

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER FAKULTÄT FÜR CHEMIE UND PHARMAZIE
DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

**Proteomics and phosphoproteomics
applied to
cell signaling and cancer**

Rochelle Caroline Justina D'souza

aus

Mangalore, Indien

2013

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 04th November 2013

Rochelle Caroline Justina D'souza

Dissertation eingereicht am **04.11.2013**

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: Prof. Dr. Peter ten Dijke

Mündliche Prüfung am **28.11.2013**

Summary

Signaling networks control and regulate outcomes in cells and organisms in both normal physiology and pathophysiological states. Signaling is traditionally represented and studied as a series of stepwise enzymatic events constituting a cascade. However, it is increasingly apparent that such representations limit understanding of signal transduction since these linear cascades function in an interconnected network that includes extensive cross talk among receptors and pathways. Mass spectrometry (MS)-based proteomics is a useful tool that allows a system-wide investigation of signaling events at the levels of post-translational modifications (PTMs), protein-protein interactions and changes in protein expression on a large scale. This technology now allows accurate quantification of thousands of proteins and their modifications in response to any perturbation.

This thesis work is dedicated to the optimization and employment of quantitative mass spectrometry to cellular signaling and an application to segregate two lymphoma subtypes at the levels of protein expression and phosphorylation, employing state of the art liquid chromatography (LC)-MS/MS technologies coupled with improved sample preparation techniques and data analysis algorithms.

In the **first project** I investigated the feasibility of a new, high accuracy fragmentation method called higher energy collisional dissociation (HCD) for the analysis of phospho-peptides. Using this method we were able to measure the phospho-proteome of a single cell line in 24h of measurement time which was a great improvement to previous capabilities. This fragmentation method that was originally thought to be slower and less sensitive than the standard method of low resolution collision induced dissociation (CID) fragmentation. However, our work proves this not to be the case and we showed that HCD outperformed the existing low resolution strategy [1].

In the **second project** I employed this HCD fragmentation technique on the LTQ-Orbitrap Velos for addressing the clinical question of segregating two subtypes of diffuse B-cell lymphoma (DLBCL). These subtypes are histologically indistinguishable but had been segregated on the basis of a gene expression signature. I employed the recently developed 'super-SILAC' approach with a 'super-SILAC mix' of multiple labeled cell lines. This heavy reference mix was spiked into several cell lines

derived from the two DLBCL subtypes and analyzed LC-MS, resulting in successful segregation based on a distinct proteomic signature [2].

The **third project** deals with the in-depth analysis of the phospho-proteome of a human cancer cell line on a quadrupole-Orbitrap mass spectrometer using a label-free quantification approach. Our analysis uncovered about 50,000 distinct phosphorylated peptides in a single cell type across a number of cellular conditions allowing assessment of global properties of this large dataset. Strikingly, we found that at least three-quarters of the proteome can be phosphorylated which is much higher than current estimates. We also analyzed phosphotyrosine events using enrichment with anti-phospho-tyrosine antibodies to identify more than 1,500 site specific phosphorylation events. Unexpectedly tyrosine phosphorylated proteins were enriched among higher abundance proteins. The observed difference in phospho-protein abundance correlated with the substrate K_m values of tyrosine kinases. For the first time we calculated site specific occupancies using label-free quantification and observed widespread full phosphorylation site occupancy during mitosis.

In the **final and main project**, I applied proteomics and phospho-proteomics to the study of signal transduction in response to transforming growth factor-beta (TGF- β), a multifunctional cytokine. TGF- β signaling regulates many biological outcomes including cell growth, differentiation, morphogenesis, tissue homeostasis and regeneration. The cellular responses to this multifunctional ligand are diverse and can even be opposed to each other, depending on the cell type and the conditions. To shed light on the reasons for the different outcomes, we analyzed the early phospho-proteome and ensuing proteome alterations in response to TGF- β treatment in a keratinocyte cell line. The early SILAC based phospho-proteome analysis uncovered over 20,000 phosphorylation events across five time points (0 to 20 min) of TGF- β treatment. Building on our recent advances in instrumentation, sample preparation, and data analysis algorithms we measured a deep TGF- β responsive proteome at six late time points (6h to 48h) with corresponding controls in only eight days of measurement time. Our label-free approach identified about 8,000 proteins and quantified more than 6,000 of them. This deep proteome covered well established pathways involved in TGF- β signaling, allowing global evaluation at the level of individual pathway members. Combining the TGF- β responsive proteome with an *in-silico* upstream regulator analysis, we correctly retrieved several known and predicted novel

transcription factors driving TGF- β induced cytostasis, de-differentiation and epithelial to mesenchymal transition (EMT). The combined analysis of transcription factor regulation with early phosphorylation changes and proteome changes enabled visualization of the intricate interplay of key transcription factors, kinases and various pathways driving cytostasis, EMT and other processes induced by TGF- β .

In summary, my thesis developed a highly efficient phospho-proteomic workflow, which was applied to the measurement of a very deep phospho-proteome of a single cancer cell line allowing analysis of its global features. The main achievement was the first in-depth and combined study of the phospho-proteome and resulting proteome changes following a defined signaling event, in this case leading to a time-resolved view of TGF- β signaling events relevant in cancer.

Abbreviations

ABC	activated B-cell
CDK	cyclin dependent kinase
CID	collision induced dissociation
Da	Dalton
DC	direct current
DHB	dihydroxy benzoic acid
DLBCL	diffuse large B-cell lymphoma
DTT	dithiothreitol
ECM	extracellular matrix
ELISA	enzyme linked immuno sorbent assay
ETD	electron transfer dissociation
EMT	epithelial to mesenchymal transition
ESI	electrospray ionization
FASP	filter aided sample preparation
FT	Fourier transformation
FT-ICR	Fourier transform ion cyclotron resonance
FWHM	full width at half mass
GCB	germinal B-cell like
HCD	higher energy collisional dissociation
HPLC	high performance liquid chromatography
iBAQ	intensity based absolute quantification
IT	ion trap
JNK	c-Jun N-terminal kinase
LFQ	label-free quantification
LIT	linear ion trap

Abbreviations

MALDI	matrix assisted laser desorption ionization
MAPK	mitogen activated protein kinase
MS	mass spectrometry
MSA	multi stage activation
m/z	mass/charge
PK	protein kinases
PP	protein phosphatases
PPI	protein-protein interactions
ppb	parts per billion
ppm	parts per million
PrEST	protein epitope sequence tag
PTM	post-translational modification
RF	radio frequency
RTK	receptor tyrosine kinase
SARA	SMAD anchor for receptor activation
SDS	sodium dodecyl sulphate
SH2	src Homology 2
SILAC	stable isotope labeling of amino acids in cell culture
SMAD	mothers against decapentaplegic homolog
STK	serine threonine kinases
TGF- β	transforming growth factor-beta
T β R	transforming growth factor-beta receptor
TFA	tri-fluoroacetic acid
TiO ₂	titanium dioxide
TGF- β	transforming growth factor-beta
ToF	time of flight
UHPLC	ultra high performance liquid chromatography

Contents

Summary.....	i
Abbreviations.....	iv
1 Introduction.....	01
1.1 <i>Basic principles of mass spectrometry (MS)-based proteomics</i>	03
1.1.1 Mass spectrometry: Instrumentation and workflow.....	04
1.1.2 PTM analysis by MS.....	20
1.1.3 Quantitative approaches in MS-based proteomics.....	24
1.1.4 Computational analysis in MaxQuant	28
1.2 <i>Clinical applications of MS-based proteomics</i>	31
1.3 <i>Signal transduction in the eukaryotic cell</i>	33
1.3.1 Phosphorylation as a PTM: The role of kinases and phosphatases.....	34
1.3.2 Phosphorylation and its relevance in cancer.....	35
1.3.3 The role of mass spectrometry to study phosphorylation as a PTM.....	36
1.4 <i>TGF-β signaling pathway</i>	38

2	Article 1: Feasibility of large-scale phospho-proteomics with higher energy collisional dissociation fragmentation.....	40
3	Article 2: Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles	50
4	Article 3: A very deep and very high accuracy phospho-proteome reveals fundamental differences between tyrosine and serine/threonine phosphorylation events.....	64
5	Article 4: In-depth and time-resolved dissection of early phospho-proteome and ensuing proteome changes in response to TGF- β signaling	95
	Conclusion and perspectives	137
	References	139
	Acknowledgements	147
	Resume	149

1. Introduction

On completion of sequencing of the human genome in the first few years of this decade [3, 4], the development of many large scale technologies ensued. The traditional 'one gene-one protein' approach was replaced by new generation 'omics' techniques such as genomics, transcriptomics and proteomics which exploited availability of sequenced genomes. These novel technologies are still being improved and have led to a hand in hand development of computational platforms and strategies. Omics studies due to their inherent unbiased nature allow researchers to look at the working of the cellular machinery starting from an 'unzoomed' approach leading to identification of novel cellular molecules involved in any given process. The large scale hypothesis free nature has enabled the application of these omics studies to better understand the complex biology of living systems. They have also been used in biomarker discovery, determination of patient response to drugs and to answer other clinical questions. They are now increasingly being applied in combination for personal omics profiling as was recently successfully applied to a single individual leading to successful prediction of predisposition to diabetes [5].

Genomics measures the genotype of an organism, the mRNA transcripts encoded by the active genes is measured by transcriptomics and proteomics measures the expressed proteins of a cell line, tissue, or organism at a given point in time. Genomics has helped researchers better understand genetic variations, varying genetic expressions, the roles of genes and their relationship with one another. The genome is mostly fixed for a given cell line while the transcriptome and proteome vary since it reflects the genes that are active at any given time and can be greatly affected by the environment making their measurements more challenging than the former. The proteome which is undoubtedly the most complex of the three reflects active gene action in the form of protein expression and transient activity based on post-translational modifications (PTMs).

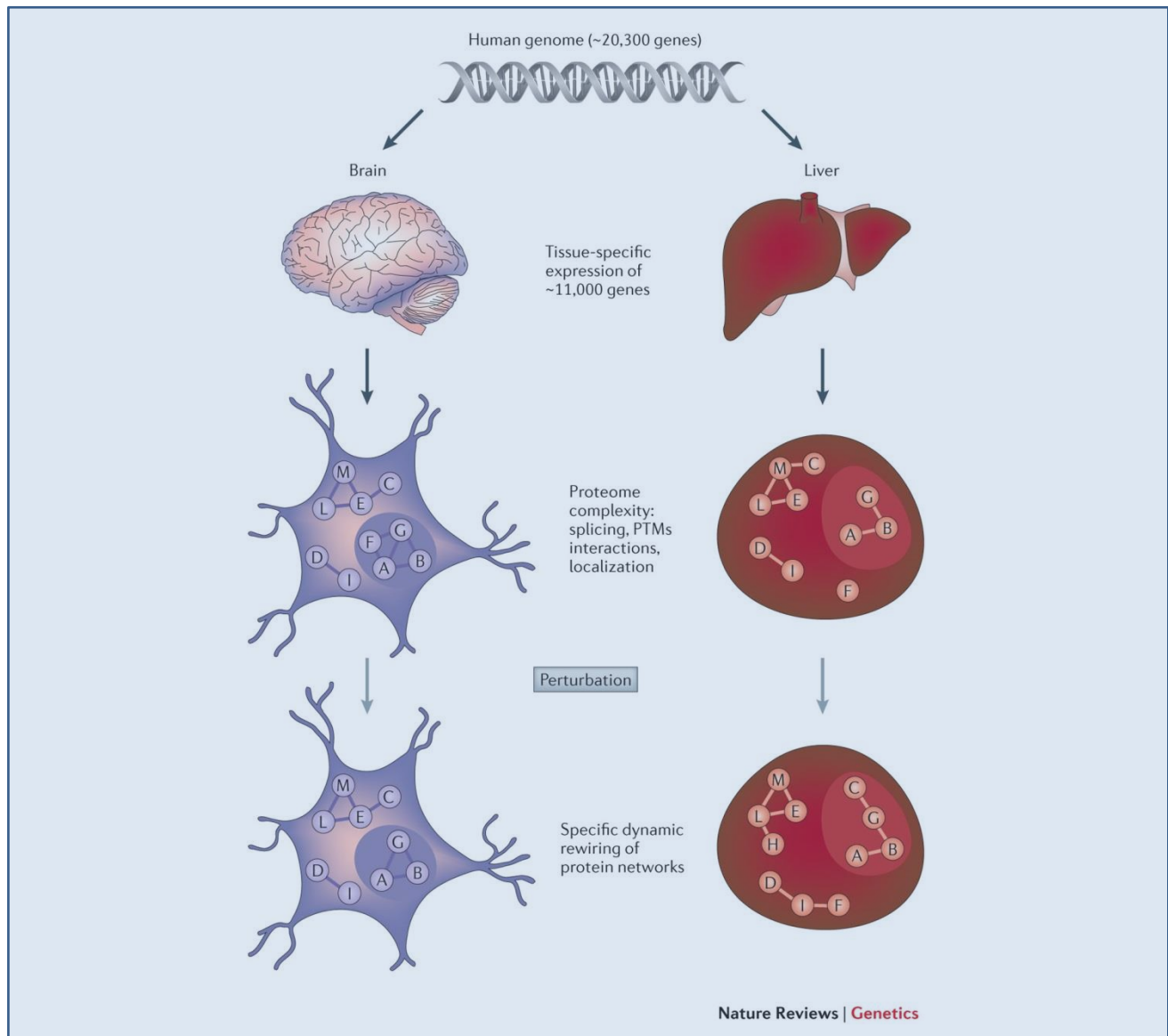


Figure1: The mechanisms of proteome regulation.

The human genome contains approximately 20,300 genes of which ~11,000 genes are predicted to be expressed in a given cell determining its phenotype. The proteome complexity is a result of many cellular mechanisms including protein regulation, including splicing variants, post-translational modifications (PTMs), protein–protein interactions (PPIs) and subcellular localization. A second layer of complexity is added as a result of differential rewiring of protein networks due to different perturbations. This figure is taken from [6].

Despite the similarities between transcriptomics and proteomics, the results are never overlapping and correlations between the two are in the range of 0.4 to 0.7 (Pearson's correlation co-efficient) [7, 8]. This is because of varying mRNA stability, varying rates of mRNA transcription in comparison to protein translation, post-translational regulation, differing protein stabilities and protein degradation mechanisms [9] as visualized in **Figure 1**. In-depth measurements of proteomes became a reality on account of recent developments and advances in mass spectrometry (MS)[10] thus enabling the routine identification of thousands of proteins [11, 12]. Despite advances in proteomics technologies, in depth analysis of proteomes is time-consuming and laborious and has not yet reached the throughput of genomics and transcriptomics. Nevertheless, proteomics technology is very attractive and is the tool of choice since it provides readout of the functional molecules rather than genetic code or mRNA abundance.

1.1 Basic principles of mass spectrometry-based proteomics

Mass spectrometry (MS)-based proteomics is today an attractive technology for the study of abundance, modification state, localization and interaction of proteins in a systematic way [13, 14]. In the last decade proteomics technologies have seen rapid advances in preparative techniques, MS-instrumentation and computational analysis, which have aided its increasing usage in all areas of basic and applied life sciences [10]. Proteomics provides a new tool that enables unbiased and global studies of cellular processes of interest with the possibility of measuring contextual relationships of proteins, such as their interactions, copy numbers, modifications and cellular localizations. This is highly desirable since researchers are realizing that protein function heavily relies on a complex, dynamic and cooperative network in contrast to the traditional molecule-centric, single directional pathway-based approach [15].

In its earlier years MS was predominantly employed by chemists for the study of small molecules. Sir Joseph Thomson, who was awarded the Nobel Prize in 1906 for his discovery of electrons, is considered the founder of the field of MS. This was then followed by research and developments in several areas of physics leading to the manufacture of many mass spectrometers including time of flight (ToF) and quadrupole mass analyzers, which allowed accurate determination of the mass/charge (m/z) of introduced molecules. Potential interest of biologists was limited by the fact that large molecules like proteins and peptides could not be vaporized and ionized, which are

necessary conditions for MS-analysis. This limitation was overcome by the invention of soft ionization methods, in particular electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI) methods by John Fenn and his co-workers at Yale University [16, 17] and Michael Karas and Franz Hillenkamp at Frankfurt University [18] (**Figure 2**).

1.1.1 Instrumentation and workflow for shotgun proteomics

A mass spectrometer basically consists of an ion source that introduces an analyte into the machine and also converts the analyte molecules into gas-phase ions, a mass analyzer that separates ions on the basis of their m/z ratios, and a detector that records the number of ions at each m/z value to generate a mass spectrum.

A) Ionization methods

One of the most important developments in instrumentation for biological mass spectrometry was the introduction of the gentle ionization methods MALDI and ESI that allowed for proteins and peptides to be analyzed by MS despite being polar, non-volatile and unstable.

In the **MALDI** approach, peptides or proteins are mixed in an excess of organic matrix and co-crystallized (**Figure 2A**). Laser energy focused on the sample is absorbed by the matrix and transferred to the peptides/proteins, causing the analyte to vaporize and ionize as singly-charged species. Since singly-charged ions are not fragmented easily by the low energy collision induced dissociation (CID) process employed in many contemporary mass spectrometers, MALDI ion sources are typically combined with ToF analyzers. The measured peptide masses of a protein can then be compared to a database containing information about the calculated peptide masses for all proteins of species.

In **ESI**, biomolecules in solution are electrosprayed at the end of a hypodermic needle into a strong electric field applied between the orifice and the entry point of the mass spectrometer (**Figure 2B**). Ions accumulate at the liquid surface upon eluting from the capillary and the liquid forms a 'Taylor cone'. Charged micro droplets are generated from which the neutral solvent quickly evaporates while the charged ions in the droplet undergo dispersion by charge repulsion resulting in desolvated analyte ions, which enter the mass spectrometer. In contrast to MALDI, this

ionization technique typically produces ions in multiple charged states which makes their fragmentation more efficient. In addition, ESI is easily interfaced with liquid chromatography (LC) thereby allowing direct on-line separation and making it the method of choice for analyzing complex peptide mixtures.

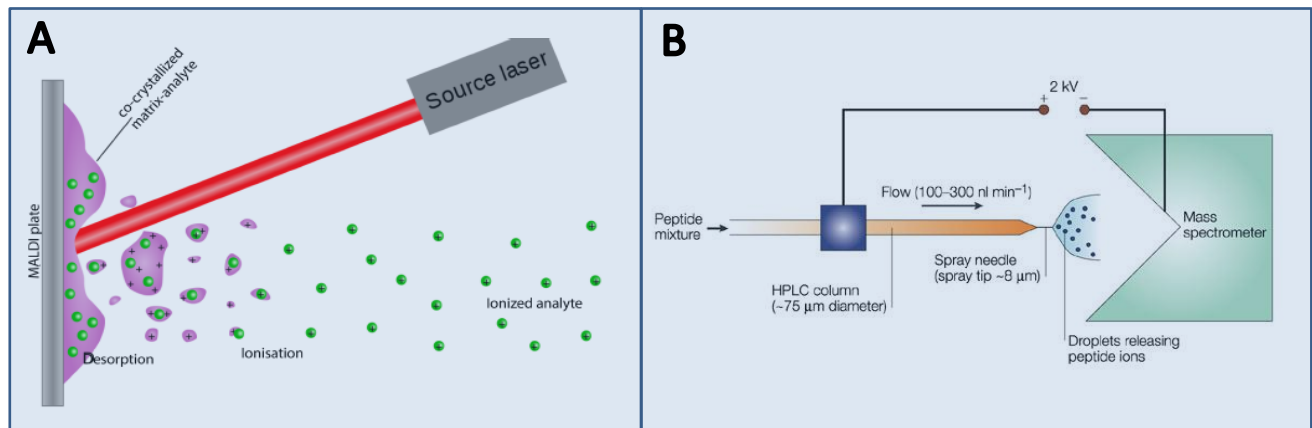


Figure 2: Soft ionization methods for biomolecules.

A: Matrix associated laser desorption ionization (MALDI) where the analyte is co-crystallized with an organic matrix and ions are generated using a laser source. Modified from [19] **B: Electrospray ionization (ESI)** in which a biomolecule solution is sprayed as a fine mist from the tip of a needle. Image from [20]

B) Mass analyzers

Principles:

Once inside the mass spectrometer the ionized peptides are channeled into the mass analyzer, the heart of the mass spectrometer. For their basic operation, classic mass analyzers use electric and magnetic fields to apply a force on charged ions. The relationship between the three factors can be summarized in the following equations:

$$F = ma \text{ (Newton's second law)}$$

$$F = e(E + v * B) \text{ (Lorentz force law)}$$

Where \mathbf{F} is the force applied to the ion, m is the mass of the ion, \mathbf{a} is the acceleration, e is the elemental charge, \mathbf{E} is the electric field and $\mathbf{v} * \mathbf{B}$ is the product of the ion velocity and the applied magnetic field. From these laws, it is apparent that the force causes an acceleration that depends on mass and that the applied force is dependent on the ionic charge. These two observations are the reason why mass analyzers separate molecule based on m/z ratios and not mass alone.

Features of mass analyzers:

For comprehensive proteome analysis, mass analyzers with the maximum possible resolution and mass accuracy, sensitivity and faster scan rates are highly desirable.

Mass accuracy is the difference between the measured mass and its calculated value and is measured in part per million (ppm), or parts per billion (ppb) or in absolute units in Dalton (Da). High mass accuracy is strongly dependent on the resolution of the mass analyzer, i.e. its ability to separate adjacent peaks. The **resolution** is mass difference of the closest mass pairs that can still be distinguished, divided by their mass ($\Delta m/m$). A related and more practical definition describes resolution as the width of a peak at a certain height, usually the height at half maximum, divided by the m/z (Full Width at Half Maximum, FWHM). In proteomics resolution is important since peptides often co-elute from the chromatography columns that are usually coupled to MS and may not be distinguishable in case of poor resolution. Similarly, accurate determination of masses is key to identifying a given peptide of given mass from a database containing all theoretically possible peptides, which contain many peptides of very similar masses. It is also key to post-translational modification (PTM) characterization.

The **sensitivity** of the mass analyzer is its ability to detect low level signals, originating from few ions. It is a key parameter in protein analysis, since the amount of biological material is limited as a rule. Furthermore, sensitivity is a precondition to achieving **high dynamic range** of the analyzer. The dynamic range in proteomics (highest abundant components compared to lowest abundant components) is a key challenge because it can span more than 10 orders of magnitude [21]. Lastly, the **scanning speed** of the analyzer is especially important because most proteomic measurements involve peptide elution from a chromatographic column which is directly sprayed into the mass spectrometer requiring very fast scanning speeds. The scanning speeds of modern

mass spectrometers cycles can enable acquisition of one MS spectrum and 10 MS/MS spectra per second [22]. ToF instruments do not have a scanning speed per se because they acquire thousands of individual spectra per second.

Types of mass analyzers:

In the past century various types of analyzers were developed including ToF, quadrupole, three dimensional ion traps (3D IT) or linear ion traps (LIT), Fourier Transform Ion Cyclotron Resonance (FT-ICR) and the Orbitrap analyzer. I will provide a short introduction to those analyzers that were used in this thesis.

Linear ion trap (LIT)

The LIT contains four orthogonally positioned conducting rods, each segmented into three sections (**Figure 3A**). The rods are paired, and a radio frequency (RF) voltage is applied to the rod pairs. To trap ions radially, two opposing rod pairs receive the same voltage, while voltages of the neighboring rods are opposite but of the same amplitude. This leads to a potential well in radial direction confining the trajectories of the ions. To trap ions axially, different direct current (DC) voltages are applied to the three sections to create an electric potential well in the center section. Both modes of trapping result in confining the trajectories of the ion to a fixed path or volume.

The behavior of the ions and their movement is explained by the *Mathieu* equations:

$$a = \frac{8zeU'}{m(x^2 + y^2)\Omega^2}$$

$$q = \frac{4zeV'}{m(x^2 + y^2)\Omega^2}$$

where m is the mass of a trapped ion, e is the ion charge, z is number of charges on the ion, V' is the potential of the DC, U' is the DC offset, Ω is the frequency of RF, x is the distance from the center of the trap to the X rods and y is the distance from the center of the trap to the Y rods. Thus, the stability of the ion motion in the trap depends on the *Mathieu* constants a and q , which are functions of the trap voltages, angular frequency (Ω), and particle m/z ratio. Plotting a against q , provides the stability diagram [23] of the quadrupole (**Figure 3B**); only ions at the intersecting

mass-scan line (a/q constant) will pass through the quad. This controls which ions will have a stable or unstable trajectory based on the specified RF and DC voltages, independent of the initial position of the ion and energy inside the device. If there is no DC offset, the equations can be simplified, because a has a value of 0, and leaves a one dimensional stability definition for ions which only depends on q . Ions are stable in an ion trap as long as their q value is below 0.908. The lowest mass-to-charge ratio that is stable in an ion trap is referred to as low mass cut-off and is a disadvantage of ITs and LITs, because it limits the lowest fragment mass that can be retained and therefore analyzed.

ITs serve as excellent mass analyzers on account of their fast scan rates, high sensitivity, small size and relatively low cost but have limitations due to lower resolution (especially in comparison to FT-ICR), the low mass cut off feature and lower resolution.

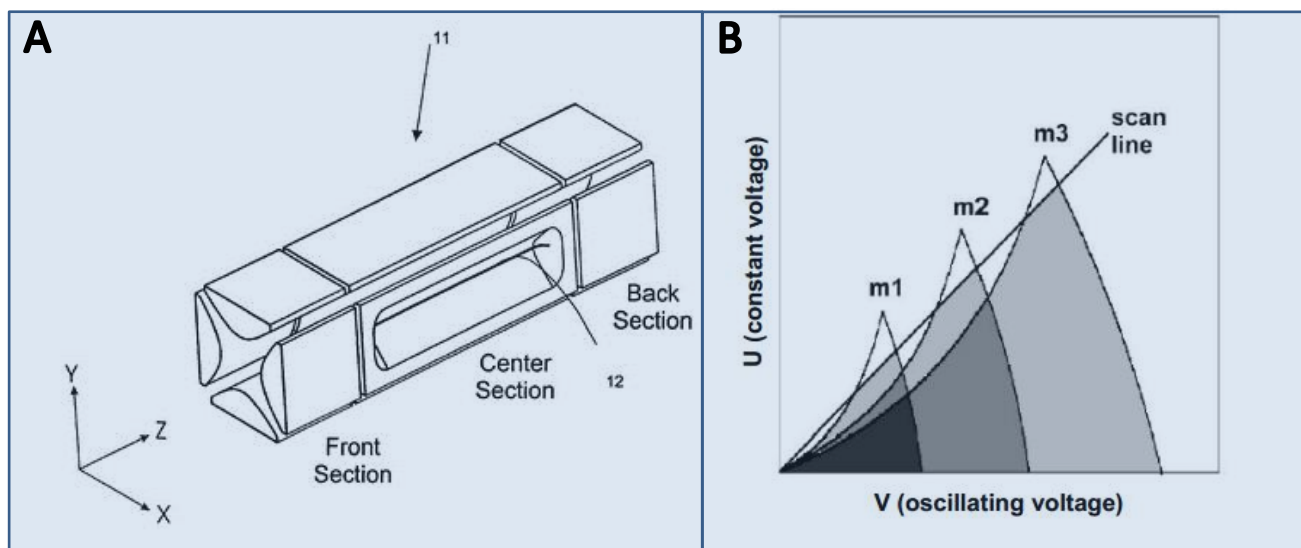


Figure 3: Linear ion trap mass analyzer. A: Schematic of the LIT. The center section has a slit that facilitates the axial ejection of ions. Figure is from [24] **B: Stability diagram** for determination of m/z ratios that are stable. m_1, m_2 and m_3 are increasing values of m/z and the triangular shape underneath each m/z is the stable region for the corresponding m/z value. At the scan line the trajectory is stable. Image from [25]

Orbitrap

The Orbitrap is a recently developed mass analyzer that has revolutionized the field of proteomics. It owes its basic design to the Kingdon trap, invented in 1923 by K. H. Kingdon [26] which was then modified by Knight in 1981 [27]. Based on this, after many improvements and novel insights, Alexander Makarov in 2000 introduced this new type of mass analyzer [28]. The Orbitrap is a purely electrostatic analyzer, in which the frequency of orbiting ions is measured as an image current. The Orbitrap is constructed from an inner spindle-like central electrode, surrounded by an outer barrel-like split electrode (**Figure 4**) resulting in a non-uniform space between the two electrodes along the z-axis. The electric field is weakest in the middle where the space between the two electrodes is largest. For efficient ion injection, before entry into the Orbitrap, ions are accumulated and stored in a quadrupole device termed 'C-trap'. Once inside the Orbitrap, the moving ions are trapped in an electrostatic field and start to orbit around the central electrode.

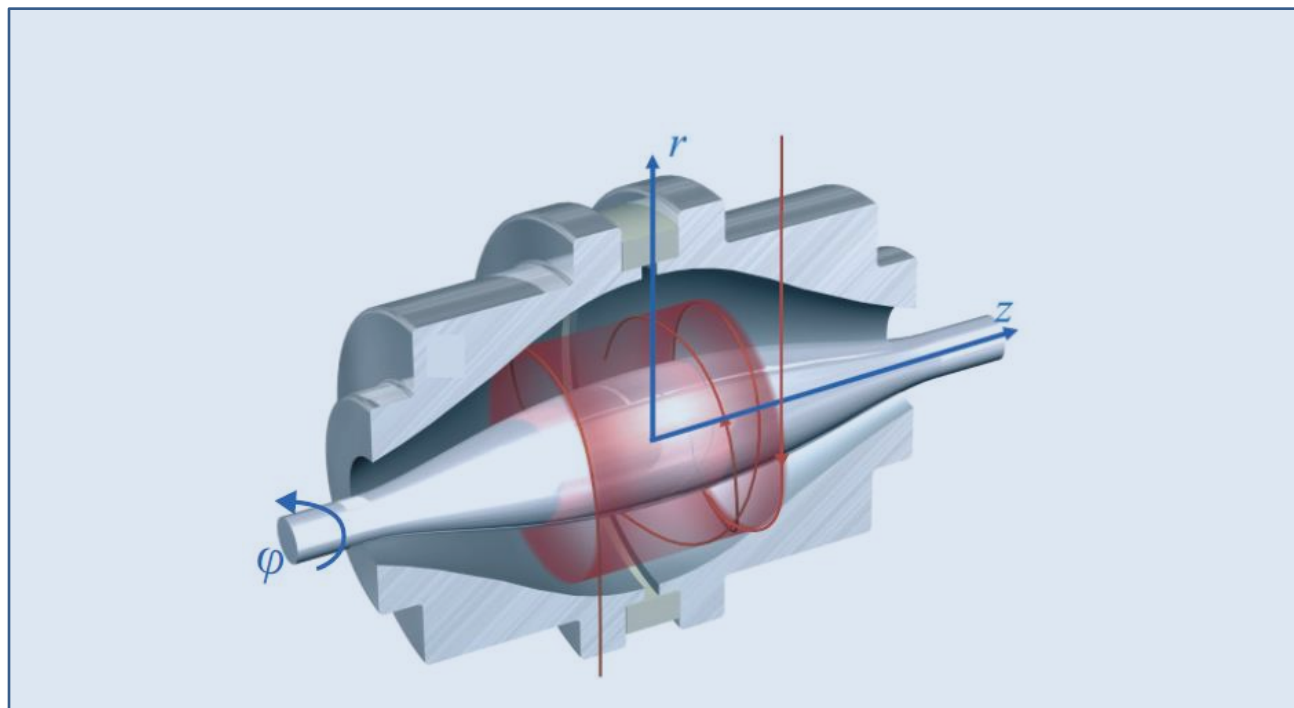


Fig 4: The Orbitrap mass analyzer.

A cross section of the Orbitrap showing the movement of ions (red) around the central electrode. Figure modified from [29].

This electrostatic attraction (axial) towards the central electrode is compensated by centrifugal forces (radial) that arise from the initial tangential velocity of ions. Since the motion along the axis of rotation is independent of rotational motion and depends only on the m/z , the ion image current in this direction is detected and employed for mass analysis. Fourier transform is employed to determine the oscillation frequencies for ions with different m/z 's very accurately. Such measurements achieve very high resolution rivaling that of FT-ICR instruments, and surpassing, by an order of magnitude [29], the resolution [30] presently obtainable with orthogonal ToF analyzers.

C) Detectors

The final element of the mass spectrometer is the detector, which records and amplifies the charge induced or the current produced from separated ions (by m/z) to generate the mass spectrum. Commonly used detectors include the electron multipliers that are effectively vacuum-tubes that multiply incident ions by a process of secondary emission leading to an avalanche of ions from a single ion. Apart from electron multipliers, other detector types use image currents as described above.

D) Fragmentation modes for data dependent acquisition

Shotgun proteomics typically involves the acquisition of a survey scan, isolation and fragmentation of peptide ions followed by acquisition of fragment spectra. This process is repeated sequentially for the top N most intense peptides. The Orbitrap generation of machines provides several modes of fragmentation and the most appropriate is chosen depending on the sample analyzed and the capabilities of the machine. The most commonly used fragmentation methods are CID and HCD, but ETD is also employed for specialized applications. The fragments resulting from fragmentation are dependent on the sequence of the peptide backbone. Different types of ions are generated based on the chemical bond cleaved. The fragment ions are classified according to the Roepstorff-Fohlmann-Biemann nomenclature[20]. The most common and informative are the b- and y-type ions, which result from cleavage of amide bonds with charge retention on the N- and C-terminus, respectively. Consecutive series of these ions spell out a partial or complete peptide sequence.

These ion types are typical in CID and HCD fragmentation. In contrast c- and z-type ions originate from cleavage of the N-C α bonds and are characteristic of electron capture dissociation (ECD) or electron transfer dissociation (ETD). Since **Article 1** of this thesis compares CID and HCD fragmentation, they will be explained in more detail in the following section.

Collision Induced Dissociation (CID) versus Higher-energy Collisional Dissociation (HCD)

In the hybrid mass spectrometers described in the chapter, CID experiments are performed in a LIT. In the first step, a precursor peptide ion selected for fragmentation is isolated, activated by an RF frequency, which accelerates it in the ion trap and leads to collisions of the peptide ions with the surrounding helium gas atoms and to fragmentation of the peptide backbone. Because the energy used for fragmentation is distributed in the peptide molecule, often only the weakest bonds are cleaved. The resulting fragments are not in resonance with the RF frequency and therefore do not fragment further. This is especially pronounced in the case of phosphorylated peptides which tend to show neutral losses and require additional strategies for comprehensive coverage of peptide fragments. Generally, in such ion trap based tandem MS experiments, the precursor masses are recorded in the Orbitrap at high resolution, the peptides are fragmented and the fragment masses are recorded in the low resolution ion trap. Such a strategy is therefore called a **‘high-low’ strategy** [31]. Advantages of CID or other ion trap fragmentation methods include high sensitivity since ion traps need fewer charges to detect a signal, faster speed of acquisition enabling fragmentation of more precursors, and – in a hybrid instrument - parallel MS and MS/MS acquisition since the two events are recorded in different analyzers. Disadvantages of CID include the relatively low mass resolution of the ion trap. Furthermore, the low mass cut off of fragments inherent to the ion trap fragmentation process [24] results in non-detection of lower mass ions, which can particularly hamper the detection of peaks diagnostic in PTM analysis. However it is possible to acquire spectra in the Orbitrap analyzer of ions fragmented by CID in ion traps to overcome a few of these limitations.

HCD which was first described in 2007, on a hybrid ion trap, Orbitrap instrument (LTQ-Orbitrap XL) [32]. It is a beam type CID method, which also generates b- and y-type fragment ions. Due to the fact that the fragment ions that initially are formed still collide with the gas in the collision chamber, HCD leads to efficient backbone cleavage, which resembles the classical triple

quadrupole fragmentation, and fewer exclusive neutral losses in the case of PTMs. In addition to the b- type and y- type ions HCD also generates a- type ions from further fragmentation [33]. Fragmentation is performed in a dedicated HCD cell and does not suffer from the 1/3 mass cut-off, another reason that produces spectra containing more information than in CID. This is especially attractive for detection of diagnostic peaks specific for some PTMs, immonium ions for all peptides and reporter ions used for TMT/iTRAQ quantification which are all in the low mass region [1, 34, 35]. HCD is almost always performed with fragment read out in the Orbitrap analyzer and is therefore a **'high-high' strategy**, because both precursor and fragment masses are acquired with high mass accuracy, leading to higher confidence in spectral matching. Disadvantages of HCD fragmentation are lower sensitivity, production of internal fragments that may make peptide matching complicated and most importantly, the consecutive acquisition of the MS and MS/MS spectra in the Orbitrap analyzer resulting in longer cycle times compared to ion trap CID. However, all these disadvantages have been greatly alleviated in the newest generation of instruments [22, 31].

E) The Orbitrap family of mass spectrometers

All experiments performed in this thesis were analyzed using the Orbitrap family of mass spectrometers. The invention of the Orbitrap analyzer, led to the development of a number of different hybrid high mass accuracy spectrometers for shotgun proteomics. The mass spectrometers belonging to this family include the LTQ-Orbitrap, LTQ-Orbitrap Velos, Orbitrap Elite and Fusion. All of these are hybrid instruments that use the Orbitrap for high accuracy measurements and all of which have an additional mass analyzer. Furthermore they also feature different fragmentation capabilities. The Exactive and Q Exactive are bench top devices that have only the Orbitrap as mass analyzer or a combination of quadrupole and Orbitrap as mass analyzers, respectively. For this thesis specifically, the LTQ-Orbitrap, LTQ-Orbitrap Velos and Q Exactive were employed and are therefore described in more detail.

LTQ-Orbitrap

The LTQ-Orbitrap was the first machine with the Orbitrap mass analyzer launched by Thermo Fischer Scientific (**Figure 5A**) [36]. This hybrid instrument uses an Orbitrap cell for measuring peptide masses at high resolution and a LIT for fragment spectrum analysis. For tandem mass spectrometry, it was equipped with CID the fragmentation mode. The LTQ-Orbitrap revolutionized proteomic analysis with researchers harnessing its ability to first analyze peptides eluting from the chromatography column with high resolution survey scans in the Orbitrap cell followed by sequential isolation and CID fragmentation of the top N most intense peptides in the LIT, which were mass measured in parallel in the LIT. One cycle, including an MS1 scan in the Orbitrap analyzer was typically configured to have a resolution of 60,000 and five fragmentation events, taking around 2.5 seconds. Measurements on a well calibrated machine accurately determine the peptide mass with an accuracy of few ppm at this high resolution. Mass accuracy was further boosted by the algorithms in MaxQuant and by injecting ambient molecules from laboratory air as internal recalibration standards [36]. This instrument was then upgraded to Orbitrap XL, essentially an LTQ-Orbitrap equipped with a dedicated collision cell for HCD fragmentation or optional ETD/ECD fragmentation [32].

LTQ-Orbitrap Velos

The LTQ-Orbitrap Velos is similar in construction to its predecessor the Orbitrap XL and was equipped with an Orbitrap cell, IT and an HCD cell (**Figure 5B**) [37]. There were several improvements however, that made it possible to perform HCD fragmentation in a high throughput fashion which was not possible before. Firstly the inlet was modified and equipped with an S lens allowing much better transmission of ions into the machine, thus increasing the sensitivity. The IT was replaced by a dual linear ion trap with the first part being operated at a higher pressure (6.7×10^{-3} mBar) than the second part. The higher pressure allows very efficient trapping, isolation and fragmentation of ions that are transferred into the second trap operated at lower pressure (5×10^{-4} mBar). This improved scanning and in combination with improved electronics and multipliers allowed recording of mass spectra at higher speeds.

HCD fragmentation scans are acquired in the Orbitrap analyzer and due to image current detection require a larger number of ions than a LIT, which has ion multipliers as detectors. The improved design (S-lens and C-trap-HCD cell combination) and electronics, enabled the analysis of up to ten-fold more ions per unit time into the HCD cell than before. This made HCD fragmentation (high-high strategy) routinely feasible and competitive with the already available CID mode (high-low strategy). A part of this thesis (Article 1) deals with assessing the application of the HCD strategy for phospho-proteomics analysis [1].

Q Exactive

The Q Exactive is a bench top instrument with an Orbitrap cell for high resolution analysis, which by design detects both precursor and fragment ions [22]. In contrast to its predecessor, the Exactive[38], it is equipped with an S-lens and a quadrupole that enables isolation of selected ions on a faster time scale (**Figure 5C**). The Q Exactive has a shorter ion path and has improved electronics. This in conjunction with ability to fill in parallel during analysis and employment of the enhanced Fourier Transform (eFT) algorithm results in a factor two increase in resolution. These improvements make the Q Exactive a more sensitive and faster instrument than any of its predecessors. A part of this thesis is dedicated to the in depth analysis of the phospho-proteome in a single cell line employing this instrument (Article 3). For the first time this instrument allowed the measurement of unfractionated proteomes in single LC-MS runs enabling near comprehensive yeast proteome analysis in a few hours of measurement time [39]. Another part of this thesis describes the unprecedented coverage using unfractionated single measurements to study cellular proteome and phospho-proteome changes in response to treatment with a growth factor on a temporal scale (Article 4).

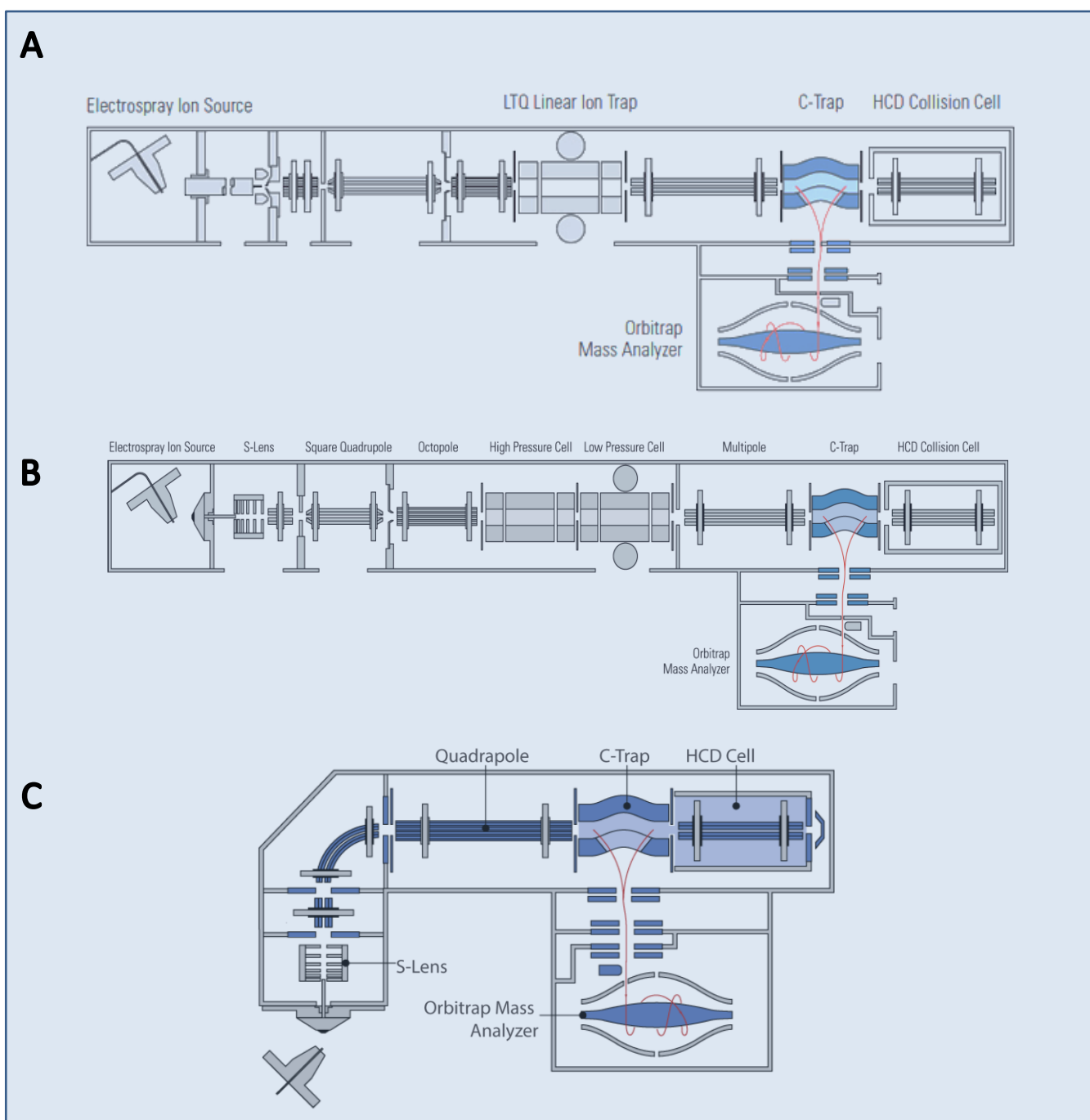


Figure 5: Schematic representation of the Orbitrap family of instruments.

