp, S. & Krieglstein, J. Phosphorylation and dephosphorylation of histidine residues in proteins. *Eur. J. Biochem.* **269**, 1067–1071 (2002).
13. Hunter, T. Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 2513–2516 (2012).
14. Potel, C. M. *et al.* Gaining Confidence in the Elusive Histidine Phosphoproteome. *Anal. Chem.* **91**, 5542–5547 (2019).
15. Unwin, R. D. *et al.* Multiple reaction monitoring to identify sites of protein phosphorylation with high sensitivity. *Mol. Cell. Proteomics* **4**, 1134–1144 (2005).
16. Karisch, R. & Neel, B. G. Methods to monitor classical protein-tyrosine phosphatase oxidation. *FEBS J.* **280**, 459–475 (2013).
17. Karisch, R. *et al.* Global proteomic assessment of the classical protein-tyrosine phosphatome and 'redoxome'. *Cell* **146**, 826–840 (2011).

Lay summary of this thesis

All cells in the human body, ranging from liver to muscle and even immune cells, share the same piece of DNA. Even though they all share the same DNA, they act in totally different ways to perform their intended function. This behavior is mainly a result of proteins, the molecules in the cells that are involved in almost all biological processes. By reading out specific proteins from the DNA, cells can change their protein profile to allow them to perform specialized functions. This protein profile is very dynamic and can give a whole range of important biological information. For example, cancer cells show a different protein profile compared to healthy cells, meaning that by finding the differences between the two you might determine and understand the cause of the disease. Additionally, cells might change their protein profile after treatment with a drug, which can then be investigated to see which processes are affected by the drug and how the cell tries to survive this treatment. As can be seen, by studying the proteins inside of a cell a treasure trove of information is released, which can lead to great new discoveries. The field which studies all the proteins inside of the cell (also called the **proteome**) is called **proteomics** and will be the focus of this thesis.

To be able to measure the proteome, multiple steps need to be taken. First, the cells need to be broken up to release the proteins inside. For every cell or organism, optimized methods need to be used. After the proteins are released, they are cut into smaller fragments by enzymes called proteases. These smaller fragments of the protein, called peptides, are then measured on a **mass spectrometer** (MS). The MS measures the total mass of peptides but also the mass of small fragments that occur after fragmentation of these peptides. Afterwards, dedicated software can analyze these masses and map them back to find which protein the peptides originally belonged to. Through these steps, proteomics can identify all the proteins present in a sample under specific conditions. Besides measuring the presence of proteins in samples, MS can also be used to measure exactly how much of the protein is present at a certain time. This is called **quantitative proteomics** and can be a very valuable tool in the investigation of diseases. For example, if a protein is ten times more present in a cancer cell compared to a healthy cell, it can be expected that it plays an important role in the disease.

Even though a lot of processes in cells are regulated by the abundance of proteins, there is another mechanism how cells can react to their environment. There are small chemical modifications, called **post-translational modifications (PTMs)**, which when attached to the protein can act as an on/off switch to totally change its behavior. For example, PTMs can change the activity of the protein, its location inside of the cell or even cause its degradation. When the processes involving the addition or removal of these PTMs are dysregulated, it can lead to disease. One of the chemical groups which can be added to proteins is the phosphate group. This phosphate is added to proteins by an enzyme called **kinase**, while it can be removed again by an enzyme called **phosphatase**. When these two are in balance, it will lead to normal cellular function. However, when these processes are dysregulated, it can lead to severe disease states such as cancer. The study of finding the presence of the phosphate modification on proteins using MS is called **phosphoproteomics**. Phosphoproteomics is one of the research subjects which are studied in this thesis.

In this thesis, we used proteomics approaches to answer fundamental questions in different sample types. We investigated the mechanism of drugs in zebrafish, determined the importance of a labile PTM in human cells and lastly researched the ease of oxidation (and the consequence thereof) of one protein.

In **Chapter 1** of this thesis the basic principles of mass spectrometry and proteomics are explained. First, a short overview of all different research strategies which can be performed using mass spectrometry are explained. Next, a detailed look is given to sample preparation in the field of shotgun proteomics, where peptides are measured to determine which proteins were present in the initial sample. Focus is given to the choice of sample type, how to lyse these, digesting the proteins in these samples and how to possibly make the sample less complex to allow for better results. After this has been established, all details regarding the mass spectrometer are discussed. Ionization methods, mass analyzers and fragmentation methods are described. Also, multiple database search algorithms are discussed. Lastly, specialized experiments which can be performed using mass spectrometry are explained. Different ways of quantifying the proteome are discussed, followed by a detailed explanation of the fields of phospho- and redox proteomics.

In **Chapter 2** of this thesis we investigated the mechanisms of the drug napabucasin using thermal proteome profiling on zebrafish embryos. In thermal proteome profiling, the proteins inside of the cell are heated to multiple temperature points. At a certain temperature, the protein will solidify and precipitate out of solution. This is similar to boiling an egg, where at a certain temperature the egg white will harden. However, when a drug binds to a protein the temperature at which the protein solidifies will change. This characteristic is employed in thermal proteome profiling, where we added the drug napabucasin to zebrafish embryos, boiled them and determined at which temperature the proteins solidify using mass spectrometry. By comparing this temperature to the temperature when no drug is added, we could see which proteins interact with the drug. Using thermal proteome profiling, we showed that aldehyde dehydrogenases are off-targets of napabucasin, which has major influences on the development of zebrafish embryos.

In **Chapter 3** of this thesis, we looked at the elusive PTM phosphohistidine in mammalian cells. Phosphohistidine is a modification that plays a large role in bacteria, but is very difficult to measure due to its sensitivity. The methods that we normally use in the lab are performed under acidic conditions, which cause phosphohistidine to be removed from the protein. However, recently methods were developed which can maintain this modification, which were used to shed light on its importance in bacteria. We were curious how important phosphohistidine is in human cells. To investigate this, multiple experiments were performed. For example, we investigated multiple cell types, but also developed a kinase assay and fractionated cellular lysate to make it less complex. In the end, many novel phosphohistidine sites were found. However, we also detected many of the same sites when the samples were treated with acid, which normally would cause the phosphohistidine to be removed. Therefore, most of the new sites that we identified are not true, leaving only a handful of interesting sites. So even though phosphohistidine plays a large role in bacteria, it does not seem to be as important in human cells.

In **Chapter 4** of this thesis, we investigated the effects of oxidation on the tyrosine phosphatase SHP2. SHP2 can remove a phosphate group from the amino acid tyrosine and through this mechanism plays an important role in signaling in the cell. However, when SHP2 is oxidized by hydrogen peroxide, it will lose its function. In this chapter, we investigated how easily hydrogen peroxide can oxidize SHP2. Additionally, we also did this for the Noonan mutant of SHP2. This mutation causes SHP2 to be in a more open state, which might cause hydrogen peroxide to oxidize the protein more easily. We showed that this is true using mass spectrometry and phosphatase assays. These findings might have important implications for syndromes associated with SHP2.

Finally, in **Chapter 5** of this thesis I share my personal view on proteomics. Additionally, a list of all my publications and my acknowledgements are shared here.

Lekensamenvatting van dit proefschrift

Alle cellen in het menselijke lichaam, van lever tot spier en zelfs immuun cellen, delen hetzelfde DNA. Ondanks dat zij hetzelfde DNA delen, hebben zij toch allemaal hun eigen gespecialiseerde functies. Dit komt vooral door eiwitten, de moleculen in de cellen die bij bijna alle biologische processen betrokken zijn. Door specifieke eiwitten uit te lezen uit het DNA kunnen cellen hun eiwit profiel veranderen en zo hun gespecialiseerde functies uitvoeren. Dit eiwit profiel is heel dynamisch en kan veel belangrijke biologische informatie geven. Kankercellen hebben bijvoorbeeld een specifiek ander eiwit profiel dan gezonde cellen, wat betekend dat wanneer je de verschillen tussen de twee profielen vindt dat je mogelijk de oorzaak van de ziekte kunt bepalen en begrijpen. Daarnaast kunnen cellen hun eiwitprofiel veranderen als ze worden behandeld met een medicijn. Door de eiwit profielen van de cel na behandeling te onderzoeken, kun je de processen vinden die veranderen door het medicijn. Zoals je kunt zien, door de eiwitten in een cel te onderzoeken kan een schatkamer vol met informatie aangeboord worden die tot grote nieuwe ontdekkingen kunnen leiden. Het onderzoeksgebied dat alle eiwitten in de cel (ook wel het **proteome** genoemd) onderzoekt heet **proteomics** en is de focus van dit proefschrift.