A: The LTQ-Orbitrap the first generation of the Orbitrap family of hybrid mass spectrometers equipped with an Orbitrap cell and a linear ion trap. (HCD cell of the Orbitrap XL is also depicted) **B: The LTQ-Orbitrap Velos**, with its improved ion source the S lens, dual ion trap and HCD capabilities. **C: The Q Exactive**, a bench top mass spectrometer combining quadrupole and Orbitrap mass analyzers.

F) Workflow for shotgun proteomics

Protein identification via MS is usually performed in the abovementioned bottom up manner, where the peptides resulting from enzymatic cleavage of the proteome are analyzed. A specialized form of proteomics omits the cleavage step and therefore analyzes intact proteins ('top-down' proteomics) [40-42].

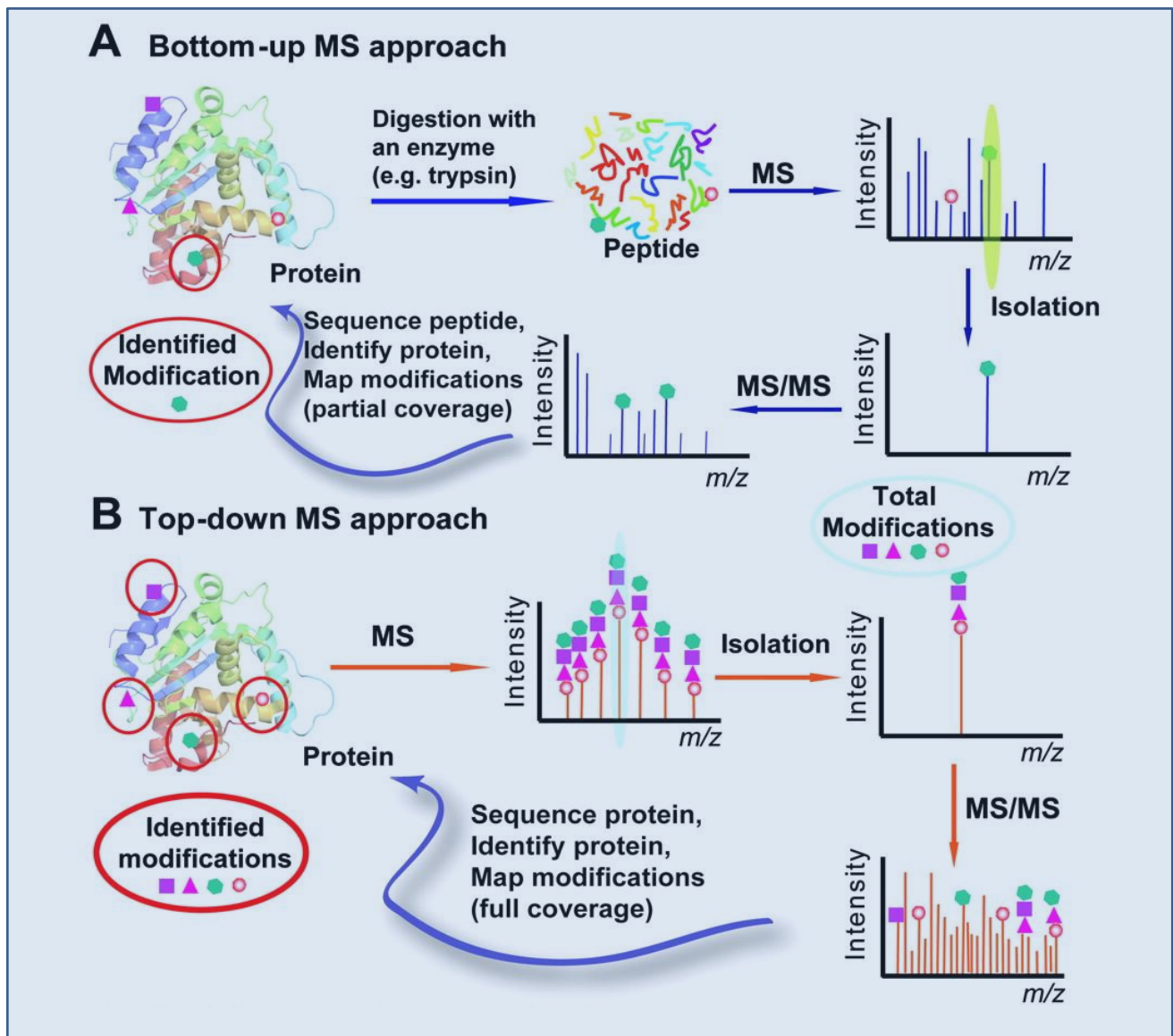


Figure 6: Two complementary modes of proteomics.

A: The bottom up approach where proteins are digested into peptides and analyzed by MS and MS/MS. **B:** The top down approach where intact proteins are analyzed in the mass spectrometer, optimally with full-sequence coverage. This figure is modified from [43]

The peptide-based bottom-up shotgun proteomics involves in-gel or in-solution proteolytic digestion of proteins with a sequence-specific protease into peptides before MS analysis (**Figure 6A**). This approach is well suited for protein identification, since it requires very little fragment information from a single or very few peptides to identify the protein in a database [20]. With the tremendous developments in instrumentation and software in the past years, bottom-up proteomics will remain the workhorse for proteomic analysis.

Top down proteomics (**Figure 6B**) measures intact proteins and in principle allows more comprehensive characterization of protein isoforms and post-translational modifications because any modification will shift the molecular weight of the protein [44, 45]. It preserves the labile structural characteristics which are lost in the bottom up mode [46]. In the form of native mass spectrometry, the top down approach can preserve non-covalent interactions with small molecules. This strategy has recently gained more prominence with its application to study protein complexes and their assembly [47-49]. This thesis exclusively uses bottom up proteomics for comprehensive proteome and phospho-proteome analysis and therefore those workflows and the data analysis strategies are discussed.

Comprehensive proteome analysis has become a reality due to availability of the genome sequence, developments in instrumentation, efficient sample preparation and separation strategies and improvements in computational analysis. MS-based proteomics has been used to study a range of organisms from prokaryotes to *Homo sapiens*. The proteome of any organism can be easily studied if its genome has been sequenced, providing the possible protein sequences as a database. Nevertheless, proteomics can also be performed on organisms whose genome is not yet or only partially available since they can still be analyzed by MS and fragmentation spectra interpreted by de novo sequencing [50, 51]. This is very challenging yet becoming more realistic.

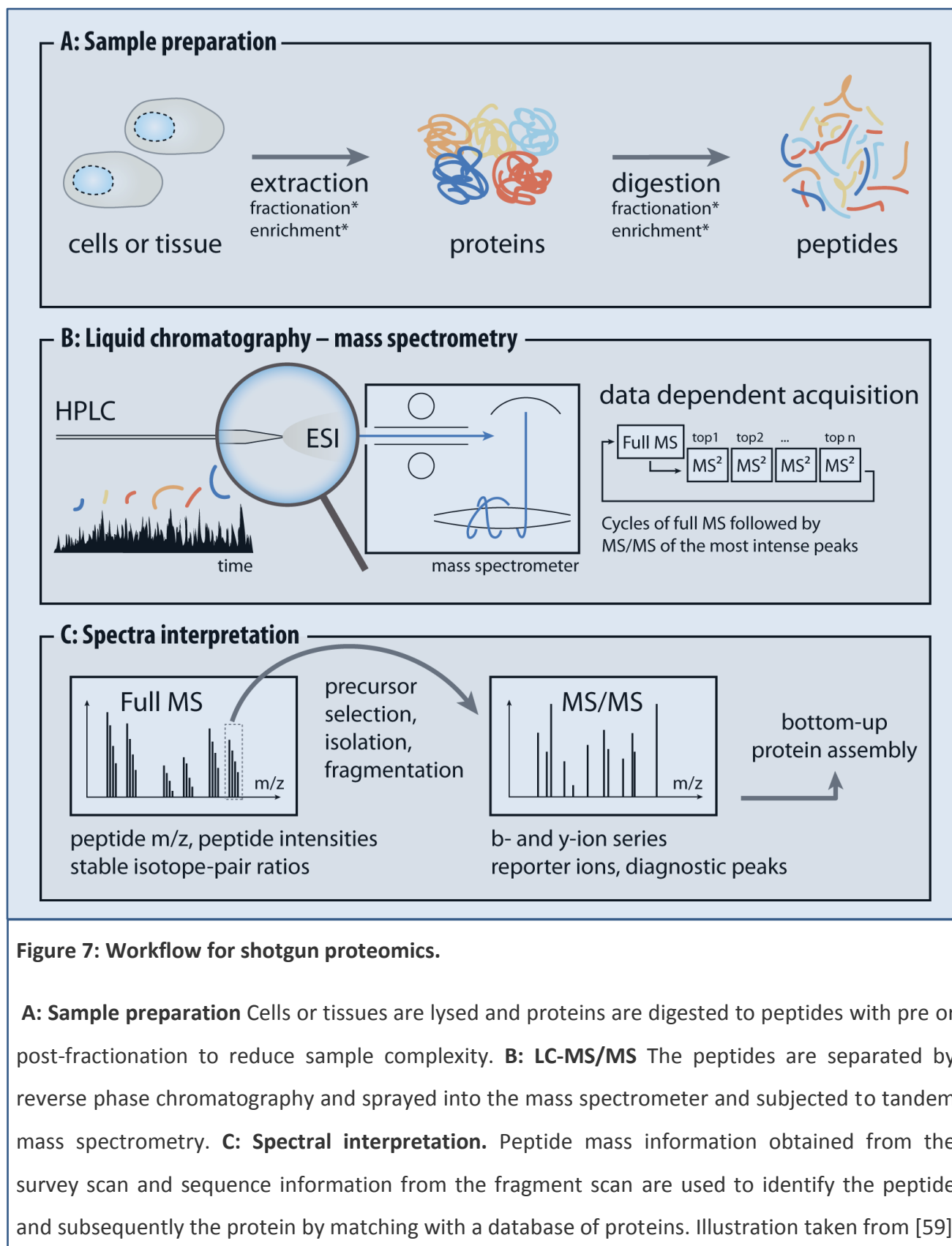
A typical shotgun proteomics workflow begins with sample preparation and digestion of the material of which the proteome needs to be analyzed (**Figure 7A**). Typically, cell lysis is the first step and it can be performed with gentle lysis buffers in cases where preserving complexes is important; or in presence of strong detergents like SDS which is used when all proteins, especially membrane proteins, need to be solubilized. The subsequent sample preparation uses in-gel

digestion where the proteins in the lysate are separated by SDS-PAGE based on their molecular weight, followed by cutting the gel small parts in which digestion can proceed [52], or alternatively by a 'gel-free' in solution digestion protocol. The method is chosen depending on the requirements of each experiment. Advantages of the gel-free method are high efficiency and simplicity. The traditional method of in-solution digestion cannot support SDS lysis and was therefore recently adapted and converted into a filter based digestion protocol, termed FASP [53]. In in-gel and in-solution methods, prior to digestion, the proteins are reduced using dithiothreitol (DTT) and alkylated using iodo-acetamide or chloro-acetamide.

The most common enzyme used for digesting proteins is the protease trypsin, which is efficient and cleaves proteins at arginine and lysine residues [54]. The resulting tryptic peptides are efficiently analyzed by MS and MS/MS in the positive mode since they carry the basic amino acids, arginines and lysines, at the C- terminus. Although trypsin the most widely used, it is also possible to use other proteases like LysC, GluC and AspN either by themselves or in combination for increased sequence coverage [12, 55].

The proteolytic peptides after digestion constitute an extremely complex mixture but can be analyzed in single measurements being separated on a reverse phase column and electrosprayed into the mass spectrometer. More often, they are subjected to fractionation, which is either performed before or after digestion. Methods of fractionation include size based SDS-PAGE [52, 56] and size exclusion chromatography, charge based cationic and anionic exchange chromatography and lastly isoelectric focusing. In certain cases where an ultra-deep coverage is desired more than one fractionation technique is used [12]. Before being sprayed into the mass spectrometer the samples can be de-salted on C18 containing membranes to remove all salts and contaminants [57].

Fractionated or unfractionated, the peptide mixture is still very complex and therefore is further separated by LC (**Figure 7B**). This is usually performed on line to the mass spectrometer and employing a narrow fused silica column packed with C18 beads [58]. Peptides are eluted from the C18 material with an increasing percentage of organic solvent. Very complex mixtures greatly benefit from longer columns and smaller bead sizes, which provide better separation. The



downside is increased back pressure which can partially be solved by a column oven to increase the working temperature and reducing the viscosity [60] or employing ultra-high pressure liquid chromatography (UHPLC) systems [39]. A recent development is the trend towards analyzing unfractionated samples as single runs employing the latest developments in reverse phase chromatography and the newest generation of mass spectrometers, which provide an appreciable proteome coverage in a short measurement time and with low sample consumption [39, 60]. The mass spectrometers described are employed for tandem MS analysis, with an initial precursor or full scan (MS1) covering a wide m/z range. This is then followed by topN (ranging from 5 to 20) full scan dependent fragment scans (MS2), which are acquired from selected peptides that are isolated and fragmented (**Figure 7B**). A dynamic exclusion list ensures that each precursor is selected only once during typical peptide elution times. Different fragmentation modes, typically CID or HCD, are applied to generate rich peptide fragment patterns, from which the amino acid sequence can be inferred by matching to a reference sequence (**Figure 7C**). This is accomplished with the aid of strong statistics-based algorithms that scan protein sequence databases (e.g. the search engines Mascot [61] or Andromeda [62]). Identified peptides are assembled into protein groups (sets of sequences that cannot be further distinguished by with the available peptide information) with FDR controls to limit the number of false positives.

Due to its largely unbiased nature, this data-dependent top N strategy is also called discovery proteomics. In contrast, it is also possible to employ hypothesis-driven, targeted approaches such as multi reaction monitoring (MRM) assays to detect and quantify proteins of prior interest [63]. This strategy can identify low abundant candidates in relatively short runs. Recently there have been efforts to combine the advantages of both approaches.

1.1.2 PTM analysis using MS

One of the challenges in the analysis of PTMs by mass spectrometry is their sub-stoichiometric abundances. Post-translationally modified peptides only constitute a minority of all peptides, and usually require enrichment for proteome-wide modification analysis prior to MS [64]. This is performed in several ways depending on the PTM under consideration. The most commonly used method employs antibody based immunoprecipitations (IPs) as is the case for ubiquitylated

peptides using the di-gly antibody [65]. This strategy actually does not analyze the PTM per se, which is a small protein modifier, but the remnant after tryptic digestion, which is shared between ubiquitin, NEDD8, ISG15 and other ubiquitin like modifiers [66, 67]. A similar strategy is also employed for lysine acetylation [68], methylation[69] and phospho-tyrosine [70] modifications. However, not all modifications have a corresponding antibody that efficiently and selectively recognizes them and therefore they require other methods. The characterization of glycosylation involves enzymatic or chemical release of the attached glycans followed by the analysis of the peptide [71, 72], from which the glycan was removed. A similar strategy of cleaving off the modification and MS analysis of the peptide or the derivatized peptide (using reporter ions) is employed for a number of less studied modifications. In addition to sample preparation, PTM analysis is computationally challenging since it not only requires identification of the peptides carrying the PTM, but also pin-pointing the actual site of modification. This along with the quantification and estimation of the pool of a peptide with the given PTM in comparison to its unmodified counterpart increases the levels of challenges faced in PTM analysis.

Phospho-proteomic analysis:

MS for proteome wide PTM analysis was first applied to phosphorylated proteins and peptides [56, 73, 74]. Phosphorylated peptides are most commonly are enriched using IMAC immobilized metal affinity chromatography (IMAC) [75, 76], strong cation exchange (SCX) chromatography and titanium dioxide (TiO₂) chromatography [43, 77, 78]. These approaches work more efficiently at the peptide level than the protein level. The TiO₂ and IMAC affinity methods mainly result in the identification of serine and threonine phosphorylation. Therefore when a deep coverage of tyrosine phosphorylated peptides is necessary, antibody affinity pull-downs are preferred (**Figure 8**). When peptides are first pre-fractionated with SCX or by the HILIC chromatography method and this is followed by a second round of enrichment with TiO₂ or IMAC [73, 75] (**Figure 8**).

The workflow used in the projects in this thesis employ Filter Aided Sample Preparation (FASP) based peptide digestion [53] to generate peptides, because experience in our laboratory has shown that this method is well suited to phospho-proteome analysis. Because of the low stoichiometry of phosphorylation, starting materials employed for phospho-proteome enrichment

are typically at least 2 to 6 mg, significantly higher than in proteome measurements. The tryptic peptides are then separated by SCX chromatography which also serves to enrich phosphopeptides [79] (**Figure 8**). SCX is employed in phospho-peptide analysis because phosphorylation results in reduced positive charges on phosphorylated peptides and in SCX this feature can be used to separate them from unphosphorylated peptides. Multiply phosphorylated peptides bind to the SCX column with minimum affinity and therefore the flow-through is also used for enrichment [79]. Next, enrichment of phospho-peptides is done with TiO_2 micro beads, achieving very high enrichment efficiency of greater than 90% [78]. To avoid interference from peptides with acidic amino acid residues (glutamic acid and aspartic acid) it is necessary to use competing agents to prevent their binding. Commonly used competing agents include 2,5-dihydroxy benzoic acid (DHB) [80] and 6% TFA [81] and these are routinely applied in large-scale phospho-proteomic analysis.

Comprehensive analysis of phospho-peptides also requires minor adjustments to the mass spectrometric methods. When phospho-peptides are fragmented by CID in the ion trap this often results in a single neutral loss of the phosphate group, due to its labile bond [82]. As a result, the fragmentation spectra of phospho-peptides can be dominated by a peak of 98 Da or 80 Da lower mass than the precursor (loss of H_3PO_4 and HPO_3 , respectively). This decreases the level of backbone fragmentation observed thus hindering identification of the peptide sequence [82]. The neutral loss is highest for serine phosphorylated peptides, followed by threonine and very rare for tyrosine phosphorylated peptides [83]. In ion traps this problem can in part be alleviated by Multi stage activation (MSA), which has now become a routine method [73, 84]. In MSA a precursor ion selected after the first MS analysis is activated at its m/z , followed by an additional activation at the anticipated m/z of the neutral loss of the phosphate group, while all the other fragments are still trapped. A second limitation of using CID-based ion trap fragmentation is the one-third rule, which in the case of phospho-tyrosine (pY) containing peptides results in the loss of the diagnostic phospho-immonium ion [34]. These limitations can be overcome by employing HCD fragmentation for large-scale phospho-proteomics, a strategy which also gains from the high mass accuracy at precursor and fragment levels. Article 1 describes the demonstration of feasibility of HCD based fragmentation for large scale phospho-proteomics on the LTQ-Orbitrap Velos. Such HCD

fragmentation based phospho-proteomic analysis is now routine on both the LTQ-Orbitrap Velos and Q Exactive platforms, identifying tens of thousands of sites. Article 2 demonstrates identification of the largest phospho-proteome of a cell line reported so far and in this project global analysis of phospho-peptides was performed on a Q Exactive instrument.

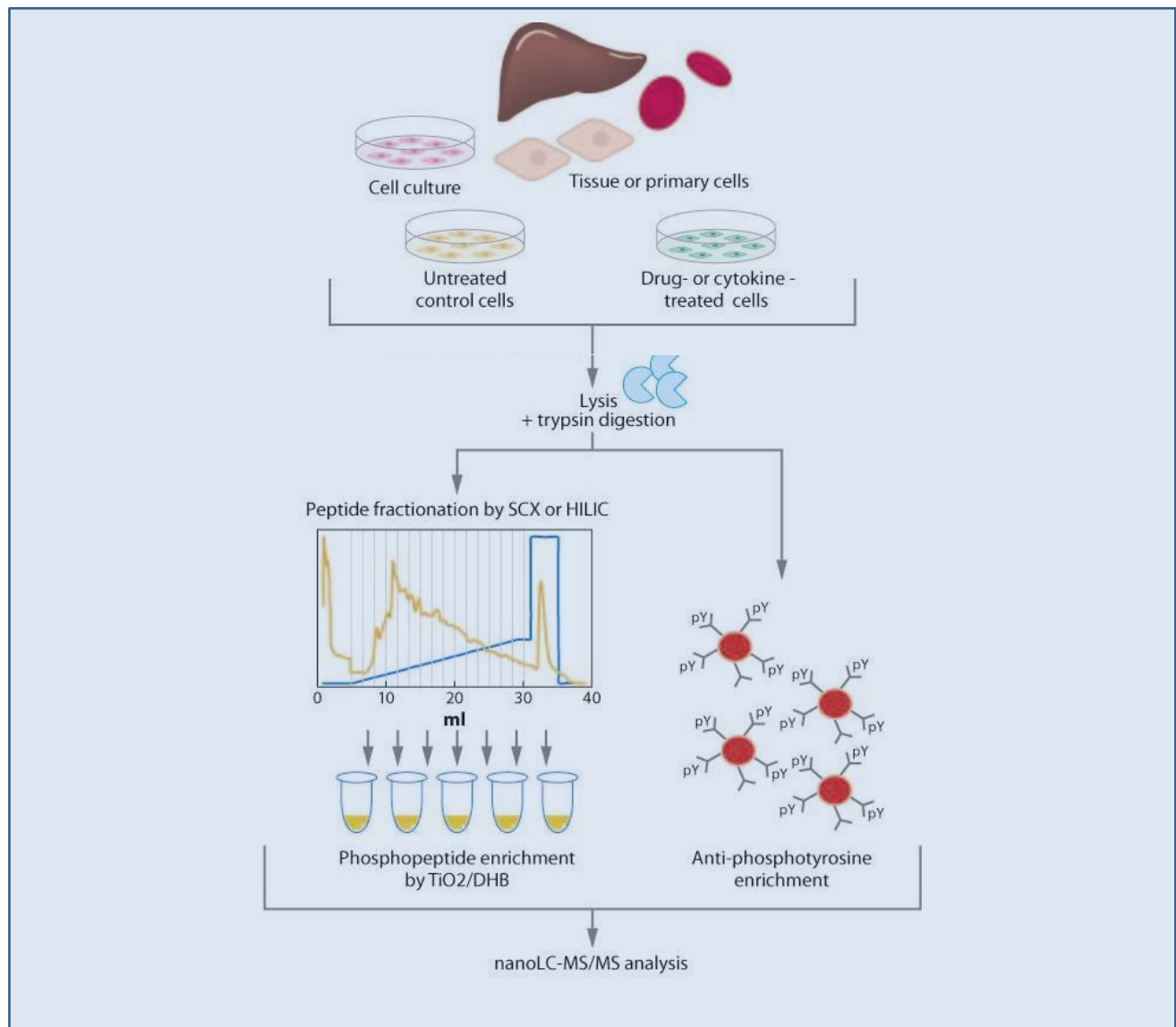


Figure 8: Phospho-proteomic workflow.

Protein lysate is digested by trypsin, and phospho-peptides are enriched by SCX and TiO₂ chromatography or by IP and measured by LC-MS. Image from [79].

1.1.3 Quantification approaches in MS-based proteomics

MS-based proteomics has matured into a technology to identify many proteins in complex mixtures in a relatively short time frame. However, identification of proteins is not sufficient. Estimation of the amounts of proteins or measuring changes in protein expression and PTM levels is also necessary. MS is not quantitative by nature due to differing behaviour of peptides during ionization, which means that concentrations of peptides cannot be determined directly from the signal intensities detected by MS. This necessitated the development of strategies to obtain quantification information for comparing of protein abundance between samples (relative quantification) or determination of concentration or of copy numbers (absolute quantification). Quantification of both types can be performed in two modes: label-based methods and label-free methods (**Figure 9**).

Label-based quantification:

Label-based approaches use isotopic labels to generate a mass difference to differentiate between proteins from the samples to be compared. The largest advantage of this methodology is that it allows multiplexing or measurement of differently labeled samples together to determine their respective amounts. Such isotopic labels can be introduced at different stages during the experiment and then combined. However, the earlier the samples are combined the better the quantification accuracy due to decreased variability from sample preparation and MS analysis [85].

Metabolic labeling employs non-radioactive isotopes and is exemplified by SILAC (stable isotope labeling with amino acids in cell culture) [86]. This is achieved by replacing essential amino acids in the growth medium with their heavier counterparts (**Figure 10**). The most commonly used stable isotope containing amino acids for differentially labeling cells are arginine and lysine. When combined with trypsin digestion (cleavage C-terminal to Arg and Lys), these labels ensure labeling of every peptide, except the C-terminal peptide of the protein (**Figure 9**). During MS analysis, two isotope clusters can be observed for every peptide, forming a SILAC pair whose intensities can be directly compared to estimate differences in peptide intensities and thus relative protein levels. Up to three biological samples with three different SILAC labels are commonly measured in a

single MS experiment, although multiplexing with many SILAC labels has recently been reported [87]. SILAC requires growing cells in media containing only these stable isotopes labeled amino acids and is sometimes not feasible when working with cells sensitive to small changes to culture media or those that cannot be cultured (e.g. primary cells). The limitation with respect to primary cells in mice has been solved with the introduction of the SILAC mouse [88] that is generated by feeding the animals with a heavy lysine diet. Today the concept has been applied to other model organisms, including bacteria [89], yeast [75] nematodes [90] and flies[91].

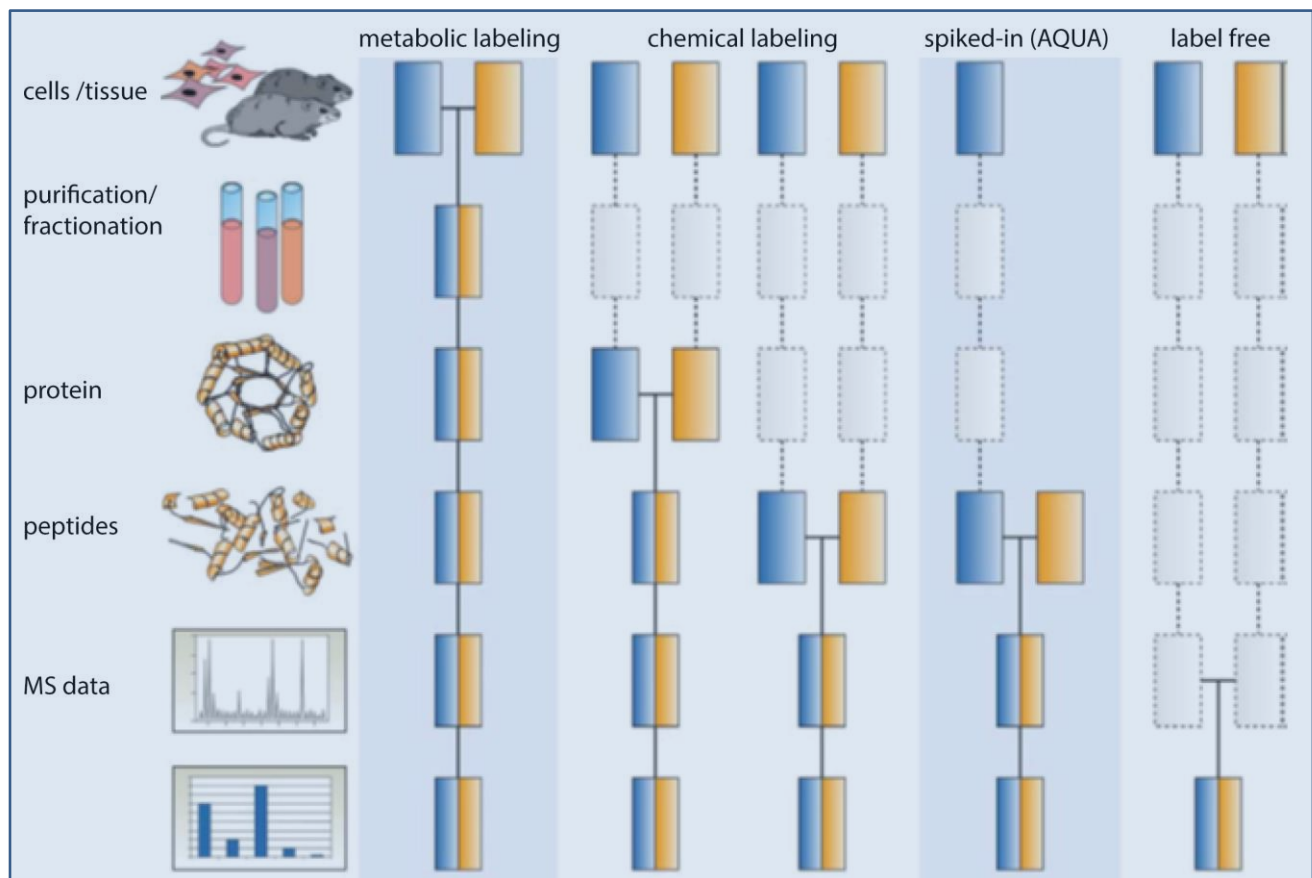


Figure 9: Labeling strategies and their impact on quantitative accuracy.

A schematic depiction of label-based and label-free workflows. Labeled samples represented by colored boxes are distinguished in the mass spectrometer while samples without a label (empty boxes) are not. Figure modified from [85].

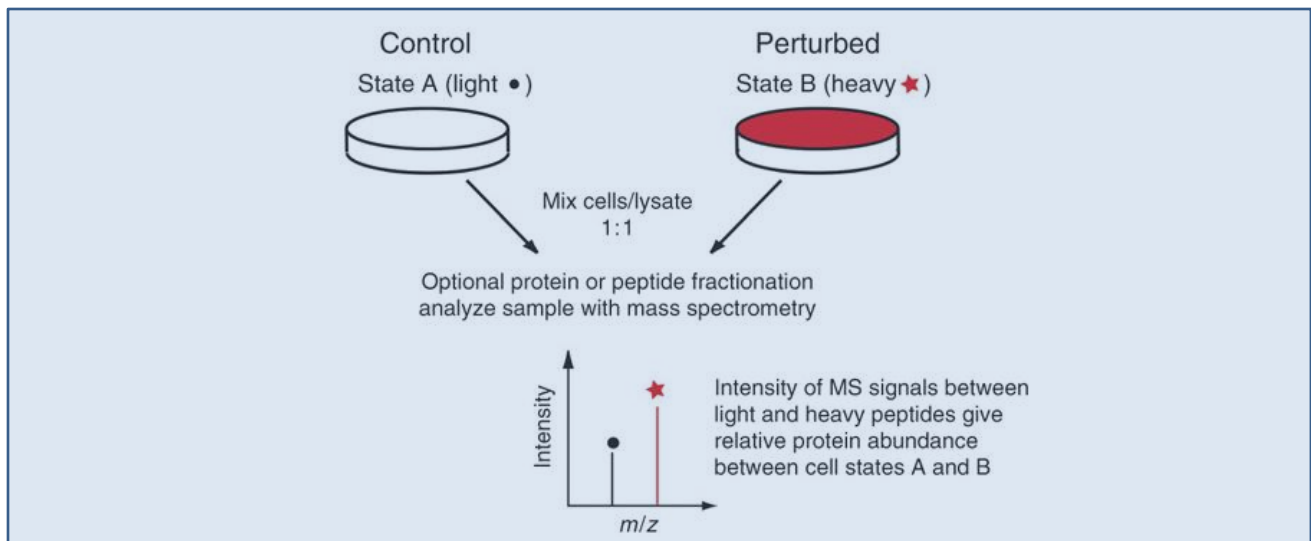


Figure 10: SILAC based relative quantification.

Cell populations are differentially labeled, mixed and analyzed together by MS-based proteomics to obtain SILAC peptide pairs, whose intensities are directly compared to assess differential protein expression between the two populations. Image modified from [92].

The limitation of SILAC to cells that can be cultured was further overcome by the development of the 'super-SILAC' method [93], in which several SILAC labeled cell lines are pooled and are spiked into any unlabeled sample including those in clinical studies. Article 2 in this thesis used the super-SILAC approach to segregate patients with two different types of lymphoma. SILAC is evolving with newer modes being developed; for example pulsed-SILAC [94] is applied to study protein turnover.

Chemical labeling is not as accurate as metabolic labeling, since the labels are introduced further downstream, either before or after digestion (**Figure 9**). Isotope-coded affinity tag (ICAT) [95] is an example of the former, where cysteine residues are covalently modified with a biotinylated label and affinity purified [109]. Dimethyl labeling [96] which involves derivatization of amino groups with light or heavy versions of formaldehyde has also been successfully employed. Other chemical labels include 'isobaric tags for relative and absolute quantification' (iTRAQ) [97] and the 'tandem mass tag' (TMT) [98]. These methods use differing low mass reporter ion masses generated by fragmentation for quantification and allow multiplexing up to ten samples. They pose challenges

for quantification due to unavoidable side chain reactions and co-fragmentation of co-eluting peptides.

The labeling strategies described so far are applicable to **relative quantification**. Some of them can additionally be applied for **absolute quantification** as is the case with 'absolute SILAC' [99] where SILAC-labeled recombinant proteins are used as internal standards and are mixed into lysates in exactly determined amounts. The Protein Epitope Signature Tags (PrEST) method [100], similarly employs accurate amounts of heavy PrESTs to measure the copy number per cell. Another method of labeling for absolute quantification is AQUA (Accurate QUAntification) [101] which employs labeled synthetic peptides (**Figure 9**). AQUA has been the most widely applied labeling strategy for absolute quantification but is not feasible when applied to many peptides and conditions due to increased costs. Furthermore, it does not account for variation in digestion efficiency and loss of AQUA peptides during storage.

Label-free quantification:

The methods for absolute quantification outlined above can be very accurate, but are limited in their throughput since they require a spike-in of every protein that needs quantification. Label-free quantitation for comparison of relative amounts between cell types or for estimation of absolute numbers is very attractive, since it is simple and applicable to any samples. It also has the added bonus of being economical and readily usable for clinical samples. However, since it compares separate measurements it is less accurate than label-based quantification and may therefore require more replicates and more MS measurement time.

Using computational analysis, it is possible to indirectly infer protein amounts from MS peptide intensities. For instance the number of peptides identifying a protein, normalized for protein length, is related to the protein amount. The exponentially modified protein abundance index (emPAI) [102], does this using the following formula:

$$emPAI = 10^{PAI} - 1$$

where PAI or protein abundance index is the ratio of the number of observed peptides divided by the number of observable peptides per protein. emPAI is directly proportional to the absolute protein amount and has been applied in many MS studies for a rough estimation of protein abundance.

A more recent method called ‘intensity-based absolute quantification’ (iBAQ) [103] calculates a ratio of the sum of intensities of all identified peptides of a protein and the number of theoretically observable peptides and log transforms this value. To calibrate this quantification, a non-labeled standard of accurately quantified proteins is spiked into the sample before sample preparation.

Label-free approaches have recently been applied to relative quantification with increasing success [104]. Differentially treated unlabeled samples are prepared and measured separately. The higher variability arising due to separate processing can be minimized and by measuring samples consecutively and with more replicates [105]. In addition the sophisticated intensity-based label-free quantification (LFQ) provided in the MaxQuant software platform normalizes intensities, aligns runs and computes label-free intensity at the protein level. This method was employed in articles 3 and 4 of this thesis [19].

1.1.4 Computational analysis in MaxQuant

In the bottom up proteomics strategy, precursor and fragment spectra and their intensities are employed for identification and quantification of peptides and proteins. This requires elaborate computational workflows that efficiently perform this process, relying on information which is either known or predicted from DNA sequencing data of the organism analyzed [25]. One such algorithm was developed in our laboratory and is called MaxQuant [106]. It reconstructs MS peaks as three dimensional (time, m/z and abundance) objects and specifically takes advantage of high resolution MS. It determines peptide masses to ppm levels with linear and non-linear mass recalibration and integration of multiple mass measurements over a liquid chromatographic peak, thus decreasing false positive identifications [107]. Most importantly, it introduced the concept of individualized mass accuracy depending on the signal and mass measurement statistics of each peptide.

In MaxQuant, the measured masses from both MS and MS/MS scans are converted to peak lists and submitted to the database search engine Andromeda [62] to perform matching. Andromeda extracts the most intense N ion peaks per 100 Da in fragment spectra and matches the peak list with theoretical spectra and its precursor mass and in addition calculates chances of the matches

arising by chance. This procedure is repeated for each peptide from the protein sequence database and results in a probability score indicative of the extent the match arising by chance.

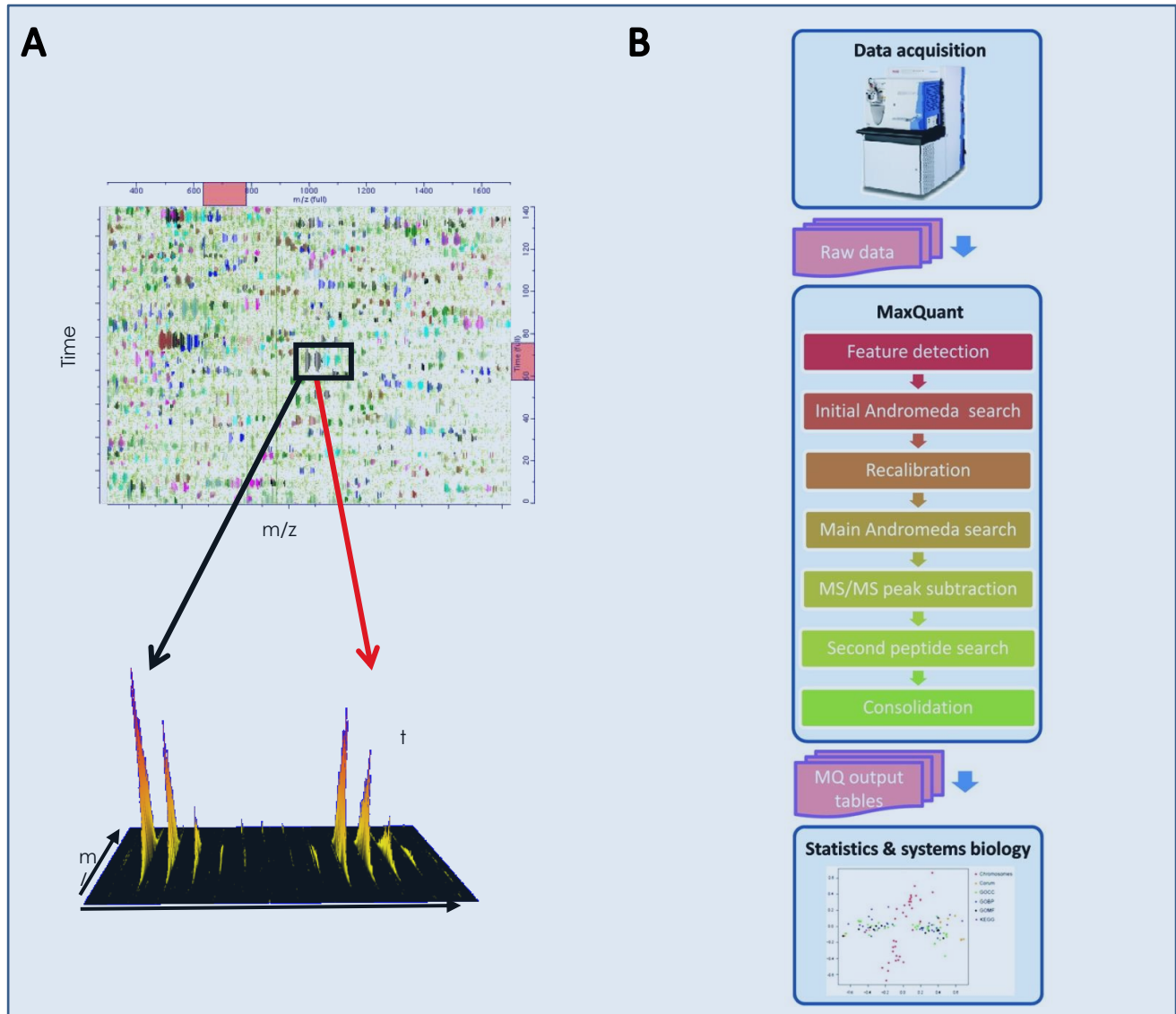


Figure 11: The MaxQuant algorithm.

A: The measured masses from both precursor and fragment scans are matched to *in silico* database of entries by the search engine Andromeda to identify peptides. Quantification is performed using label-based (SILAC in this figure) or label-free methods to report protein expression changes. **B:** The data acquired is subjects to many steps of analysis in the MaxQuant environment to obtain protein identification and quantification, which is then analyzed in Perseus. Figure adapted from [106]

Due to non-ideal and random matching between theoretical and experimental spectra, database searching programs inevitably produce both correct and incorrect identifications. Therefore afterwards, a score cutoff is applied in MaxQuant to guarantee sufficient evidence for the assignment of peptide sequences using a target decoy strategy [108] with a database containing both the regular and the reversed protein sequences of the organism analyzed. By assuming that incorrect identification of peptides are equally likely in the target and decoy database, it is possible to obtain an estimate of the number of false positives by doubling the number of hits found in the decoy portion of the database, which are incorrect identifications by definition. In the second step, the list of identified peptides is filtered according to user-specified criteria and the False Discovery Rate (FDR) is estimated from the number of decoy hits. The same basic principle is applied during the reassembly step from identified peptide sequences back to proteins. The software has additional features that make it an ideal solution for the quantitative analysis of a large number of raw data, such as methods for computing statistics at peptide and protein levels and statistically robust methods for quantifying proteins.

MaxQuant is also particularly suited to the identification and quantitation of PTMs. To identify these during MS/MS database searches, specific mass shifts are consecutively considered on the residues, on which the modification of interest may be localized. MaxQuant in addition to FDR thresholding, uses a special score cutoff (calibrated based on a synthetic phospho-peptide library [109]) and probability based scoring method to identify phospho-peptides and to pinpoint the site of modification with high confidence [73]. A part of this thesis deals with application of computational approaches to label-free phospho-proteomics (Article 3).

Once identifications and quantifications of peptides and proteins have been performed, the resulting datasets can be subjected to computational analysis. Specialized software environments, such as the Perseus framework, which is part of the MaxQuant environment, offer a range of algorithms to perform this task and thus extract informative results.

1.2 Clinical applications of MS-based Proteomics

MS-based proteomics has revolutionized cell biology and signal transduction research. However, one of the ultimate goals of proteomics is the application of the technology to a clinical setting. The protein technologies currently employed in the clinic are limited to monitoring a few proteins in plasma, serum or urine in low throughput assays or targeted ones such as ELISA. The proteins measured in these assays make up a miniscule percentage of proteins that could be studied [110, 111]. Proteomics is increasing being applied to discovery of biomarkers of which a few have been validated to some extent (Table 1). There are very few studies that have gone further to assess clinical utility and implications for clinical practice especially cost-effectiveness in the clinic [112].

Table 1: PubMed items for concepts in different stages of the roadmap among items that are retrieved by “Proteomic*” ($n=31\,686$ as of August 25, 2010) taken from [112]

Analytical tools	Mass spectrometry	14,097
	Electrophoresis	10,125
Clinically oriented Discovery	Clinical	3,815
	Discovery	2,791
	Biomarker	4,684
	Biomarker discovery	1,224
	Potential biomarker	1,758
Validation	Validation	1,525
	Independent validation	148
	External validation	22
Clinical application	Clinical practice	246
	Clinical utility	171
	Clinical outcomes	169
	Cost-effectiveness	41
	Commercialization	4
	Reclassification	2
Post-clinical application appraisal	Audit	4

MS-based proteomics, with its ability to measure and quantify thousands of proteins in a few hours, holds great promise to identify and monitor new biomarkers in body fluids, cells and tissues. This could be used in diagnosis at a very early stage of disease or for stratification of patients for specific treatment [113, 114]. The prospect of quantitative analysis of thousands of

proteins simultaneously and the inherent sensitivity of mass spectrometry holds great promise for application in clinical diagnosis [115].

One of the most attractive and probably closest applications of MS-based proteomics is in patient classification in cancer[116]. Transcript-based profiling technology has been applied to segregation of cancer subtypes based on their gene expression signatures [117, 118]. Interpretation of such gene based signatures with respect to the biology of the disease is challenging especially because they do not provide information on the extent to which changes on the transcript level are transferred to the next level of the gene expression program, i.e., proteins. It also ignores the effects of post-translational modifications. Because of the ability of quantitative proteomics to study expression both at the level of proteins and PTMs, it can be employed to not only segregate patients based on their protein profiles but also to determine the activity of signaling pathways directly. This may help in estimating the risk of progressing to other stages of cancer, especially metastasis.

Article 2 in this thesis employs the super-SILAC approach to show that proteomic methods can accurately segregate two diffuse large B-cell lymphoma (DLBCL) subtypes; germinal-center B-cell-like (GCB) and the activated B-cell subgroup (ABC) the latter of which has a signature characteristic of tumor cells activated via their B-cell receptor.

1.2 Signal transduction in the eukaryotic cell

The function and working of eukaryotic cells is controlled by a variety of elaborate and intertwined mechanisms. These mechanisms may have an effect intracellularly or in the extracellular environment. Intracellular mechanisms that ensure precise activity of cell components are governed mainly by processes that control protein expression at several different stages. They comprise regulation at levels of transcription of genes and translation of mRNA into proteins. An additional layer of complexity in the regulation circuitry is mediated by several PTMs. These control mechanisms work in tandem to allow fine-tuning of regulatory networks.

In addition, cells cannot exist by themselves, especially in the context of multicellular organisms. Cell to cell communication can be mediated by several types of signaling molecules that can act locally (exocrine or paracrine signaling), in neighboring cells or are transported to tissues in another part of the organism (endocrine signaling). The released signaling molecule or ligand can elicit a response in only those cells that have the corresponding receptor. Signaling molecules can be chemicals, peptides, soluble proteins and proteins bound on cell surfaces or the extracellular matrix (ECM). These signaling molecules can either diffuse through the cell membrane and bind intracellular receptors or bind to cell surface receptors. The binding of the ligand to its receptor induces a conformational change and activates or inhibits the proteins in the cytoplasm thus converting the extracellular signal into a cascaded cellular response, a process that is termed signal transduction. Signaling events also involve intricate networks, which encompass feedback loops, crosstalk with signals that regulate and can be regulated by other cellular regulatory mechanisms, such as transcriptional networks. These intricate networks are responsible for key processes such as growth, development, differentiation, apoptosis and repair while deregulated signal transduction is a well-established cause of diseases such as cancer, diabetes, obesity, heart failure etc. Consequently, studying the nature and mechanisms of signaling events is a large and crucial part of biological and medical research.

Although the nature of signals and their processing can be very diverse, they share general characteristics. Signals need to be specific to lead to a prompt cellular response and quenched once the required response has been elicited. Signaling cascades are therefore usually controlled

by highly dynamic post-translational modifications (PTMs) that are kinetically fast and mostly reversible. PTM of a protein is accomplished by modification of a chemical group (e.g.: phosphorylation, acetylation) or the addition of another protein (e.g.: ubiquitylation, sumoylation) on one or more amino acids. An intricate interplay of these modifications regulates signaling processes by altering basic protein properties such as structure, stability, localization, activity and interaction with other proteins.

1.3.1 Phosphorylation as a PTM: The role of kinases and phosphatases

Phosphorylation was the first identified and extensively studied reversible PTM and is a fast and transient mode of regulation of protein function [119]. The opposing mediators of phosphorylation are protein kinases (PK) which catalyze the addition of the phosphorylation moiety and protein phosphatases (PP) which catalyze the removal of the phosphate group [120]. The human genome encodes 518 PKs [121] and approximately 150 PPs [122, 123], which together make up about 3.5% of the proteome. Interestingly, despite generally being low-level, regulatory proteins, many kinases are nevertheless in the top 25% of the proteome by expression value [12].

In eukaryotes phosphorylation predominantly occurs on side chains of serine (Ser/S), threonine (Thr/T) and tyrosine (Tyr/Y) residues, and very rarely on histidine (His/H), arginine (Arg/R) or lysine (Lys/K) side-chains [124]. Accordingly, in humans, of the 513 PKs, 90 are Tyr kinases while the remaining two thirds are serine threonine kinases (STKs) [121]. A given STK may simultaneously or sequentially phosphorylate multiple Ser and Thr residues of a target. Similarly, several Tyr may be phosphorylated by a Tyr kinase, as is the case, for instance, on the activation loop of the insulin receptor [125]. In addition, a given protein can be modified by more than one type of PTM allowing for crosstalk with other signals [126]. The phosphorylation mediated ubiquitylation in many Skp, Cullin, F-box containing complex (SCF) E3 ligase substrates is an example of this. Approximately one of every three proteins has been estimated to be phosphorylated at some point in its life cycle [127]. It has recently been shown that many substrate proteins of cyclin dependent kinases (CDKs) are completely phosphorylated at particular stages of the cell cycle [128]. Researchers attribute this high stoichiometry of phosphorylation on these proteins as a means to ensure effective inactivation of an entire substrate population. Most other proteins

show low phosphorylation stoichiometries, which is thought to be sufficient for ensuring sufficient activity. Phosphorylation on Tyr residues is the most common mode of activation of transmembrane receptors that control cellular proliferation and differentiation as exemplified by receptor tyrosine kinases (RTKs) [129]. Phosphorylation of the Tyr residue in these RTKs induces a conformational change on the receptor 'turning on' enzymatic activity. The structural changes associated with the addition of the phosphate also generate docking sites for non-covalent interaction with other proteins via specific protein domains (such as the SH2 domain and the PTB domain that recognize phosphorylated Tyr) on interacting proteins [130, 131]. In comparison to Tyr, phosphorylation on Ser and Thr residues is ubiquitous. Pathways predominantly regulated by STK phosphorylation include cell-cycle progression [132] and the DNA damage response.

1.3.2 Phosphorylation and its relevance in cancer

Signaling networks are predominantly controlled by PTMs on member proteins to transduce signals in a precise and orderly sequence to amplify signals from the extracellular environment. Deregulation of these phosphorylation-driven processes by activating or inactivating mutations causes many pathological disorders including endocrine disorders, immunodeficiencies, cardiovascular diseases and especially cancer. More than 400 kinases are implicated in disease of which at least 180 have roles in cancer[133]. One such example is the Philadelphia chromosome [134], which generates the breakpoint cluster region (BCR)-Abelson (ABL) fusion protein common in many leukemias. This results in an active BCR-ABL tyrosine kinase [135] that drives proliferation and inhibits apoptosis of the mutated white blood cells. The mutation results in constitutively phosphorylated ABL kinase, which is targeted by a tyrosine kinase inhibitor Imatinib [136]. Imatinib became the first kinase inhibitor to be approved as a drug and heralded a new era of targeting kinases with small molecules to combat diseases. To date there are 18 kinase-targeting therapeutics, including targeting receptor kinases with antibodies, that have been approved for use by the FDA and several hundred are estimated to be in development and in clinical trials [137-139].

Similar to the BCR-ABL tyrosine kinase, several diseases are a result of mutations in particular protein kinases and phosphatases and are therefore considered druggable targets. In addition, other proteins and PTMs are also gaining prominence.

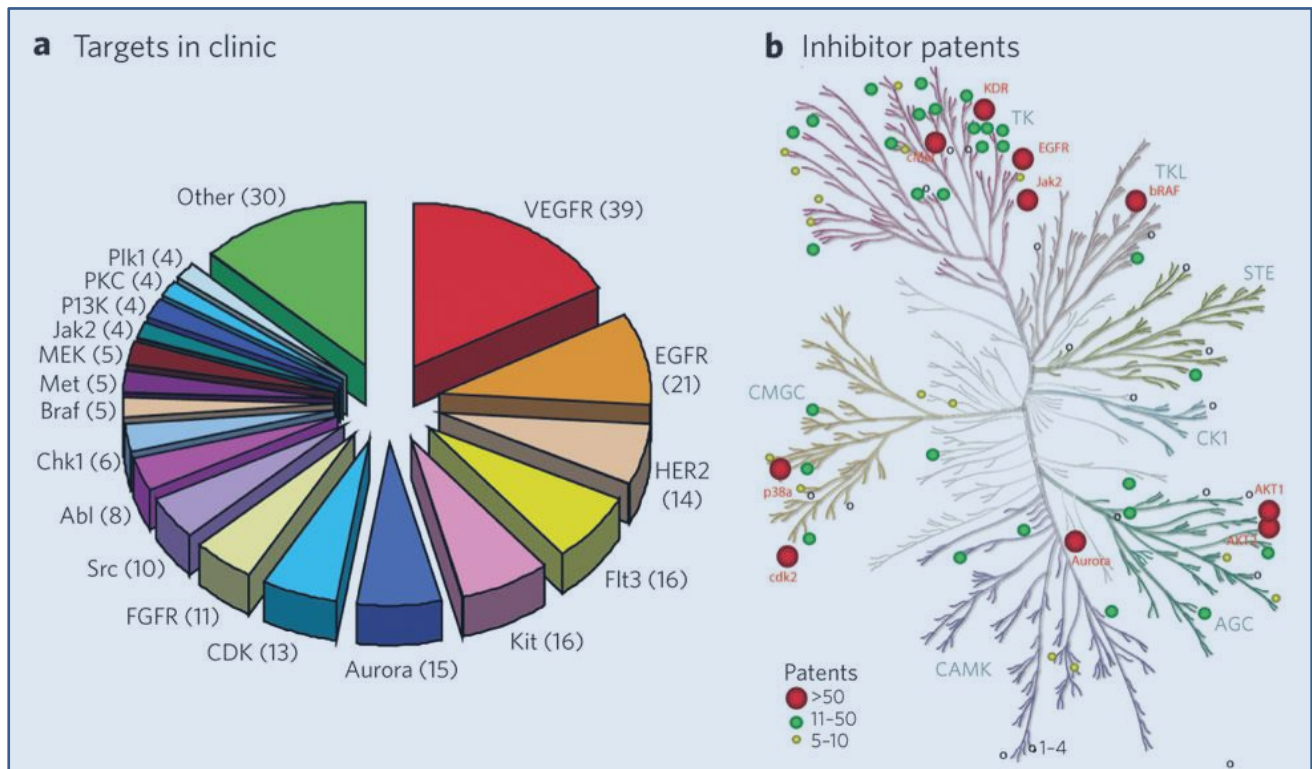


Figure 12: Drugs and their targeted kinases.

(a) Kinases currently targeted in clinical trials. (b) Kinases for which inhibitors have been claimed as patent literature. Number of patents per target is reflected by the radius of the dot used. This figure is taken from Fedorov et al [140].

1.3.3 The role of mass spectrometry to study phosphorylation

Traditionally, signaling processes are studied as individual discrete pathways that amplify a signal via modification of key sites on member proteins. However, as our understanding of signal transduction increases it become clear that our current picture of signaling as a directional cascade is overly simplistic. Researchers have realized that signaling constitutes a complex, dynamic and cooperative network [141][15]. The complexity of post-translational modifications of signaling networks and their downstream effects therefore requires that new technologies and tools be developed to handle such multifaceted data. Since addition of phospho moieties in these complex networks is a principal mode of regulation, the phosphorylation sites should be identified

and quantified in an unbiased and global manner. In the past decade, quantitative mass spectrometry (MS) has proven itself as an ideal platform for such studies, due to its inherent properties of being precise yet generic. MS-based mapping and quantification of PTMs has therefore revolutionized signaling research and has become the method of choice in large scale phosphorylation studies as well as studies of other PTMs. Phospho-proteomics has been employed to study a variety of systems and perturbations including epidermal growth factor (EGF) signaling, cell cycle, kinase profiling of drugs, immune system response, amongst many others [73, 128, 142, 143]. MS will also remain the key technology in the future as there are only a limited number of phospho-related reagents and tools available. The most comprehensive phosphorylation databases such as PhosphoSitePlus [144], PhosphoELM [145] and Phosida [146] altogether comprise more than a 150,000 phosphorylation sites and for some of these sites their regulation in response to perturbations, almost all of which have been identified from MS-based studies.

1.3 The TGF- β signaling pathway

Transforming growth factor beta (TGF- β) signaling drives several key cellular processes like proliferation, differentiation, migration and apoptosis and is also implicated in diseases like cancer and auto-immune disorders [147, 148]. Signal transduction begins with TGF- β 1 binding to its transmembrane receptor T β R-I or ALK5, which then dimerizes with T β R-II on the cell membrane. Upon dimerization, TGF β R-I is phosphorylated in the juxtamembrane region, which then transduces the signal into the cell via binding and phosphorylation of Smad2/3 assisted by SARA [149]. The activated Smad dimerizes with a regulatory Smad (Co-Smad) [150], which then shuttles to the nucleus, binds to target genes such as ID1 and SKI to induce or repress their transcription [151]. Transcriptional activation generally requires co-operating DNA binding factors like activating protein (AP)-1 and activating transcription factor (ATF)-2 as well as co-repressors such as TG-interacting factor (TGIF), Ski and SnoN . In recent years there has been an increasing interest in TGF β R-induced non-SMAD signaling, which is typically mediated by p38, Jun N-terminal kinases (JNKs) and the extracellular signal regulated (ERK) mitogen activated kinases (MAPKs) [152, 153].

The cellular responses to this multifunctional ligand are diverse and can even be opposed to each other, depending on the cell type and the conditions [154]. For example, TGF- β can promote cell growth but also have anti-proliferative effects, and it can contribute to maintain stem cell pluripotency but also to differentiation. Further, TGF- β suppresses pre-malignant cells by inhibiting cell proliferation, but it does not do so in metastatic ones, which nevertheless remain responsive to TGF- β induced migration and invasion [155, 156]. A key mode of action of TGF- β in cancer progression is the induction of epithelial to mesenchymal transition (EMT), a process wherein epithelial cells acquire mesenchymal characteristics [157]. EMT is an indispensable process in normal tissue development and organogenesis, as well in tissue remodeling and wound healing. However, inappropriate reactivation of EMT crucially contributes to the development of a variety of human pathologies, particularly those associated with tissue fibrosis and cancer cell invasion and metastasis, for instance in breast cancer [158, 159].

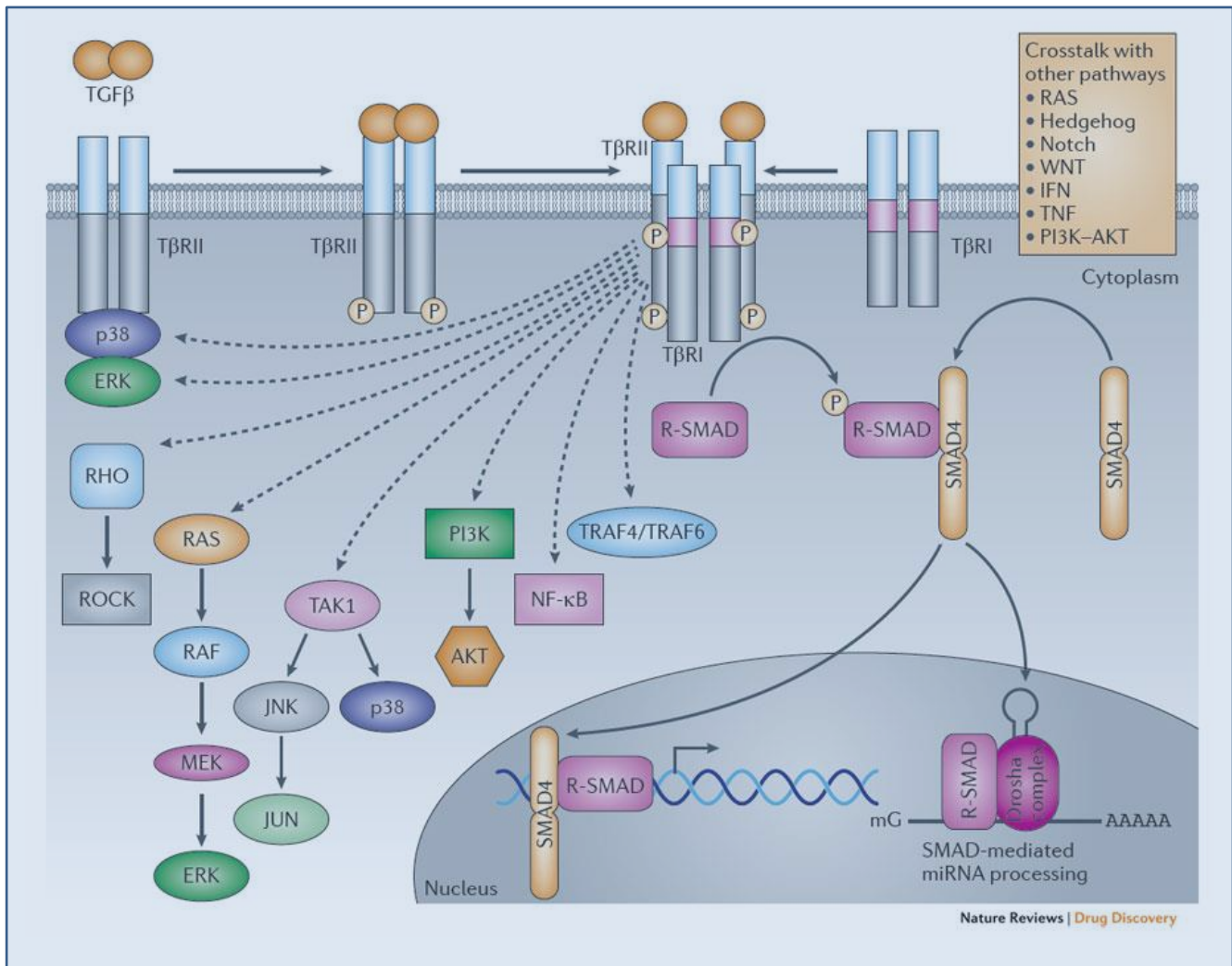


Figure 13: The scheme of TGF- β signal transduction.

Signal transduction begins when TGF- β binds to the receptor leading to dimerization and phosphorylation of the receptor. The activated receptor allows the binding of a regulatory SMAD(R-SMAD) which is phosphorylated and binds to SMAD4. The SMAD complex translocates into the nucleus, binds to target genes and turns on gene transcription. Alternatively phosphorylation of the receptor also activates non-SMAD signaling pathways including RAS-RAF, JNK, P38, PI3K signaling pathways are also triggered. The depiction is taken from [160].

2. Feasibility of large-scale phospho-proteomics with higher energy collisional dissociation fragmentation (Article 1).

Phosphorylation is one of the key PTMs on proteins and is a prime effector of cell signaling. Among a wide range of analytical approaches, the combination of high resolution MS scans in an Orbitrap analyzer, fragmentation by collision induced dissociation (CID) and acquisition of fragment scans with low resolution in a linear ion trap has proven to be particularly successful. Measurement of fragment scans in the linear ion trap is attractive because of its high sensitivity, fast scan speeds and parallel operation with the Orbitrap or FT instrument.

However, when analyzing phospho-peptides, CID in the ion trap results in significant neutral loss for phospho-serine (pS) and phospho-threonine (pT) containing peptides, and needs multiple activation steps to efficiently fragment them. Furthermore, due to the “one-third rule cutoff”, informative low molecular weight reporter ions often cannot be measured, as is the case for the phospho-tyrosine (pY) immonium-ion.

As described in more detail in the introduction, Higher energy collisional dissociation (HCD) is an alternative beam type fragmentation technique that can be used on the LTQ-Orbitrap. In HCD, peptide ions are fragmented in a collision cell at the far side of the C-trap and analyzed at high resolution and mass accuracy in the Orbitrap mass analyzer. This strategy not only overcomes the problem of the low mass cutoff of ion trap fragmentation but also results in high accuracy of both precursor and fragments masses.

Here I investigated the feasibility of routine large-scale phospho-proteomics by HCD fragmentation on the LTQ-Orbitrap Velos. We analyzed key parameters such as sensitivity and fill times, cycle times for MS/MS experiments, identification success rates, and depth of phospho-proteome coverage which, unexpectedly, in comparison with a high-low strategy, showed similar or superior performance. As a result of this project, HCD is routinely employed to analyze phospho-peptides.

This article was published in the year 2010 in the Journal of Proteome Research [1].

Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation

Nagarjuna Nagaraj,^{*,†} Rochelle C. J. D'Souza,^{*,†} Juergen Cox,[†] Jesper V. Olsen,[‡] and Matthias Mann^{*,†}

Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, 82152 Martinsried, Germany, and Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark

Received June 21, 2010

Abstract: Mass spectrometry (MS)-based proteomics now enables the analysis of thousands of phosphorylation sites in single projects. Among a wide range of analytical approaches, the combination of high resolution MS scans in an Orbitrap analyzer with low resolution MS/MS scans in a linear ion trap has proven to be particularly successful ("high-low" strategy). Here we investigate if the improved sensitivity of higher energy collisional dissociation (HCD) on an LTQ-Orbitrap Velos instrument allows a "high-high" strategy. A high resolution MS scan was followed by up to 10 HCD MS/MS scans, and we achieved cycle times of about 3 s making the method compatible with chromatographic time scales. Fragment mass accuracy increased about 50-fold compared to the "high-low" strategy. Unexpectedly, the HCD approach mapped up to 16 000 total phosphorylation sites in one day's measuring time — the same or better than the standard high-low strategy. Reducing the target values from a standard of 30 000 to 5000 ions did not severely affect identification rates but did decrease identification and localization scores for phosphorylation sites. We conclude that HCD in the new configuration is now a viable method for large-scale phosphoproteome analysis alongside collisional induced dissociation, (CID) and electron capture/transfer dissociation (ECD/ETD).

Keywords: phosphoproteomics • HCD • mass spectrometry LTQ-Orbitrap Velos • mass accuracy

Introduction

Global mapping and localization of post-translational modifications (PTMs) such as phosphorylation are crucial for understanding the activity of the cell. Phosphorylation acts as a molecular switch in various signaling pathways and plays a pivotal role in many biological processes.¹ Mass spectrometry (MS)-based proteomics has emerged as a powerful technique for studying PTMs.^{2–4} Among many different versions of MS-

based phosphoproteomics, hybrid instruments with quadrupole and time-of-flight analyzers (quadrupole - TOF) or with two different types of ion traps have gained popularity in the past decade.⁵ In particular, the combination of high mass accuracy for the precursor ion and low mass accuracy for the fragment ions from linear ion trap-Orbitrap instruments (LTQ-Orbitrap) is a widely applied instrumental configuration. Employing this "high-low" strategy in large-scale phosphoproteomic approaches has led to the identification and quantification of several thousand phosphosites in single projects.^{6–8} However, analyzing phosphopeptides by collision induced dissociation (CID) in the ion trap (resonant excitation mediated collision) results in significant neutral loss for phosphoserine (pS) and phosphothreonine (pT) containing peptides, and this can require multiple activation steps to efficiently fragment them.^{9,10} Furthermore, in ion trap fragmentation the "one-third rule" (loss of low mass ions depending on the fragmentation *q* value)¹¹ precludes the analysis of low molecular weight reporter ions that are very informative, for example, in the case of phosphotyrosine (pY) ions.^{12,13} A different class of fragmentation techniques, electron capture dissociation (ECD)¹⁴ or electron transfer dissociation (ETD),¹⁵ complements CID, particularly for labile phosphopeptides.¹⁶

Higher energy collisional dissociation (HCD) is an additional fragmentation technique that can be used on the LTQ-Orbitrap.¹⁷ HCD is performed by injecting peptide ions from the ion trap into a collision cell at the far side of the C-trap. Fragment ions are transferred back to the C-trap and analyzed at high resolution and mass accuracy in the Orbitrap analyzer. HCD fragmentation is similar to the fragmentation in triple-quadrupole or quadrupole - TOF instruments, and it overcomes the problem of low mass cutoff of ion trap fragmentation. In addition, high accuracy at both the precursor mass and fragment levels (a "high-high" strategy) should be very desirable as it would dramatically improve the quality of fragmentation spectra. For example, charge states can easily be distinguished in high resolution spectra, and fragments are less likely to be assigned to the wrong peptide sequence. High-high strategies based on HCD fragmentation have been adopted for proteomic and phosphoproteomics studies, but they have previously been restricted to specialized applications because of lower sensitivity compared to CID, resulting in extended injection times and a slower duty cycle.^{18–20} The newly developed LTQ Orbitrap Velos features an S-lens, improving

* To whom correspondence should be addressed. E-mail: mmann@biochem.mpg.de.

[†] Both authors contributed equally.

[‡] Max-Planck Institute for Biochemistry.

[§] University of Copenhagen.

HCD-Based Large-Scale Phosphoproteomics

ion current into the instrument by at least 10-fold, as well as a more efficient HCD cell.²¹ This improved sensitivity could make routine HCD measurements of phosphopeptides feasible within the short MS and MS/MS duty cycles required in large-scale phosphoproteomics.

Here we investigate the feasibility of routine large-scale phosphoproteomics by HCD fragmentation on a LTQ Orbitrap Velos instrument. We analyze key parameters such as sensitivity and fill times, cycle times for MS/MS experiments, identification success rates, and depth of phosphoproteome coverage that are of prime importance for designing a large-scale phosphoproteome study. Comparison with a high–low strategy showed similar or superior performance.

Experimental Procedures

Cell Culture and Peptide Preparation. HeLa S3 cells (ATCC) were cultured in roller bottles in RPMI 1640 (Gibco) supplemented with 10% fetal bovine serum (Invitrogen) and 1% penicillin/streptomycin (Invitrogen). On reaching sufficient confluence in suspension, the cells were centrifuged at 1000 rpm. The HeLa cell pellet was lysed in a buffer of 4% SDS and 100 mM DTT in 100 mM tris-HCl pH 7.5. The lysate was processed by the FASP method.²² Briefly, the lysate was sonicated and heated in the SDS buffer to ensure complete homogenization and denaturation. The protein concentration was measured and lysate was loaded onto 15 mL Amicon filter units (10 kDa MWCO) (Millipore) and washed with Tris buffer (UA buffer) containing 8 M urea to remove SDS.²³ Proteins were alkylated with 50 mM iodoacetamide in urea buffer and incubated for 20 min followed by removal of excess iodoacetamide by multiple washes with urea buffer. After reduction and alkylation, the proteins were equilibrated in 20 mM ammonium bicarbonate and digested with trypsin (Promega) in a protein to enzyme ratio of 100:1 at 37 °C overnight. After digestion, the peptides were eluted by centrifugation with an additional elution with 50 μ L of water. Elution with water avoids the desalting step for further processing of the peptides. Two large-scale phosphoproteomics experiments were performed for HCD. Peptide yields were 4 mg for the first large-scale HCD experiment and 2.5 mg for the second experiment.

Fractionation of Peptides by Strong Cation Exchange (SCX) Chromatography. Peptides were concentrated into a volume of 7 mL, and the pH was adjusted to 2.7 and then adjusted to 10 mL with 100% ACN. The peptides were separated by strong cationic exchange chromatography (SCX) as described.^{6,7} The peptide mixture was loaded onto a cation exchanger column equilibrated with 30% ACN containing 5 mM KH_2PO_4 . The flow-through which predominantly contains multiply phosphorylated peptides was collected. The peptides bound to the column were eluted in an increasing salt gradient with buffer containing 5 mM KH_2PO_4 and 150 mM KCl. The fractions generated by SCX were then pooled to seven fractions based on UV absorbance. The flow through from the SCX column was subjected to three TiO_2 incubations (see below), and thus each sample resulted in 10 LC-MS/MS runs.

Enrichment of Phosphopeptide by TiO_2 Beads. The flow through and the seven SCX fractions were subjected to TiO_2 enrichment with three consecutive incubations for the flow through and one incubation each for the remaining seven fractions as described. First, the UV absorbance of the fractions was measured, and they were incubated with TiO_2 (MZ-Analysentechnik, Germany) with a peptide to bead ratio of 1:2 to 1:8.²⁴ Before mixing with the fractions, the TiO_2 beads were

technical notes

resuspended in 30 mg/mL solution of dihydrobenzoic acid (Sigma) to prevent nonspecific binding. Next the phosphopeptide bound beads were washed with 30% ACN and 3% TFA twice followed by two washes with 75% ACN and 0.3% TFA. The phosphopeptides were then eluted under basic conditions using 25% ammonium hydroxide and ACN. Finally, the eluted phosphopeptides were loaded on C_{18} StageTips.²⁵

Reverse Phase Chromatography and Mass Spectrometry. Peptides were separated in a 15 cm column (75 μ m inner diameter) packed in-house with 3 μ m C_{18} beads (Reprosil-AQ Pur, Dr. Maisch) on a Proxeon EASY-nLC system (Proxeon Biosystems, Odense, Denmark) using a binary gradient provided by buffer A (0.5% acetic acid) and buffer B (0.5% acetic acid and 80% ACN). The peptides (4 μ L) were loaded directly without any trapping column with buffer A at a flow rate of 500 nL/min. Elution was carried out at a flow rate of 250 nL/min, with a linear gradient from 10% to 35% buffer B in 95 min followed by 50% B for 15 min. At the end of the gradient, the column was washed with 90% B and equilibrated with 5% B for 10 min. The LC system was directly coupled in-line with a LTQ-Orbitrap Velos instrument (Thermo Fisher Scientific) via the Proxeon Biosystems nanoelectrospray source. The source was operated at 2.1–2.25 kV, with no sheath gas flow, with the ion transfer tube at 200 °C.

The mass spectrometer was programmed to acquire in a data-dependent mode. For the high–high strategy, survey scans were acquired in the Orbitrap mass analyzer with resolution 30 000 at m/z 400 with lock mass option enabled for the 445.120025 ion.²⁶ However, the target lock mass abundance was set to 0% instead of 5–10% in order to save the injection time for the lock mass. For the full scans, 1E6 ions were accumulated within a maximum injection time of 250 ms in the C trap and detected in the Orbitrap analyzer. The 10 most intense ions with charge states ≥ 2 were sequentially isolated (signal threshold of 10 000) to a target value of 3E4 or 4E4 with a maximum injection time of 150 ms and fragmented by HCD in the collision cell (normalized collision energy of 40%) and detected in the Orbitrap analyzer at 7500 resolution. For the HCD based method, the activation time option in the Xcalibur file was set to 0.1 ms. For the high–low strategy, full scans were acquired in the Orbitrap analyzer at 60 000 resolution as parallel acquisition is enabled in the high–low mode. Up to the 20 most intense peaks with charge state ≥ 2 were selected for sequencing (signal threshold of 1000) to a target value of 5000 with a maximum injection time of 25 ms and fragmented in the ion trap by collisional induced dissociation with normalized collision energy of 35%, activation $q = 0.25$ and activation time of 10 ms. For CID “wideband activation” and “multistage activation” options were enabled with the appropriate neutral loss mass list for singly, doubly, and triply phosphorylated peptides. The fragmentation spectra were acquired in the ion trap at normal scan rate by lateral ejection and recorded by the dynode-multiplier system.

For all sequencing events, dynamic exclusion was enabled to minimize repeated sequencing. Peaks selected for fragmentation more than once within 30 s were excluded from selection (10 ppm window) for 60 s.

Data Processing and Analysis. The raw data acquired were processed with the MaxQuant software version 1.0.14.10 and processed as per the standard workflow. Since the HCD spectra were acquired in profile mode, deisotoping was performed similar to the survey MS scans to obtain singly charged peak lists. Peaks lists generated from the “quant” module were

technical notes

searched against IPI Human version 3.46 database using the Mascot search engine version 2.2 (Matrix Science, UK) with initial precursor mass tolerance of 7 ppm and fragment mass deviation of 0.02 Da for the “high–high” strategy. For the “high–low” strategy, these values were 7 ppm and 0.5 Th, respectively. The search included cysteine carbamidomethylation as a fixed modification and N-acetylation of protein, oxidation of methionine, and phosphorylation of STY as variable modifications. Up to two missed cleavages were allowed for protease digestion and peptide had to be fully tryptic. The “identify” module in MaxQuant was used to filter identifications at 1% false discovery rate (FDR) at three levels namely, site, peptide, and protein. As such, there is no fixed cutoff score threshold, but instead spectra are accepted until the 1% FDR rate is reached.²⁷ Only peptides with a minimum six amino acid length were considered for identification. The identified phosphosites are listed in Supplementary Table 1, Supporting Information, and annotated spectra can be visualized as described at the end of the document.

Results and Discussion

Although the phosphoproteome is very complex, phosphopeptides only constitute a small minority of all peptides after enzymatic degradation of mammalian cell lysates. We digested 20 mg of HeLa cell lysate using the FASP method²² and divided the resulting peptides into two parts to avoid overloading the SCX column. Both were separated by SCX into 10 fractions (three flow-through and seven SCX fractions). Although it has been shown that strong anion exchanged based fractionation leads to a similar depth of coverage as the SCX method,²⁸ the latter was applied for easier comparison to widely applied workflows for phosphoproteomics. Each fraction was enriched by TiO₂ beads using DHB,²⁹ providing a rich and diverse source of phosphopeptides. To minimize the variations owing to sample processing, the enriched fractions were pooled after StageTip purification and for each injection half of the samples were used. Peptides were loaded onto a reverse phase column and separated by 140 min chromatographic runs (100 min gradient time). For mass spectrometric analysis, the eluent of the column was electrosprayed into the LTQ Orbitrap Velos instrument. Making use of the more than 10-fold increased ion current and more efficient HCD of this instrument,²¹ we devised a top10 method consisting of an MS scan with 30 000 resolution at 400 *m/z* (0.5 s scan) in the Orbitrap analyzer, followed by up to 10 MS/MS scans at 7500 resolution (0.95 s) also in the Orbitrap analyzer. Total measurement time for phosphoproteome analysis of HeLa cells was one day. In this experiment, through analysis by MaxQuant²⁷ we identified 16 559 distinct phosphorylation sites with 99% confidence (1% FDR). Of all identified peptides, 76% were phosphorylated. As in previous analyses, we define the sites with at least 0.75 localization probability as Class I sites. For the 9668 Class I sites mean localization probability⁷ was higher than 0.997 (Supplementary Table 1, Supporting Information).

The experiment was then repeated with the other half of the sample but with the difference that CID and a top20 method was used. Because of parallel operation, MS resolution was set to 60 000 at 400 *m/z* (1 s scan). In this “high–low” mode and within one day of measuring time, we identified 11 893 sites (9016 Class I sites). For assessing overlap of identifications, we also repeated the HCD experiment with another preparation of HeLa cell lysate.

Proportions of pS/pT/pY were similar between HCD and CID (Supplementary Table 1, Supporting Information). However, HCD was somewhat more efficient at identifying doubly and more highly phosphorylated peptides than CID (singly/doubly/higher phosphorylated peptides were 51.4%:43.4%:5.2% for HCD, whereas they were 64.6%:32.6%:2.8% for CID). Apart from these three large-scale data sets, several other data sets were acquired with parameters described below.

General Features of Phosphopeptides in High–High Strategy.

For phosphopeptide detection in HCD mode, peptides are fragmented in the collision cell and the fragment ions are sent back to the C-trap from where they are injected into the Orbitrap analyzer for detection. In CID mode, in contrast, phosphopeptides are isolated and fragmented by “pseudo MS³”, meaning that both the precursor mass and the mass of potential dominant fragments due to neutral loss of one or more phospho groups are excited together. Furthermore, potential water loss fragments of the precursor are also excited. The resulting fragment spectrum is recorded in the ion linear ion trap at normal scan speed (33 400 Th/s).

From the large-scale experiments described above, we inspected dozens of identified spectra to obtain a qualitative view of the differences between the HCD and the CID spectra. Figure 1 shows representative examples visualizing these differences. The most striking distinction is the mass accuracy of the fragments, which generally deviated 0.1–0.3 Da from calculated values for ion trap measurements, whereas they deviated only a few ppm for the Orbitrap measurements. Figure 1A shows the HCD spectrum of the doubly phosphorylated peptide PlpSPSPpSAILER, which features the typical characteristics of extensive sequence coverage by y ions and a few low mass b ions (Figure 1A). The a₂b₂ ion pair is among the intense peaks. In the corresponding spectrum of this peptide for CID fragmentation acquired in the ion trap (Figure 1B), the a₂ ion is not recorded owing to the low mass cutoff, and the b₂ ion is less than 2% of the base peak intensity and has a mass deviation of 0.104 Da. The noise level in the spectrum is high compared to HCD spectra, which consequently has more well-defined peaks of low abundance. (Note that exact noise levels are difficult to compare because the Orbitrap data system employs noise filtering.) For peptides of typical length, we observe similar amino acid coverage in both CID and HCD. CID spectra have more b-type ions compared to the HCD. This is because b-ions are less stable and fragment further in HCD.^{21,30}

For the longer peptides, the high resolution of the HCD method allowed unambiguous assignment of charge states of multiply charged fragment ions. Because of the many possible fragmentation pathways, charge states and the low resolution, the CID spectra tend to become crowded and unambiguous peak assignment becomes difficult. This is illustrated by the singly phosphorylated peptide AAAAAALSGAGTPPAGGGAGG-GAGGGGpSPPGGWAVAR, the MS/MS spectra of which appear much cleaner in HCD than in CID. The dynamic range of the HCD spectrum was much higher than in CID as exemplified with one of the fragments annotated in the HCD but not the CID spectrum (Figure 1C,D). Together this led to better amino acid sequence coverage of 84% for HCD as compared to 53% in CID. Of the total ion current in the MS/MS spectra 78% corresponds to annotated peaks in HCD versus 44% in CID. As a consequence, the Mascot score is 115.8 for the HCD and 34.4 for the CID spectrum. Furthermore, the Mascot search engine does not take account of mass accuracies better than

HCD-Based Large-Scale Phosphoproteomics

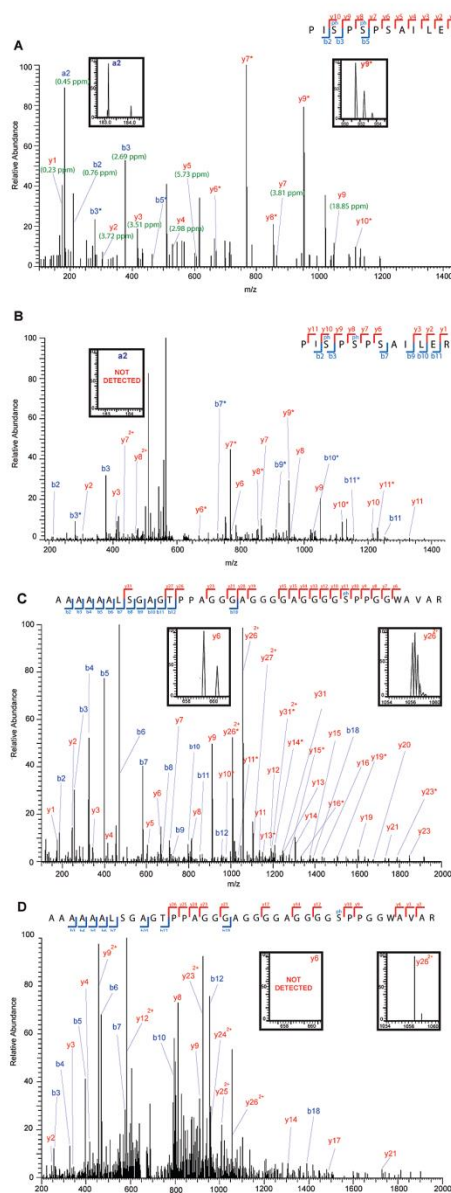


Figure 1. Comparison of CID and HCD fragment spectra. (A) HCD spectrum of the double phosphorylated peptide "PiSphPSPSPhAIL-ER". The peptide has near complete coverage by y ions and low mass b ions. The characteristic a_2, b_2 ion pair is clearly visible (low mass inset) and all peaks are clearly isotope resolved (high mass inset). (B) CID spectrum for the same peptide. Y-ion coverage is less but the spectrum shows high mass b-ions absent from the HCD spectrum. (C) HCD spectrum of the longer peptide AAAAAALSGAGTTPAGGGAGG-GAGGGGpSPGGWAVAR and the corresponding CID spectrum in (D). Note that HCD but not CID allows confident charge state assignment of the fragments as seen in the inset for the y_{26}^{2+} ion. The y_6 ion is not present in CID because it was fragmented by pseudo MS³ of the neutral phospho-loss of the precursor.

technical notes

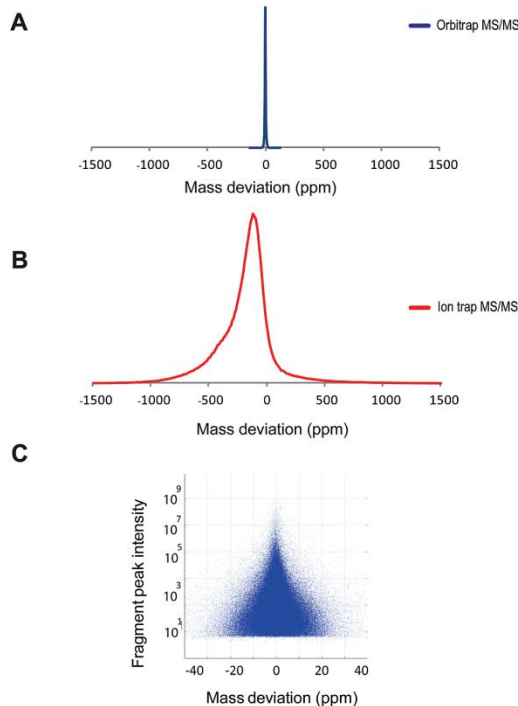


Figure 2. Distribution of fragment-ion mass deviations. Mass deviation of fragment ions from calculated values in ppm is plotted as a histogram. HCD data are in blue (A) and CID data are in red (B). (C) Mass deviations for 1.8 million fragments of identified phosphopeptides. The two-dimensional density distribution of mass deviation and number of fragment peaks is shown as a function of the fragment peak intensity.

± 0.25 Da in the score. If this would be done, the database identification score of the HCD Orbitrap spectra in general would be several fold higher than the scores for CID linear ion trap spectra. Note, however, that the high mass accuracy is still incorporated into database searching, because decoy hits at high mass accuracy are less likely, which has the effect of lowering the score for statistically significant database hits.

High Mass Accuracy of Fragment Ions. To assess the overall mass accuracy of in the HCD experiments, we plotted the mass deviation between calculated and measured fragments for all identified peptides (1.6 million data points; Figure 2A,C). Almost all the fragment peaks were identified within 20 ppm with 95% of the peaks falling within a 12.5 ppm window. In contrast, the fragment mass deviations resulting from ion trap detection were spread much more widely and 95% of the peaks fall within 542 ppm (Figure 2B). Thus, HCD fragment ion mass accuracy was about 50-fold improved over CID with detection in the ion trap.

While fragment mass accuracy in HCD was much higher than in CID, it was nevertheless much lower than the sub-ppm mass accuracy that can be achieved for precursors with Orbitrap measurements.²⁷ This can mainly be attributed to the fact that precursor ions are measured several times across the elution peak and with higher resolution. To test if the low abundance of phosphopeptides and their fragments influences

technical notes

Table 1. Direct Comparison of HCD and CID Fragmentation for Phosphopeptide with High Mass Accuracy at Both Precursor and Fragmentation Scan

experiment	total sites	Class I sites	Non-Class I sites[%]	mean Mascot score	mean localization probability
CID_1	1320	949	28	40.9	0.985
CID_2	1293	941	27	40.5	0.985
CID_3	1207	861	29	40.5	0.985
CID_4	1539	1179	23	42.8	0.987
HCD_1	1735	1252	28	42.6	0.987
HCD_2	1448	1030	29	45.6	0.986
HCD_3	1702	1224	28	46.6	0.985
HCD_4	1960	1460	26	48.1	0.988

the mass accuracy, we plotted measured mass deviations against fragment peak intensity (Figure 2C). Mass accuracy is indeed dependent on intensity. However, as the Orbitrap analyzer has a large dynamic range also for mass measurement accuracy,³¹ mass deviations do not generally exceed 15 ppm even for very weak peaks.

Direct Comparison of Phospho HCD and Phospho CID at High Resolution. To directly compare HCD and CID fragmentation for phosphopeptides, we analyzed the peptides by HCD and CID both in high–high mode. Enriched phosphopeptides were analyzed in replicate runs without any prefractionation in a 140 min LC-MS/MS method. To keep all parameters the same except for the fragmentation, 30 000 charges were accumulated within a maximum injection time of 150 ms for the two fragmentation techniques and the spectra were acquired in the Orbitrap detector. In this way, any differences in efficiency between ion-trap and triple-quadrupole like fragmentation for phosphopeptides at the same high resolution should become apparent. The HCD fragmentation yielded 1700 ± 200 sites, a value that was somewhat higher than the CID which yielded 1300 ± 140 sites (Table 1). Furthermore, the Mascot scores identified by HCD were 13% higher than the CID Mascot scores. This indicates that HCD fragmentation by itself contributes to better peptide identification, even without any mass accuracy advantages compared to CID.