Om het proteome te kunnen meten, moeten meerdere stappen genomen worden. Ten eerste moeten de cellen opgebroken worden om de eiwitten vrij te maken. Voor elk type cel of organisme zijn geoptimaliseerde methodes nodig. Nadat de eiwitten zijn vrijgemaakt worden zij opgeknipt in kleinere stukken door enzymen genaamd proteases. Deze kleinere stukjes eiwit, ook wel peptiden genoemd, worden dan gemeten op een **massa spectrometer** (MS). De MS meet de totale massa van de peptiden maar ook de massa van kleinere fragmenten van de peptiden die ontstaan na fragmentatie hiervan. Deze massa's worden dan geanalyseerd door speciale software om zo het eiwit te vinden waar de peptiden origineel onderdeel van uit maakten. Door het volgen van deze stappen kan proteomics alle eiwitten onderzoeken die zich onder een bepaalde conditie in een cel bevinden. Naast het meten van de aanwezigheid van eiwitten in monsters kan MS ook gebruikt worden om de exacte hoeveelheid van eiwitten in een cel te bepalen. Dit noemen ze **kwantitatieve proteomics** en is een waardevol hulpmiddel in het onderzoek naar ziekten. Zo kan verwacht worden dat wanneer een bepaald eiwit tien keer meer aanwezig is in een kankercel als een gezonde cel, dat deze een belangrijke rol speelt in dit ziektebeeld.

Naast de hoeveelheid van eiwitten veranderen, hebben cellen nog een mechanisme hoe zij om kunnen gaan met signalen uit de omgeving. Er zijn kleine chemische stoffen, ook wel **post-translationele modificaties** (PTMs) genoemd, welke wanneer zij op een eiwit geplaatst worden kunnen dienen als een aan/uit schakelaar om belangrijke processen te beïnvloeden. PTMs kunnen bijvoorbeeld de activiteit, locatie binnen de cel en zelfs de degradatie van eiwitten beïnvloeden. Wanneer de processen die het plaatsen of verwijderen van deze PTMs ontregeld zijn, kan het tot zware ziektebeelden leiden. Een van de chemische groepen die op een eiwit geplaatst kunnen worden is de fosfaat groep. Het enzym **kinase** plaatst een fosfaat groep op het eiwit, terwijl het enzym **fosfatase** het weer kan verwijderen. Wanneer deze twee in balans zijn, zal de cel zich normaal gedragen. Echter wanneer deze twee ontregeld zijn, kan het tot zware ziekten zoals kanker leiden. Het onderzoeken van fosfaat op eiwitten met behulp van MS wordt ook wel **fosfoproteomics** genoemd en is een van de onderzoeksgebieden die wordt besproken in dit proefschrift.

In dit proefschrift hebben wij proteomics methoden gebruikt om fundamentele vraagstukken te beantwoorden in verscheidene soorten monsters. Wij onderzochten het mechanisme van een medicijn in zebravis, onderzochten het belang van een labiele PTM in menselijke cellen en tot slot hebben wij het oxidatie gedrag (en de gevolgen daarvan) van een eiwit bepaald.

In **Hoofdstuk 1** van dit proefschrift worden de basisprincipes van massa spectrometrie en proteomics uitgelegd. Allereerst wordt een kort overzicht gegeven van alle verschillende onderzoeksstrategieën die gevolgd kunnen worden met behulp van massa spectrometrie. Vervolgens wordt het vakgebied van shotgun proteomics, waar peptiden worden gemeten om de oorspronkelijke eiwitten te vinden, gedetailleerd besproken. De focus wordt gelegd op de keuze van monster, hoe deze op te breken, hoe de eiwitten gedigesteerd worden en hoe de monsters mogelijk minder complex gemaakt kunnen worden om zo betere data te genereren. Nadat dit besproken is, worden alle details met betrekking tot de massa spectrometer uitgelicht. Ionisatie methodes, massa analysatoren en fragmentatie methodes worden behandeld. Daarnaast worden ook meerdere database zoekalgoritmes besproken. Tot slot worden gespecialiseerde experimenten die met massa spectrometrie uitgevoerd kunnen worden uitgelegd. Zo worden verschillende manieren om het proteome te kwantificeren uitgelegd en volgt een gedetailleerde bespreking van de onderzoeksgebieden fosfo- en redox proteomics.