Cycle Time for the Top 10 HCD Method. One of the obstacles to using high resolution MS/MS for phosphopeptide analysis at a large scale has been their very low abundance, which leads to long fill times for peptide fragmentation. In complex peptide mixtures that are separated by HPLC, the resulting MS and MS/MS cycle times were previously too long compared to LC peak widths. Our top10 method with 30 000 resolution for MS and 7500 for MS/MS uses 2.5 s measurement time in the Orbitrap analyzer²¹ and should therefore be compatible with the LC time scale. However, fill times to reach the desired target values have to be added to the scan time.

To investigate the actual cycle times in our experiment, we classified all cycles into 10 groups according to number of sequencing events. As this number increased from 1 through 10, the cycle times increased (Figure 3). Maximum fill time for MS/MS was set to 150 ms; therefore, the maximum total time for filling of the C-trap and for acquiring all transient should be 4 s. The median cycle time for the full 10 MS/MS events was 3.4 s (the longest cycle time was 4.24 s), indicating that relatively fast cycles were achieved even in this case.

With our parameters, only 12% of scan cycles with MS/MS events had all 10 fragmentation events (Supplementary Figure 1A, Supporting Information). Many such cycles had only

Nagaraj et al.

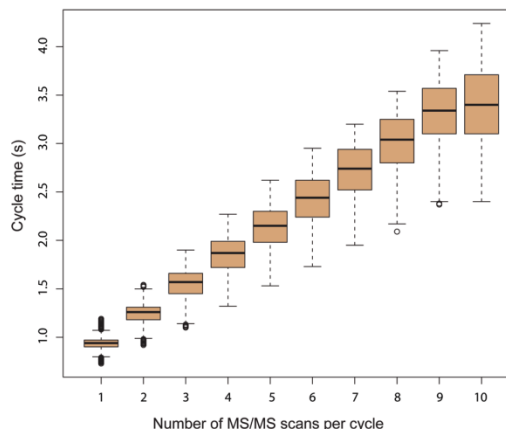


Figure 3. Duty cycle for different number of MS/MS events. The box plots show the distribution of cycle times for scan cycles with different numbers of MS/MS events per cycle. Cycles with no MS/MS events are not shown. Even when 10 peptides were sequenced in one cycle, total cycle time is within the chromatographic width of most phosphopeptides. Unlike standard complex peptide mixtures of nonmodified peptides, the phosphopeptides are not abundant, and this is reflected in the distribution of cycle times using up to 150 ms fill times per MS/MS event. The black bars in the plot represent the median and the circles represent the outliers.

one MS/MS event (35%), and the median number was two. We expected that this number would vary as a function of elution time. Indeed, the median number of MS/MS events per cycle was zero until the first peptides eluted from the column at 20 min, then rose to 10 at 42 min where the density of phosphopeptides was highest and gradually dropped toward zero at the end of the gradient (Supplementary Figure 1B, Supporting Information). This indicates that the top10 method was a good choice under these conditions because at all points in the gradient there were sufficient MS/MS events to target the peptides recognized by the instrument data system. However, this does not imply that all peaks were chosen for fragmentation. While the number of MS/MS events peaks at 40–45 min, the total ion current (TIC) was spread evenly across the gradient for the first but not for later fractions (Supplementary Figure 1C, Supporting Information).

The data system had been programmed to accumulate 30 000 ions or — if calculated injection times were longer than 150 ms — to accumulate for that time interval (“maxing out”). We found that due to the low intensity of the phosphopeptides overall a substantial proportion of all MS/MS events reached the 150 ms maximum injection time (Figure 4A). This is in contrast to complex mixture analysis of unmodified peptides on the same instrument, for which we previously determined a mean injection time 8.7 ms with the same HCD target values.²¹ Thus, sensitivity remains a critical parameter in phosphoproteome analysis, much more so than in the analysis of unmodified peptides. Fortunately, underfilling of the Orbitrap analyzer has no deleterious effects on the mass accuracy apart from reducing the signal of the fragments.

Given the very large number of phosphopeptide identifications of our experiment, we suspected that identification rates were high despite the underfilling. In total, 38.8% of all MS/

HCD-Based Large-Scale Phosphoproteomics

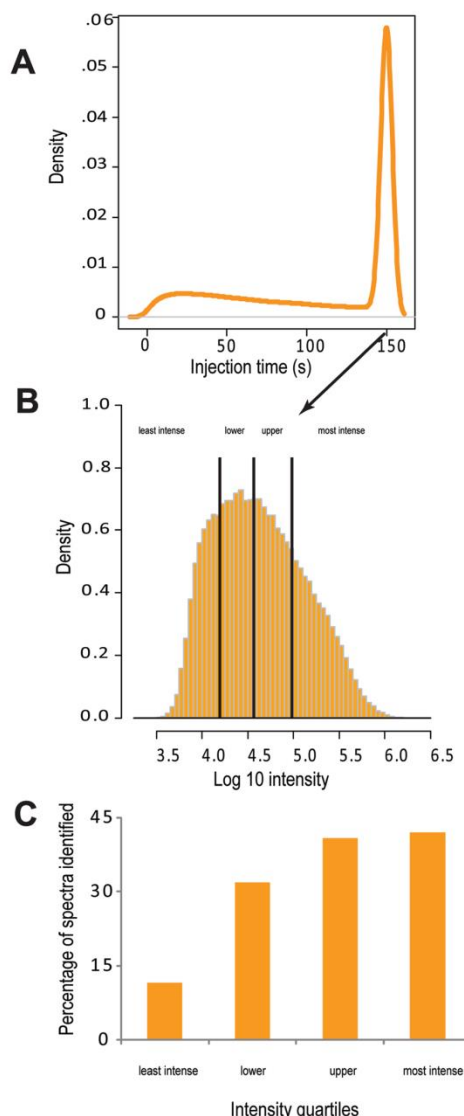


Figure 4. Injection times and MS/MS spectra identification rates. (A) Density distribution of injection time for MS/MS events for scan cycles. (B) Distribution of total ion signal for all the MS/MS spectra that maxed out the injection time. The distribution is divided into four quantiles. (C) Identification rates in each quantiles of the histogram are shown. The two most intense quantiles have identification rates of about 40%. The identification rate drops drastically only for the two least intense quantiles of the distribution, which have 1 or 2 orders of magnitude fewer ions.

MS events led to peptide identifications and 84% of these were phosphopeptides. We first determined the identification percentage of the SCX flow-through fractions, which have more concentrated phosphopeptide populations and therefore had a median injection time of only 118 ms. These fractions indeed had somewhat higher identification rates (47.8%) than non-

technical notes

Table 2. Measurement of Phosphosites with Different Target Values for MS/MS Fragmentation

ion target value	number of sites identified	median Mascot score	median PTM score	median localization probability
30000	1661	41.91	114.15	0.997
15000	1457	35.69	96.33	0.993
15000	1548	39.02	96.33	0.995
10000	1461	30.33	80.46	0.978
10000	1516	32.01	81.30	0.985
5000	1361	22.15	57.67	0.903
5000	1304	21.59	56.73	0.912

flowthrough fractions (29.9%). This clearly indicated that increased peptide intensity was beneficial but also that it was possible to identify phosphopeptides with fewer than 30 000 accumulated ions.

Next, we divided the MS/MS spectra into those for which the injection time was within 150 ms and those for which injection time maxed out at 150 ms. Approximately 27.83% of the maxed out fragmentation spectra were still assigned to a peptide sequence, and these were binned into four quantile based on the total ion signal in the MS/MS spectra (Figure 4B). Interestingly, the apparent total ion signal varied over more than 2 orders of magnitude and the identification rates were not reduced significantly at least for the two most intense quartiles. These have similar identification rates as compared to the least intense quantile of the spectra that did reach their target values within 150 ms. For the second least intense quantile, the identification rate went down to 30%, and for the least intense quantile it dropped to less than 15%. This suggested that far less than 30 000 ions were still sufficient for spectra identification in many cases. To investigate this phenomenon in more detail, we performed experiments with different target values for each run.

Phosphopeptide Identification Rates at Different Target Values for MS/MS. To test the identification success rate at defined lower target values, we enriched the HeLa lysates for phosphopeptides by TiO₂ beads without any prior separation by SCX. Peptides were separated and analyzed with a 90 min reversed phase LC MS/MS method (70 min gradient). The target value of 30 000 ions served as a reference value for comparison. In duplicate experiments, we decreased target values to 15 000, 10 000, and 5000 ions. Because the phosphopeptides were concentrated into one fraction median fill times ranged from 6.3 ms for the 30 000 target value to 1.2 ms for the 5000 target value. Accordingly, only a few percent of the spectra maxed out. The identification rate of MS/MS spectra for the reference target value run was 66%, which is comparable to overall HCD identification rates for non-phosphopeptides in our experience. Remarkably, the number of identified phosphosites dropped only slightly from 1661 for the reference run at the 30 000 target value to an average of 1332 for the runs with 5000 target value (about a 20% reduction in both the number of identifications and the MS/MS identification percentage). For the 10 000 ion target value, the reduction was less than 10% (Table 2). However, the quality of the tandem mass spectra was affected as the target values were reduced. For all different target values and runs, 752 sites were identified in common. Mean Mascot or PTM scores^{7,32} reduced to 90% of the reference value for these sites already at the 15 000 ion target value and further reduced to about half at the 5000 ion target value.

In each of the identified phosphopeptides, the site of phosphorylation was assigned and the quality of this assign-

technical notes

ment quantified by the localization probability, which is the probability value for a site to be phosphorylated among other possible sites in the same peptide.⁷ While the Mascot or PTM scores are a measure of quality of identification of phosphopeptides, localization probability is a quality measure of how well the site of phosphorylation can be pinpointed with the available information in the fragmentation spectrum. The median localization probability was larger than 0.99 for the reference run at the 30 000 target value. It did not change appreciably at the 15 000 target value but then declined to about 0.9 for the run with 5000 target value. This implies that at the lower target values fragmentation peaks that are important to pinpoint the site of phosphorylation start to be lost. This apparently happens before the quality of the spectrum deteriorates such that the peptide cannot be identified with high confidence any more. Cycle times did not improve markedly between the reference and the low target values (median of 2.71 s vs 2.67 s).

These findings explain why the overall number and the identification percentage were very high in our large-scale experiment. Despite the fact that many spectra did not reach their target values, their reliable identification was generally not impaired even though the localization confidence dropped for a subset of the spectra.

Together, the above results suggest that top5 to top10 methods with about 30 000 target values and about 150 ms maximum injection times are a good choice for HCD phosphopeptide experiments. Values larger than 10 MS/MS events per cycle or much larger than 150 ms injection times risk extending cycle times to more than 4 s in complex mixtures. In less complex mixtures, however, a larger maximum injection time at the expense of the number of MS/MS events may be advantageous. In any case, many phosphopeptides are confidently identified with one tenth or even fewer ions than the target value of 30 000. As a practical point, we have found it important to regularly clean the S-lens as otherwise ion currents can drop drastically (presumably due to charging effects) leading to very long fill times.

Comparison of HCD and CID for Large-Scale Phosphoproteomics. Next we analyzed the second half of the sample on the same machine and with the same methods, except that we fragmented phosphopeptides by CID with analysis in the ion trap instead of HCD and detection in the Orbitrap analyzer. We used a top20 method with a target value of 5000 ions. Distribution of MS/MS events actually performed by the instrument was similar to the HCD case, with a similar number of full scans (HCD: 91 245; CID: 87 338) and MS/MS scans (HCD 129 468; CID 145 776). In particular, the instrument only performed 10 or more CID scans in 1% of the cases and in most cases only fragmented none or one precursor ion (Supplementary Figure 2, Supporting Information).

Interestingly, both fragmentation methods identified a similar number of Class I sites (9668 for HCD and 9016 for CID), but HCD identified more phosphorylation sites overall (16 559 vs 11 893). We therefore checked the quality of identification in both methods. For all phosphorylation states of peptides with Class I sites, the high-high strategy yields better identification quality (Supplementary Figure 3A,B, Supporting Information) as the Mascot scores are 30% higher as compared to the high-low strategy (Figure 6). As mentioned above, this is despite the fact that the Mascot score does not increase with high mass accuracy. For the non-Class I phosphorylation sites, the mean peptide Mascot score is still higher in HCD compared

Nagaraj et al.

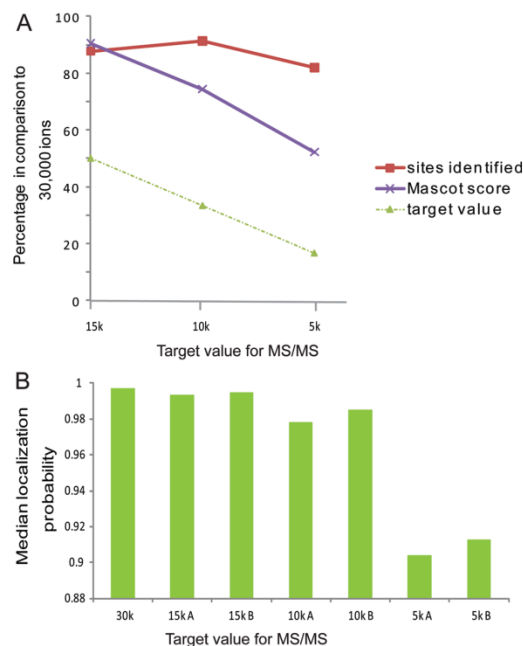


Figure 5. Identification and localization scores as a function of different target values. (A) Percentage reduction in the number and quality of identification with respect to the reference target value used (30 000). While peptides are still efficiently identified, the identification score decreases below a target value of 10 000. (B) In tandem with the identification scores median localization probability is considerably decreased below 10 000 target value.

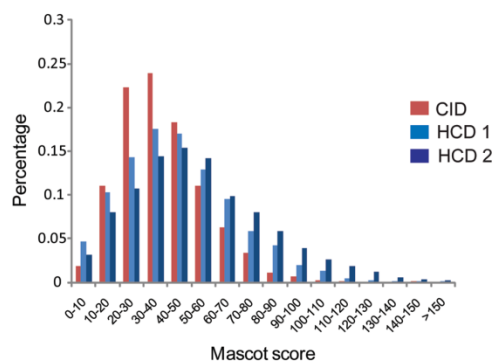


Figure 6. Comparison of HCD and CID database identification scores. The histogram of Mascot scores for all Class I sites of HCD (blue) and CID (red) shows a higher average and mean value for HCD. Very high identification scores were only achieved with HCD even though Mascot scores are insensitive to mass accuracy.

to CID, but the localization probabilities are not significantly different. Further, the CID and HCD runs have near identical peptide intensity distribution indicating that HCD detection in Orbitrap analyzer is not biased against low abundant peptides compared to CID measurement in ion trap (Supplementary Figure 4, Supporting Information).

HCD-Based Large-Scale Phosphoproteomics

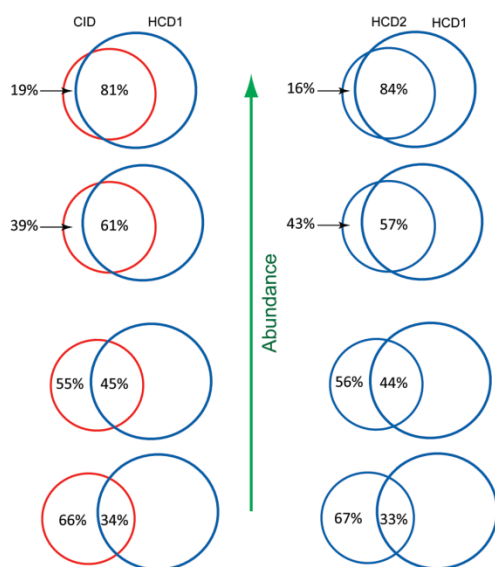


Figure 7. Overlap of phosphosites identified with HCD and CID. (Left) All phosphorylation sites were divided into four quantiles based on total intensity of the precursor phosphopeptide ion. 85% of the peptides in the most intense quantile identified in CID were also identified by HCD, whereas only 34% of the ones in the lowest quantile overlapped. (Right) Overlap determined as in the left panel but between the two HCD experiments.

We next compared both data sets with our previous phosphoproteomic study of EGF signaling in HeLa cells where we identified 5849 Class I sites in a high-low strategy using CID.⁷ Overlap with the HCD data was even higher than that with the CID data as we covered 2985 (51.0%) and 2408 (41.2%) sites, respectively. Even though the experimental situations are not directly comparable because one experiment employed growth factor stimulation, these results suggest that HCD does not miss specific subsets of phosphosites previously identified in large-scale phosphoproteomics studies.

To address the question of overlap in phosphosites between CID and HCD in more detail, we classified phosphopeptides into four quantiles by intensity. As shown in Figure 7 for the most intense quantile, more than 80% of the sites identified in the CID experiment were also identified by HCD measurement, whereas in the least intense quantile the overlap is only 35%. When we compared the overlap between two large-scale HCD experiments (see Materials and Methods) we found the same result. This demonstrates that the degree of overlap is given by the probability of the peptide to be “picked for sequencing” by the instrument rather than any difference between the fragmentation methods.

We also combined the results of the two large-scale HCD experiments, and this resulted in 13 529 Class I sites which covered all but 2691 sites of the CID experiment. We were not able to identify any attribute that distinguishes these peptides from the other peptides. Together our results demonstrate that HCD is capable of identifying phosphopeptides to at least the same depths as CID and suggest that it does not discriminate against any particular phosphopeptide populations.

technical notes

Conclusion and Outlook

Our experiments show that the data quality of HCD makes it an attractive method for phosphopeptide characterization. In contrast to its implementation on previous LTQ-Orbitrap instruments, HCD in conjunction with the S-lens and higher efficiency HCD on the Velos instrument now make it a viable method for large-scale phosphoproteomics. Interestingly, despite the somewhat slower measurement cycle of our “high-high” strategy HCD showed similar or even better performance compared to CID. Furthermore, there was no indication of subpopulations of phosphopeptides that were preferentially better suited to CID than to HCD. Therefore, there appears to be no reason against using HCD in a routine manner on the Velos instrument.

Advantages of HCD with the Orbitrap analyzer detection compared to CID with ion trap detection include ready determination of charge states, assignment of neutral loss ions, and the presence of reporter ions such as the 216.041 ion for phosphotyrosine. HCD spectra generally have a higher dynamic range and are less noisy. Conversely, CID spectra often retain high mass b-ions that are absent in HCD because they fragment further.

The nominal number of ions requested for HCD spectra was 30 000 as opposed to 5000 for CID. This should have given an advantage to CID for the preferential detection of low abundance phosphopeptides. However, we did not observe such a trend, and precursor peptide intensity distributions were identical between HCD and CID. Indeed, phosphopeptides were still efficiently identified with as little as 5000 ion target values (even though identification and localization scores started to deteriorate). Nevertheless, we found that it is important to load a relatively high amount of phosphopeptides as this increased the percentage of MS/MS spectra identified and the total number of detected phosphosites. Therefore, improvement of sensitivity for phosphopeptide analysis remains an important research and development goal.

Here we have not systematically compared CID with fragment detection in the Orbitrap analyzer to HCD. That approach should also gain from the sensitivity improvement due to the S-lens and should offer many of the advantages of HCD that are related to the excellent fragment mass accuracy. Likewise, comparison of HCD to ETD with low or high resolution³³ would be of interest on the Velos platform. Another important area for future study is the scoring and localization of phosphogroups using algorithms specifically designed for high mass accuracy data. Database search scores should increase several fold for high resolution MS/MS spectra once this high mass accuracy is taken into account. Additionally, we expect that HCD will further improve relative to CID if MS/MS speed, which is currently only half as fast for HCD as for CID in the ion trap, could be increased.

Abbreviations: MS, mass spectrometry; MS/MS, tandem mass spectrometry; HCD, higher energy collisional dissociation; CID, collisional induced dissociation; PTM, post-translational modification; HPLC, high performance liquid chromatography; IPI, international protein index; FDR, false discovery rate.

Acknowledgment. We thank other members of the department for proteomics and signal transduction, for sharing insights. This work was partially supported by PROSPECTS, a seventh Framework grant by the European Directorate (grant agreement HEALTH-F4-2008-201648/PROSPECTS) and by the Max Planck Society for the advancement of Science.

technical notes

Nagaraj et al.

Supporting Information Available: Figure S1: MS/MS events over the gradient for HCD. Figure S2: MS/MS events over the gradient for CID. Figure S3: Localization probabilities for phosphorylation sites in HCD. Figure S4: Density plot of phosphopeptide precursors. Supplementary Table 1_A: Information about all phosphopeptides from the large-scale experiment HCD1. Supplementary Table 1_B: Information about all phosphopeptides from the large-scale experiment CID1. Supplementary Table 1_C: Information about all phosphopeptides from the large-scale experiment HCD2. This material is available free of charge via the Internet at <http://pubs.acs.org>. Viewing annotated spectra for any phosphorylation site: Annotated phosphorylation spectra can be viewed at TRANCHE (www.proteomecommons.org) using the hash key below. For any phosphorylation site of interest, note the raw file name and scan number from Supplementary Table 1 and open the file with this information as the concatenated name. An annotated MS/MS spectrum in png format will appear (png files can be viewed in almost all imaging applications, for example Adobe Illustrator or Microsoft picture manager). Accession codes for raw mass spectrometry files: The data associated with this manuscript may be downloaded from ProteomeCommons.org Tranche using the following hash: /Gyf6Cxs8Xlx8aUTof4/OcFDVdL3TDb-6J4UPLceZSTXL2kdZr9OUb5j6NdlduK6+ ehqHJ3Td9GZSQTka-DtUM4/gMsNYAAAAAATLQ==. For the annotated phospho HCD spectra (more than 30 000): w/ouOjs+Thet/KE1R91+8O3zh/4xQ2X815AvTsiHFC3yvcC9tF/5qIOJhDDhKYNuL.XfLwCOBF1tv-bav4Y5eG7hLPZ8E8AAAAAFKvRQ==.

References

- Cohen, P. The regulation of protein function by multisite phosphorylation - a 25 year update. *Trends Biochem. Sci.* **2000**, *25* (12), 596–601.
- Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255–261.
- Witte, E. S.; Old, W. M.; Resing, K. A.; Ahn, N. G. Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* **2007**, *4* (10), 798–806.
- Grimsrud, P. A.; Swaney, D. L.; Wenger, C. D.; Beauchene, N. A.; Coon, J. J. Phosphoproteomics for the masses. *ACS Chem. Biol.* **2010**, *5* (1), 105–119.
- Nita-Lazar, A.; Saito-Benz, H.; White, F. M. Quantitative phosphoproteomics by mass spectrometry: past, present, and future. *Proteomics* **2008**, *8* (21), 4433–4443.
- Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (33), 12130–12135.
- Olsen, J. V.; Blagoev, B.; Gnäd, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635–648.
- Olsen, J. V.; Vermeulen, M.; Santamaria, A.; Kumar, C.; Miller, M. L.; Jensen, L. J.; Gnäd, F.; Cox, J.; Jensen, T. S.; Nigg, E. A.; Brunak, S.; Mann, M. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal* **2010**, *3* (104), ra3.
- Schroeder, M. J.; Shabanowitz, J.; Schwartz, J. C.; Hunt, D. F.; Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal. Chem.* **2004**, *76* (13), 3590–3598.
- Boersema, P. J.; Mohammed, S.; Heck, A. J. R. Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* **2009**, *44* (6), 861–878.
- Louris, J. N.; Cooks, R. G.; Syka, J. E. P.; Kelley, P. E.; Stafford, G. C.; Todd, J. F. Instrumentation, Applications, and Energy Deposition in Quadrupole Ion-Trap Tandem Mass-Spectrometry. *Anal. Chem.* **1987**, *59* (13), 1677–1685.
- Annan, R. S.; Huddleston, M. J.; Verma, R.; Deshaies, R. J.; Carr, S. A. A multidimensional electrospray MS-based approach to phosphopeptide mapping. *Anal. Chem.* **2001**, *73* (3), 393–404.
- Steen, H.; Kuster, B.; Fernandez, M.; Pandey, A.; Mann, M. Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway. *J. Biol. Chem.* **2002**, *277* (2), 1031–1039.
- Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **2000**, *72* (3), 563–573.
- Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (26), 9528–9533.
- Stensballe, A.; Jensen, O. N.; Olsen, J. V.; Haselmann, K. F.; Zubarev, R. A. Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun. Mass Spectrom.* **2000**, *14* (19), 1793–1800.
- Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709–712.
- Zhang, Y.; Askenazi, M.; Jiang, J.; Luckey, C. J.; Griffin, J. D.; Marto, J. A. A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol. Cell. Proteomics* **2010**, *9* (5), 780–790.
- Zougman, A.; Pilch, B.; Podtelejnikov, A.; Kiehn, M.; Schnabel, C.; Kumar, C.; Mann, M. Integrated analysis of the cerebrospinal fluid peptidome and proteome. *J. Proteome Res.* **2008**, *7* (1), 386–399.
- Zhang, Y.; Ficarro, S. B.; Li, S. J.; Marto, J. A. Optimized Orbitrap HCD for quantitative analysis of phosphopeptides. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (8), 1425–1434.
- Olsen, J. V.; Schwartz, J. C.; Griep-Raming, J.; Nielsen, M. L.; Damoc, E.; Denisov, E.; Lange, O.; Remes, P.; Taylor, D.; Splendore, M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **2009**, *8* (12), 2759–2769.
- Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6* (5), 359–360.
- Nagaraj, N.; Lu, A.; Mann, M.; Wisniewski, J. R. Detergent-based but gel-free method allows identification of several hundred membrane proteins in single LC-MS runs. *J. Proteome Res.* **2008**, *7* (11), 5028–5032.
- Li, Q. R. N. Z.; Tang, J. S.; Nie, S.; Zeng, R. Effect of peptide-to-TiO₂ beads ratio on phosphopeptide enrichment selectivity. *J. Proteome Res.* **2009**, *8* (11), 5375–5381.
- Rappsilber, J.; Ishihama, Y.; Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **2003**, *75* (3), 663–670.
- Olsen, J. V.; de Godoy, L. M. F.; Li, G. Q.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4* (12), 2010–2021.
- Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–1372.
- Wisniewski, J. R.; Nagaraj, N.; Zougman, A.; Gnäd, F.; Mann, M. Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. *J. Proteome Res.* **2010**, *9* (6), 3280–3289.
- Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* **2005**, *4* (7), 873–886.
- Sleno, L.; Volmer, D. A. Ion activation methods for tandem mass spectrometry. *J. Mass Spectrom.* **2004**, *39* (10), 1091–1112.
- Makarov, A.; Denisov, E.; Lange, O.; Horning, S. Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.* **2006**, *17* (7), 977–982.
- Olsen, J. V.; Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (37), 13417–13422.
- Wenger, C. D.; McAlister, G. C.; Xia, Q.; Coon, J. J.; Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an ETD-enabled orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2010**, *9* (5), 754–763.

PR100637Q

3. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles (Article 2).

Correct classification of cancer patients into subtypes is a prerequisite for acute diagnosis and effective treatment and currently relies mainly on histological assessment. Transcript-based profiling technologies have enabled the segregation of subtypes based on their gene expression signatures. However, their usage in the clinic is limited so far because such signatures do not provide information if or to what extent the detected transcript is translated into proteins, and it ignores the effects of post-translational modifications. An in-depth, high accuracy quantitative proteomics approach capable of revealing common and distinct functional features between tumor entities would provide valuable insights into cancer subtypes of potential clinical relevance.

We chose two known subtypes of diffuse large B-cell lymphoma (DLBCL) to evaluate if proteomic methods can accurately subtype cancer subtypes. While the germinal-center B-cell-like (GCB) subgroup is similar to normal germinal center B-cells, the activated B-cell subgroup (ABC) harbors characteristics of tumor cells activated via their B-cell receptor and are histologically indistinguishable.

Using a mix of multiple SILAC-labeled cell lines, a technique termed “super-SILAC”, we quantitatively analyzed cell lines derived from both subtypes and determined a proteomic signature. This set of proteins partially overlapped with a gene expression signature obtained previously from the same system. However, it also revealed a new set of proteins that could explain functional differences between the two subtypes that may be clinically useful in diagnosis and chemotherapeutic development. Our results show that high resolution shotgun proteomics combined with super-SILAC-based quantification is a promising new technology for tumor characterization and classification.

This article was published in the journal Molecular and Cellular proteomics in 2012 [2].

Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles*[§]

Sally J. Deeb[‡], Rochelle C. J. D'Souza[‡], Jürgen Cox[‡], Marc Schmidt-Suppran^{§¶}, and Matthias Mann^{‡||}

Correct classification of cancer patients into subtypes is a prerequisite for acute diagnosis and effective treatment. Currently this classification relies mainly on histological assessment, but gene expression analysis by microarrays has shown great promise. Here we show that high accuracy, quantitative proteomics can robustly segregate cancer subtypes directly at the level of expressed proteins. We investigated two histologically indistinguishable subtypes of diffuse large B-cell lymphoma (DLBCL): activated B-cell-like (ABC) and germinal-center B-cell-like (GCB) subtypes, by first developing a general lymphoma stable isotope labeling with amino acids in cell culture (SILAC) mix from heavy stable isotope-labeled cell lines. This super-SILAC mix was combined with cell lysates from five ABC-DLBCL and five GCB-DLBCL cell lines. Shotgun proteomic analysis on a linear ion trap Orbitrap mass spectrometer with high mass accuracy at the MS and MS/MS levels yielded a proteome of more than 7,500 identified proteins. High accuracy of quantification allowed robust separation of subtypes by principal component analysis. The main contributors to the classification included proteins known to be differentially expressed between the subtypes such as the transcription factors IRF4 and SPI1/PU.1, cell surface markers CD44 and CD27, as well as novel candidates. We extracted a signature of 55 proteins that segregated subtypes and contained proteins connected to functional differences between the ABC and GCB-DLBCL subtypes, including many NF- κ B-regulated genes. Shortening the analysis time to single-shot analysis combined with use of the new linear quadrupole Orbitrap analyzer (Q Exactive) also clearly differentiated between the subtypes. These results show that high resolution shotgun proteomics combined with super-SILAC-based quantification is a promising new technology for tumor characterization and classification. *Molecular & Cellular Proteomics* 11: 10.1074/mcp.M111.015362, 77–89, 2012.

From the [‡]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany and the [§]Department of Molecular Immunology and Signal Transduction, Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany

* Author's Choice—Final version full access.

Received October 24, 2011, and in revised form, February 22, 2012

Published, MCP Papers in Press, March 21, 2012, DOI 10.1074/mcp.M111.015362

Clinical heterogeneity in terms of patient survival rates and response to therapy is a major challenge in cancer treatment. This difficulty partly stems from grouping together molecularly distinct tumor entities as one clinical type and treating them in the same manner. Transcript-based profiling technology enables the segregation of subtypes based on their gene expression signatures (1, 2). However, it is often difficult to interpret such signatures with respect to the biology of the disease (1). In addition, gene expression signatures do not provide information if or to what extent the detected transcript is translated into proteins, and it ignores the effects of post-translational modifications. An in-depth, high accuracy quantitative proteomics approach capable of revealing common and distinct functional features between tumor entities may provide valuable insights into cancer subtypes of potential clinical relevance.

MS-based proteomics has recently evolved into an important tool in mining deregulated signaling pathways in cancer because of its ability to move one step closer toward the cancer phenotype and because of substantial progress in technology and methodology (3, 4). These advances in MS now allow the identification of thousands of proteins in a single experiment as a result of enhanced sensitivity, accuracy, and speed of analysis (5–7). In addition, a variety of quantitative proteomic approaches can monitor expression changes of thousands of proteins and post-translational modifications in a reproducible manner (8, 9). Stable isotope labeling with amino acids in cell culture (SILAC)¹ is a particularly accurate method of quantitative proteomics (10, 11), but until recently it was limited to cell lines or animals that could be metabolically labeled with heavy amino acids. This limitation of SILAC in studying patient tumor samples has been overcome through the use of a mix of multiple SILAC-labeled cell lines as an internal standard, a technique called super-SILAC

¹ The abbreviations used are: SILAC, stable isotope labeling with amino acids in cell culture; ABC-DLBCL, activated B-cell-like diffuse large B-cell lymphoma; BCR, B-cell receptor; DLBCL, diffuse large B-cell lymphoma; GCB-DLBCL, germinal-center B-cell-like DLBCL; PCA, principal component analysis; FASP, filter-aided sample preparation; SAX, strong anion exchange; STAT, signal transducers and activators of transcription; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Super-SILAC Distinguishes Lymphoma Subtypes

(12). This mix achieved superior quantification accuracy compared with a single SILAC-labeled cell line (13). In particular, a narrow ratio distribution was obtained with 90% of proteins contained within an easily quantifiable 4-fold range between the tumor and the SILAC mix. We reasoned that this ability to quantify several thousand proteins with high accuracy might enable confident proteomic classification of tumors in different subtypes.

The subclassification of diffuse large B-cell lymphoma (DLBCL), the most common lymphoma in adults, by gene expression profiling was a major breakthrough because it resulted in the identification of two histologically indistinguishable subtypes that differ in their outcomes after multiagent chemotherapy (14). The germinal-center B-cell-like (GCB) subgroup has a gene expression signature characteristic of normal germinal center B-cells, whereas the activated B-cell-like (ABC) subgroup, being the one with worse prognosis, has a gene expression signature characteristic of B-cells activated through their B-cell receptor. One of the key pathways that are differentially activated between DLBCL subgroups is signaling to NF- κ B family transcription factors, which are constitutively active in the ABC subgroup (15). In B-cells, NF- κ B controls the expression of genes necessary for both proliferation and survival in response to stimulation, including antigen recognition by the B-cell receptor (BCR). The IRF4 transcription factor, an NF- κ B target, plays multiple roles in B lymphocyte development and function and is critical for plasma cell differentiation. Its high expression in ABC-DLBCL reflects constitutive NF- κ B activity and plasmacytic differentiation. Recently, mutations leading to "chronically active" BCR signaling have been described as a mechanism providing aberrant cellular survival signals in ABC-DLBCL (16). In these cases, the constitutive NF- κ B activation in ABC-DLBCL depends on the multiprotein CARD11-BCL10-MALT1 (CBM) complex (17–19). Such findings may open the door for new therapeutic modalities that target components of BCR signaling upstream of NF- κ B. Furthermore, improvements in DNA sequencing technologies have paved the way to the discovery of novel aspects of DLBCL pathology, such as impairments in chromatin methylation and evasion of T cell immune surveillance (20). This shows that the deployment of novel methodologies continuously enhances our understanding of the complex biology of lymphomas.

Despite the success of gene expression profiling in differentiating between tumor subtypes, the extracted transcriptional signatures do not always suffice to identify biological drivers of tumor pathogenesis. Furthermore, their adoption in the clinic, where protein-based assays are more commonly used, has been slow. A long standing aim of the proteomics community is to directly study human cancer at the protein rather than the transcript level (3). Here, we use high resolution shotgun proteomics combined with a super-SILAC quantitative approach in an attempt to segregate DLBCL subtypes. If applicable, the super-SILAC technology should be particu-

larly accurate, robust, and reproducible because it provides an entire reference proteome consisting of thousands of heavy labeled proteins for comparison of a large number of tumor samples. We evaluate the super-SILAC spike-in approach for distinguishing cell lines derived from ABC- and GCB-DLBCL patients. Choosing such closely related disease entities sets a high bar for our quantitative proteomics technology. Furthermore, the fact that specific biological differences between ABC and GCB are already known allows us to evaluate proteomics results in light of those differences.

EXPERIMENTAL PROCEDURES

Cell Culture Sample Preparation—ABC-DLBCL cell lines (HBL1, OciLy3, RIVA, TMD8, and U2932) and GCB-DLBCL cell lines (BJAB, DB, HT, SUDHL-4, and SUDHL-6) were grown in RPMI medium (Invitrogen) supplemented with 20% fetal bovine serum. Cell lysis was performed using a buffer consisting of 4% SDS, 0.1 M DTT, and 0.1 M Tris-HCl pH 7.5 followed by incubation at 95 °C for 5 min. The lysates were sonicated using a Branson type sonicator and then centrifuged at $16,100 \times g$ for 10 min.

Cell lines selected for inclusion in the super-SILAC mix were grown in RPMI medium containing $^{13}\text{C}_6$ $^{15}\text{N}_2$ -lysine (Lys⁸) and $^{13}\text{C}_6$ $^{15}\text{N}_4$ -arginine (Arg¹⁰) (Cambridge Isotope Laboratories) instead of the natural amino acids and supplemented with 20% dialyzed fetal bovine serum. The cells were cultured for at least six passages until they were fully labeled as assessed by quantitative mass spectrometry. Less than 1% of tryptic peptides contained unlabeled arginine or lysine in the nine labeled cell lines (Ramos, Mutu, BL-41, U2932, OciLy3, BJAB, L1236, L428, and DB) and less than 0.3% of identified peptides showed evidence of Arg to Pro conversion. Equal amounts of the heavy lysates were mixed to generate the super-SILAC mix.

Protein Digestion and Fractionation—The super-SILAC mix (100 μ g) was combined with an equal amount of the unlabeled cells and further processed by the filter-aided sample preparation (FASP) method (21). In short, the sample was loaded on Microcon filters with a 30-kDa cutoff (Millipore, Billerica, MA), which allows the replacement of SDS with a urea containing buffer. The proteins were then alkylated with iodoacetamide followed by overnight trypsin digestion at 37 °C in 50 mM ammonium bicarbonate. Peptides were collected from the filter after centrifugation and elution with water (2 \times).

Using strong anion exchange chromatography, 40 μ g of the peptide mixture from each replicate was fractionated (22). In summary, the strong anion exchange (SAX) was performed in tip columns prepared from 200- μ l micropipet tips stacked with six layers of a 3M Empore anion exchange disk (1214-5012; Varian, Palo Alto, CA). We used Britton & Robinson universal buffer composed of 20 mM acetic acid, 20 mM phosphoric acid, and 20 mM boric acid and titrated with NaOH to the desired pH for column equilibration and fraction elution. After loading the peptides at pH 11 and collecting it, five additional fractions were collected consecutively with buffers of pH 8, 6, 5, 4, and 3. The eluted fractions were desalted on reversed phase C₁₈ Empore disc StageTips (23). Peptide elution was performed twice with 20 μ l of buffer B containing 80% ACN in 0.5% acetic acid. Organic solvents were removed by a SpeedVac concentrator to prepare the samples for analysis by LC-MS/MS.

Liquid Chromatography and MS for Fractionation Experiments—Eluted peptides were separated on an in-house-made 15-cm column with a 75- μ m inner diameter packed with ReproSil-Pur C₁₈-AQ 3 μ m resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) using an Easy nanoflow HPLC system (Proxeon Biosystems, now Thermo Fisher Scientific). The HPLC was coupled via a nanoelectrospray ion source (Proxeon Biosystems) to an LTQ-Orbitrap Velos mass spec-

trometer (Thermo Fisher Scientific) (24). Approximately 2 μg of peptides were loaded in buffer A (0.5% (v/v) acetic acid) with a flow rate of 500 nl min^{-1} and eluted with a 200-min linear gradient at a flow rate of 200 nl min^{-1} . Four different gradients were applied for optimal separation based on average peptide hydrophobicity. A gradient of 2–25% buffer B to separate the pH 11 fraction; 7–25% buffer B for the pH 8 fraction; 7–30% buffer B for the pH 6 and 5 fractions; and 7–37% buffer B for the pH 4 and 3 fractions. After each gradient, the column was washed, reaching 90% buffer B followed by re-equilibration with buffer A.

The mass spectra were acquired with an automatic switch between a full scan and up to 10 data-dependent MS/MS scans. Target value for the full scan MS spectra were 1,000,000 and resolution was 30,000 at m/z 400. Up to the 10 most intense ions (minimum signal threshold of 5,000) were sequentially isolated and accumulated to a target value of 40,000 with a maximum injection time of 150 ms and were fragmented by higher energy collisional dissociation (25). For a subset of measurements, MS/MS target values were set to 50,000. The spectra of the fragmented ions were acquired in the Orbitrap analyzer with resolution of 7,500 at m/z 400.

Liquid Chromatography and MS for Single-shot Experiments—The peptides were separated on an in-house-made 50-cm column with a 75- μm inner diameter packed with 1.8 μm C_{18} resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). The Thermo EASY-nLC 1000 system with a binary buffer system consisting of 0.5% formic acid (buffer A) and 80% acetonitrile in 0.5% formic acid (buffer B) was used for reverse phase chromatography. Peptides (~4 μg) were eluted with a 220-min linear gradient of buffer B up to 30% at a flow rate of 250 nl min^{-1} . The column temperature was kept at 40 $^{\circ}\text{C}$ by an in-house designed oven with a Peltier element (26). The LC was coupled to a Q Exactive mass spectrometer (27) (Thermo Fisher Scientific) via the nanoelectrospray source (Proxeon Biosystems, now Thermo Fisher Scientific). Mass spectra were acquired on the Q Exactive in a data-dependent mode with an automatic switch between a full scan and up to 10 data-dependent MS/MS scans. Target value for the full scan MS spectra was 3,000,000 with a maximum injection time of 20 ms and a resolution of 70,000 at m/z 400. The 10 most intense ions with charge two or more from the survey scan were selected with an isolation window of 1.6 Th and fragmented by higher energy collisional dissociation (25) with normalized collision energies of 25. The ion target value for MS/MS was set to 1,000,000 with a maximum injection time of 60 ms and a resolution of 17,500 at m/z 400. These settings lead to constant injection times of 60 ms, fully in parallel with transient acquisition of the previous scan, ensuring fast cycle times. Repeat sequencing of peptides was kept to a minimum by dynamic exclusion of the sequenced peptides for 25 s.

Data Analysis—The acquired raw files were analyzed by MaxQuant (28) (version 1.2.0.34). Andromeda, a probabilistic search engine incorporated into the MaxQuant framework (29), was used to search the peak lists against the IPI human database version 3.68 which contains 87,083 entries. Common contaminants were added to this database. The search included cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. The second peptide identification option in Andromeda was enabled (29). For statistical evaluation of the data obtained, the posterior error probability and false discovery rate were used. The false discovery rate was determined by searching a reverse database. A false discovery rate of 0.01 for proteins and peptides was required. Enzyme specificity was set to trypsin allowing N-terminal cleavage to proline. Two miscleavages were allowed, and a minimum of six amino acids per identified peptide were required. Peptide identification was based on a search with an initial mass deviation of the precursor ion of up to 6 ppm, and the allowed fragment mass deviation was set to 20 ppm. The mass accuracy of

the precursor ions was improved by retention time-dependent mass recalibration (28). To match identifications across different replicates and adjacent fractions, the “match between runs” option in MaxQuant was enabled within a time window of 2 min. Quantification of SILAC pairs was performed by MaxQuant with standard settings using a minimum ratio count of 2. Bioinformatics analysis was done with Perseus tools available in the MaxQuant environment.

When needed for the analysis, the missing values were replaced using data imputation. The idea of our algorithm for imputation of missing values is that they should simulate signals of low abundant proteins. To accomplish this, we first determine the mean and standard deviation of all valid values in the matrix. Then we draw numbers for the missing entries from a suitable probability distribution in an independently, identically distributed way. For that purpose, we use a normal distribution with a mean and standard deviation adjusted in such a way as to simulate signals of low abundant proteins. This is necessary because the missing values are biased toward the detection limit of the LC-MS/MS measurement. Optimal values for the down shift parameter were adjusted in a way that the distribution of imputed values adjusts smoothly to the lower end of the distribution of measured values. We iteratively adjusted the values to avoid too large or too small down shifts. The former would result in a separation of imputed and measured values (a bi-modal total distribution), whereas the latter would introduce too much noise into the system and would potentially destroy protein signatures. The two values for downshifting and width adjustment are determined once but then apply to all the cell lines. These optimal values were different for the label-free and SILAC reference cases. For label-free data, we employed a width of 0.3 and a downshift of 1.8; in the super-SILAC data, the width was 0.3, and the downshift was 0.5, each in units of the standard deviation of the distribution of present values.

RESULTS AND DISCUSSION

Development of a Lymphoma Super-SILAC Mix—To accurately quantify proteome differences between lymphoma subtypes, we set out to generate a super-SILAC mix that would be optimally suited as an internal standard for a broad range of B-cell lymphomas. We considered commonly used cell lines derived from patients with different types of the disease. First we selected two lines, L428 and L1236, to represent Hodgkin's lymphoma. Of the non-Hodgkin's lymphomas, we selected three cell lines of patients with Burkitt's lymphoma, which is characterized by a c-Myc t(8;14) translocation. For DLBCL, we started out with the five ABC type cell lines and five GCB type cell lines that we wished to segregate by proteomics. From these, we chose two ABC type cell lines (Oci-Ly3 and U2932), as well as two GCB type cell lines (BJAB and DB).