In **Hoofstuk 2** van dit proefschrift hebben wij het mechanisme van het medicijn napabucasin onderzocht in zebravis embryo's met behulp van thermal proteome profiling. In thermal proteome profiling worden de cellen in een cel verhit tot verschillende temperaturen. Bij een bepaalde temperatuur zal een eiwit vast worden en uit de vloeistof precipiteren. Dit is vergelijkbaar met het koken van een ei, waar bij een bepaalde temperatuur het eiwit hard zal worden. Echter, wanneer een medicijn bindt aan een eiwit zal de temperatuur waarbij het eiwit vast wordt veranderen. Dit gedrag van eiwitten wordt gebruikt in thermal proteome profiling. Wij voegden het medicijn napabucasin toe aan zebravis embryo's, kookten deze en bepaalden de temperatuur waarbij de eiwitten precipiteerden via massa spectrometrie. Door deze temperatuur te vergelijken met de temperatuur wanneer je geen medicijn toevoegt, hebben wij de eiwitten gevonden waar het medicijn aan bindt. Zo hebben wij laten zien dat de eiwitten aldehyde dehydrogenases off-target doelwitten van het medicijn zijn, wat grote gevolgen heeft op de embryonale ontwikkeling van zebravis embryo's.

In **Hoofstuk 3** van dit proefschrift hebben wij de gevoelige PTM phosphohistidine onderzocht in menselijke cellen. Phosphohistidine is een modificatie welke een belangrijke rol speelt in bacteriën, maar tegelijkertijd heel moeilijk te meten is. De standaardmethodes in het lab worden uitgevoerd onder zure omstandigheden, wat ervoor zorgt dat phosphohistidine van het eiwit af valt. Recentelijk zijn er methodes ontwikkeld die wel phosphohistidine kunnen meten, waarmee het belang ervan in bacteriën bewezen is. Nu waren wij geïnteresseerd of het ook een belangrijke rol in menselijke cellen speelt. Om dit te onderzoeken hebben wij meerdere experimenten uitgevoerd. Zo hebben wij meerdere celtypen onderzocht, een kinase assay ontwikkeld maar ook cel lysaat gefractioneerd om het minder complex te maken. Door deze experimenten hebben wij veel nieuwe phosphohistidine sites gevonden. Echter vonden wij vaak dezelfde sites wanneer het eiwit met zuur was behandeld, wat er normaal voor zorgt dat phosphohistidine verdwijnt. Hieruit blijkt dat de gevonden phosphohistidine sites niet waar kunnen zijn. Dus ondanks dat phosphohistidine een grote rol speelt in bacteriën, is het niet zo belangrijk of veelvuldig aanwezig in menselijke cellen.

In **Hoofdstuk 4** van dit proefschrift hebben wij de invloed van oxidatie op de tyrosine fosfatase SHP2 onderzocht. SHP2 kan de fosfaat groep van het aminozuur tyrosine verwijderen en via deze eigenschap een belangrijk effect hebben op signalering in de cel. Echter wanneer SHP2 geoxideerd wordt door waterstofperoxide zal het zijn activiteit verliezen. In dit onderzoek hebben wij onderzocht hoe makkelijk SHP2 geoxideerd wordt door waterstofperoxide. Daarnaast hebben wij dit ook gedaan voor de Noonan mutant van SHP2. Deze mutatie zorgt ervoor dat het eiwit zich in een meer open toestand bevindt, wat er mogelijk voor zorgt dat het sneller geoxideerd zal raken. Wij hebben laten zien dat dit daadwerkelijk het geval is met behulp van massa spectrometrie en fosfatase assays. Deze vondst kan mogelijk grote implicaties hebben voor syndromen die veroorzaakt worden door mutaties in SHP2.

Tot slot deel ik in **Hoofdstuk 5** van dit proefschrift mijn visie op de toekomst van proteomics. Daarnaast kunnen daar een lijst van mijn publicaties en mijn dankwoord gevonden worden.