Next, we wished to select an optimal subset of these nine representative cell lines (green in Fig. 1A). Instead of empirically testing different combinations, we reasoned that an in-depth proteome of each of the nine cell lines should be sufficient to mathematically determine the best combination. For this purpose, we performed a six-fraction FASP-SAX-based analysis, with 4-h gradients on an LTQ-Orbitrap Velos and higher energy collisional dissociation-based fragmentation (Fig. 1A and “Experimental Procedures”). This involved a single day of measurement time for each of the nine proteomes.

Super-SILAC Distinguishes Lymphoma Subtypes

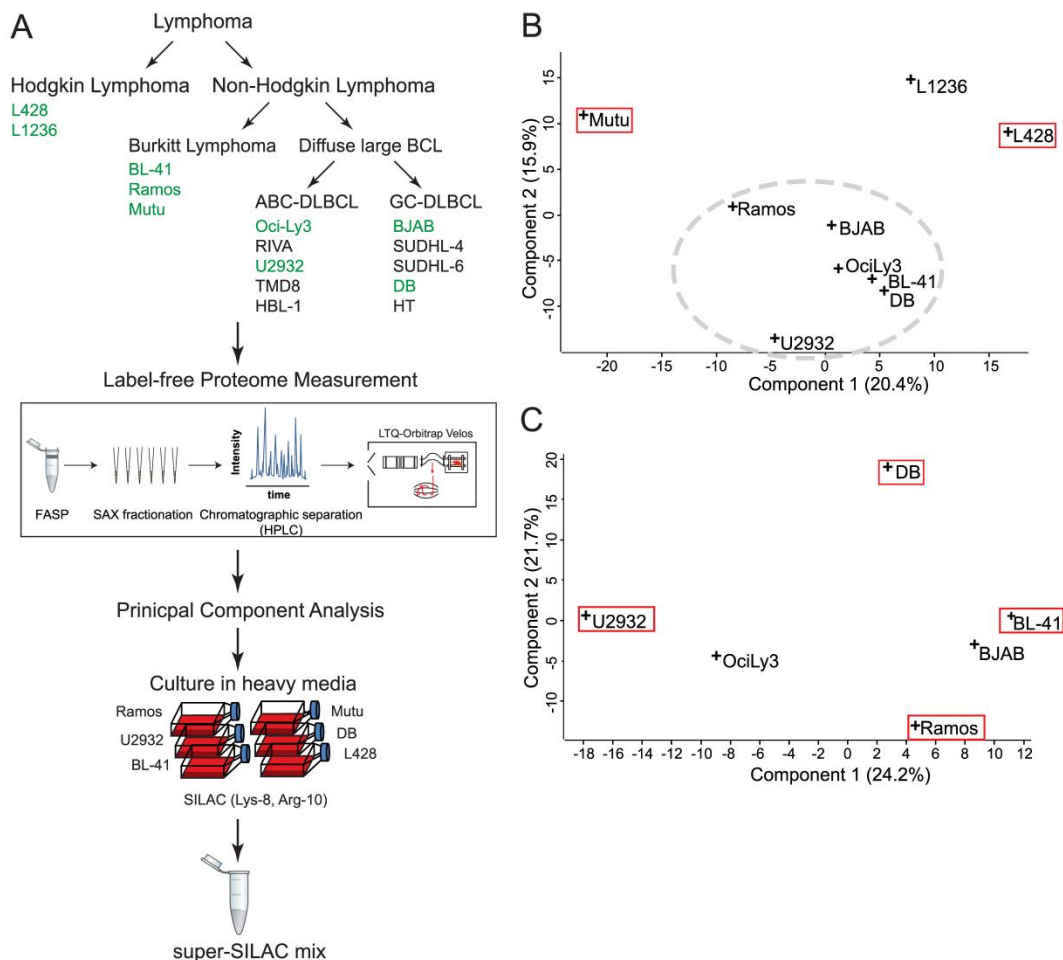


FIG. 1. Rational construction of lymphoma super-SILAC mix. A, label-free proteomics of nine B-cell lymphoma cell lines was performed after FASP-SAX processing and analyzed using high resolution precursor and fragment measurements on an Orbitrap Velos. They included two Hodgkin lymphoma cell lines (L428 and L1236) and seven non-Hodgkin lymphoma cell lines (Ramos, Mutu, BL-41, OciLy3, U2932, BJAB, and DB). B, PCA of nine B-cell lymphoma cell lines based on their protein expression profiles. The red boxes indicate cell lines selected for the super-SILAC mix. The gray dashed ellipse groups non-Hodgkin lymphoma cell lines to be further analyzed by a second PCA. C, PCA of the six non-Hodgkin lymphoma cell lines encircled in B. The red boxes indicate cell lines selected for the super-SILAC mix.

To compare the label-free proteomes of the cell lines to each other, we performed principal component analysis (PCA). PCA transforms large data sets into points in a data space of orthogonal components, such that the first component captures most of the variability. Because PCA analysis requires a complete data set (in this case label-free protein intensities for all identified proteins in all samples), we employed “data imputation” as described in “Experimental Procedures.” We aimed to create a mixture of cells that capture the largest diversity. Therefore, we searched for those that were most distant from one another.

Mutu(–), one of the Burkitt-derived cell lines, was the furthest outlier (Fig. 1B). L1236 and L428, the only Hodgkin’s lymphoma cell lines, were also outliers. We therefore selected Mutu(–) and one of the two Hodgkin’s lymphoma cell lines (L428). We then performed a second round of PCA on the remaining seven non-Hodgkin cell lines and selected the four outermost in the resulting PCA space (U2932, DB, BL-41, and Ramos) (Fig. 1C).

To produce the super-SILAC mix from the selected six cell lines, we grew them in heavy SILAC media and mixed them in equal proportions. For a first evaluation, we spiked the mix

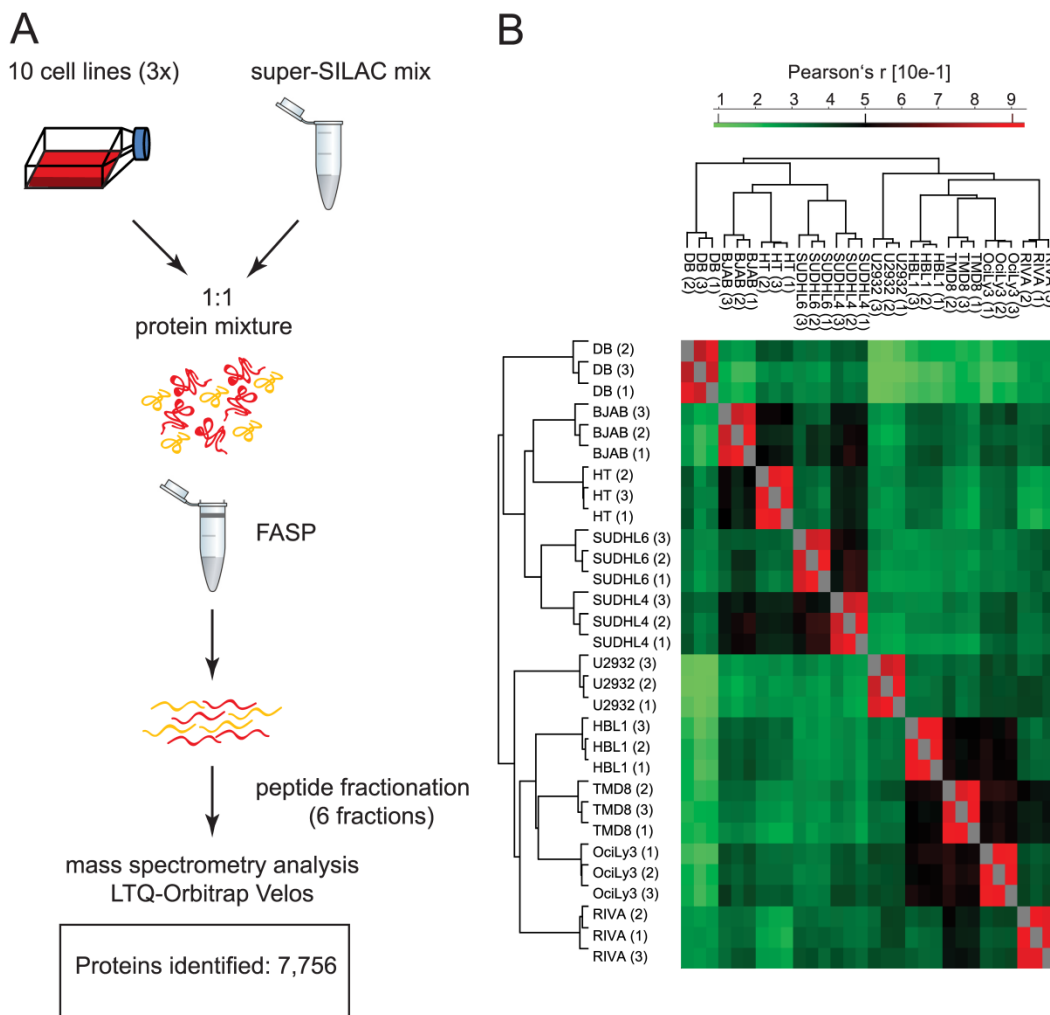


FIG. 2. **Proteomic workflow and overall results.** A, the super-SILAC mix developed on the basis of label-free proteome comparison was used as an internal standard for 10 different DLBCL cell lines. The samples were processed by FASP-SAX followed by triplicate 1-day proteome analyses. B, heat map of Pearson correlation coefficients showing reproducibility between replicates.

into lysate of an unlabeled ABC-DLBCL cell line (HBL-1) that was not part of the original selection. The histogram of fold changes between cell line proteins and super-SILAC proteins was narrow, with 96% of the values within a 4-fold range. To check the overall selection procedure, we also performed this experiment with a mix of all nine initially selected cell lines. The width of the distribution was essentially unchanged, indicating that the six-cell line mix already adequately represented the proteome. Finally, a three-cell line mix of only the largest outliers of the PCA analysis (Mutu, L428, and U2932;

Fig. 1B) also performed surprisingly well, attesting to the usefulness of our selection procedure (supplemental Fig. 1).

In-depth Proteome Coverage Using the Lymphoma Super-SILAC Mix—We spiked the super-SILAC mix into five unlabeled ABC and five GCB cell lines and analyzed them as described above for the label-free experiment, except that each proteome was measured in triplicate (Fig. 2A). Joint analysis of the resulting 180 LC MS/MS files (30 days measuring time) in MaxQuant identified a total of 7,756 different protein groups, by far the largest B-cell lymphoma proteome

Super-SILAC Distinguishes Lymphoma Subtypes

reported to date. Of these proteins, 6,263 were quantified in at least two replicates of the same cell line (supplemental Tables I and II).

At this depth of proteome coverage, we identified and quantified a large number of known members of the B-cell receptor-initiated signal transduction pathway (supplemental Fig. 2). Likewise, many transcription factors relevant to B-cell biology were measured. Altogether, we identified 285 proteins annotated by Gene Ontology to have sequence-specific DNA-binding transcription factor activity (supplemental Table II). This list included many transcription factors playing important roles in B-cells, such as basic leucine zipper transcription factor ATF-like (BATF), B-cell lymphoma 3 protein (BCL3), B-cell lymphoma 6 protein (BCL6), immunoglobulin transcription factors 1 and 2 (ITF1 and ITF2), Ets domain-containing PU.1, and the B-lineage specifying transcription factor PAX-5.

Next, we quantified all 30 proteome measurements against each other based on the ratios to the super-SILAC mix and calculated their Pearson correlation coefficients (r). Unsupervised clustering of the rows and columns of the matrix of the 30×30 coefficients co-clustered the triplicates in each case (Fig. 2B). Good reproducibility is further indicated by the high average Pearson coefficients of the triplicates ($r = 0.87$).

Segregation between DLBCL Subtypes—To investigate whether our proteomics measurements can segregate ABC from GCB proteomes and to determine an optimal data analysis strategy, we started by performing unsupervised hierarchical clustering of all proteome measurements. We required that proteins were present in at least 50% of the 30 measurements and filled any missing values by “data imputation” (“Experimental Procedures”). Again, replicate measurements were always clustered together. Intriguingly, the two major branches of the dendrogram precisely grouped all the ABC and all the GCB subtypes together and apart from each other. This indicates that these subtypes have quite different protein expression patterns at a global level that are capable of defining them as distinct entities.

The cluster indicated with arrow B in Fig. 3A, consists of 107 proteins, 70 of which are annotated as ribosomal, 12 of which are components of the 20 S proteasome, and 14 of which are components of the 26 S proteasome (CORUM annotation) (15). As shown in Fig. 3B, their expression varies little across the cell lines; thereby they serve as “loading controls” and validate correct normalization and imputation of the proteome samples by MaxQuant. This ensures that the variation of protein expression values between ABC and GCB can directly be attributed to biological differences between these cell types rather than experimental artifacts. Fig. 3C shows the differences in expression of two clusters in the upper part of Fig. 3A (indicated with arrows C and D) with large differential expression patterns between the two main branches of the dendrogram. The first cluster consists of 16 proteins that are up-regulated in the ABC subtype relative to GC. This cluster

includes proteins such as CD44, FOXP1, IL4I1, VAV2, and BID (supplemental Table IV). The second cluster consists of 19 proteins that are up-regulated in the GCB subtype and includes proteins such as CD81, KIND3, WIP, INPP5B, PAG, and BRDG1 (supplemental Table V).

Principal component analysis was performed to project the SILAC-based proteome measurements into a two-dimensional data space. We first applied PCA for the subgroup of proteins that were quantified in each of the 30 proteome measurements (100% valid values; 3,007 protein groups). Component 1 of the PCA, which accounts for 20.5% of total variability (*horizontal axis* in the two-dimensional plot of Fig. 4A), clearly separates GCB (group on the *left side*) from ABC (group on the *right side*). Furthermore, Fig. 4A shows that the distance between the replicates is much smaller than the separation between the groups, supporting the robustness of the segregation.

The proteins that are most responsible for separating the proteomes in the PCA can be seen in the “loadings.” The loadings of component 1, which capture the differences between the two groups, include the transcription factor IRF4, mentioned above as one of the main drivers of the functional differences between GCB and ABC lymphomas (Fig. 4B). In fact, high expression of IRF4 in ABC-DLBCL is tied to the constitutive activity of NF- κ B that is required for survival of this subtype of lymphoma cells (15). This transcription factor, which was quantified in 30 of 30 proteomes, is the strongest differentiator in this unbiased large scale analysis. PTP1B was another one of the strongest loadings of component 1. PTP1B is a key tyrosine phosphatase implicated in the regulation of JAK/STAT signaling. The preferential expression of PTP1B in ABC-DLBCL is already known, and its overexpression has been suggested to contribute to the enhanced STAT6 dephosphorylation that is observed in these tumors upon IL-4 stimulation (30, 31).

The above analysis required quantification of the proteins in every proteome measurement, which could exclude many interesting proteins, such as those exclusively expressed in only one subtype. We therefore employed imputation of missing values to make a larger subset of the proteome amenable to PCA analysis. We first filtered for at least 50% valid values (4,991 proteins) and imputed the missing values. Incorporation of the information from these additional proteins led to an even stronger separation of the subtypes (Fig. 4C). The GCB cell lines appear to cluster more tightly together, whereas the ABC cell lines U2932 and RIVA are somewhat separated from the other ABC cell lines. The loadings in Fig. 4D reveal additional known markers such as the cell surface markers CD44 for ABC-DLBCL (quantified exclusively in ABC) and CD27 for GCB-DLBCL (Fig. 4D). The above analysis demonstrates that requiring less than 100% valid values and imputing missing values is a valid and robust strategy for segregation of subtype groups, as well as for finding individual differentiators by proteomics.

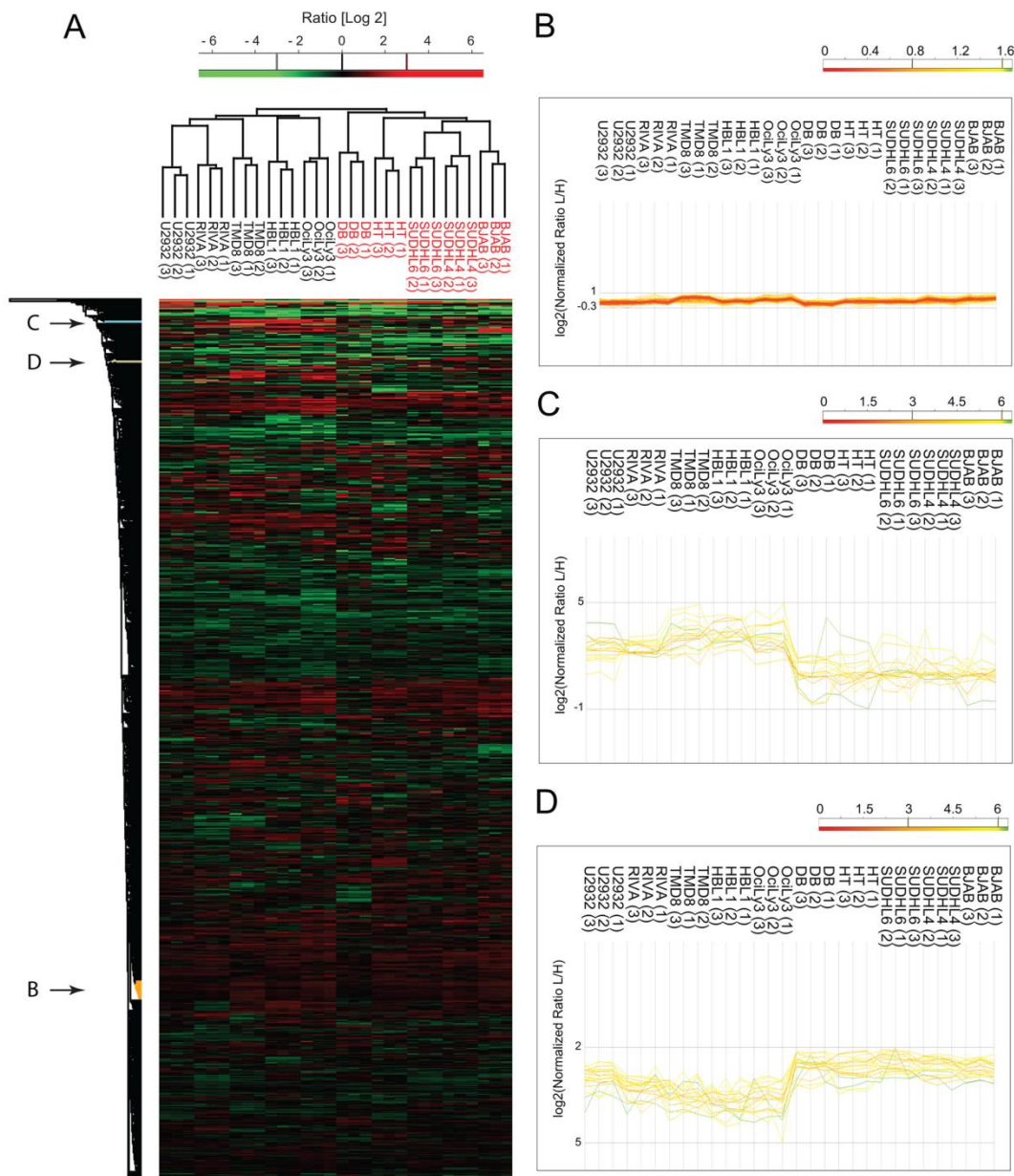
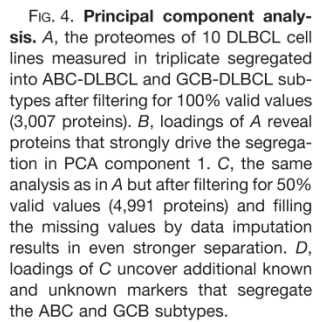


FIG. 3. **Unsupervised hierarchical clustering.** A, unsupervised clustering of protein expression profiles of 10 DLBCL cell lines after filtering for 50% valid values and imputation of missing values. B, expression patterns for a cluster enriched for ribosomal and proteasomal proteins. C, expression patterns for a cluster of proteins with higher expression levels in ABC relative to GCB. D, expression patterns for a cluster of proteins with higher expression levels in GCB relative to ABC.

Super-SILAC Distinguishes Lymphoma Subtypes



Super-SILAC Distinguishes Lymphoma Subtypes

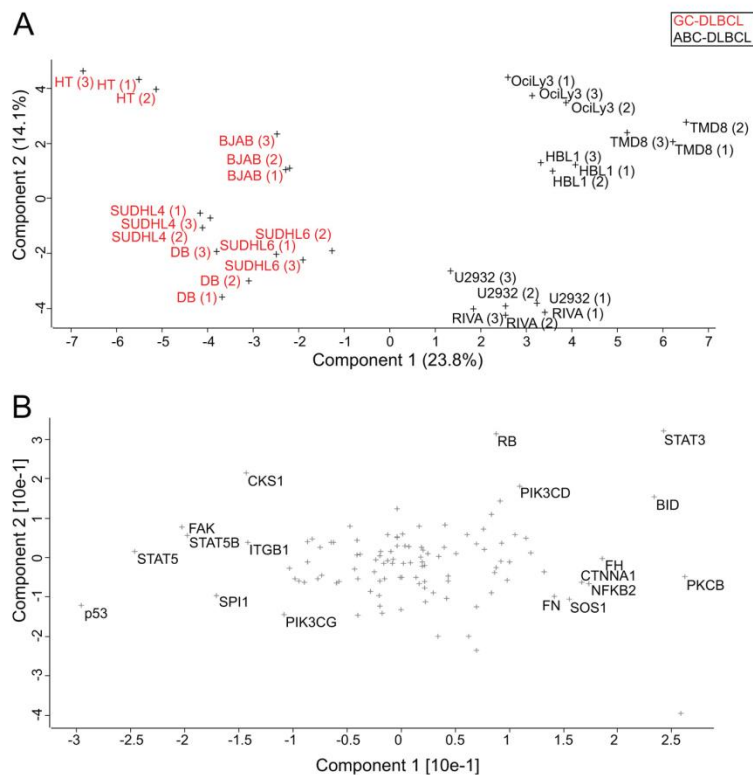


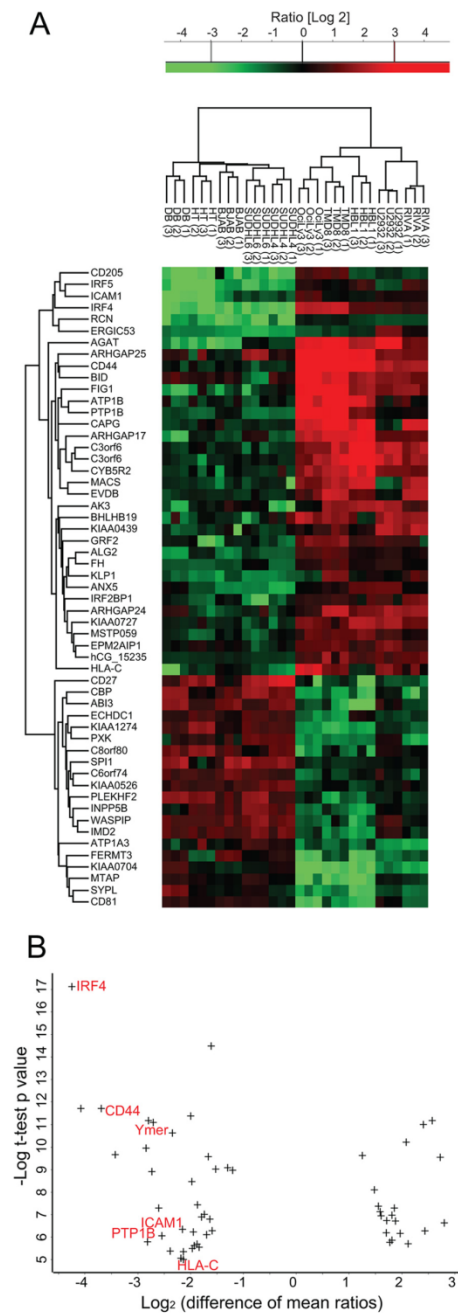
FIG. 5. **Category-based analysis of subtype differences.** A, PCA of 10 lymphoma cell lines after filtering for proteins annotated by KEGG to be involved in cancer (KEGG category: pathways in cancer). B, loadings of PCA in A.

Category-based Analysis of Subtype Differences—The above analyses were global and unbiased in that they considered the entire proteome. To determine whether specific groups of proteins by themselves could differentiate the subtypes, we extracted the proteins belonging to specific KEGG categories from the quantified proteome. We then performed PCA analysis on this subset as described above. Interestingly, the category “pathways in cancer” (108 quantified proteins) was able to clearly separate the groups, albeit to a lesser degree than the full proteome (Fig. 5A). The strongest loadings preferential for GCB in this category were p53, STAT5, STAT5B, and SPI1/PU.1 (Fig. 5B). SPI1 has a major role in maintaining germinal center B-cells through repressing the expression of plasma cell transcriptional regulators and thus blocking plasma cell differentiation (32). The strongest loadings preferential for ABC included the anti-apoptotic protein BCL2, overexpression of which is a known mechanism by which NF- κ B driven tumors evade apoptosis. Surprisingly, the pro-apoptotic protein BID was also in this group. The loadings preferential for ABC include STAT3. It has been shown that NF- κ B signaling in ABC induces the expression of IL-6 and

IL-10, which act through JAK kinases and STAT3 as autocrine signals (33). The constitutive activity of STAT3 promotes proliferation and cell survival in the ABC subtype (34). This explains the synergistic effect of blocking JAK signaling and NF- κ B signaling in killing ABC cells (33). PKCB is another interesting driver of the ABC subtype because its overexpression is a strong marker for refractory or fatal DLBCL and a recognized drug target (35). Our finding that it is preferentially expressed in the aggressive ABC subtype compared with the GCB subtype may therefore be of clinical interest. The observation that a small group of proteins can separate the subtypes prompted us to search for such groups in the entire quantified data set.

***t* Test Signature**—To identify in a supervised manner a set of proteins that significantly distinguishes the ABC from the GCB subtype, we performed a *t* test between the cell lines using a permutation-based false discovery rate (0.05). This analysis resulted in a set of 55 proteins (Fig. 6A) that strongly segregated the subtypes as seen after PCA analysis (supplemental Fig. 3). Interestingly, cell lines of the GCB subtype collapse into a single cluster, indicating that the proteins most

Super-SILAC Distinguishes Lymphoma Subtypes



strongly differentiating ABC and GCB subtypes do not distinguish different GCB cell lines from each other (variation between cell lines is equal to the variation between replicates). In contrast, replicates of ABC cell lines remained distinguishable, which indicates a larger degree of heterogeneity in ABC-type cell lines as we observed previously.

In the signature obtained by unbiased proteomic analysis, there are at least six proteins whose gene expression levels are already described to be different between the subtypes (ABC-like: IRF4, CD44, and PTP1B; GCB-like: CD27, SPI1, and WIP). Because the total number of signature proteins is small, this already validates our proteomic signature and encouraged us to further investigate the new proteins in our signature. These include the recently described GTPase Speckled-like pattern in the germinal center (SLIP-GC), whose expression is limited to germinal center B-cells and to lymphomas derived from the germinal center including diffuse large B-cell lymphomas (36). This finding supports the potential use of SLIP-GC as a potential marker that can be used to differentiate the two subtypes from each other. Another member of the signature set is the surface marker CD81, which has also very recently been reported to be highly expressed in normal germinal B-cells. Further assessment of the role of this cell surface marker in the risk stratification of patients with DLBCL has already been recommended (37). A further novel protein that has a higher expression level in our GCB-DLBCL signature is the signaling adaptor Cbp/PAG. In B-cell non-Hodgkin lymphomas, PAG and Lyn kinase constitute the core of an oncogenic signalosome that results in proliferative and pro-survival signals. The Lyn and PAG signalosome can interact with downstream kinases to mediate these signals in different lymphoma cell lines (38). Our finding that PAG is up-regulated in GCB suggests investigating the modality associated with PAG in this subtype in particular. Ymer, a protein that we previously identified as an effector of EGF signaling (10, 39), is relatively up-regulated in the ABC subtype. Ymer is also known as CCDC50 or C3orf6, and although not studied in the context of DLBCL, this protein has been shown to be required for cell survival in chronic lymphocytic leukemia and mantle cell lymphoma cells (40). It is involved in the control of NF- κ B signaling, which is a characteristic of the ABC subtype where it is up-regulated (15). Therefore, in addition to validating differentiators known from gene expression profiling, our proteomic signature reveals a novel set of proteins,

FIG. 6. t test signature. A, t test analysis of the proteins from the two groups of cell lines resulted in a signature of 55 proteins that are most significantly different. The panel depicts a heat map of the ratios of these proteins after clustering. B, plot of the difference of mean ratios versus the significance of signature proteins. The proteins on the left are significantly up-regulated in the ABC relative to GCB subtype. The protein names highlighted in red indicate NF- κ B regulated genes.

Super-SILAC Distinguishes Lymphoma Subtypes

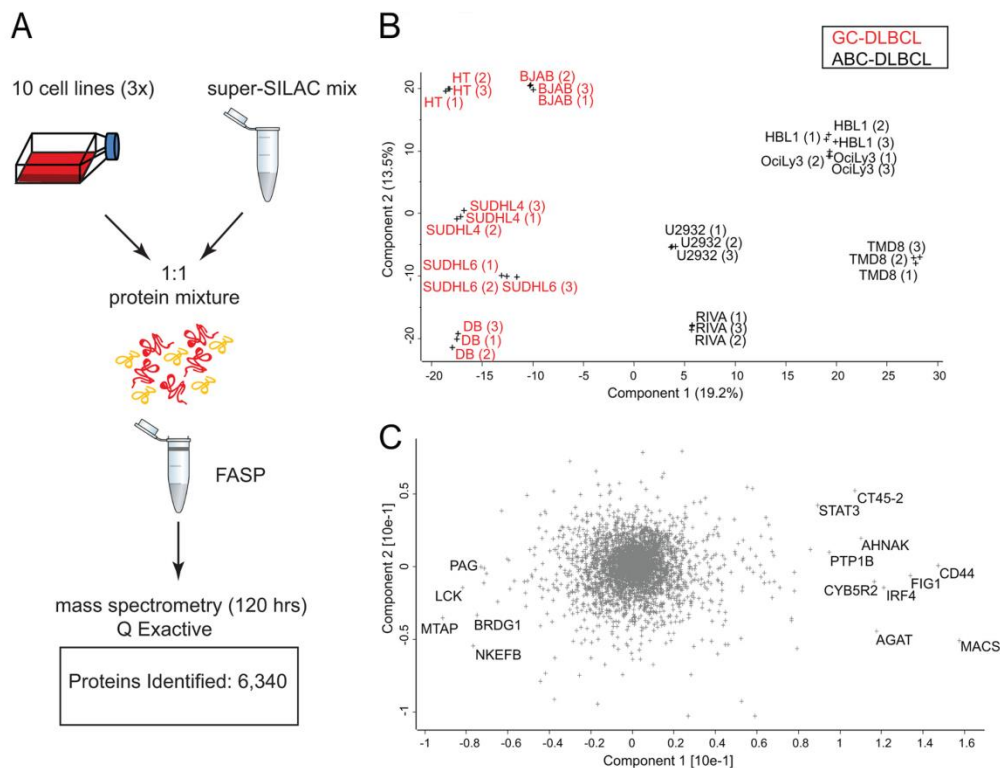


Fig. 7. Single-shot proteome measurements to distinguish ABC from GCB. A, unfractionated, FASP-processed peptide mixtures were directly loaded onto a relatively long column (50 cm) after StageTipping. The proteomes were analyzed in triplicates in 4-h runs by an UHPLC (EASY nLC 1000) system coupled to a benchtop quadrupole Orbitrap mass spectrometer (Q Exactive). B, principal component analysis of the single-shot measurements. C, loadings of PCA in B highlighting the proteins that strongly drive the segregation in PCA component 1.

some of which have been shown to be involved in lymphomagenesis and might be of clinical relevance.

In an attempt to identify annotated protein categories that are significantly and exclusively up-regulated in one lymphoma subtype relative to the other, we plotted the difference of mean ratios *versus* the statistical significance of our signature proteins. This revealed that all NF- κ B-regulated proteins in the signature are significantly up-regulated in ABC relative to GCB subtype (Fig. 6B). They included IRF4, CD44, Ymer, PTP1B, ICAM, and HLA-C. Thus, our proteomic findings, similar to previous results of gene expression profiling, are consistent with high NF- κ B activity in ABC-DLBCL as a hallmark that distinguishes this subtype from GCB-DLBCL.

BCR signaling has been shown to play an important role in lymphomagenesis where malignant B-cells exploit the normal regulatory roles of this pathway for their own purposes (41). We extracted proteins from our data set that are KEGG annotated to be involved in BCR signaling and found that we covered all but eight proteins in this category (59 of 67;

supplemental Fig. 2). To better investigate small but reproducible protein changes, we normalized their expression levels by z-scoring across replicates and cell lines. Taking the median values for every subtype revealed four large clusters. The proteins highlighted in red in supplemental Fig. 4A consist of the BCR signaling proteins that exhibit the largest expression differences between the two subtypes and that are higher in the ABC subtype. Interestingly, this cluster included NF- κ B1, as well as the two upstream regulatory proteins MALT1 and CARD11 (supplemental Fig. 4B). This is consistent with the role of the multiprotein CARD11-BCL10-MALT1 (CBM) complex in driving the constitutive NF- κ B activity in the ABC subtype (17–19). Conversely, the proteins (highlighted in green in supplemental Fig. 4A) are BCR signaling proteins that are higher in GCB (supplemental Fig. 4C).

Rapid Lymphoma Classification in Single-shot Runs— Above, we have demonstrated that quantitative proteomics can readily segregate cell lines derived from patients in a robust manner. However, sample amounts and measurement

Super-SILAC Distinguishes Lymphoma Subtypes

time of our workflow (Fig. 1) would still be an obstacle to clinical application. We therefore wanted to investigate the possibility of making the approach more practical by reducing the measurement time and the amount of sample consumed. Taking advantage of the higher speed and sensitivity of the newly introduced quadrupole Orbitrap mass spectrometer (Q Exactive) (27), we investigated whether we could reach the depth required to segregate the cell lines in a single-shot experiment, that is, without fractionation. The samples were prepared as before, except that FASP-prepared peptides were directly loaded on StageTips and eluted into the autosampler device. Single 4-h gradient runs were performed in triplicates for each of the 10 cell lines, and data files were processed together in MaxQuant. This resulted in the identification of 6,340 proteins and the quantification of 4,611 in at least two replicates of the same cell line (Fig. 7A) (supplemental Tables VI and VII). Filtering for 50% valid values resulted in 3,566 quantified proteins. Upon PCA analysis of the single measurements, we obtained a similar segregation of the two subtypes, and the loadings responsible for the PCA segregation showed a very strong overlap with the previously obtained loadings (Fig. 7B). Interestingly, the data obtained from singlets was sufficient to segregate the two subtypes. This shows that single-shot measurements can reach the depth required for robust separation of lymphoma subtypes, opening up for the analysis of several patient samples per day with sample requirements in the low microgram range.

Conclusion and Outlook—Here we have shown that high accuracy, quantitative proteomics based on a super-SILAC approach can robustly segregate closely related cancer subtypes directly at the level of expressed proteins. We developed and used a super-SILAC mix of labeled B-cell lymphoma cell lines as an internal standard to segregate subtypes of DLBCL. We selected the cell lines with the most distinct protein expression profiles to obtain the best coverage of different lymphoma-specific proteins. The mix was spiked into five ABC-DLBCL and five GCB-DLBCL cell lines, which allowed robust, unsupervised segregation of these two histologically indistinguishable lymphomas based on their protein expression profiles. We found that requiring protein quantification values to be present in half of the samples and replicates and imputing the remaining values led to the most robust segregation. The data also revealed a protein expression signature that differentiates the two subtypes. This signature confirmed known markers previously discovered by gene expression studies and highlighted novel ones. Interestingly, our straightforward PCA analysis of the proteome differences revealed proteins such as IRF4, CD44, STAT3, PTP1B, and CD27 as the strongest differentiators between subtypes. The fact that these and a number of other proteins, which all have a strong biological rationale to drive subtype differences, emerge as the top hits in an unbiased analysis, is very intriguing. Furthermore, unbiased and supervised segregation revealed a number of novel

proteins, which can now be studied for their involvement in these lymphoma subtypes.

To our knowledge, this is the first high accuracy, quantitative proteomics study that unequivocally classified tumor cell lines on par with microarray-based methods. This ability of the super-SILAC proteomic approach to readily segregate between tumor subtypes now opens up the possibility of employing proteomics in many situations that have previously been studied with transcript-based approaches. Toward this goal, we already combined the super-SILAC quantitative approach with single-shot measurements on a benchtop quadrupole Orbitrap instrument. These measurements attained the depth and accuracy required to segregate the two subtypes as exemplified by a number of representative cell lines. Considering the significant reduction in measuring time and in required sample amount, it is conceivable that this workflow could be employed in routine settings to answer practical clinical questions such as tumor classification or drug efficacy.

Acknowledgments—We thank Daniel Krappmann, Berit Jungnickel, Ralf Küppers, Martin Janz, Stephan Mathas, and Georg Lenz for the provision of cell lines. We thank Tamar Geiger and Maria Robles (Charo) for helpful discussions.

The acquired raw data for the fractionated proteome is uploaded to Tranche (<https://proteomecommons.org/tranche/>) with the hash code: 1b9WXUuYMjCxiCk40HNWo3YM+ xvP31WEgYg6RWM-jm5295exN5kmoMbyekZXqMvXpG8rvLgDELXBKcceeScd+R2WqDkAAAAAACQyQ==.

* This work was supported by European Commission's 7th Framework Program PROSPECTS Grant Agreement HEALTH-F4-2008-201648 and a Deutsche Krebshilfe Onconet2 grant. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ This article contains supplemental materials.

¶ Supported by Deutsche Forschungsgemeinschaft Grant TRR54.

|| To whom correspondence should be addressed. Tel.: 49-89-8578-2557; E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Quackenbush, J. (2006) Microarray Analysis and Tumor Classification. *New Engl. J. Med.* **354**, 2463–2472
2. McDermott, U., Downing, J. R., and Stratton, M. R. (2011) Genomics and the Continuum of Cancer Care. *New Engl. J. Med.* **364**, 340–350
3. Hanash, S., and Taguchi, A. (2010) The grand challenge to decipher the cancer proteome. *Nat. Rev. Cancer* **10**, 652–660
4. Choudhary, C., and Mann, M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11**, 427–439
5. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207
6. Nilsson, T., Mann, M., Aebersold, R., Yates, J. R., 3rd, Bairoch, A., and Bergeron, J. J. (2010) Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nat. Methods* **7**, 681–685
7. Cox, J., and Mann, M. (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80**, 273–299
8. Vermeulen, M., and Selbach, M. (2009) Quantitative proteomics: A tool to assess cell differentiation. *Curr. Opin. Cell Biol.* **21**, 761–766
9. Mallick, P., and Kuster, B. (2010) Proteomics: A pragmatic perspective. *Nat. Biotechnol.* **28**, 695–709
10. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H.,

- Pandey, A., and Mann, M. (2002) 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
11. Mann, M. (2006) Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7**, 952–958
 12. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7**, 383–385
 13. Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., Kruger, M., Ishihama, Y., and Mann, M. (2011) Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nat. Protoc.* **6**, 147–157
 14. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511
 15. Davis, R. E., Brown, K. D., Siebenlist, U., and Staudt, L. M. (2001) Constitutive nuclear factor κ B activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. *J. Exp. Med.* **194**, 1861–1874
 16. Davis, R. E., Ngo, V. N., Lenz, G., Tolar, P., Young, R. M., Romesser, P. B., Kohlhammer, H., Lamy, L., Zhao, H., Yang, Y., Xu, W., Shaffer, A. L., Wright, G., Xiao, W., Powell, J., Jiang, J. K., Thomas, C. J., Rosenwald, A., Ott, G., Muller-Hermelink, H. K., Gascoyne, R. D., Connors, J. M., Johnson, N. A., Rimsza, L. M., Campo, E., Jaffe, E. S., Wilson, W. H., Delabie, J., Smeland, E. B., Fisher, R. I., Braziel, R. M., Tubbs, R. R., Cook, J. R., Weisenburger, D. D., Chan, W. C., Pierce, S. K., and Staudt, L. M. (2010) Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* **463**, 88–92
 17. Ngo, V. N., Davis, R. E., Lamy, L., Yu, X., Zhao, H., Lenz, G., Lam, L. T., Dave, S., Yang, L., Powell, J., and Staudt, L. M. (2006) A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441**, 106–110
 18. Ferch, U., Kloos, B., Gewies, A., Pfänder, V., Düwel, M., Peschel, C., Krapmann, D., and Ruland, J. (2009) Inhibition of MALT1 protease activity is selectively toxic for activated B cell-like diffuse large B cell lymphoma cells. *J. Exp. Med.* **206**, 2313–2320
 19. Hailfinger, S., Lenz, G., Ngo, V., Posvitz-Fejfar, A., Rebeaud, F., Guzzardi, M., Penas, E. M., Dierlamm, J., Chan, W. C., Staudt, L. M., and Thome, M. (2009) Essential role of MALT1 protease activity in activated B cell-like diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19946–19951
 20. Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V. A., Grunn, A., Messina, M., Elliot, O., Chan, J., Bhagat, G., Chadburn, A., Gaidano, G., Mullighan, C. G., Rabadan, R., and Dalla-Favera, R. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.* **43**, 830–837
 21. Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362
 22. Wiśniewski, J. R., Zougman, A., and Mann, M. (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* **8**, 5674–5678
 23. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670
 24. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8**, 2759–2769
 25. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712
 26. Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Frohlich, F., Cox, J., and Mann, M. (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10**, M110.003699
 27. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wiegand, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, M111.011015
 28. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372
 29. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805
 30. Lu, X., Malumbres, R., Shields, B., Jiang, X., Sarosiek, K. A., Natkunam, Y., Tiganis, T., and Lossos, I. S. (2008) PTP1B is a negative regulator of interleukin 4-induced STAT6 signaling. *Blood* **112**, 4098–4108
 31. Lu, X., Nechushtan, H., Ding, F., Rosado, M. F., Singal, R., Alizadeh, A. A., and Lossos, I. S. (2005) Distinct IL-4-induced gene expression, proliferation, and intracellular signaling in germinal center B-cell-like and activated B-cell-like diffuse large-cell lymphomas. *Blood* **105**, 2924–2932
 32. Schmidlin, H., Diehl, S. A., Nagasawa, M., Scheeren, F. A., Schotte, R., Uittenbogaart, C. H., Spits, H., and Blom, B. (2008) Spi-B inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression. *Blood* **112**, 1804–1812
 33. Lam, L. T., Wright, G., Davis, R. E., Lenz, G., Farinha, P., Dang, L., Chan, J. W., Rosenwald, A., Gascoyne, R. D., and Staudt, L. M. (2008) Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor- κ B pathways in subtypes of diffuse large B-cell lymphoma. *Blood* **111**, 3701–3713
 34. Ding, B. B., Yu, J. J., Yu, R. Y., Mendez, L. M., Shaknovich, R., Zhang, Y., Cattoretti, G., and Ye, B. H. (2008) Constitutively activated STAT3 promotes cell proliferation and survival in the activated B-cell subtype of diffuse large B-cell lymphomas. *Blood* **111**, 1515–1523
 35. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuber, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74
 36. Richter, K., Brar, S., Ray, M., Pisitkun, P., Bolland, S., Verkoczy, L., and Diaz, M. (2009) Speckled-like pattern in the germinal center (SLIP-GC), a nuclear GTPase expressed in activation-induced deaminase-expressing lymphomas and germinal center B cells. *J. Biol. Chem.* **284**, 30652–30661
 37. Luo, R. F., Zhao, S., Tibshirani, R., Myklebust, J. H., Sanyal, M., Fernandez, R., Gratzinger, D., Marinelli, R. J., Lu, Z. S., Wong, A., Levy, R., Levy, S., and Natkunam, Y. (2010) CD81 protein is expressed at high levels in normal germinal center B cells and in subtypes of human lymphomas. *Human Pathol.* **41**, 271–280
 38. Taubin, S., Ding, H., Burdette, D., Borisch, B., and Hoessli, D. C. (2011) Membrane-associated signaling in human B-lymphoma lines. *Exp. Cell Res.* **317**, 151–162
 39. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **308**, 1472–1477
 40. Farthing, A., Engel, F., Seifert, M., Hartmann, E., Ott, G., Rosenwald, A., Stilgenbauer, S., Döhner, H., Boutsos, M., Lichter, P., and Pscherer, A. (2009) Gene knockdown studies revealed CCDC50 as a candidate gene in mantle cell lymphoma and chronic lymphocytic leukemia. *Leukemia* **23**, 2018–2026
 41. Lenz, G., and Staudt, L. M. (2010) Aggressive lymphomas. *New Engl. J. Med.* **362**, 1417–1429

4. A very deep and high-accuracy phosphoproteome reveals fundamental differences between tyrosine and serine/threonine phosphorylation events (Article 3).

Protein phosphorylation, the most extensively studied PTM, is often utilized by signaling systems for transient alteration of protein function ranging from regulation of enzyme activity, protein interactions and protein conformations to targeted destruction of proteins. Despite being one of the most studied PTMs, the total number of the phosphorylation events that can occur in a cell and their location in protein sequences is not known in its entirety.

Using high resolution mass spectrometry and very high statistical stringency we mapped more than 50,000 distinct phosphorylated peptides in a single cell type across a number of cellular conditions allowing us to analyze global properties of this very large dataset. By combining protein abundance measurements with phosphorylation changes across mitosis and epidermal growth factor (EGF) stimulation, we determined for the first time the occupancy of thousands of phosphorylation sites using label-free quantification.

In contrast to common estimates, we found that at least three-fourths of the proteome can be phosphorylated. Our analysis revealed that phosphorylation events on tyrosine tend to be present on more abundant proteins in comparison to serine and threonine sites. The observed difference in phospho-protein abundance correlated with the substrate K_m values of tyrosine kinases.

This manuscript is in preparation for submission.

A very deep and high-accuracy phosphoproteome reveals fundamental differences between tyrosine and serine/threonine phosphorylation events

Kirti Sharma, Rochelle C.J D'Souza, Stefka Tyanova, Jacek R. Wiśniewski, Juergen Cox and Matthias Mann

Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry Am Klopferspitz 18, D-82152 Martinsried, Germany

Summary

Despite being one of the most studied post translational modifications, the total number and site-specific localizations of phosphorylation events that occur in a cell are not known. Using high resolution mass spectrometry and high statistical stringency we mapped more than 50,000 distinct phosphorylated peptides in a single cell type across a number of cellular conditions. Strikingly and in contrast to commonly cited estimates, we find that least three-quarters of cellular proteins can be phosphorylated. As depth of coverage increased, phospho-tyrosine becomes proportionately less of the total phosphoproteome, whereas the increase in ser/thr sites only appeared to saturate for technical reasons. Inhibition of tyrosine phosphatases *in vivo* resulted in deep coverage of phosphotyrosine sites and unexpectedly these phosphorylation events were enriched on higher abundance proteins. We find that the observed difference in phospho-proteins abundance correlates with the substrate K_m values of tyrosine kinases. Additionally, we demine phosphorylation site specific occupancies from label free quantification data. Our findings demonstrate that phosphotyrosine is a separate functional regulatory post-translational modification of eukaryotic proteomes.

Introduction

Different components of signaling pathways act in concert as ‘writers’, ‘readers’ and ‘erasers’ of post translational modifications (PTMs) for transmitting cellular regulatory information (Hunter, 2009). The first essential step in understanding the complex molecular circuitry behind cellular signal transmission is to find practicable methods for measuring the extent and nature of the PTMs that occur in a cell. For this purpose, modern quantitative mass spectrometry (MS) has proven to be an ideal platform because it is a highly precise yet generic method for the identification and quantification of proteins and their modifications in an unbiased and global manner (Choudhary and Mann, 2010). Protein phosphorylation, one of the most extensively studied PTM, is often utilized by signaling systems for transient alteration of protein function ranging from regulation of enzyme activity, protein interactions and protein conformations to targeted destruction of proteins. It occurs primarily on serine, threonine, and tyrosine residues which represent approximately 17% of the total amino acid content in an average protein (Echols et al., 2002). Based on this fact it has been calculated that in an average eukaryotic cell, there exist nearly 700,000 different potential phosphorylation sites (Ubersax and Ferrell, 2007). Recent MS based studies have reported the identification of tens of thousands of phosphorylation sites in tissues and cultured cells (Humphrey et al., 2013; Lundby et al., 2012; Zhou et al., 2013). The increasing number of the identified phosphosites and sheer number of potential ones across the proteome raises fundamental question about the scale of the phosphoproteome and the biological relevance of measured phosphorylation events.

The vast majority of cellular protein phosphorylation events reported in MS based studies occurs on serine and threonine residues while phosphotyrosines generally account for <1% of these events. Many years ago the S/T/Y phosphorylation ratios had already been estimated as 90:10:0.05 (Hunter and Sefton, 1980). Despite the relatively large number of tyrosine kinases there are several reasons for proportionately fewer tyrosine phosphorylation events (Hunter, 2009). Firstly, most tyrosine kinases are only activated in specific circumstances and otherwise remain stringently negatively regulated. Secondly, unless protected by binding to SH2/PTP domains (Sadowski et al., 1986), phosphotyrosine residues have a very short half-life owing to

high activity of phosphotyrosine phosphatases (PTPs). Finally, unlike phospho-serine/threonine, phospho-tyrosine rarely plays a structural role in proteins, and is primarily regulatory. Phosphotyrosine based signaling pathways seem to be a comparatively recent molecular innovation system in the evolutionary history of living organisms and appear to be a hallmark of more complex organisms (Lim and Pawson, 2010). With increasing organismal complexity, the number of protein tyrosine kinases increases. Furthermore, species that have more tyrosine kinases have proportionally less tyrosines encoded in their genome (Tan et al., 2009). This may reduce the amount of possible nonfunctional or deleterious phospho-tyrosine events and therefore reduces noise and enhances regulation in phospho-tyrosine-dependent signaling systems. However, no specific observation about tyrosine phosphorylation at the proteome level has been made.

Here, we present an analysis of the nature of the phosphoproteome to great depth by combining protein abundance measurements with phosphorylation changes across mitosis and epidermal growth factor (EGF) stimulation. We employed a highly stringent approach towards phosphopeptide identification and phosphosite localization. For the first time, we use a label free approach to quantify the modified peptides and determine their fractional occupancy. Our parallel and in-depth investigation of proteome and phosphoproteome resolved to serine, threonine and tyrosine residues revealed many interesting aspects of protein phosphorylation as a cellular control process and its potential crosstalk with other modifications. Our findings suggest that phosphotyrosine is a separate functional regulatory post-translational modification of eukaryotic proteomes. While our data appear to represent close to complete coverage of tyrosine phosphorylation, the serine/threonine phosphoproteome is accompanied by a large number of variable 'background events' and is probably far from completeness even though we do cover a core serine/threonine phosphoproteome.

Results

Protein and phosphopeptide identification from a cancer cell line

While the proteome of model organisms has been identified and quantified essentially to completion (Beck et al., 2011; de Godoy et al., 2008; Nagaraj et al., 2011), the total number of

the phosphorylation events that occur in these organisms is not yet known. To survey proteome and phosphoproteome at maximum possible depth, HeLa S3 cells were left either untreated, mitotically arrested, pervanadate treated or stimulated with EGF for five or fifteen minutes (Fig1A). A double thymidine block in combination with nocodazole arrest was used to obtain a mitotic phase population and synchronization of HeLa cells was confirmed by fluorescence-activated cell sorting (FACS) analysis (Fig1A, left panel). Epidermal growth factor stimulation of cells was confirmed by increased phosphorylation of the activation loop Thr/Tyr residues of MAPK1 and MAPK3. For robust statistics, we analyzed biological quadruplicates for each cellular condition and six replicates for control cells (untreated and asynchronous). For each replicate analysis, HeLa cell lysates were digested in solution with trypsin using the filter-aided sample preparation method with endoproteinase Lys-C and trypsin (Wisniewski et al., 2009b) (Fig. 1B). For global phosphoproteome analysis, resulting peptides were subjected to SCX based fractionation and phosphopeptide enrichment by titanium dioxide beads (Macek et al., 2009; Zhou et al., 2011). To obtain a deep coverage of tyrosine phosphorylation, we performed immuno-affinity enrichment of tyrosine phosphorylated peptides from HeLa phosphopeptides (Kettenbach and Gerber, 2011) obtained from cells in which tyrosine hyperphosphorylation was induced by inhibition of phosphatases by pervanadate. In addition, tyrosine phosphorylated peptides were also enriched from untreated, mitotic and EGF stimulated cells where early EGF signaling (5 min) is known to be dominated by Tyr phosphorylation events (Fig1A, right panel). For quantitative analysis of the corresponding proteomes, 25 µg of peptides were fractionated by pipette tip-based strong anion exchange chromatography into six fractions (Wisniewski et al., 2009a).

Subsequently, all peptide and phosphopeptide mixtures were resolved by nanoLC (Fig1C) and measured on a quadrupole-Orbitrap mass spectrometer (Q Exactive) (Michalski et al., 2011), where both MS and MS/MS spectra were recorded in the Orbitrap analyzer with high-resolution using the HCD technology (Nagaraj et al., 2010). This resulted in parts-per-million-range mass accuracy for both precursor peptides (Fig1C, right panel) and their fragments, which aided unambiguous identification of peptides and sites. The total data set over proteome and phosphoproteome replicates comprises 273 LC-MS/MS experiments with 4h or 2h gradients

with a total measuring time of about 40 days during which about 20 million MS/MS scans were acquired. The observed average absolute mass deviation for the corresponding precursor ions was 500 ppb.

A stringent data analysis pipeline for phosphoproteomics

In Figure 2 we provide an overview of the computational workflow that is utilized in MaxQuant for the identification and positioning of amino acid modifications. Here we apply this analysis pipeline phosphorylation but it is in principle applicable to all posttranslational modifications and modifications introduced during sample preparation.

Prior to the peptide identification, MS/MS spectra are processed in order to improve the quality of the search engine results. First, isotope patterns are detected in the MS/MS spectra as in the conventional MaxQuant spectral processing pipeline. Fragment peaks that are assigned to isotope patterns are collapsed to the monoisotopic peak. Furthermore, multiply charged isotope patterns are transformed to charge state one. Fragment peaks that are not assembled into isotope patterns are left unchanged. If multiple peaks fall into a mass tolerance window of 20 ppm they are combined into a single peak with the summed intensity and the intensity weighted average mass. Precursor masses and charge states are calculated according to the standard MaxQuant workflow. In particular m/z and retention time dependent mass calibration is performed on the precursor masses with the help of a ‘first search’ approach using the Andromeda peptide search engine, which is part of MaxQuant.

To identify modified peptides we perform an Andromeda search with individualized mass tolerances for the precursor ions. The best-scoring peptide-spectrum match (PSM) is retained for every spectrum. These are further filtered for an Andromeda score of 40 and an Andromeda delta score of 17. The delta score measures the difference between the best spectrum match and the next best match with a different amino acid sequence.

Posterior error probability (PEP) is used to rank best PSMs and control the false discovery rate on PSM level. The PEP is calculated based on a peptide level target-decoy strategy, where the decoy part is generated according to the newly developed ‘reward’ strategy. Reward stands for

a combination of reverse and forward. One half of the peptide is reversed with respect to the original peptide while the other half is retained. The amino acids which cause the cleavage – in this case lysine and arginine – remain in place. This mixed forward-reverse strategy makes sure that the decoy peptides, while being false positives, have at the same time partial matches to homologous sequences in the forward proteome. The PEP is adjusted in peptide classes with specific combinations of number of modifications per peptide, number of missed cleavages and charge. A distribution of search engine scores of identified phosphopeptides with a minimum score cutoff of 40 is shown in Figure S1. A typical spectrum and its annotation for a low scoring phosphopeptide just above the filtering threshold shows that there are sufficient peaks to support phosphopeptide identification even in such a case. Using a recently developed “Expert system” for computer assisted annotation of MS/MS spectra based on literature derived rules we further increased confidence in the identification of phosphopeptides as it allows annotation of previously unannotated peaks (Neuhauser et al., 2012). . The expert system based additional peak annotation becomes particularly beneficial with high mass-accuracy spectra acquired in modern MS instrumentation, as a very high percentage of fragment ions can now be accounted for.

For the localization of the modification site we iteratively explore all possible distributions of the modifications on the peptide onto the amino acids capable of carrying this modification- here serine, threonine and tyrosine. For each such positioning an Andromeda score is calculated and exponentiated to obtain a probability. The localization score for a site is then defined as the normalized sum of the probabilities for the cases where the site is carrying a modification as shown in Figure 2C.

Site occupancies are calculated based on similar principles as were previously applied to SILAC cell cycle data (Olsen et al., 2010), but without SILAC ratios. The advantage of label-free quantification with multiple conditions is that for the occupancy calculation in one sample one can select a reference sample that is most suitable for the calculations involved in the formula (‘Proportion’ in Figure 2D). Problematic reference samples are those with equal or very similar expression compared to the sample of interest. These lead to near ‘0/0’-cancellations in the

formula which would result in large statistical errors. In the case of label free quantification over many samples, instead, the reference sample can be chosen such as to bring the formula as far away as possible from this singularity. We define an occupancy error estimate based on error propagation, which can be used to filter out occupancies for which the value is expected to be unreliable due to closeness to the singularity.

Coverage of Proteome and Phosphoproteome in HeLa cells

Following analysis with the MaxQuant software environment (Cox and Mann, 2008), 43% of the acquired MS/MS scans were unambiguously identified. This led to 145,340 identified non-redundant peptide sequences and 10,826 proteins (or protein groups) at a protein false discovery rate of less than 1%. Applying the Andromeda score and delta score filters for modified peptides as described above, we identified more than 50,000 unique phosphopeptides corresponding to more than 40,000 phosphorylation events from 8,014 proteins that could be localized with high confidence to single amino acid sequence locations (Fig. 3A, left panel). This number constitutes a lower limit on the size of a human cancer cell line phosphoproteome. Based on the observed evidence of phosphorylation on these proteins, we show that in contrast to commonly cited estimates (Hunter, 2007), at least three-quarters of the proteome can be phosphorylated. Our very deep phosphoproteome dataset measured in conjunction with proteome abundance is a valuable resource for further system level analysis of phosphorylation events. In addition, it also serves as a catalogue for regulatory protein expression and phosphorylation events associated with mitotic and growth factor signaling. Tables with all identified proteins, phosphopeptides and phosphosites are provided in Supplementary Data available online (table S1-S3). In addition, all the identified phospho-sites, phospho-peptides and phospho-proteins are easily accessible through the MaxQB database (Schaab et al., 2012). When comparing site specific phosphorylation events in different replicate analyses, we found that more than 22,000 phosphosites of the 40,000 sites (55%) were identified in all the biological conditions (termed ‘core phosphoproteome’) and 20% were found exclusively in the mitotic stage (Fig. 3A, right panel). Given the array of regulatory events that control this stage of cell cycle the preponderance of sites in this stage is not surprising.

Our dataset contained about 50% of phosphosites usually followed by antibodies in cellular signaling research based on the PhosphositePlus database and antibodies available from Cell Signaling Technology (CST) (Fig. 3B, left panel). Since protein phosphorylation is tightly regulated, a fifty percent coverage of functional phosphosites across only three biological states and in a single cell line implies that our analysis reached considerable depth. Further, we found the sites usually followed by antibodies present to be enriched on more abundant proteins with median intensity about half an order of magnitude higher than those of all identified sites (Fig. 3B, middle panel). Next, we represented the coverage of these 261 functional phosphosites ranked by their abundance as a function of phosphopeptide signal intensity. While they span the entire dynamic range, 90% of the functional phosphopeptides are present within four orders of magnitude. This suggests that the dynamic range achieved in our phosphoproteome is in principle sufficient to sample most of the cellular phosphoproteome.

Label-free quantification of proteome and phosphoproteome changes during growth factor and mitotic signaling

Our dataset contains accurate quantification information of many of the key phosphosites that are known to be activated by epidermal growth factor (EGF) treatment. For instance the activating phosphorylations on tyrosine 1197 and tyrosine 1172 on epidermal growth factor receptor (EGFR) that are upstream to activation of RAS/RAF/MAPK signaling show the correct profile (Fig. 4A, upper panels and table S3) as the phosphorylation of these sites significantly increased in cells treated with either EGF or pervanadate. Similarly, we observed inhibiting phosphorylation on threonine14/tyrosine15 on cyclin dependent kinase 1/2/3 and activating phosphorylation of polo like kinase-1 (PLK1) only in the mitotic samples (Fig. 4A, upper panels). To extract phosphosites significantly regulated upon either EGF treatment or mitotic arrest, we subjected the normalized dataset to a stringent multiple-sample ANOVA test using a permutation based FDR of 0.01. This allowed the identification of regulated phosphorylation sites (table S3), which we then clustered based on their profiles (Fig. 4A, Heatmap). We obtained cluster specific footprints of kinase activation using a Fisher exact test for kinase-

substrate motifs. CDK1, proline directed substrates and polo box domain containing motifs showed the strongest enrichments in the regulated dataset that was specifically up-regulated in the mitotic arrest group in comparison to the total quantified dataset (Fig. 4B). Analysis of GO annotation terms and KEGG pathways revealed strong and highly significant enrichment of categories that are characteristic of mitosis - regulation of exit from mitosis, septin ring assembly and kinetochore organization (Fig. 4A, table S3).

To functionally understand the differences in protein expression levels between two phosphoproteomes, one can divide a histogram of quantitative expression changes into quantiles. Each quantile can then be tested separately for enrichment of every annotation term (such as KEGG pathways and GO biological processes) to find whether the corresponding expression values have a preference to be systematically larger or smaller than the global distribution of the values for all proteins (Pan et al., 2009). Here we used a generalized version of the above quantile analysis, termed 1D enrichment analysis (Cox and Mann, 2012). Although similar to the quantile-based enrichment calculations, the 1D annotation enrichment employed here to analyze the functional differences between mitotically arrested and asynchronous cell population has the advantage that it is not necessary to define an arbitrary positioning of quantiles or regulation cutoffs beforehand. Instead, the distribution of values is scanned for interesting sub-categories in an unbiased way. For those categories that are significant, a position score (termed 's'; a number between -1 and 1) is calculated indicating where the center of the distribution of values for the protein category is located relative to the overall distribution of values. A value near 1 indicates that the protein category is strongly concentrated at the high end of the ratio distribution while a value near -1 means that the values are all at the low end of the distribution. This analysis revealed a significant enrichment of terms such as cell cycle phase, cell cycle process G2 phase/ mitotic cell cycle and establishment of chromosome localization with a positive score indicating increased phosphorylation on member phosphoproteins (Fig 4A, lower panel, inset).

Phosphosite occupancy across biological conditions

Large-scale phosphoproteomics has successfully been used for the relative quantitation of phosphorylation sites between different cellular states. As described above, our measurements of the phosphorylated peptide, its unmodified counterpart, and the corresponding total protein abundance enabled the determination of absolute stoichiometry of phosphosites. This enabled determination of site occupancies for 8,264 phosphorylation sites with high confidence, where calculated fractional occupancy is represented on a scale of 0 to 1 (table S4). In accordance with SILAC cell cycle data (Olsen et al., 2010), we determined a high fractional occupancy for sites in mitosis (Fig. 4B, upper panel). For half of these serine/threonine sites we determined a fractional occupancy of 75% or higher. In contrast, most EGF treated and control phosphosites show less than 25% fractional occupancy (Fig. 4B, upper panel, Fig. S2). We reasoned that the high occupancy sites observed during mitosis should make them detectable even without any phospho-enrichment. This was indeed the case, as we identified more than 2,000 phosphosites in the total proteome measurements of mitotic cells in contrast to tens of low occupancy phosphosites identified in EGF stimulated proteome measurements.

The in-depth coverage of phosphotyrosines in our dataset enabled for the first time site stoichiometry estimations for tyrosine specific phosphorylation events on a large scale. We reliably estimated occupancies for 260 tyrosine specific phosphorylation events for at least one cellular condition (Fig. 4B, lower panel). Distribution of occupancies for phospho-Tyr sites across different conditions revealed that most mitotic and control phosphosites have less than 10% fractional occupancy while half of those from EGF treated samples show fractional occupancy of 50% or higher. For pervanadate treated cells, we determined that three quarters of tyrosine phosphorylation events have a fractional occupancy greater than 50%. In general, site occupancies correlated with the cellular signaling state. Based on the distribution of occupancies, we conclude that in untreated cells, phosphorylation tends occur with low fractional occupancy and that tyrosine phosphorylation is maintained at especially low stoichiometric levels in the absence of specific signaling events.

Global properties of protein phosphorylation

Not only the identity but also the relative abundance is an important property of the measured phosphopeptides. The identified phosphopeptides span six orders of magnitude in their MS signals (Fig. 5A, left panel) revealing the sensitivity with which MS based proteomics can now capture these modified peptides. The ranked distribution of all individual phosphopeptides revealed that 70% of the quantified phosphopeptides are contained within only one order of magnitude above or below the median phosphopeptide abundance. We plotted the cumulative contribution of phosphopeptides to determine a total phosphopeptide signal (Fig. 5A, right panel). The 232 most abundant phosphopeptides comprised 25% while the most abundant 3,715 phosphopeptides constituted 75% of the cumulative phosphopeptide signal. These numbers represent the amount of ATP transferred as phosphate moieties onto proteins in the form of Ser, Thr and Tyr modifications and they may be useful for performing system level analysis of proteins, their phosphorylation and the cellular ATP pool.

We first analyzed the distribution of the number of phosphorylation events on each protein and found that about 15% of proteins were phosphorylated on just one residue, while the remaining 85% are phosphorylated at multiple sites (Fig. 5B, left panel). Remarkably, half of the phosphorylated proteins had six or more detected phosphorylation sites. The observed multiple phosphorylation events on so many proteins may either be functional and reflect cross talk downstream of multiple signaling pathways or may also reflect background phosphorylation due to low level kinase activity. Our large-scale proteomic analyses encompassing close to 11,000 proteins enabled the estimation of the abundance of individual proteins in cells. A parallel measurement of phosphoproteome and proteome for each biological condition allowed us to investigate if there is a correlation between a protein's abundance and its propensity to be phosphorylated. When, we analyzed the number of phosphorylated residues identified on a protein as a function of its abundance (Fig 5B, right panel), we observed that there is a tendency for a greater number of identified phosphosites with increasing protein abundance. The observed correlation is low, but highly statistical significant ($p < 1e-13$).

A majority of the more than 40,000 instances of phosphorylation were on serine (81.7%) and threonine (14.8 %) residues as opposed to tyrosine residues (0.4%) (Fig. 5C, left panel), and these proportions are close to classical radioisotope based estimates of this distribution (Hunter and Sefton, 1980). Based on comparison of this distribution to our previous phosphoproteome investigations, we conclude that analyzing phosphoproteome in greater depth results in more identified Ser and Thr phosphorylation events, but proportionally fewer phosphotyrosine sites. To ensure maximal representation of Tyr phosphorylation events in our comparative analysis, apart from the three cellular states (EGF stimulation, mitosis and untreated control cells) we immuno-enriched tyrosine phosphorylated peptides from pervanadate treated cells. Inhibition of tyrosine phosphatases by pervanadate treatment resulted not only in an increase in the number of identified phosphotyrosine residues but also increased their overall abundance by an order of magnitude (Fig. 5C, middle panel). We identified about 2,000 tyrosine phosphorylated peptides on 1,300 phosphoproteins. The distribution of phosphorylation events across S, T and Y changed as we covered ten-fold more phosphotyrosines. A residue resolved comparison of our dataset with PhosphositePlus database (Hornbeck et al., 2012) revealed that while 60% of the identified Ser and Thr phosphorylation events were novel, only 18% were novel for tyrosine residues (Fig. 5C, right panel). Therefore, we conclude that as depth of coverage increases, detectable tyrosine sites are covered relatively rapidly, whereas the increase in serine/threonine sites only appeared to saturate for technical reasons.

Tyrosine versus serine/threonine phosphorylation

To further investigate residue specific features of protein phosphorylation, we compared the abundance distribution of the entire proteome with those of proteins that were found to be phosphorylated on serine/threonine and tyrosine residues, respectively (Fig. 6A). This analysis revealed that proteins carrying phosphotyrosines were significantly more abundant than other proteins ($p < 1e-16$). However, proteins with phosphorylated threonine or serine residues were similar in abundance to the total proteins. We therefore asked whether the observed difference as shown in **Fig. 6A** is just a reflection of the absence of lower abundant tyrosine

phosphorylated peptides in the proteome dataset. However, the new and higher intensity phosphotyrosine peptides identified after *in vivo* inhibition of tyrosine phosphatase (Fig 5C) were also found on more abundant proteins. The abundance of phosphopeptides is determined not only by the abundance of protein but also by the stoichiometry of site phosphorylation. If our data was technically biased in detecting only highly abundant phosphopeptides, we would identify tyrosine phosphopeptides with high fractional occupancies. However, we did not find evidence for such a bias. In fact, the occupancy of tyrosine specific phosphosites was rather determined by the signaling state of the cell. We observed low site occupancy for phosphotyrosine sites in control and mitotic conditions, and when we activated RTK signaling or treated the cells with tyrosine phosphatase inhibitor, higher occupancy sites were clearly evident (Fig 6A, S2). These observations further support our finding that tyrosine phosphorylation on average occurs on more abundant proteins compared to serine/threonine protein phosphorylation. To investigate this further, we analyzed the data by kinase motifs. Based on sequence motifs, again the abundance of Tyr kinase substrates are more abundant than those of serine/threonine kinases (Fig. 6B). A possible explanation for this very clear observation could be the difference in the substrate K_m values, an important factor defining activity of a kinase. To ensure efficient phosphorylation, substrates of a kinase should be present at concentrations above K_m . The abundance of the phospho-Tyr and phospho-serine/threonine proteins correlates well with the relatively high K_m values of the tyrosine kinases compared to those for serine/threonine kinases (Fig 6B, Table 1). Tyr kinases have high K_m values for their substrates than serine/threonine kinases and therefore their substrates should be present in higher amounts for efficient phosphorylation. In fact, difference in median site occupancy for Tyr and serine/threonine phosphorylation events in untreated cells can also be explained by differences in K_m values for serine/threonine and tyrosine kinase (Fig. S2).

Next we tested our dataset for evidence of cross-talk between phosphorylated residues and lysine modifications. We overlaid our sites with those on acetylation, ubiquitination and sumoylation obtained from the PhosphoSitePlus resource (see Materials and Methods). We tested if a preference existed for phosphorylated residues and modified lysine residues to lie within close sequential proximity. The fractions of modified to non-modified lysine residues at

even amino acid intervals at different distances from a phosphorylation-site were computed and compared to randomized sets. The distribution of modified lysine residues over the protein sequence differs strongly from random (Fig. 6C). For all three phospho-acceptor residues the measured data (solid lines) lies above the randomized data (dashed lines), suggesting a tendency of modified lysine residues to occur with higher preference in the surroundings of phosphorylated residues. Furthermore, the fraction of modified to non-modified lysine residues increased as the distance to the phosphorylation site decreased. For all phospho-acceptor residues the highest proportion of modified lysine residues was found within five amino acids distance from the phosphorylation site, regardless if N- or C-terminus direction was considered. The strongest effect was observed for phospho-tyrosine residues, where the fold difference between the measured and the randomized data was 1.70, whereas this difference for phospho-serines was 1.27 and for phospho-threonines 1.26.

Discussion

Here, we present an analysis of the nature of the phosphoproteome at the level of more than 50,000 distinct phosphopeptides and about 11,000 proteins within a single cell type. We demonstrate that at least three-fourths of the proteome (8,014 out of 10,826 proteins) can be phosphorylated. By combining protein abundance measurements with phosphorylation changes across mitosis and epidermal growth factor (EGF) stimulation, we determined the occupancy of thousands of phosphorylation sites using a label free quantification approach. As depth of coverage increased, phospho-tyrosine became proportionately less of the total, whereas the increase in serine/threonine sites only appeared to saturate for technical reasons. Based on the occupancies distribution, we conclude that in unstimulated cells, serine/threonine phosphorylation occurs with low fractional occupancy while tyrosine phosphorylation is maintained at even lower levels. Our data clearly demonstrate that phospho-tyrosine has different properties from phospho- serine/threonine on multiple levels. Despite the large number of tyrosine kinases in the genome, phosphotyrosines account for <1% of total phosphorylation events. This and many other lines of evidences suggest that specificity in tyrosine phosphorylation is strictly maintained in cellular systems (Hunter, 2009). To meet this

specificity, the three part tyrosine signaling system (kinase, PTP and phospho-tyrosine binding proteins) is kept under strict biochemical controls. These are thought to be designed over evolution where erasers came into existence before the writers of this modification (Lim and Pawson, 2010). The observed low fractional occupancy of phosphotyrosine sites in unstimulated cells indeed correlate with cellular control mechanisms determining specificity of these events in response to specific signals.

Interestingly, we found that proteins phosphorylated on tyrosine residues are on average more abundant than the bulk of proteins and the observed difference in phosphoproteins abundance correlates with the substrate K_m values of tyrosine kinases. K_m differences have important consequences in the cell, where the protein kinases are exposed to varying substrate concentrations. We speculate that high abundance of substrate proteins coupled with low efficiency of tyrosine kinases buffer against harmful effects of the occasional stray phosphorylation of functionally important sites. Further, in contrast to serine/threonine kinases, because of their high K_m values for general substrates the tyrosine kinases might not be significantly inhibited by competition from general substrates allowing them efficiently phosphorylate a subset of low K_m substrates. Consequently, the low activity of tyrosine kinases towards general targets, combined with high abundance of their specific targets, helps to explain the specificity in the tyrosine based signaling events. In fact, a recent study revealed that an ensemble of fine-tuned weak interactions control cellular decisions as exemplified by cell fate control by RTKs (Findlay et al., 2013). The tailored protein abundances of tyrosine kinase substrates add yet another layer of complexity that adds specificity to the signaling systems. Further, we also found that the tyrosine phosphorylated residues are preferred hotspots for PTM crosstalk at least with modified lysines.

In summary, our data revealed that while we are close to complete coverage of tyrosine phosphorylation, serine/threonine based phosphoproteome is accompanied by a large number of variable 'background events' and is probably far from completeness. Our findings highlight the nature of phosphotyrosine as a separate functional regulatory post-translational modification of eukaryotic proteomes.

Experimental Procedures

Cell culture and treatment

HeLa S3 cells (ATCC) were cultured in roller bottles in RPMI 1640 (Gibco) supplemented with 10% fetal bovine serum (Invitrogen) and 1% penicillin/streptomycin (Invitrogen). For the mitotic experiment, cells were synchronized in G1-S overnight with a thymidine block at a concentration of 4 mM (Sigma). Cells were then released from thymidine and subsequently arrested overnight with nocodazole, after the 7.5-hour release from thymidine and lysed. Western blotting and FACS analyses were performed to monitor the efficiency of the cell cycle arrest (Fig. 1A). For EGF treatment cells were suspended in 1x PBS and treated with EGF for 5 and 15 min. For pervanadate treatment HeLa cells were pre-incubated with 1xPBS for 30 min, and treated with pervanadate for 20 min. A 4x concentrated SDS lysis buffer (final concentration 4% SDS 100 mM DTT in 100 mM tris-HCl pH 7.5) was added to the EGF and pervanadate samples.

Total proteome and phosphoproteome sample preparation and MS analyses

The lysate was processed by the FASP method (Wisniewski et al., 2009b). Briefly, the lysate was sonicated, heated and lysate was loaded onto 15 ml Amicon filter units (10 kDa MWCO) (Millipore). Proteins were alkylated and equilibrated in 20 mM ammonium bicarbonate and digested with trypsin (Promega) in a protein to enzyme ratio of 100:1 at 37 °C overnight. For proteome analysis 30 µg of the peptides were separated on a tip based SAX column as described (Wisniewski et al., 2009a). For phosphopeptide enrichment the peptides obtained from FASP were fractionated by strong cation exchange (SCX) chromatography (Macek et al., 2009) and subjected to phosphopeptide enrichment using TiO₂ beads as described (Zhou et al., 2011). The phospho-peptides were then eluted under basic conditions using 25% ammonium hydroxide and ACN. Finally, the eluted phospho-peptides were loaded on C₁₈ StageTips (Rappsilber et al., 2003). For phosphotyrosine enrichment the digested peptides were first subject to TiO₂ enrichment and then subjected to immunoprecipitation with Y99 anti-tyrosine antibody as described (Kettenbach and Gerber, 2011) and then desalted (Fig. 1).

Reverse phase chromatography and mass spectrometry

Peptides were separated in a 50 cm column packed in-house with 1.8 μm C₁₈ beads (Reprosil-AQ Pur, Dr. Maisch) on a Proxeon Ultra EASY-nLC system using a binary gradient provided by buffer A (0.5 % formic acid) and buffer B (0.5% formic acid and 80% ACN) for 265 min for proteome measurements and 120 min or 245 min for phosphopeptide measurements. The LC system was directly coupled on-line with a Q Exactive instrument (Thermo Fisher Scientific) via a nano-electrospray source. The mass spectrometer was programmed to acquire in a data dependent mode using a fixed ion injection time strategy (Kelstrup et al., 2012). Full scans were acquired in the Orbitrap mass analyzer with resolution 70,000 at 200m/z. For the full scans, 3E6 ions were accumulated within a maximum injection time of 20 ms and detected in the Orbitrap analyzer. The ten most intense ions with charge states ≥ 2 were sequentially isolated to a target value of 1e6 with a maximum injection time of 60 ms or 80 ms and fragmented by HCD in the collision cell (normalized collision energy of 25%) and detected in the Orbitrap analyzer at 17,500 resolution (Fig. 1).

Data processing and analysis

Raw mass spectrometric data was analyzed in the MaxQuant environment (Cox and Mann, 2008), versions 1.3.10.15, and employed Andromeda for database search (Cox et al., 2011). The MS/MS spectra were matched against the human International Protein Index sequence database (IPI version 3.37). The MS/MS spectra were matched against the human International Protein Index sequence database (IPI version 3.37). Enzyme specificity was set to trypsin, allowing for cleavage N-terminal to proline and between aspartic acid and proline. The search included cysteine carbamidomethylation as a fixed modification and N-acetylation of protein, oxidation of methionine and/or phosphorylation of serine, threonine tyrosine residue (STY) as variable modifications. For phosphopeptide identification, an Andromeda minimum score and minimum delta score threshold of 40 and 17 were used, respectively (Fig. 2). Up to two missed cleavages were allowed for protease digestion and peptide had to be fully tryptic. The 'identify' module in MaxQuant was used to filter identifications at 1% FDR at the peptide and protein level.

Downstream bioinformatics analyses

Bioinformatic analysis was performed in the Perseus software environment, which is part of MaxQuant. Hierarchical clustering of proteins or phosphosites was performed on logarithmized intensities or ratios for the data that was quantified in at least 50% of the time points studied. For multiple sample t-tests analysis (ANOVA), replicates were grouped and the statistical test was performed with a permutation-based FDR cutoff of 0.01.

Categorical annotation was supplied in the form of GO biological process, molecular function, and cellular component, KEGG pathways for pathway annotation and human protein reference database (HPRD) for kinase substrate motifs. Enrichment for these categories was evaluated by Fisher exact test to obtain a p-value. The annotation matrix algorithm was used to compute the difference of any significant protein annotation term from the overall intensity distribution as described (Cox and Mann, 2012). The specific test we used is a two-dimensional version of the nonparametric Mann-Whitney test. Multiple hypothesis testing was controlled by using a Benjamini-Hochberg FDR threshold of 0.05.

Determination of cross talk: Ubiquitination, acetylation and sumoylation data sets were obtained from the public repository PhosphoSitePlus (Hornbeck et al., 2012). The modified lysine residues were mapped to the phosphorylation data set. The fraction of modified to non-modified lysine residues at the flanking regions of each phospho-site was computed. Flanking regions of different lengths (from 5 to 40 amino acids in steps of 5) and directions (C-terminus and N-terminus) were analyzed. Next, the positions of the modified lysine residues were randomized over all lysine residues in the corresponding proteins and the resulting fractions of modified to non-modified lysines were computed. The randomization was repeated 1,000 times, creating a background distribution of random distances. The measured and randomized fractions were plotted for each phospho- serine, - threonine and tyrosine residue.

Immuno-blotting and fluorescence activated cell sorting

EGF treated lysates prepared for mass spectrometry were blotted with anti-phospho ERK1/2 (Cell signaling technology) and equal loading was controlled with anti-tubulin (Cell signaling

technology). Mitotic arrest assay was by fluorescence activated cell sorting (FACS). To detect a nocodazole arrest, HeLa cells were labeled with 20 μ M bromodeoxyuridine (BrdU; Roche) for 2 h, harvested, and fixed in 70% ethanol. After RNase A treatment (50 μ g/mL for 30min) and DNA denaturation (5 M HCl/0.5% Triton X-100 for 20 min), cells were stained with anti-BrdU-FITC (Boehringer Mannheim) and propidium iodide (Sigma-Aldrich). The cells were analyzed for FITC (BrdU incorporation) and propidium iodide (total DNA content) fluorescence by a BD LSR II flow cytometer (BD Biosciences).

Acknowledgments

We thank co-workers at the MPI Martinsried for helpful discussions. We are thankful to Martin Dodel, Igor Paron and Korbinian Mayr for technical assistance. This work was partially supported by European Union 7th Framework project PROSPECTS (Proteomics Specification in Time and Space, grant HEALTH-F4-2008-201645) at the MPI, Martinsried.

References

- Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. *Mol Syst Biol* **7**, 549.
- Choudhary, C., and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Bio* **11**, 427-439.
- Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372.
- Cox, J., and Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC bioinformatics* **13 Suppl 16**, S12.
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794-1805.
- de Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251-U1260.

Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N.M., Bertone, P., Zhang, Z.L., and Gerstein, M. (2002). Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* **30**, 2515-2523.

Findlay, G.M., Smith, M.J., Lanner, F., Hsiung, M.S., Gish, G.D., Petsalaki, E., Cockburn, K., Kaneko, T., Huang, H.M., Bagshaw, R.D., *et al.* (2013). Interaction Domains of Sos1/Grb2 Are Finely Tuned for Cooperative Control of Embryonic Stem Cell Fate. *Cell* **152**, 1008-1020.

Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261-D270.

Humphrey, S.J., Yang, G., Yang, P., Fazakerley, D.J., Stockli, J., Yang, J.Y., and James, D.E. (2013). Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. *Cell metabolism* **17**, 1009-1020.

Hunter, T. (2007). The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell* **28**, 730-738.

Hunter, T. (2009). Tyrosine-phosphorylation: thirty years and counting. *Curr Opin Cell Biol* **21**, 140-146.

Hunter, T., and Sefton, B.M. (1980). Transforming Gene-Product of Rous-Sarcoma Virus Phosphorylates Tyrosine. *P Natl Acad Sci-Biol* **77**, 1311-1315.

Kelstrup, C.D., Young, C., Lavalley, R., Nielsen, M.L., and Olsen, J.V. (2012). Optimized Fast and Sensitive Acquisition Methods for Shotgun Proteomics on a Quadrupole Orbitrap Mass Spectrometer. *Journal of Proteome Research* **11**, 3487-3497.

Kettenbach, A.N., and Gerber, S.A. (2011). Rapid and Reproducible Single-Stage Phosphopeptide Enrichment of Complex Peptide Mixtures: Application to General and Phosphotyrosine-Specific Phosphoproteomics Experiments. *Anal Chem* **83**, 7635-7644.

Lim, W.A., and Pawson, T. (2010). Phosphotyrosine Signaling: Evolving a New Cellular Communication System. *Cell* **142**, 661-667.

Lundby, A., Secher, A., Lage, K., Nordsborg, N.B., Dmytriiev, A., Lundby, C., and Olsen, J.V. (2012). Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nat Commun* **3**.

Macek, B., Mann, M., and Olsen, J.V. (2009). Global and Site-Specific Quantitative Phosphoproteomics: Principles and Applications. *Annu Rev Pharmacol* **49**, 199-221.

Michalski, A., Damoc, E., Hauschild, J.P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011). Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol Cell Proteomics* **10**.

Nagaraj, N., D'Souza, R.C.J., Cox, J., Olsen, J.V., and Mann, M. (2010). Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. *Journal of proteome research* 9, 6786-6794.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7.

Neuhauser, N., Michalski, A., Cox, J., and Mann, M. (2012). Expert System for Computer-assisted Annotation of MS/MS Spectra. *Mol Cell Proteomics* 11, 1500-1509.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., *et al.* (2010). Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Sci Signal* 3.

Pan, C., Olsen, J.V., Daub, H., and Mann, M. (2009). Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. *Molecular & cellular proteomics : MCP* 8, 2796-2808.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* 75, 663-670.

Sadowski, I., Stone, J.C., and Pawson, T. (1986). A Noncatalytic Domain Conserved among Cytoplasmic Protein-Tyrosine Kinases Modifies the Kinase Function and Transforming Activity of Fujinami Sarcoma-Virus P130gag-Fps. *Mol Cell Biol* 6, 4396-4408.

Schaab, C., Geiger, T., Stoehr, G., Cox, J., and Mann, M. (2012). Analysis of High Accuracy, Quantitative Proteomics Data in the MaxQB Database. *Mol Cell Proteomics* 11.

Tan, C.S.H., Pasculescu, A., Lim, W.A., Pawson, T., Bader, G.D., and Linding, R. (2009). Positive Selection of Tyrosine Loss in Metazoan Evolution. *Science* 325, 1686-1688.

Ubersax, J.A., and Ferrell, J.E. (2007). Mechanisms of specificity in protein phosphorylation (vol 8, pg 530, 2007). *Nat Rev Mol Cell Bio* 8, 665-665.

Wisniewski, J.R., Zougman, A., and Mann, M. (2009a). Combination of FASP and StageTip-Based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome. *Journal of proteome research* 8, 5674-5678.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009b). Universal sample preparation method for proteome analysis. *Nat Methods* 6, 359-U360.

Zhou, H.J., Di Palma, S., Preisinger, C., Peng, M., Polat, A.N., Heck, A.J.R., and Mohammed, S. (2013). Toward a Comprehensive Characterization of a Human Cancer Cell Phosphoproteome. *Journal of proteome research* 12, 260-271.

Zhou, H.J., Low, T.Y., Hennrich, M.L., van der Toorn, H., Schwend, T., Zou, H.F., Mohammed, S., and Heck, A.J.R. (2011). Enhancing the Identification of Phosphopeptides from Putative Basophilic Kinase Substrates Using Ti (IV) Based IMAC Enrichment. *Mol Cell Proteomics* 10.