Curriculum vitae

I studied the bachelor biomedical engineering at the technical university Eindhoven, where I focused on biochemistry and cell biology. Afterwards, I followed the master medical engineering, where I specialized myself in mass spectrometry in the field of clinical chemistry and proteomics. I conducted 2 research projects. In my first project, under the guidance of prof. dr. ir. Luc Brunsveld and prof. dr. Volkher Scharnhorst, I developed a mass spectrometry method on a triple quadrupole mass spectrometer. Using this method, the peptide hormone hepcidin could be measured to concentrations in the range of ng/ml in patient sera, which gave an added benefit to the diagnosis of anemia. At the end of my master I performed an internship at the Karolinska institute in Stockholm. Here, I performed my research in the group of prof. dr. Roman Zubarev, where I conducted thermal proteome profiling experiments to determine the global stability of proteins in cells with different grades of stemness. After witnessing the great potential of mass spectrometry and proteomics, I decided to further my career in this discipline.

After my studies, I came to Utrecht, where I started my PhD in the biomolecular mass spectrometry group of prof. dr. Albert Heck. I performed my PhD under the guidance of dr. Simone Lemeer, where I focused on multiple research topics in the field of proteomics. I tried to elucidate the role of phosphohistidine in mammalian systems, investigated the mode of action of the STAT3 inhibitor napabucasin by performing thermal proteome profiling on zebrafish embryos and lastly studied the oxidative behavior of the tyrosine phosphatase SHP2 and its mutants. The resulting work of the last 4.5 years are described in this thesis.

List of publications

Leijten, N., Bakker, P., Spaink, H.P., Den Hertog, J., Lemeer, S. Thermal proteome profiling in zebrafish reveals effects of napabucasin on retinoic acid metabolism. *Molecular & cellular proteomics* **20***,* 100033 (2021)

Jarzab, A., Kurzawa, N., Hopf, T., Moerch, M., Zecha, J., **Leijten, N.** *et al.* Meltome atlas – thermal proteome stability across the tree of life. *Nature methods* **17**, 495-503 (2020)

Mulder, C., **Leijten, N.**, Lemeer, S. Proteomic tools to study drug function, *Current Opinion in Systems Biology* **10**, 9-18 (2018)

Schmitz, E.M.H., **Leijten, N.** *et al*. Optimizing charge state distribution is a prerequisite for accurate protein biomarker quantification with LC-MS/MS, as illustrated by hepcidin measurement, *Clinical Chemistry and Laboratory Medicine,* **56** (9), 1490-1497 (2018)

Sabatier, P., Beusch, C., Saei, A., Aoun, M., Moruzzi, N., **Leijten, N.** *et al.* Plurifaceted proteomics method identifies key regulators of translation during stem cells maintenance and differentiation. [Under revision]

Acknowledgements

By writing this chapter, my four-year PhD journey has come to an end. I learned a lot about myself and how to conduct research during these four years, lessons that I will always cherish. However, I could not have made this journey all by myself, I had a lot of help. Therefore, I want to thank a whole bunch of people that have helped me along the way.

First and foremost I want to thank **Simone**. I want to thank you for all the guidance which you have given me over the last four years. Your optimism and insights always gave me new energy when we discussed projects, which led to new and exciting results. Even though some projects did not go as planned, you always taught me to see the possibilities and make the most out of it. During these four years, you have made me a better scientist and person, for which I am extremely grateful. I also want to thank **Albert**, who invited me for an interview four years ago, which led to me being a part of the group. I am extremely grateful for the opportunity to work in your group and really appreciated the honest feedback that you gave.

I also want to thank all the people of the phospho subgroup. It was always nice to unwind after a hectic week in the lab during the Friday meetings. Also, I will never forget the trip to the Zugspitze, where I had my first experience skiing. **Nadine**, I really enjoyed working together with you. All the time working together on the Princess and Horst, troubleshooting the finicky protocol and worrying about sticky peptides were nicer with you around. Thanks for showing me around Berlin, which I really enjoyed! I wish you the best in the last part of your PhD, which I am sure you will finish nicely! I am very grateful that you are standing beside me as paranimf during my defense. **Celine** and **Sander,** thank you for guiding me in the first year of my PhD! You really taught me a lot and were always willing to help if I had silly or stupid questions. **Miao** and **Clement**, thanks for the help while I was learning the phosphohistidine protocol. Thanks to **Grete** and **Justin** for their help on the phosphohistidine chapter.