Fig 1

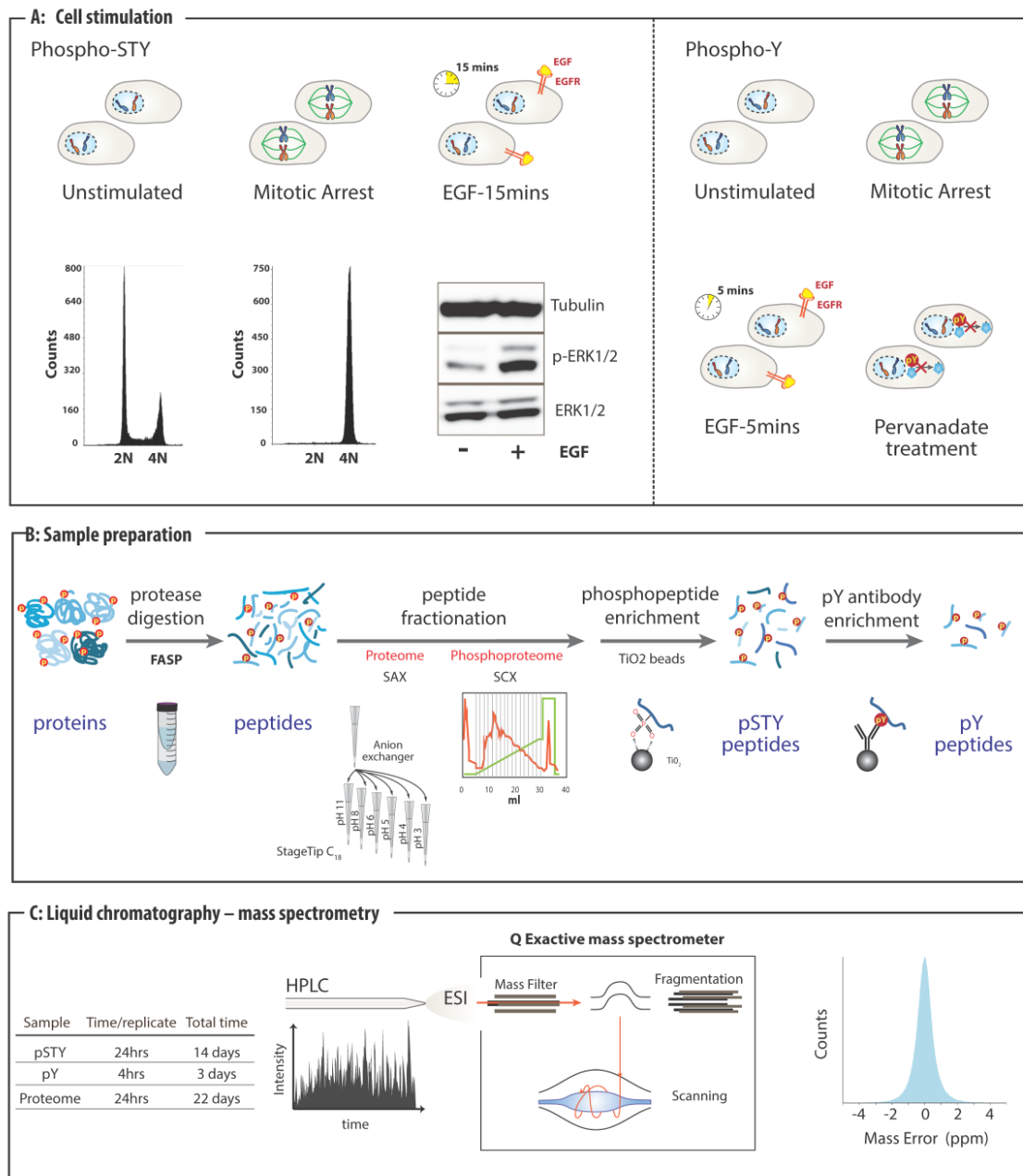


Figure 1: Workflow for large-scale phosphoproteome. **A. Cell stimulation.** HeLa S3 cells were synchronized to arrest with nocodazole, and released from arrest for 7.5h when they were in mitosis. Mitosis was monitored by FACS analysis. Two other sets of HeLa S3 cells were treated with EGF for 5 and 15 min. A separate population of cells was treated with sodium pervanadate for inhibition of tyrosine phosphatases. EGF induced ERK1/2 activation was monitored by immunoblotting. **B. Sample preparation.** Cell lysates from the above treatments were lysed in SDS buffer and digested by FASP method. For proteome measurements peptides were separated into 6 fractions on tip based strong anion exchange (SAX) discs. For total phosphoproteome analysis, phosphopeptides were enriched by strong cation exchange (SCX chromatography) and TiO₂ microbeads. Phosphotyrosine peptides were immuno-precipitated with anti-phospho-tyrosine antibodies from phosphopeptides. **C. LC-MS** All fractions were separated on a reverse-phase column and electrosprayed into a Q Exactive mass spectrometer which was operated in a data dependent mode. High resolution measurements in the Orbitrap analyzer resulted in the acquisition of high accuracy spectra with average absolute mass deviation of 500 ppb for precursor ions.

Fig 2

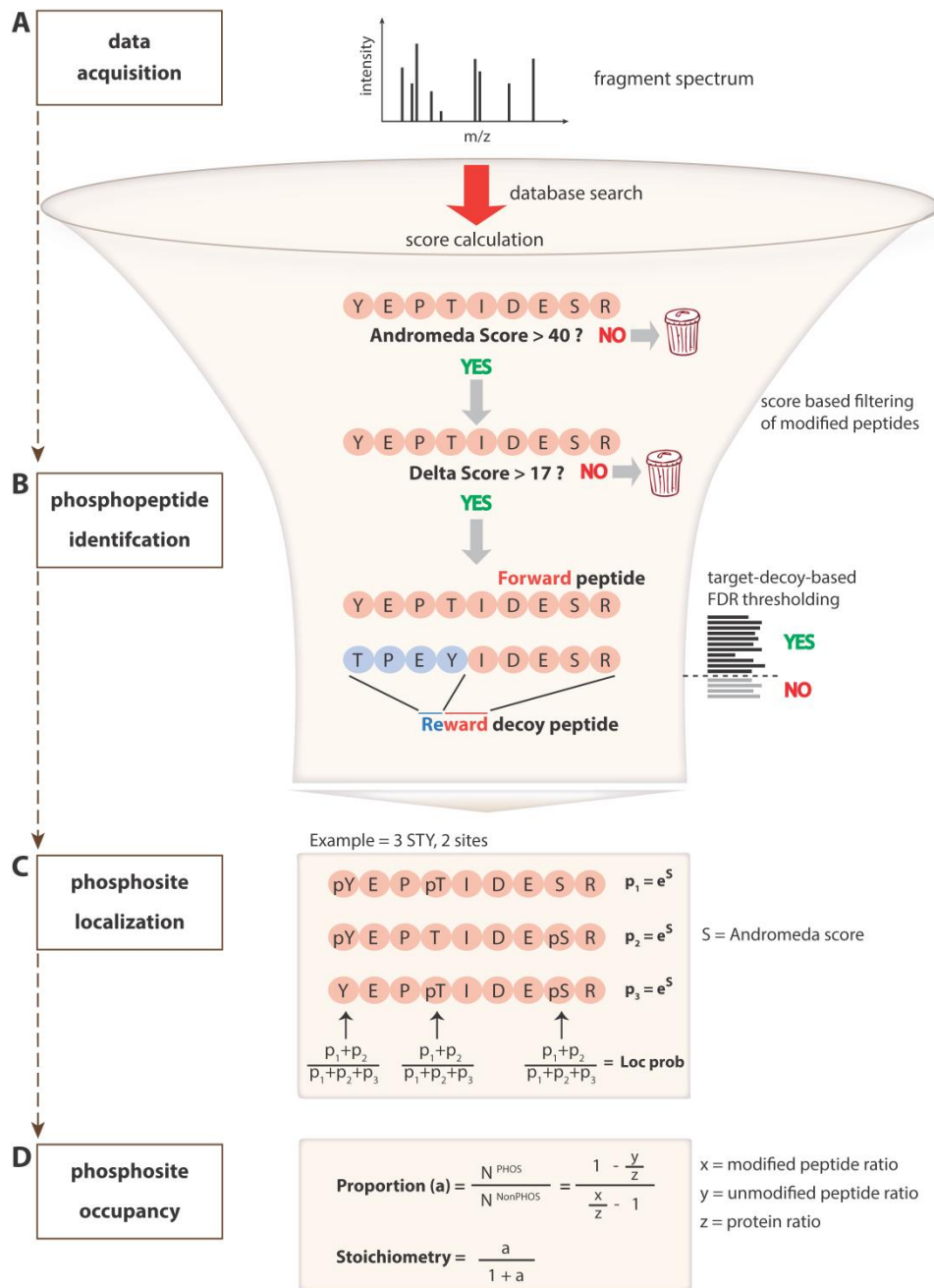


Figure 2: Computational pipeline for high stringency identification and quantification of phosphopeptides in MaxQuant. **A. Data acquisition.** High quality fragment spectra were acquired by LC-MS/MS. **B. Phosphopeptide identification.** Spectra were converted to peak lists that were matched by the Andromeda search engine to obtain identification scores. Only those identified with an Andromeda score >40 and a delta score of >17 were retained. Posterior error probability (PEP) was calculated based on the ‘reward’ strategy to control the false discovery rate. **C. Phosphosite localization.** Localization of the modification site was assigned by looping through possible combinations for the phosphorylation on individual amino acid residues on the peptide for which the Andromeda score is calculated and exponentiated to obtain the localization probability. **D. Phosphosite occupancy.** The proportion of phosphorylated peptides was calculated based on the ratios of modified peptide, unmodified ratio and corresponding protein ratios.

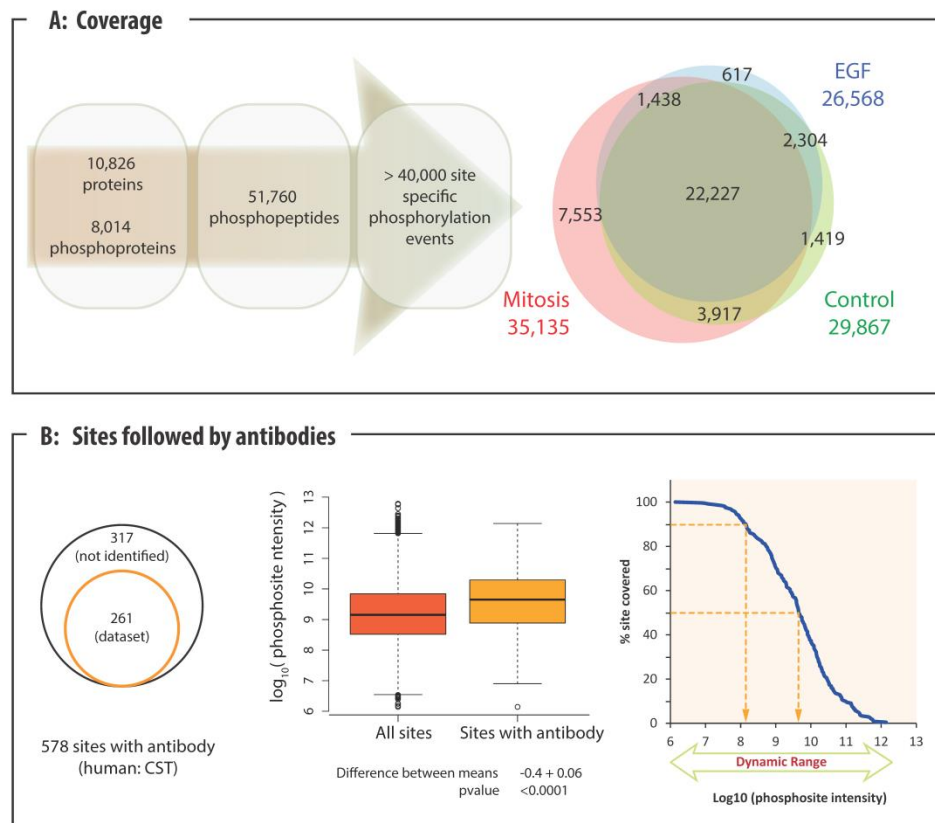
Fig 3

Figure 3: Overview of the identified phosphoproteome. A. Coverage. The large scale MS-based analysis resulted in the identification of 10,826 proteins of which >8,000 were phosphorylated. A total of about 52,000 phosphopeptides were identified of which >40,000 were localized to specific S/T/Y residues (left panel). Venn diagram of site specific phosphorylation events across the conditions studied. **B. Sites with antibodies.** Venn diagram of functional sites followed by antibodies from the CST database which were identified in our dataset (Left panel). Distribution of phosphopeptide intensities of all sites identified in comparison to those with antibodies (Middle panel). Coverage of intensity ranked phosphosites (followed by antibodies) plotted as a function of their abundance.

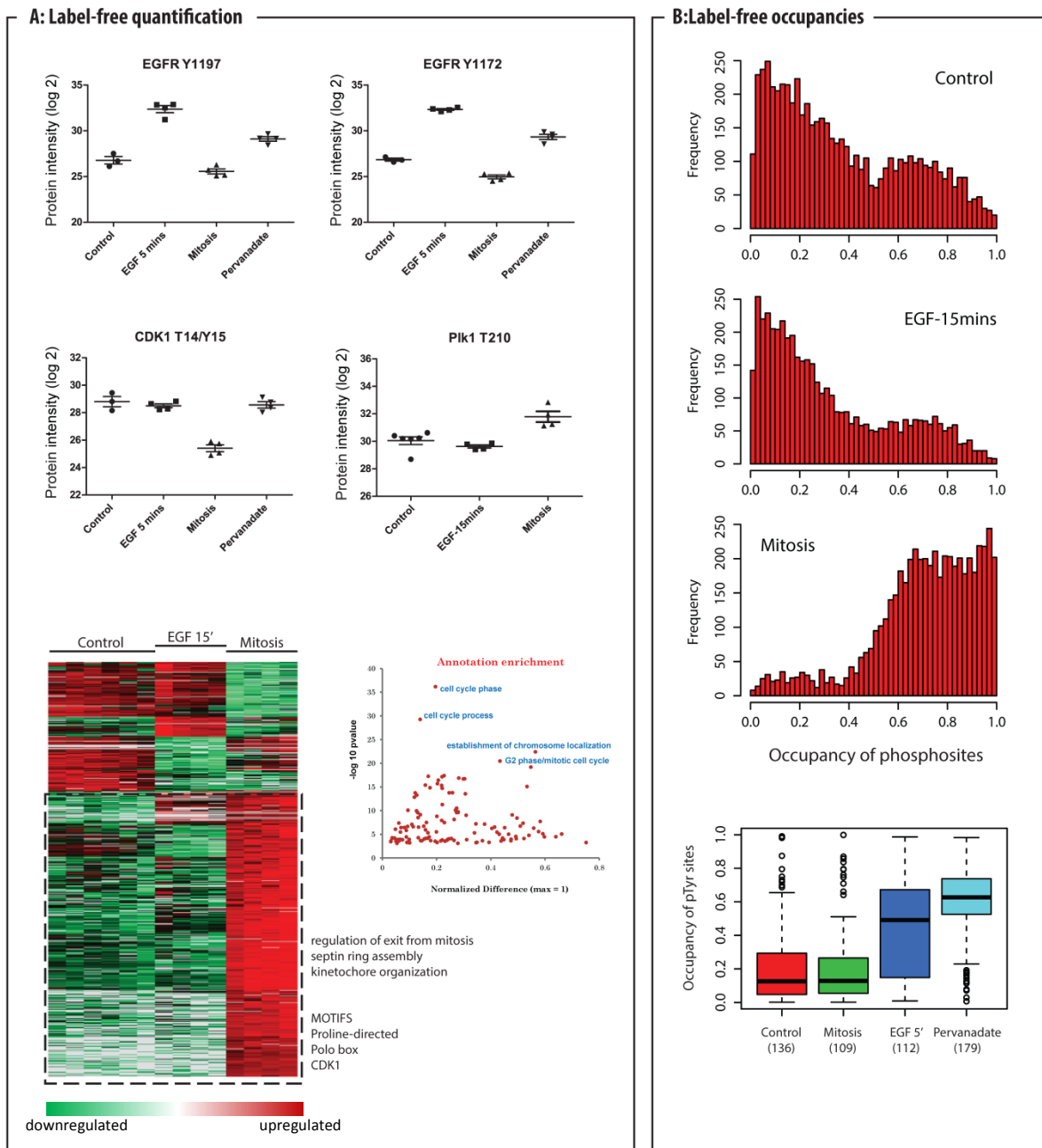
Fig 4

Figure 4: A. Label-free quantification of regulated phosphopeptides. Label-free quantification of individual replicates (mean \pm SD of replicates) for the conditions studied is shown for Y1197 and Y1172 on EGFR, Y15/Y16 on CDK1/2/3 and T210 on PLK1 (upper panel). Heatmap of phosphosites regulated across different conditions (color scale from green to red indicating decreased and increased phosphorylation) with kinase motifs and categories enriched (lower panel). Inset in lower panel shows 1-D annotation analysis of annotation terms of proteins significantly up-regulated during mitosis. The data points corresponding to annotation terms whose members are regulated with a very high significance are labeled ($p < 1e-20$). **B. Label-free occupancies.** Histogram of phosphosite occupancies for control, EGF treated and mitotic samples (upper panel). Distribution of phosphotyrosine occupancies across the conditions measured (lower panel).

Fig 5

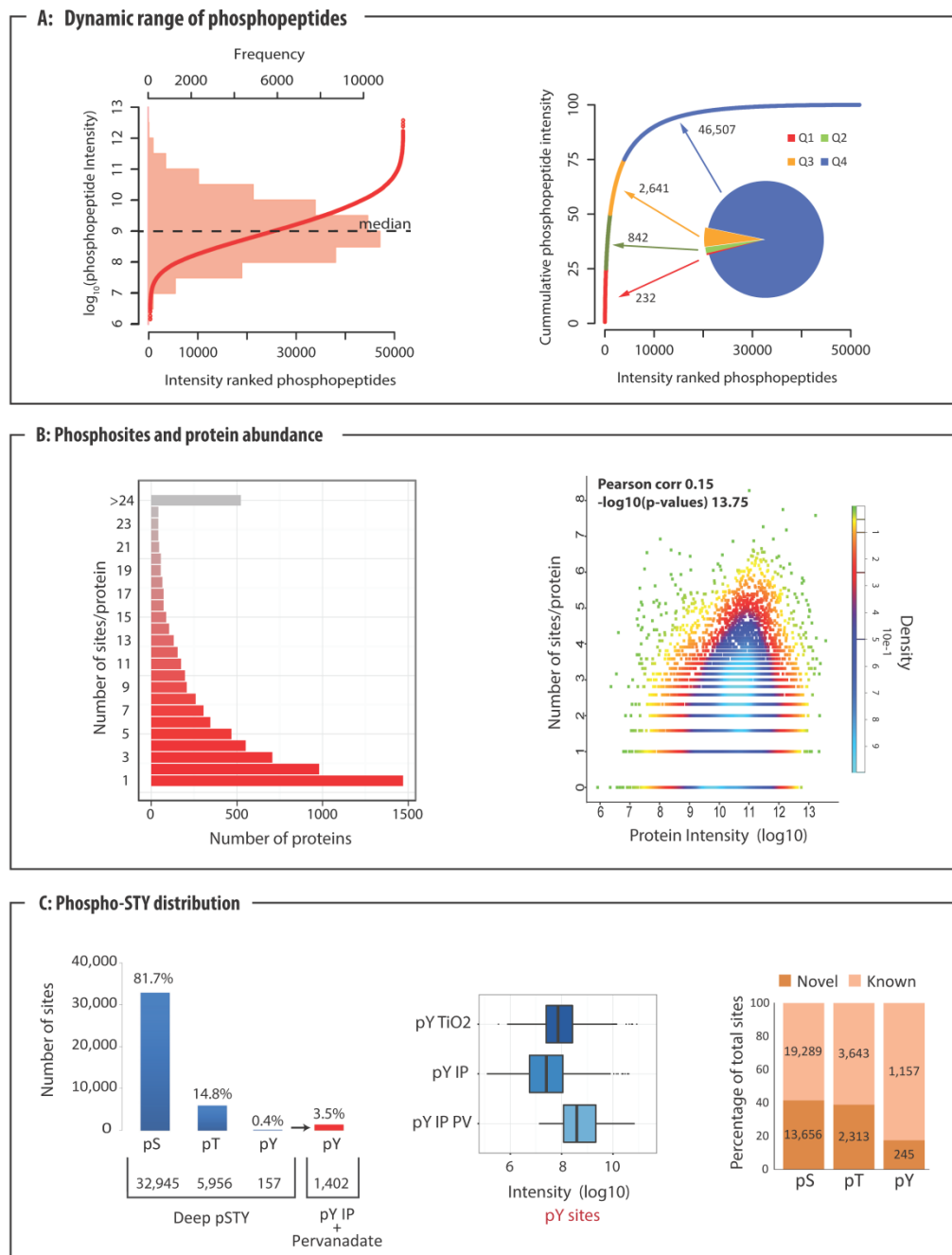


Figure 5: Specific properties of the phosphoproteome. A. Dynamic range of the phosphoproteome. Histogram of phosphopeptide intensities showing median intensity on which ranked phosphopeptide abundances from decreasing to increasing abundance are overlaid (Left panel). Cumulative phosphopeptide abundance from the highest to the lowest abundance with pie chart separating the abundances into four intensity quantiles (Right panel). **B. Overview of phosphorylation sites per protein.** Distribution of phosphoproteins based on number of phosphorylation sites per protein (Left panel) and density scatter plot of protein abundance versus number of sites per protein. The color code indicates the percentage of points that are included in a region of a specific color (Right panel). **C. Phosphosite distribution across S/T/Y residues.** Distribution of the S/T/Y phosphorylation events by global phosphoproteomics and tyrosine IP (Left panel). Box plots of phosphotyrosine peptide intensities from global (TiO₂-based), phosphotyrosine immunoprecipitation (IP) and IP+ pervanadate treated samples (Middle panel). Distribution of the known and novel phospho-tyrosine/pS/pT sites after matching with PhosphositePlus database (Right panel).

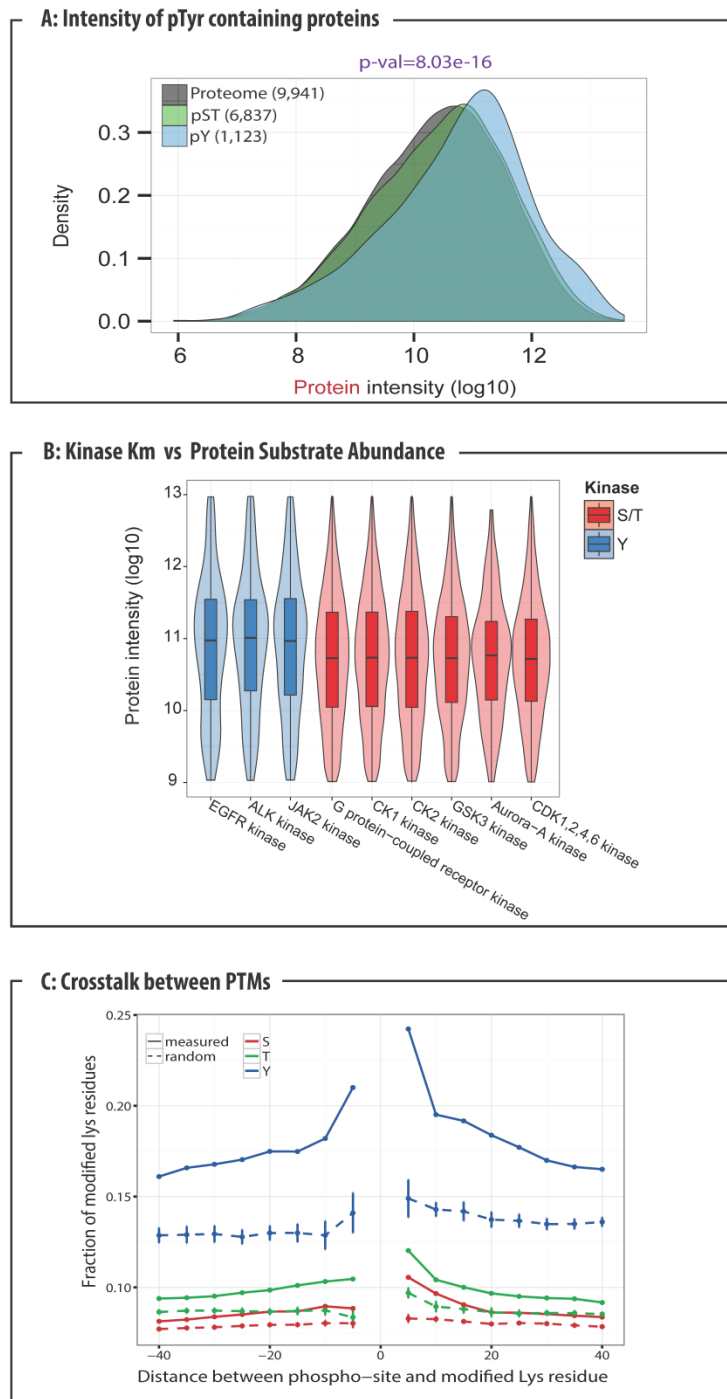
Fig 6

Figure 6: Comparison of phosphorylation on S/T versus Y residues. A. Density distribution of intensities of all proteins identified (black) compared to proteins with serine/threonine phosphorylation (green) and tyrosine phosphorylation (blue). **B.** Violin plot distributions of protein intensities of substrates phosphorylated by indicated kinases. Tyrosine kinases substrates are depicted in blue while those for S/T kinases in red. **C.** Distribution of modified lysine residues over the protein sequence compared to random occurrences (dashed lines) plotted separately for S, T and Y.

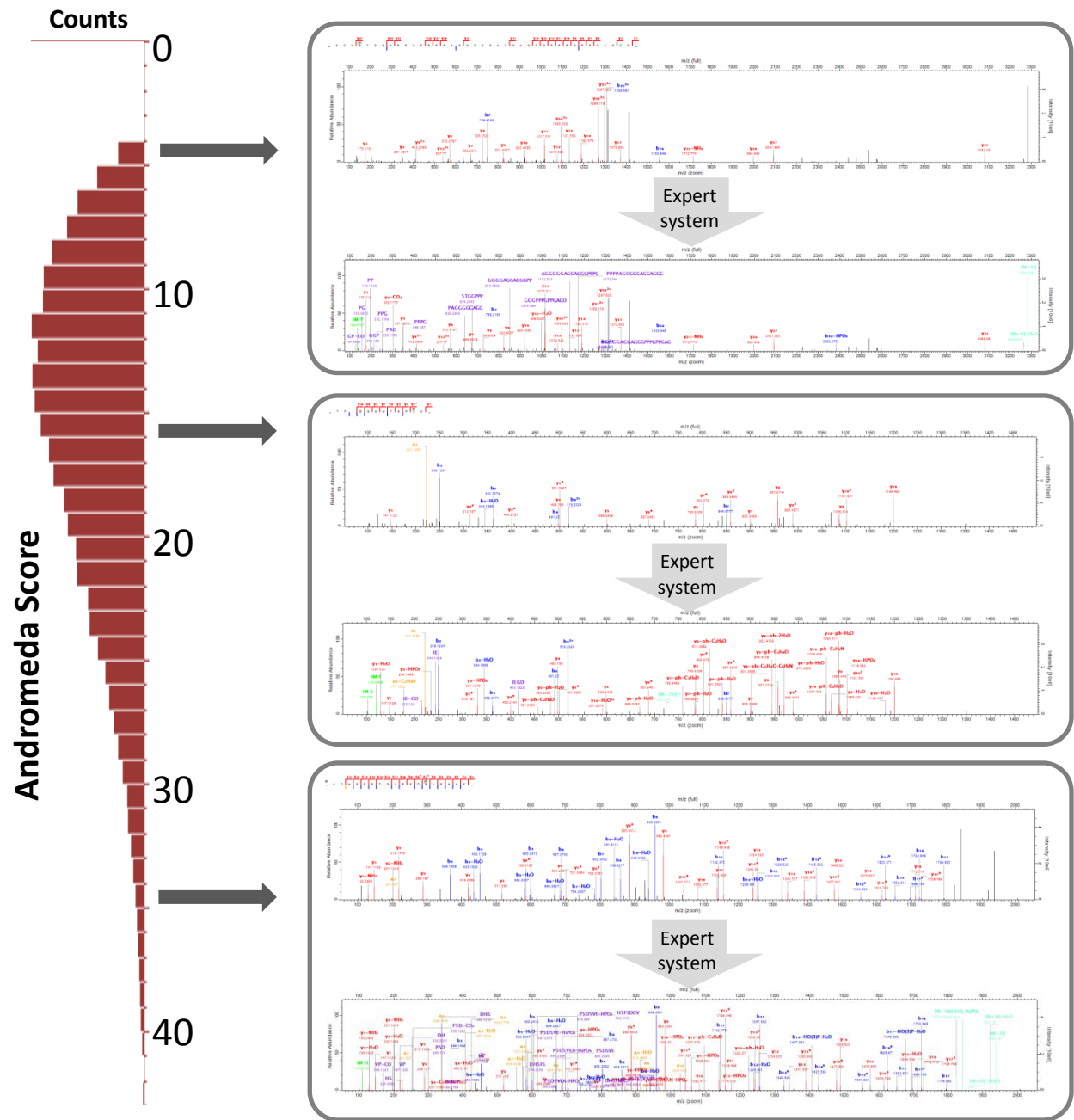


Fig S1: Distribution of Andromeda scores of identified phosphopeptides. Additional peak annotation by 'expert system' for phosphopeptides over increasing Andromeda scores.

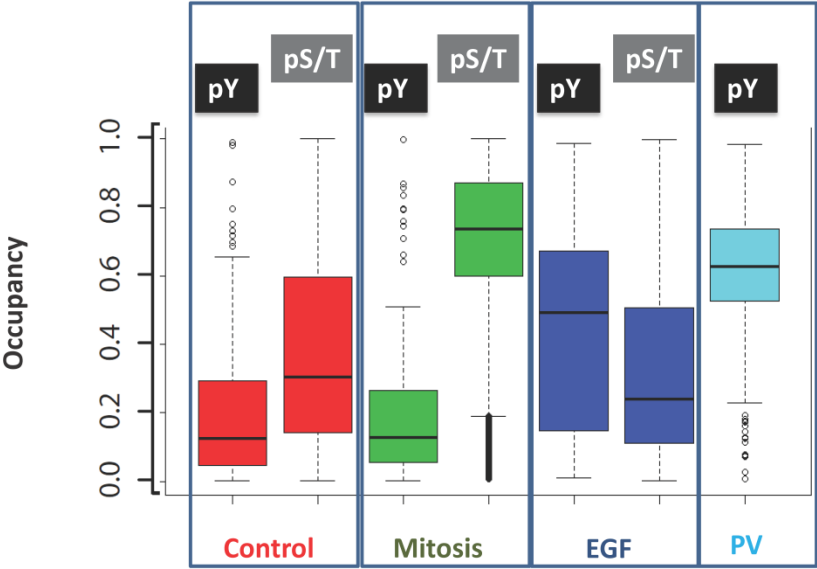


Fig S2: Box-plots of occupancies for phosphoserine/threonine versus phosphotyrosine across different biological conditions.

5. In-depth and time-resolved dissection of early phospho-proteome and ensuing proteome changes in response to TGF- β (Article 4).

Transforming growth factor beta (TGF- β) is a multifunctional cytokine that controls cell growth and differentiation and promotes cell invasion through the induction of epithelial to mesenchymal transition (EMT) in both normal physiology and development as well as in pathology, such as cancer. The effects TGF- β exerts on its target cells are dictated by the cellular context and these effects can even be opposed to each other. Despite the fact that TGF- β mediated signal transduction has been very extensively studied since the 1980s, the mechanisms and mediators that bring about its diverse downstream effects have not been characterized at a systems level.

I therefore chose to investigate early signaling by phospho-proteomics and combine this with proteome changes at a much longer time scale employing advances in MS instrumentation and data analysis platforms. I obtained an unprecedented in depth of dynamic changes in protein and phospho-protein expression induced by TGF- β .

My results not only confirm known critical pathways of TGF- β induced cytostatic response, ECM remodeling and epithelial dedifferentiation but greatly extend them and also uncovered involvement of novel effector pathways for TGF- β . An in-depth combined analysis of transcription factor regulation with early phosphorylation changes and ensuing proteome regulation driving processes induced by TGF- β enabled, for the first time, the visualization of the intricate interplay and network of the various pathways that mediate these biological responses. Intriguingly the analysis also revealed that the mediators of the contextual differences in TGF- β signaling on cell growth are acting on a temporal scale where early events constitute a mixed signal that is tailored into a pro or anti-proliferative response at later time points.

This article has been submitted to Science Signaling in October 2013.

In-depth and time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF- β

Rochelle C. J. D'souza^{1#}, Kirti Sharma^{1#}, Anna Korhonen^{2,3}, Nagarjuna Nagaraj¹, Chunaram Choudhary^{1,4}, Peter ten Dijke^{2*} and Matthias Mann^{1,4*}

One Sentence Summary: Large-scale and time-resolved proteome and phosphoproteome of the TGF- β induced growth arrest and EMT switch

Affiliations:

¹ Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

² Department of Molecular Cell Biology, Cancer Genomics Centre and Centre for Biomedical Genetics, Leiden University Medical Center, Postbus 9600, 2300 RC, Leiden, The Netherlands

³ Department of Medical Biochemistry and Genetics, and Turku Doctoral Program of Biomedical Sciences, University of Turku, FI-20520Turku, Finland

⁴ Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark

* To whom correspondence may be addressed

Matthias Mann: mmann@biochem.mpg.de

Peter ten Dijke: P.ten_Dijke@lumc.nl

These authors contributed equally to this work.

Abstract

Transforming growth factor beta (TGF- β) is a multifunctional cytokine that controls cell growth and differentiation and promotes cell invasion through the induction of epithelial to mesenchymal transition (EMT) in both normal physiology and development as well as in pathology, such as cancer. Here, we report a time-resolved proteome and early phosphoproteome upon TGF- β stimulation in unprecedented depth. Known critical pathways of TGF- β induced cytostatic response, ECM remodeling and epithelial dedifferentiation are confirmed and greatly extended, and there is evidence for the involvement of multiple novel effector pathways for TGF- β such as vesicular trafficking. We observe rapid phosphoproteomic kinetics in response to TGF- β demonstrating a key role of non-SMAD signaling pathways. Combined analysis at the levels of signaling pathway, transcription factors and the proteome changes provides novel regulators and a detailed molecular picture of mechanisms of TGF- β action. Early TGF- β signaling appears to be a mixture of pro and anti-proliferative signals, with cellular context determining the final outcome.

Introduction

Transforming growth factor- β (TGF- β) is a cytokine that belongs to a large family of 33 mammalian members, which includes TGF- β s, bone morphogenetic proteins (BMPs), activins, inhibins and nodals. TGF- β signaling regulates various biological outcomes including cell growth, differentiation, morphogenesis, tissue homeostasis and regeneration. This family of proteins is increasingly implicated in diseases such as cancer and auto-immune disorders (1-3). The cellular responses to this multifunctional ligand are diverse and can even be opposed to each other, depending on the cell type and the conditions. For example, TGF- β can promote cell growth but also have anti-proliferative effects, and it can contribute to maintain stem cell pluripotency but also to differentiation (4). Further, TGF- β suppresses pre-malignant cells by inhibiting cell proliferation, it does not do so in metastatic ones, which nevertheless remain responsive to TGF- β induced migration and invasion (5, 6). A key mode of action of TGF- β in cancer progression is the induction of epithelial to mesenchymal transition (EMT), a process wherein epithelial cells acquire mesenchymal characteristics (7, 8). This is evidenced by loss of epithelial markers such as E-cadherin from the plasma membrane and induction of EMT related proteins including plasminogen activator inhibitor (PAI)-1, fibronectin (FN1) and α -smooth muscle actin (9-11). EMT is an indispensable process in normal tissue development and organogenesis, as well in tissue remodeling and wound healing. However, inappropriate reactivation of EMT crucially contributes to the development of a variety of human pathologies, particularly those associated with tissue fibrosis and cancer cell invasion and metastasis, for instance in breast cancer (12, 13).

TGF- β initiates signal transduction by binding to its transmembrane type II receptor (T β R-II), which then recruits and phosphorylates type I receptor (T β R-I) in the juxtamembrane region (2). The activated T β R-I transduces the signal into the cell via binding and carboxy-terminal phosphorylation of SMAD2/3 assisted by an adaptor protein termed SMAD anchor for receptor activation (SARA) (14). The heteromeric complex formed by activated SMAD2/3 and the regulatory SMAD (Co-SMAD4) and the resulting complex is shuttled to the nucleus, where it binds to target genes (15, 16). Transcriptional activation generally requires co-operating DNA

binding factors like activating protein (AP)-1 and activating transcription factor (ATF)-2 (17) as well as co-repressors such as TG-interacting factor (TGIF), Ski and SnoN (18, 19). Well established targets genes of the SMAD complex include *cyclin-dependent kinase inhibitor 1A* (*p21CIP1*) and *cyclin-dependent kinase 4 inhibitor B* (*p15/ INK4B*) (mediating cell cycle arrest)(20, 21) and fibronectin (FN1) and PAI-1 (conferring extracellular matrix (ECM) remodeling) (23). In recent years there has been an increasing interest in TGF β R-induced non-SMAD signaling, which is typically mediated by p38, Jun N-terminal kinases (JNKs) and the extracellular signal regulated (ERK) mitogen activated kinases (MAPKs) (24-26).

Despite the fact that TGF- β mediated signal transduction has been very extensively studied since the 1980s, the mechanisms and mediators that bring about its diverse downstream effects have not been characterized at a systems level. Mass spectrometry (MS)-based proteomics is an attractive technology in signal transduction research due to its unbiased nature (27) and has already been successfully applied to study many pathways (28, 29). Advances in MS instrumentation and data analysis platforms enable routine identification of close to complete proteomes in several organisms ranging from bacteria to vertebrates (30-32). Importantly, deep proteome coverage can now be attained in single liquid chromatography (LC)-MS runs (33), which enables sophisticated experimental designs, such as time courses, to be performed.

There have been a few attempts to study TGF- β induced EMT by proteomics in the recent years. However, they reported only a few hundred identified and many fewer regulated proteins since they either employed 2D gel electrophoresis (34, 35), which tends to be biased towards abundant proteins, or otherwise did not yet use the high resolution MS-based technologies available today (36, 37). To obtain a global view of the TGF- β pathway, we employ high resolution MS-based proteomics for an in-depth temporal investigation of proteome at extended time scales up to 48h in a human keratinocyte (HaCaT) cell line after treatment with TGF- β . To capture the preceding, upstream phosphoproteome changes following receptor kinase activation, these were studied at early time points up to 20 min.. Our deep proteome covers the critical pathways involved in TGF- β signaling, allowing global evaluation at the level

of individual pathway members. Based on the TGF- β responsive proteome we correctly retrieve several known transcription factors driving the EMT process and also predict several novel ones. The early TGF- β induced phosphoproteome includes known and novel substrates identifying multiple kinases involved in this process. The combined analysis of transcription factor regulation with early phosphorylation changes and ensuing proteome regulation enables visualization of the intricate interplay of key transcription factors, kinases and various pathways driving cytoskeleton, EMT and other processes induced by TGF- β .

Results

Proteomic analysis of TGF- β induced EMT on a temporal scale

When stimulated with TGF- β , several epithelial cell lines, including the human keratinocyte derived HaCaT cells, are growth inhibited and undergo an EMT-like switch (38). We characterized effects of TGF- β treatment in HaCaT cells and observed arrest at the G1 cell cycle phase, loss of E-cadherin from the plasma membrane, replacement of cortical actin filaments by actin stress fibers in response to TGF- β at 40h of treatment (Fig. 1A). To analyze proteomic changes in HaCaT cells we stimulated them with TGF- β 1 for 0h, 6h, 12h, 24h, 36h and 48h. To account for autocrine secretion of TGF- β and other growth factors, which has been shown to be proportional to cell density (39, 40), we used corresponding untreated controls for normalizing the proteome at each time point after TGF- β treatment. We performed quadruplicate single-run analysis of the time points for robust statistics, leading to a total of 48 proteome measurements. This took a total of eight days of LC-MS measurement time on a quadrupole-Orbitrap mass spectrometer with high sequencing speed and resolution (41) (Materials and Methods). Where possible, the identity of peptides present but not sequenced in a given run was obtained by transferring identifications across liquid chromatography (LC)-MS runs. This is a feature in the MaxQuant software used to analyze all data (42, 43) that was also used to improve overall depth of coverage from a fifth replicate per time point in which peptides were fractionated into three strong cation exchange (SCX) fractions (Fig. 1B).

Joint analysis of all raw data resulted in the identification of about 7,500 proteins at a false discovery rate (FDR) of 1%. The label-free algorithm in MaxQuant (44) enabled robust

quantification of 6,113 protein groups (table S1). This is by far the most comprehensive coverage of proteome wide expression changes in response to TGF- β treatment. Fig. 1C illustrates the extraction of peptide signals using the example of the changing intensity of a fibronectin (FN1) peptide over the time-course. Identification and quantitation of identified proteins spans seven orders of magnitude in the MS-signals, demonstrating the power and sensitivity of our proteomics workflow. The biological quadruplicate measurements showed excellent quantitative reproducibility with a Pearson's correlation coefficient of at least 0.96 (Fig. S1). The correlation of TGF- β treated samples to the untreated control - while remaining high overall - gradually diminished over the time course, indicating a systematic and reproducible proteome change during ligand induced growth arrest and epithelial dedifferentiation.

Extraction of regulated proteins and bioinformatic analysis

The major cellular effects exerted by TGF- β in HaCaT cells are a potent cell cycle arrest, ECM remodeling and an EMT like switch. We observed significant regulation of well-established proteins (table S2) including up-regulation of p21CIP1 and p15/ INK4B involved in cell cycle arrest (21); up-regulation of FN1, transgelin (TAGLN), transglutaminase2 (TGM2) and PAI-1 involved in ECM remodeling (45, 46); and up-regulation of smooth muscle actin and integrins involved in EMT (47). Among the top five most proteins regulated in magnitude FN1 had the largest increase of around 25 fold (Fig 2A). FN1 production and deposition into the ECM is a hallmark of TGF- β induced EMT (23, 48). In contrast, Antigen KI-67, a prototypic cell cycle related nuclear protein commonly employed as a marker for proliferating cells especially cancerous cells (43), was the most down-regulated in magnitude with a 16-fold decrease (Fig. 2A). This corroborates TGF- β as one of the most effective suppressors of epithelial cell proliferation (49). The other most up-regulated proteins included TAGLN and TGM2 whereas kinesin family member 4A (KIF4A), anillin (ANLN) and minichromosome maintenance complex component 6 (MCM6) were among the most down-regulated proteins, most of these providing further positive controls for our dataset (50). However even amongst the most regulated proteins, there were novel proteins such as proline rich protein 9 (PRR-9), which have never

been never been implicated in TGF- β biology. Our dataset captured regulation of various categories of proteins. For example, integrin $\beta 6$ (ITGB6) is a plasma membrane protein which significantly increased throughout the time-course (Fig. 2B, right panel). It is a critical activator of latent TGF- β and its induction may serve as a positive feed forward loop (51). Similarly, we quantified the levels of a low abundant kinase, cyclin dependent kinase 1 (CDK1). Its expression was constant until 12h after stimulation, after which it gradually decreased to about a tenth of its value (Fig. 2B, left panel). Regulation of CDKs is an important mechanism of the anti-proliferative effects of TGF- β (52).

The temporal profile of all TGF- β responsive proteins reveals that substantial proteome-wide changes occur earliest at 24h of treatment (Fig. 2C). Most proteins showed gradual regulation over the time points, with up-regulation typically induced earlier than down-regulation. A total of 2,079 proteins showed reproducible and significant differences in expression in at least one of the six time points after TGF- β treatment (ANOVA, FDR <0.05) (table S2).

To obtain a global view of biological processes and pathways involved in the TGF- β induced responses including cell cycle arrest, ECM remodeling and EMT, we first clustered these significantly regulated proteins based on their temporal profiles. The two resulting clusters corresponded to up- and down-regulation, respectively (Fig. 2C). Cluster one was most enriched for the gene ontology (GO)(53) and KEGG (54) terms ECM, focal adhesion, actin cytoskeletal rearrangement and other processes that are essential for the transition of immotile, polarized epithelial cells into motile mesenchymal cells (Fig. 2C). Interestingly we also observed a strong enrichment for annotation terms related to the regulation of vesicle-mediated transport including protein glycosylation, COPI vesicle, functions that are not generally associated with TGF- β action. Cluster two was significantly enriched for categories involved in cell cycle, DNA replication, RNA processing, spliceosome and chromatin modification some of which are beginning to be implicated in TGF- β signaling (55, 56) (Fig. 2C, table S6).

In an orthogonal approach we assessed the enrichment of biological pathways in a time-point specific manner instead of across the time-profile. We used a recently developed algorithm that relates proteome expression levels to any protein annotation categories (57). For those

categories that are statistically significant, we calculate a position score (termed 's'; a number between -1 and 1), specifying where the mean of the distribution of expression values for the protein category is located relative to the overall distribution of expression values. A position score near 1 indicates that the protein category is strongly concentrated at the high end of the expression value distribution ('up-regulation' of the category) while a score near -1 means that the expression values are all at the low end of the distribution (down-regulation). We used this algorithm to create an annotation matrix of biological pathways (based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (54)) corresponding to the six time points. The z-scored matrix was then analyzed by hierarchical clustering (Fig. 2D). About 50 KEGG pathways were significantly enriched in at least one time point. The quantitative profiles of specific proteins belonging to these pathways are listed in table S3. Down-regulated categories included cell cycle, DNA replication and repair, RNA polymerase and ribosome biogenesis. In general, both GO and KEGG analyses yielded very similar regulated functional protein categories. However, time-point based enrichment of KEGG terms facilitated higher-resolution analysis of regulated pathways. For instance, in addition to vesicle mediated traffic, exocytosis, Soluble NSF Attachment Protein (SNARE) interactions in vesicular transport and phagosome were clearly apparent. Furthermore, sphingolipid metabolism and mammalian target of rapamycin (mTOR) pathway were specifically enriched in the distributions of later time-points. These later time points also showed up-regulation for various pathways such as focal adhesion, ECM-receptor interaction and regulation of actin cytoskeleton, which are in agreement with the established effects of TGF- β on the ECM remodeling during EMT (Fig. 2D). In contrast, the changes in DNA repair pathway and RNA processing observed in both analyses as well as vesicle mediated transport are unexpected categories.