I want to thank the whole of the Heck lab, both past and present members. There are too many people to name, but I want to address some of you in particular. **Arjan** and **Tobi**, I really liked sharing the (manliest) office with you guys! Even if I sometimes felt like Arjan's secretary, I really liked the laughs we shared. Z609!!! **Donna**, i always enjoyed our "koffieslurp" momentjes at the Gutenberg, where we could always share frustrations and motivate each other again. I will also never forget the summer school in Brixen, with the Rindfleisch bus and me struggling to get on top of the mountain. The summer school was nicer with you around. I wish you all the best in the final stretches of your PhD! I also want to thank you for standing beside me as paranimf during my defense. **Kelly D**, my FAQ co-founder, I really enjoyed setting these meetings up together with you and seeing how everyone benefitted from it. I also really enjoyed going to Brixen together with you. **Henk,** thanks for always being patient and helping me when I ran into issues with R. All the (past) members of team Fusion, **Suzy**, **Domenico**, **Theo**, **Kelly D**, **Karli and Johannes**, thanks for the many hours we spent together fixing the most problematic mass spectrometer of the lab. Even though at some point it was more broken than running, I liked working together with you to get it into the shape were in it is now.

I also want to extend my thanks to all the technicians of the lab. **Mirjam,** thank you for always taking care of the ordering and answering my stupid questions. You really helped me get familiar in the lab and be able to perform my experiments successfully. **Harm**, thanks for keeping order in the mass spec labs and always helping if something went wrong. **Dominique,** thanks for always helping me when I had LC issues that I could not fix myself. **Ceri**, **Pieter** and **Soenita,** thanks for taking care of everything cell culture related. **Geert**, thank you for keeping all the computers in the lab working and quickly responding if something went wrong. Lastly, thank you **Corine** for keeping everything in the lab running smoothly.

Besides working with colleagues in the Heck lab, I also had the pleasure of working together with people from the group of Jeroen den Hertog at the Hubrecht institute, which I would like to acknowledge here. **Jeroen**, thank you for being my second promotor. I really enjoyed working together with you on the TPP and SHP2 projects. I greatly value your insightful feedback and positive mindset. When I thought that data was not that great, you always distilled worthwhile information out of it! Thank you for guiding me during these four years. **Petra**, thanks for all the time that we worked together grinding zebrafish down to gain insight into drug interactions. I learned a lot from you and wish you well in the rest of your career. **Jelmer**, thank you for the nice collaboration on SHP2 in the beginning of my PhD! Even though we did not get the results we necessarily wanted, I learned a lot from it. Many thanks for stepping in in the final weeks of my PhD to work on the CoIPs! Lastly, thank you **Maaike** for continuing the research on the oxidation of SHP2 and getting the nice results that are shown in this thesis.

Naast veel steun van collega's, heb ik ook heel veel steun gehad van mijn familie en vrienden. Het was altijd fijn om te kunnen relaxen na een zware week op werk om zo weer met frisse moed aan de slag te gaan de volgende maandag. De abies, **Sander**, **Martin**, **Ivo**, **Perry** en **Thomas**, er was niks ontspannenders als in het weekend een biertje met jullie te pakken. Alle stress smolt weg tijdens de stap avonden en potjes 30 seconds, dank hiervoor!

Daarnaast heb ik ook heel steun gehad van mijn familie. **Pa** en **ma**, bedankt voor de onconditionele steun die jullie voor mij hadden tijdens dit traject! Jullie enorme interesse in wat ik doe, ondanks dat jullie er veel niet van begrepen, vond ik enorm fijn. Door jullie support heb ik dit allemaal tot een einde kunnen brengen. **Lars** en **June**, bedankt voor alle steun en interesse. **Rinus**, **Patricia** en **Mick**, vanaf het begin heb ik mij zeer welkom bij jullie gevoeld. Bedankt voor alle interesse en steun gedurende deze 4 jaar.

Last but not least, **Rosali**. Natuurlijk krijg jij een plek in mijn acknowledgements! Zonder jou had ik deze PhD niet tot een einde kunnen brengen. Je kwam in mijn leven precies op het moment dat ik met dit traject startte, dus je hebt alle highs en lows meegemaakt. Desondanks heb jij mij van het begin tot eind enorm gesteund. Altijd wanneer ik bij jou thuiskwam, of het nu in Den Haag was of in Dordrecht, smolt alle stress van mij af en werd ik weer een beetje meer mijzelf. Ook al volg je soms niet helemaal wat ik nu precies doe, “is dat dan eiwit?”, vond ik het enorm fijn om mijn verhaal bij jou kwijt te kunnen. Jij hebt mij een beter mens gemaakt en ik kan niet wachten op wat de toekomst ons brengt. Ik hou van jou.