Independent validation of regulated candidate proteins

We chose several proteins for independent assessment by immuno-blotting or quantitative real-time PCR (qRT-PCR), based on extent of regulation, availability of other tools for analysis and possible connections to TGF- β signaling: adhesion molecule with Ig-like domain 2 (AMIGO2), inositol polyphosphate-4-phosphatase type II (INPP4b), programmed cell death

protein 4 (PDCD4), serine protease 23 (PRSS23), transmembrane protein 2 (TMEM2), cyclin dependent kinase 17 (CDK17) and OCIA domain containing protein 2 (OCIAD2). For all of these candidate proteins, we confirmed their TGF- β dependent up-regulation by immuno-blotting and/or qRT-PCR, demonstrating the power of quantitative proteomics to identify new pathway players in an unbiased manner (Fig. 3A).

Further validation was performed by qRT-PCR for PRSS23 which was found to be induced over the examined time course (Fig. 3B, table S2). We knocked down PRSS23 using three different shRNA constructs and confirmed its down-regulation of mRNA expression by qRT-PCR (Fig. 3B). We tested the effect of PRSS23 depletion on TGF- β induced PAI-1 expression, an inhibitor of urokinase-type plasminogen activator (uPA) that mediates extracellular matrix proteolysis. While we observed that knockdown had no effect on SMAD2 phosphorylation, a strong reduction of PAI-1 mRNA and protein expression was found after 6h and 24h of treatment with TGF- β (Fig. 3C). Thus, PRSS23 could be a mediator of TGF- β induced ECM remodeling. In fact, a recent study on endothelial to mesenchymal transition also showed that PRSS23 is essential for this process and its knockdown had no effect on the phosphorylation of SMAD2 but affected SNAIL transcription downstream to SMAD activation by TGF- β (58).

Analysis of upstream transcription regulators induced by TGF- β treatment

As a final outcome of TGF β signaling is transcriptional regulation, we sought to analyse transcriptional regulators that potentially account for our observed proteome changes. To identify transcriptional regulators that are either activated or inhibited in response to TGF- β treatment and act upstream of the observed proteome wide changes, we analyzed the regulated proteins by the upstream regulator analysis module in Ingenuity Pathway Analysis (IPA, Materials and Methods). This resulted in 253 upstream regulators with significant p-values, which included 50 transcription factors and numerous kinases, miRNAs and translational regulators and growth factors such as TGF- β itself (table S4). At each of the time points, we ordered the regulated transcription factors by the strongest regulation (IPA activation score). We clustered the 40 highest scoring transcription factors to identify temporally regulated factors (table S4, Fig. 4A). This analysis validated our approach by correctly predicting key direct

and indirect transcriptional regulators of TGF- β signaling, including SMAD2, SMAD3, SNAIL2, SMAD7 and MYC (17), solely from our measured proteomic expression profiles of their target genes (Fig. 4B). Several of these canonical TGF- β signaling mediators had a constant activation profile (SMAD2/3/4 and SNAIL2), whereas the activity of other activated regulators appeared to be delayed in comparison (Fig. 4A). Our analysis also supports the involvement of several transcription factors, some of which have previously been implicated in TGF- β signaling, including p53 and SMARCA4 (59, 60), while others are not usually or weakly represented as key components of canonical TGF- β induced signaling such as Vitamin D receptor (VDR), TATA box binding protein (TBP)-associated factor 4 (TAF4), SAM pointed domain containing ETS transcription factor (SPDEF), CUT-like homeobox 1 (CUX1), T-box 2 (TBX2), specificity protein 1 (SP1) (61-65). Finally, the analysis revealed novel transcriptional regulators including myotrophin (MTPN) and huntingtin (HTT), whose roles in the context of TGF- β signaling have not been explored.

Phospho-proteomic analysis of early signaling events

Immediate events following TGF- β binding to its receptor are mediated by protein phosphorylation. To study phosphorylation changes and their kinetics at very early stages, we analyzed the phosphoproteome within 20 min of TGF- β stimulus, in contrast to the proteome changes, which were measured from 6h onwards. At such early time points expression level changes will be minimal and observed quantitative changes in phosphopeptides can be attributed to changes at the modification site level. To this end we employed a 'double triple' SILAC approach (stable isotopes labeling by amino acids in cell culture (66)). In one experiment the 0 time point control was encoded as the 'light' SILAC state, the 5 min as the 'medium' encoded time point and the 15 min as the 'heavy' encoded time point. In the other experiment, we compared 0 min (light), 10 min (medium) and 20 min (heavy). Combining the two experiments using the common time point then generated a full time course profile (Fig. 1B). The entire experiment was done in biological triplicates with a SILAC label swap in the third replicate. We employed phospho-peptide enrichment using titanium dioxide microbeads followed by reverse phase chromatography and tandem MS using a quadrupole Orbitrap

instrument (Materials and Methods). The acquired data from the resulting total of 60 fractions were analyzed in the MaxQuant environment with a FDR of 1% at the peptide and protein levels (28). We identified 22,388 different phosphopeptides from 5,345 proteins. A total of 18,841 phosphorylation events could be localized with high confidence to single amino acid sequence locations (class I sites, average localization probability 0.99, table S5). A majority of 16,416 sites were on serine residues (87.2%) while 2,342 were on threonine (12.4 %) and 82 on tyrosine (0.4%). In contrast the previous studies assessing TGF- β response reported identification of 140 (67) and 111 phospho-peptides (68), respectively. To assess the coverage of the measured phosphoproteome, we mapped our sites onto those reported in the comprehensive PhosphositePlus database (69). We were able to match about 12,000 class I sites to existing entries, indicating that we identified approximately 5,000 novel sites.

Of all phosphorylation sites identified, 14,010 were quantified with at least six measurements in response to TGF- β treatment and these were used for all further analyses (table S5). We identified phosphorylation of activating residues on the carboxy terminus of SMAD2 (Ser 467 and S467) and SMAD3 (Ser 423 and S425), direct substrates of the receptor (table S5). The dataset contains many other key phosphosites that are known to be activated by TGF- β treatment, including inhibitory phosphorylation on tyrosine 15 on CDK1/2/3, activating phosphorylation events on TGF- β activating kinase 1 (TAK1), p21 protein (Cdc42/Rac)-activated kinase 1 (PAK1) (Fig. 5A), eukaryotic translation initiation factor 4E binding protein (EIF4EBP1)(Fig. 5A), as well as phosphorylation of key residues of mTOR pathway members (67, 70, 71). To extract phosphosites significantly regulated upon TGF- β treatment, we subjected the normalized dataset to a stringent multiple sample ANOVA test using a permutation based FDR of 0.01. This allowed the identification of 2,892 regulated phosphorylation sites (table S5) which also contained 371 transcription regulators. A Fisher exact test identified kinase substrate motifs that were enriched in the regulated dataset in comparison to the total quantified dataset to obtain cluster specific footprints of kinase activation in response to TGF- β . ERK1/2, CDKs, AKT, ataxia-telangiectasia mutated (ATM) and glycogen synthase kinase 3 (GSK3) substrate motifs had the strongest enrichments (Fig. 5B, $p < 10^{-10}$). Analysis of GO annotation terms and KEGG pathways revealed enrichments of the categories cell cycle, regulation of actin

cytoskeleton, spliceosome, DNA replication, mTOR signaling pathway, adherens/tight junction and focal adhesion, which are mediated by TGF- β (Fig. 5B, table S6). Interestingly, additional categories not generally associated with early TGF- β signaling, but similar to those identified in our expression proteomics experiments, included DNA mismatch repair and endocytosis.

The regulated dataset was clustered in an unsupervised fashion, which identified three major clusters with distinct time course profiles (Fig. 5C). Cluster 1, which consists of phosphosites that peak at earlier time points, was composed of 538 phosphosites. This cluster was enriched for casein kinase II substrate motifs and β -adrenergic receptor kinase motif, suggesting that TGF- β signaling also triggers early up-regulation of this type of phosphorylation. Cluster 3, which had an opposing time profile (increased phosphorylation at 15 and 20 min time points) had 1,386 phosphosites and was enriched for proline directed phosphorylation events (Fig. 5C). Such motifs are specific to MAPKs and CDKs and are otherwise disfavored by the vast majority of Ser/Thr kinases and phosphatases (72). The enriched MAPK motif is in accordance with the known activation of this signaling module by TGF- β through a non-SMAD route (25). The latter is indicative of active signaling by CDKs (fig. S2) that control cell cycle progression. Cluster 2 consisted of 782 phosphosites, which showed an interesting cycling behavior over the time course and was enriched for polo-like kinase 1 (PLK1) polo box domain (PBD) domain binding and GSK3 substrate motifs. While the role of phosphatidylinositol 3 kinase (PI3K)/AKT/GSK3 in response to TGF- β stimulation has been well documented (26), enrichment of PBD domain binding indicates CDK1 priming phosphorylation of PLK1 substrates.

Pathway specific view of TGF- β signaling in cell cycle

We next focused on the cell cycle pathway because one of the major cellular effects of TGF- β in normal cells is growth arrest in G1 and was strongly enriched in all our analysis (73). Despite the low abundance of many of its members we obtained excellent pathway coverage of the cell cycle pathway as mapped in KEGG, with 83 of 98 members identified (Fig 6). Interestingly, this key pathway was one of the most regulated ones, with 48 out of the 83 proteins identified by MS significantly regulated at the phosphoproteome (24 proteins) or proteome levels (31 proteins) or both (7 proteins). The early phosphorylation events (within the first 20 minutes)

tended to occur on cyclins and their downstream effectors, whereas proteome changes (measured from 6 to 48h) tended to take place at all the levels of the pathway. Our data thus highlight the usefulness of measuring both the proteome and phosphoproteome to measure the impact of signaling on downstream pathways. A number of proteins driving the cell cycle were down-regulated in their abundance after 24h by TGF- β treatment, eventually leading to cell cycle arrest (Fig. 6). In contrast, temporal profiles corresponding to early TGF- β signaling showed mixed profiles of phosphosites on individual proteins (or protein groups) at key residues. These were inhibitory phosphorylation Tyr 15 on CDK1/2/3 and Ser 64 on S-phase kinase-associated protein 2 (SKP2), which exhibited increasing phosphorylation over time, while activating phosphorylations on Ser 1068 and Ser 1112 on p130/RBL2 which were decreased. While some phosphorylation events are indicative of proliferative signaling (such as RB1 phosphorylation), others indicated inhibition of this process (inhibitory phosphorylation on CDK1/2/3, activation of SMADs). Checkpoint activation is not normally associated with TGF- β signaling but was clearly evident by two-fold decrease of phosphorylation S120 on checkpoint kinase 2 (CHK2).

Mechanistic network of the regulated proteome and phosphoproteome

Analyzing all pathways, instead of only the cell cycle, also showed enrichment of similar categories in the phosphoproteome and proteome, including cell cycle, focal adhesion, regulation of actin cytoskeleton and DNA repair pathways (table S6). To investigate these in greater depths we employed BioCarta (<http://www.biocarta.com/>), which offers detailed signaling pathway information. We observed strong enrichments of specific pathways like vitamin D3 receptor signaling, ATM signaling pathway, chromatin remodeling, cell cycle, RHO cell motility signaling pathway and mechanism of gene regulation by peroxisome proliferator-activated receptors α (PPAR α). Our data set provides extensive quantitative details for each individual pathway player (table S6).

TGF- β mediated signal transduction is usually envisioned as a linear process driven by canonical SMAD signaling. However, there is increasing interest in non-SMAD pathways and accompanying cross-talk to other signaling pathways. Therefore, for a systems perspective, we

sought to extract a functional signaling network driving TGF- β mediated EMT and cell cycle arrest by combining our data with literature knowledge on protein-protein interactions and kinase-substrate relations. We constructed a graph using (a) the top 100 most significantly regulated proteins from the proteome analysis, (b) the 271 phosphoproteins containing significantly regulated phosphosites including 40 regulatory phosphosites usually followed by antibodies in small-scale studies and (c) the direct transcriptional mediators of TGF- β induced response (SMAD2, SMAD3, SMAD7 and SNAIL2), which were also identified in our analysis of upstream regulators. This list of TGF- β responsive proteins and phosphoproteins was mapped onto high confidence protein-protein interactions in STRING (74). Remarkably, this analysis identified a well-connected core network of 150 proteins. This network was rendered in Cytoscape (75) with additional mapping of kinase substrate relationships as annotated in PhosphoSitePlus (69) and is illustrated in Fig. 7. It turned out to be enriched for proteins associated with focal adhesion (29 members; $p < 3e-12$) and cell cycle (18 members; $p < 1e-9$). Hubs within the network comprised mTOR signaling pathway members, small GTP-binding proteins, nucleoporins, MCM complex, ECM/adhesion components and proteins regulating the actin cytoskeleton. The temporal data and cellular location annotations of the network clearly confirmed that TGF- β leads to temporal accumulation of specific ECM components, an essential process for cell migration during EMT. Multiple regulated phosphorylation events on the plasma membrane indicated potential crosstalk between different receptors including integrins and tight junction proteins. Proteins associated with plasma membrane also include proteins that concentrate in areas of cell-cell and cell-matrix contacts as well as scaffolding proteins such as TLN1, BCAR1, NF2 and SPTBN1 linking plasma membrane to actin cytoskeleton, focal adhesions and eventually cell migration. Additionally, the depicted protein tyrosine kinases (including tyrosine protein kinase SRC and epidermal growth factor receptor (EGFR)) contribute to non-canonical TGF- β signaling in mesenchymal or dedifferentiated epithelial cells (48).

The network is clearly centered on key transcription factors and kinases. In fact, we identified several transcription factors with regulated phosphorylation events after TGF- β treatment including SMADs and SMARCA4, mediator complex subunit 1 (MED1), catenin β 1 (CTNNB1), RB1 and p53 (marked by asterisks in Fig. 4). The major TGF- β responsive kinases present in the

network were CDK1/2, SRC, protein kinase C δ (PRKCD), PAK1, PAK2, Rho-associated, coiled-coil containing protein kinase 2 (ROCK2) and mitogen-activated protein kinase kinase kinase 4 (MAP4K4). These proteins are central to the network, connecting phosphorylation signaling to proteomic changes associated with the TGF- β induced cellular response.

Overall, our proteomics-derived signaling network covering TGF- β regulated proteins, early phosphorylation events, upstream kinases and signaling output mediating transcription factors highlights the multi-functionality of this cytokine (Fig. 7). It supports the notion that its biological output is a result of extensive crosstalk with other signaling pathways such as EGFR and integrins, as well as direct signaling through both canonical SMAD and non-SMAD pathways (26). Such a global view also sheds light on individual mechanisms of signaling. For example, TGF- β activation of AKT is reflected in inhibitory phosphorylation of GSK3 α (Ser21), activating AKT1S1 phosphorylation and downstream mTOR pathway members. GSK3 α phosphorylates β -catenin, decreasing its stability. Our analysis revealed increased phosphorylation of β -catenin, not on the site downstream to GSK3 but rather on the site phosphorylated by PAK1, a kinase that was also activated in response to TGF- β . Phosphorylation of this PAK1 site is known to increase β -catenin stability resulting in accumulation of this key transcription regulator (76). Opposite regulation of upstream kinases acting on both these sites provides an explanation for up-regulation of β -catenin in our proteomic data. This observation illustrates that while a network view shows multiple pathways activated by TGF- β , a site resolved view allows visualization of how these pathways cross talk.

Discussion

Utilizing state of the art technology, the near complete yeast proteome and a very large percentage of the mammalian proteome can now be captured in a short time in 'single-run' analysis. The attraction of this approach, especially when combined with label-free quantification, is that it avoids any upfront protein or peptide fractionation and provides a convenient systems-wide view of the proteome (77). Here we employed this strategy to study TGF- β signaling induced proteome changes across six different time points and corresponding controls in quadruplicates in just eight days of measurement time. In comparison, a classical

fractionation based approach would have taken almost two months of measuring time. Additionally, we analyzed early time-resolved phosphorylation based signaling events that had never been studied at this scale. Our dataset contains many protein and phosphorylation changes that have already been described in TGF- β signaling in the HaCaT cell system, such as those associated with robust cell cycle inhibition, extracellular matrix formation and EMT. We also uncovered many novel regulated protein and phosphorylation changes that should represent a valuable resource for the TGF- β community. We validated several of these novel TGF- β induced proteins and identified PRSS23 as a regulator of ECM remodeling (Fig. 3).

To our knowledge this is the first study to obtain a deep time resolved molecular snapshot of both early and late events involved in the cellular signaling response to a ligand. Widespread phosphorylation changes preceded massive proteomic changes, encompassing a third of the quantified proteome. Upstream transcription regulator analysis and network analysis integrated these two systems level data sets and pointed to several phosphorylated transcription factors or cofactors as candidate regulators of TGF- β -induced gene expression (Fig. 5,7). The core SMAD and MYC transcription factors were regulated throughout the entire time course and our analysis provided downstream targets of these regulators (Fig. 5, table S4). We detected additional transcriptional regulators, which may act independently or in conjunction with the SMAD complex to achieve high affinity and selectivity for specific subsets of target genes. One such candidate is the vitamin D3 receptor (VDR), a ligand-dependent transcription factor that is activated upon binding to vitamin D or analogs. VDR was a significant hit in our upstream regulator analysis and “control of gene expression by vitamin D receptor” was the most enriched category in our pathway analysis (Fig 4A, table S6). VDR has recently been implicated in the regulation of keratinocyte proliferation(78) and differentiation as well as EMT (79) and in a negative feedback loop in global regulation of SMAD signaling (65). Myotrophin (MTPN) is another interesting regulator that we found to be activated via the above analysis and to be induced by TGF- β (Fig. 4A, B). Its role in classical TGF- β signaling remains unexplored but its overexpression has been reported to lead to increased TGF- β and FN1 expression in cardiac hypertrophy (80). Thus MTPN and TGF- β could be part of a positive feedback loop.

Epithelial cells respond to TGF- β by cell cycle inhibition while gaining a motile phenotype through EMT. These well-described signaling outputs were prominent in our GO enrichment and KEGG pathway analyses (Fig. 2C,D), which showed strong up-regulation of focal adhesion, cytoskeleton and ECM proteins, and down-regulation of proteins that drive the cell cycle. Our measurements provide a temporal view of these processes, indicating major adjustments of biological pathways after about 24h stimulation. Analyses of this time-resolved proteomic data allowed us to identify new TGF- β regulated pathways, which are likely to contribute to the induced biological processes.

One such pathway is vesicular transport, with endocytosis, SNARE interactions, and lysosomes significantly enriched (Fig 2C, table S6). The prominent upregulation of endocytic pathways could represent feed-forward and negative regulatory loops that are activated upon TGF- β signaling to modulate signaling capacity. For example, internalization of activated TGF- β receptor complexes via the caveolin route leads to receptor degradation, and terminates signaling (81). The up-regulation of vesicle mediated transport pathways along TGF- β induced EMT could also suggest its role in this process. During EMT, tight junction proteins are internalized via clathrin mediated endocytosis, which together with recycling of focal adhesion components such as integrins is necessary for migration (82, 83). Finally, along with the induction of extracellular matrix components such as fibronectin, collagens and laminins, the observed TGF- β mediated induction of secretory pathway can be explained by the increased demand for secretion of these ECM components.

Investigation of cell cycle pathway members revealed that long term proteome changes are clearly tailored to mediate cell cycle arrest. Early phosphorylation based signaling indicates increased phosphorylation of SMAD transcription factors, direct TGF- β substrates known to induce transcription of the CDK inhibitors p21CIP1 and p15/INK4B (20, 84), as also evidenced in our data. Likewise in accordance with earlier reports (67, 85), we observed induced inhibitory tyrosine phosphorylation of CDK1/2/3 as an early inactivating signaling response much before a decrease in protein level (Fig 6). However, many other phosphorylation events suggest CDK activation. For example, we observed an activating phosphorylation (Ser375) of upstream

phosphatase, CDC25B, which in turn activates CDK1 (86) and downstream phosphorylation of classical CDK substrates including SKP2, DNA Ligase 1 (LIG1), high mobility group protein HMGA1, Nuclear Casein Kinase And Cyclin-Dependent Kinase Substrate (NUCKS) and Retinoblastoma 1 (RB1) (Fig. 6,7, S2) (87). This was also recapitulated in our kinase motif analysis, which shows enrichment of CDK substrates among sites peaking at 15 to 20min after TGF- β addition (Fig. 5, Cluster 3).

Direct phosphorylation of RB1 by CDK1/2 inactivates this otherwise tumor-suppressive protein. During cell cycle arrest TGF- β inhibits CDK1/2 and thereby preventing this inactivation (88). However, in some cell types where TGF- β has growth stimulatory effects, it activates CDK1/2 leading to increased RB1 phosphorylation (89). At the early time points studied in our phosphoproteomics experiment, we observed up-regulation of phosphorylation of this protein at multiple sites in response to TGF- β stimulation (Fig 5, 7). At these time points, TGF- β may exert both stimulatory and inhibitory effects on the cell cycle, whereas at later time points the inhibitory signals predominate, leading to the very prominent cell cycle inhibition visible in the proteomic data.

Supporting this hypothesis, our data on inactivating hyperphosphorylation site of EIF4EBP1, a translation-inhibitory protein, at early time-points is suggestive of proliferation, because it frees eIF4E and consequently stimulates protein synthesis leading to cell growth and proliferation (90, 91) (Fig. 7). At later time points, induction of EIF4EBP1 has an established role in mediating the antiproliferative effects of TGF- β (92). The non-canonical TGF- β induced signaling pathways (mainly the PI3K/AKT/mTOR cascade) that control EIF4EBP1 phosphorylation are likely responsible for this early mTOR pathway regulation. (25). A similar early mitogenic cross talk and/or activation of CDKs can explain regulation of RB1 and multiple other CDK substrates. Later signaling events accompanied by protein level adjustments then override the proliferative signaling leading to cell cycle arrest. In this regard it is pertinent that imbalances in the activation status of canonical and noncanonical TGF- β signaling systems may underlie the pro-tumorigenic effects of TGF- β (4).

In conclusion, this study provides a new, global perspective on TGF- β signaling. To elucidate how TGF- β signal travels from the membrane as phosphorylation events to the transcription regulators in nucleus, we weaved together phosphoproteome and ensuing proteome events. We identified several TFs acting at the intersection of TGF- β induced early phosphorylation events and transcription activation of proteins involved in ECM remodeling and epithelial dedifferentiation. Finally, our analysis suggests that the mediators of such contextual effects on cell growth are encoded on a temporal scale where early events constitute a mixed signal that is tailored into a pro or anti-proliferative response at later time points. In this way, it sheds new light on the well-known propensity of TGF- β responses to be dictated by cellular context.

Materials and Methods

Cell culture and TGF- β treatment

HaCaT cells were pre-incubated with DMEM containing 2% serum for 12h and treated with 5 ng/mL TGF- β 1 (Peprotech) for 0, 6, 12, 24, 36 and 48h (Fig. 1B). Untreated controls were also collected at all time points. For phosphoproteomics studies we employed a double triple SILAC based approach as described earlier (28). SILAC encoded HaCaT cells were incubated in serum free medium for 6h and treated with TGF- β 1 for 0, 5 and 15 min (Fig. 1B). A second, identically labeled set of cells was treated for 0, 10 and 20 min.

Total proteome and phosphoproteome sample preparation and MS analyses

Cells were lysed in buffer containing 6 M guanidium chloride, CAA, TCEP and 100 mM Tris-HCl pH 8.5 (93). The lysate was sonicated, incubated for 5 min at 95°C, quantified and processed by in-solution digestion. Briefly, 20 μ g of cell lysate was digested overnight at 37°C in buffer containing 10% ACN, 25 mM Tris-HCl pH 8.5, 0.03 μ g trypsin (Promega). For phosphoproteome analysis, samples were prepared as described earlier (94). Briefly, cells were lysed in buffer containing SDS based buffer, sonicated and centrifuged for 15 min at 14,000 rpm. Cell lysates from light, medium and heavy conditions were mixed in a 1:1:1 ratio and a total of 10 mg sample was digested according to the FASP method (95). The resulting peptides were fractionated by strong cation exchange (SCX) chromatography and subjected to phosphopeptide enrichment using TiO₂ beads (28, 96). The peptides or phosphopeptides were

desalted on StageTips (97) and separated on a reverse phase column (packed in-house with 1.8- μm C₁₈- Reprosil-AQ Pur reversed-phase beads) over 270 minutes (single-run proteome analysis) or 120 to 240 mins (fractionated phosphoproteome). The peptides eluting at the tip were electrosprayed and analyzed by tandem mass spectrometry on a Q Exactive (41) using HCD based fragmentation, which was set to alternate between a full scan followed by up to 5 or 10 fragmentation scans.

Data processing and analysis

Raw mass spectrometric data was analyzed in the MaxQuant environment (42), versions 1.3.10.15 and 1.3.10.18, and employing Andromeda for database search (98). The MS/MS spectra were matched against the human International Protein Index sequence database (IPI version 3.37). Enzyme specificity was set to trypsin, allowing for cleavage N-terminal to proline and between aspartic acid and proline. The search included cysteine carbamidomethylation as a fixed modification and N-acetylation of protein, oxidation of methionine and/or phosphorylation of serine, threonine tyrosine residue (STY) as variable modifications. For phosphopeptide identification, an Andromeda minimum score and minimum delta score threshold of 40 and 17 was used, respectively. Up to two missed cleavages were allowed for protease digestion and peptide had to be fully tryptic. The 'identify' module in MaxQuant was used to filter identifications at 1% FDR at the peptide and protein level. The mass spectrometry proteomics data are deposited at the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD000496 (for reviewing process the data is accessible with the username: review12691 and password: xsRpTFzK).

Downstream bioinformatics analyses

Bioinformatic analysis was performed in the Perseus software environment, which is part of MaxQuant. Hierarchical clustering of proteins or phosphosites was performed on logarithmized intensities or ratios for the data that was quantified in at least 50% of the time points studied. For multiple sample t-tests analysis (ANOVA), replicates were grouped and the statistical test was performed with a permutation-based FDR cutoff of 0.01 and 0.05 for the proteome and

phosphoproteome, respectively. The phosphoproteome dataset was normalized by z-scoring the log2 ratios across SILAC experiments.

Categorical annotation was supplied in the form of GO biological process, molecular function, and cellular component, KEGG pathways for pathway annotation and human protein reference database (HPRD) for kinase substrate motifs. Enrichment for these categories was evaluated by Fisher exact test to obtain a p-value or EASE score p-values (99). The annotation matrix algorithm was used to compute the difference of any significant protein annotation term from the overall intensity distribution as described (43, 57). The specific test we used is a two-dimensional version of the nonparametric Mann-Whitney test. Multiple hypothesis testing was controlled by using a Benjamini-Hochberg FDR threshold of 0.05. Fisher exact test for motif enrichment was performed with a FDR value of 0.02.

Upstream regulator analysis: This analysis was performed in IPA (<http://www.ingenuity.com/products/ipa>). Upstream regulator analysis employed previous knowledge of expected effects of transcriptional regulators (TRs) and their target genes as stored in the Ingenuity® Knowledge Base. Two statistical measures, standard in IPA, were used to detect potential TRs: an overlap p-value and an activation Z-score. First, the analysis examined how many known targets of each TR were present in our data set, resulting in an estimation of an overlap p-value. Second, the known effect (activation or suppression) of a TR on each target gene was compared with observed changes in gene expression. Based on concordance between them, an activation Z-score was assigned, showing whether a potential TR was in 'activated' (Z-score > 2) or 'inhibited' (Z-score < -2).

Immuno-blotting and quantitative real-time (qRT)-PCR: TGF- β dependent induction of novel proteins identified by proteomics was validated by immuno-blotting of lysates prepared for mass spectrometry with anti-INPP4b (Cell signaling technology), anti-PDCD4 (Cell signaling technology), anti-AMIGO2 (R&D systems), anti-CDK17 (Prestige antibodies), anti-TMEM2 (Abcam), and anti-OCIAD2 (Sigma Aldrich). Equal loading was controlled with anti- β -Actin (Cell signaling technology).

shRNA experiments: For knockdown experiments, lentiviral shRNA constructs targeting PRSS23, and the non-targeting control shRNA were obtained from Sigma-Aldrich (MISSION® shRNA; shPRSS23#1 TRCN0000047039, shPRSS23#2 TRCN0000047040, shPRSS23#3 TRCN0000047042), and produced as described (100). Cells were infected for 12 h with lentiviral supernatants in the presence of 5 ng/mL polybrene (Sigma-Aldrich). After selection with 1 µg/mL puromycin (Sigma-Aldrich) for 48 h, the cells were pre-incubated overnight in DMEM containing 5 % FBS, and treated with 5 ng/mL TGF-β (generously provided by Kenneth K. Iwata, OSI pharmaceuticals, New York, USA) for the hours indicated in Figure 3. To determine PRSS23 knockdown efficacy by qRT-PCR, total RNA was isolated by NucleoSpin® RNA II kit (Macherey-Nagel), reverse-transcribed by RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific), and analyzed using primers Fw 5'-GAGCCGAAGCCAAATTAGAA-3' and Rv 5'-AGGATGTAGATGCCCCACCTG-3'. Expression was normalized to the expression of β-Actin (Fw 5'-AATGTCGCGGAGGACTTTGATTGC-3', Rv 5'-AGGATGGCAAGGGACTTCCTGTAA-3'). The effect of PRSS23 knockdown on PAI-1 expression was determined by qRT-PCR (Fw 5'-CACAAATCAGACGGCAGCACT-3', Rv 5'-CATCGGGCGTGGTGAACCTC-3') and immuno-blotting (anti-PAI-1, BD Biosciences), and on SMAD2 phosphorylation (101) by immuno-blotting. Equal loading was controlled with anti-β-Actin (Sigma-Aldrich).

Confocal microscopy: E-Cadherin and F-actin were visualized as described (100).

Growth arrest assay by fluorescence activated cell sorting (FACS): To detect TGF-β induced cell cycle arrest, HaCaT cells were pre-treated with DMEM containing 5% FBS for 16 h and stimulated with or without 5 ng/mL TGF-β3 (OSI pharmaceuticals, New York, USA) for 40 h. Cells were labeled with 20 µM bromodeoxyuridine (BrdU; Roche) for 2 h, harvested, and fixed in 70% ethanol. After RNase A treatment (50 µg/mL for 30min) and DNA denaturation (5 M HCl/0.5% Triton X-100 for 20 min), cells were stained with anti-BrdU-FITC (Boehringer Mannheim) and propidium iodide (Sigma-Aldrich). The cells were analyzed for FITC (BrdU incorporation) and propidium iodide (total DNA content) fluorescence by a BD LSR II flow cytometer (BD Biosciences).

List of Supplementary data

fig. S1: A Pearson correlation matrix for protein quantification across biological quadruplicates measured for six different time points after TGF- β treatment.

fig. S2: Interaction map showing high confidence interaction of CDK1/2 with various substrates that were regulated by TGF- β treatment.

table S1: Protein groups table (supplied separately via ftp because of Science Signaling file size restriction, please contact the editor.)

table S2: Protein groups with ANOVA 0.01 (supplied separately via ftp because of Science Signaling file size restriction, please contact the editor.)

table S3: Protein groups corresponding to KEGG pathways up-regulated or down-regulated and cluster specific enrichments for proteome measurements

table S4: IPA all time points TF table and other upstream regulators

table S5: Phosphosites total and regulated (supplied separately via ftp because of Science Signaling file size restriction, please contact the editor.)

table S6: Enrichment tables

table S7: Explanation for proteins and phosphosites tables from MaxQuant

References

1. J. Massague, TGFbeta in Cancer. *Cell* **134**, 215 (2008); published online EpubJul 25 (10.1016/j.cell.2008.07.001).
2. A. Moustakas, K. Pardali, A. Gaal, C. H. Heldin, Mechanisms of TGF-beta signaling in regulation of cell growth and differentiation. *Immunology letters* **82**, 85 (2002); published online EpubJun 3
3. G. C. Blobe, W. P. Schiemann, H. F. Lodish, Mechanisms of disease: Role of transforming growth factor beta in human disease. *New Engl J Med* **342**, 1350 (2000); published online EpubMay 4 (Doi 10.1056/Nejm200005043421807).
4. J. Massague, TGF beta signalling in context. *Nat Rev Mol Cell Bio* **13**, 616 (2012); published online EpubOct (Doi 10.1038/Nrm3434).

5. B. Bierie, H. L. Moses, Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. *Nature reviews. Cancer* **6**, 506 (2006); published online EpubJul (10.1038/nrc1926).
6. R. J. Akhurst, R. Derynck, TGF-beta signaling in cancer--a double-edged sword. *Trends in cell biology* **11**, S44 (2001); published online EpubNov
7. C. H. Heldin, M. Vanlandewijck, A. Moustakas, Regulation of EMT by TGFbeta in cancer. *FEBS letters* **586**, 1959 (2012); published online EpubJul 4 (10.1016/j.febslet.2012.02.037).
8. J. Zavadil, L. Cermak, N. Soto-Nieves, E. P. Bottinger, Integration of TGF-beta/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition. *The EMBO journal* **23**, 1155 (2004); published online EpubMar 10 (10.1038/sj.emboj.7600069).
9. M. M. Verbeek, I. Otte-Holler, P. Wesseling, D. J. Ruiter, R. M. de Waal, Induction of alpha-smooth muscle actin expression in cultured human brain pericytes by transforming growth factor-beta 1. *The American journal of pathology* **144**, 372 (1994); published online EpubFeb
10. R. A. Imitola, J. Massague, Transforming growth factor-beta stimulates the expression of fibronectin and collagen and their incorporation into the extracellular matrix. *The Journal of biological chemistry* **261**, 4337 (1986); published online EpubMar 25
11. K. Miyazono, Transforming growth factor-beta signaling in epithelial-mesenchymal transition and progression of cancer. *P Jpn Acad B-Phys* **85**, 314 (2009); published online EpubOct (Doi 10.2183/Pjab.85.314).
12. T. Tsuji, S. Ibaragi, G. F. Hu, Epithelial-mesenchymal transition and cell cooperativity in metastasis. *Cancer research* **69**, 7135 (2009); published online EpubSep 15 (10.1158/0008-5472.CAN-09-1618).
13. C. Scheel, R. A. Weinberg, Cancer stem cells and epithelial-mesenchymal transition: Concepts and molecular links. *Semin Cancer Biol* **22**, 396 (2012); published online EpubOct
14. T. Tsukazaki, T. A. Chiang, A. F. Davison, L. Attisano, J. L. Wrana, SARA, a FYVE domain protein that recruits Smad2 to the TGF beta receptor. *Cell* **95**, 779 (1998); published online EpubDec 11 (Doi 10.1016/S0092-8674(00)81701-8).
15. Y. Zhang, X. Feng, R. We, R. Derynck, Receptor-associated Mad homologues synergize as effectors of the TGF-beta response. *Nature* **383**, 168 (1996); published online EpubSep 12 (10.1038/383168a0).
16. G. Lagna, A. Hata, A. HemmatiBrivanlou, J. Massague, Partnership between DPC4 and SMAD proteins in TGF-beta signalling pathways. *Nature* **383**, 832 (1996); published online EpubOct 31 (Doi 10.1038/383832a0).
17. J. Massague, D. Wotton, Transcriptional control by the TGF-beta/Smad signaling system. *The EMBO journal* **19**, 1745 (2000); published online EpubApr 17 (10.1093/emboj/19.8.1745).
18. K. Luo, S. L. Stroschein, W. Wang, D. Chen, E. Martens, S. Zhou, Q. Zhou, The Ski oncoprotein interacts with the Smad proteins to repress TGFbeta signaling. *Genes & development* **13**, 2196 (1999); published online EpubSep 1
19. S. Akiyoshi, H. Inoue, J. Hanai, K. Kusanagi, N. Nemoto, K. Miyazono, M. Kawabata, c-Ski acts as a transcriptional co-repressor in transforming growth factor-beta signaling

- through interaction with smads. *The Journal of biological chemistry* **274**, 35269 (1999); published online EpubDec 3 (
20. K. Pardali, A. Kurisaki, A. Moren, P. ten Dijke, D. Kardassis, A. Moustakas, Role of smad proteins and transcription factor Sp1 in p21(Waf1/Cip1) regulation by transforming growth factor-beta. *Journal of Biological Chemistry* **275**, 29244 (2000); published online EpubSep 22 (DOI 10.1074/jbc.M909467199).
 21. G. J. Hannon, D. Beach, P15(Ink4b) Is a Potential Effector of Tgf-Beta-Induced Cell-Cycle Arrest. *Nature* **371**, 257 (1994); published online EpubSep 15 (Doi 10.1038/371257a0).
 22. X. Hua, Z. A. Miller, G. Wu, Y. Shi, H. F. Lodish, Specificity in transforming growth factor beta-induced transcription of the plasminogen activator inhibitor-1 gene: interactions of promoter DNA, transcription factor muE3, and Smad proteins. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 13130 (1999); published online EpubNov 9
 23. Y. Watanabe, S. Itoh, T. Goto, E. Ohnishi, M. Inamitsu, F. Itoh, K. Satoh, E. Wiercinska, W. Yang, L. Shi, A. Tanaka, N. Nakano, A. M. Mommaas, H. Shibuya, P. Ten Dijke, M. Kato, TMEPAI, a transmembrane TGF-beta-inducible protein, sequesters Smad proteins from active participation in TGF-beta signaling. *Molecular cell* **37**, 123 (2010); published online EpubJan 15 (10.1016/j.molcel.2009.10.028).
 24. B. A. Hocevar, T. L. Brown, P. H. Howe, TGF-beta induces fibronectin synthesis through a c-Jun N-terminal kinase-dependent, Smad4-independent pathway. *The EMBO journal* **18**, 1345 (1999); published online EpubMar 1 (10.1093/emboj/18.5.1345).
 25. Y. E. Zhang, Non-Smad pathways in TGF-beta signaling. *Cell research* **19**, 128 (2009); published online EpubJan (10.1038/cr.2008.328).
 26. Y. Mu, S. K. Gudey, M. Landstrom, Non-Smad signaling pathways. *Cell and tissue research* **347**, 11 (2012); published online EpubJan (10.1007/s00441-011-1201-y).
 27. N. Awai-Kasaoka, T. Inoue, T. Kameda, T. Fujimoto, M. Inoue-Mochita, H. Tanihara, Oxidative stress response signaling pathways in trabecular meshwork cells and their effects on cell viability. *Mol Vis* **19**, 1332 (2013); published online EpubJun 12
 28. J. V. Olsen, B. Blagoev, F. Gnäd, B. Macek, C. Kumar, P. Mortensen, M. Mann, Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635 (2006); published online EpubNov 3 (10.1016/j.cell.2006.09.026).
 29. C. Choudhary, M. Mann, Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews. Molecular cell biology* **11**, 427 (2010); published online EpubJun (10.1038/nrm2900).
 30. N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Paabo, M. Mann, Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548 (2011)10.1038/msb.2011.81).
 31. L. M. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther, M. Mann, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251 (2008); published online EpubOct 30 (10.1038/nature07341).
 32. F. He, A. Kumar, Z. H. Song, Heat shock protein 90 is an essential molecular chaperone for CB2 cannabinoid receptor-mediated signaling in trabecular meshwork cells. *Mol Vis* **18**, 2839 (2012); published online EpubNov 29

33. N. Nagaraj, N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm, M. Mann, System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Molecular & cellular proteomics : MCP* **11**, M111 013722 (2012); published online EpubMar (10.1074/mcp.M111.013722).
34. D. Vergara, P. Simeone, P. del Boccio, C. Toto, D. Pieragostino, A. Tinelli, R. Acierno, S. Alberti, M. Salzet, G. Giannelli, P. Sacchetta, M. Maffia, Comparative proteome profiling of breast tumor cell lines by gel electrophoresis and mass spectrometry reveals an epithelial mesenchymal transition associated protein signature. *Molecular bioSystems* **9**, 1127 (2013); published online EpubJun (10.1039/c2mb25401h).
35. J. Milosevic, P. Bulau, E. Mortz, O. Eickelberg, Subcellular fractionation of TGF-beta 1-stimulated lung epithelial cells: A novel proteomic approach for identifying signaling intermediates. *Proteomics* **9**, 1230 (2009); published online EpubMar (DOI 10.1002/pmic.200700604).
36. N. A. Ali, M. J. McKay, M. P. Molloy, Proteomics of Smad4 regulated transforming growth factor-beta signalling in colon cancer cells. *Molecular bioSystems* **6**, 2332 (2010); published online EpubNov (10.1039/c0mb00016g).
37. K. E. Bollinger, J. S. Crabb, X. L. Yuan, T. Putliwala, A. F. Clark, J. W. Crabb, Quantitative Proteomics: TGF beta(2) Signaling in Trabecular Meshwork Cells. *Invest Ophth Vis Sci* **52**, 8287 (2011); published online EpubOct (Doi 10.1167/lovs.11-8218).
38. J. Zavadil, M. Bitzer, D. Liang, Y. C. Yang, A. Massimi, S. Kneitz, E. Piek, E. P. Bottinger, Genetic programs of epithelial cell plasticity directed by transforming growth factor-beta. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 6686 (2001); published online EpubJun 5 (DOI 10.1073/pnas.111614398).
39. M. Kato, A. Ishizaki, U. Hellman, C. Wernstedt, M. Kyogoku, K. Miyazono, C. H. Heldin, K. Funo, A human keratinocyte cell line produces two autocrine growth inhibitors, transforming growth factor-beta and insulin-like growth factor binding protein-6, in a calcium- and cell density-dependent manner. *The Journal of biological chemistry* **270**, 12373 (1995); published online EpubMay 26
40. A. Capone, V. Visco, F. Belleudi, C. Marchese, G. Cardinali, M. Bellocchi, M. Picardo, L. Frati, M. R. Torrisi, Up-modulation of the expression of functional keratinocyte growth factor receptors induced by high cell density in the human keratinocyte HaCaT cell line. *Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research* **11**, 607 (2000); published online EpubNov
41. A. Michalski, E. Damoc, J. P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, S. Horning, Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & cellular proteomics : MCP* **10**, M111 011015 (2011); published online EpubSep (10.1074/mcp.M111.011015).
42. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367 (2008); published online EpubDec (10.1038/nbt.1511).
43. T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins.

- Molecular & Cellular Proteomics* **11**, (2012); published online EpubMar (DOI 10.1074/mcp.M111.014050).
44. C. A. Lubber, J. Cox, H. Lauterbach, B. Fancke, M. Selbach, J. Tschopp, S. Akira, M. Wiegand, H. Hochrein, M. O'Keefe, M. Mann, Quantitative Proteomics Reveals Subset-Specific Viral Recognition in Dendritic Cells. *Immunity* **32**, 279 (2010); published online EpubFeb 26 (DOI 10.1016/j.immuni.2010.01.013).
 45. M. D. George, T. M. Vollberg, E. E. Floyd, J. P. Stein, A. M. Jetten, Regulation of transglutaminase type II by transforming growth factor-beta 1 in normal and transformed human epidermal keratinocytes. *The Journal of biological chemistry* **265**, 11098 (1990); published online EpubJul 5
 46. S. Akiyoshi, M. Ishii, N. Nemoto, M. Kawabata, H. Aburatani, K. Miyazono, Targets of transcriptional regulation by transforming growth factor-beta: Expression profile analysis using oligonucleotide arrays. *Jpn J Cancer Res* **92**, 257 (2001); published online EpubMar
 47. E. Honda, K. Yoshida, H. Munakata, Transforming Growth Factor-beta Upregulates the Expression of Integrin and Related Proteins in MRC-5 Human Myofibroblasts. *Tohoku J Exp Med* **220**, 319 (2010); published online EpubApr (Doi 10.1620/Tjem.220.319).
 48. R. A. Ignatz, T. Endo, J. Massague, Regulation of fibronectin and type I collagen mRNA levels by transforming growth factor-beta. *The Journal of biological chemistry* **262**, 6443 (1987); published online EpubMay 15
 49. A. Sethi, R. J. Wordinger, A. F. Clark, Gremlin utilizes canonical and non-canonical TGF beta signaling to induce lysyl oxidase (LOX) genes in human trabecular meshwork cells. *Exp Eye Res* **113**, 117 (2013); published online EpubAug (DOI 10.1016/j.exer.2013.05.011).
 50. S. Diskin, W. S. Chen, Z. Y. Cao, S. Gyawali, H. Y. Gong, A. Soza, A. Gonzalez, N. Panjwani, Galectin-8 Promotes Cytoskeletal Rearrangement in Trabecular Meshwork Cells through Activation of Rho Signaling. *Plos One* **7**, (2012); published online EpubSep 4 (ARTN e44400DOI 10.1371/journal.pone.0044400).
 51. J. S. Munger, X. Z. Huang, H. Kawakatsu, M. J. D. Griffiths, S. L. Dalton, J. F. Wu, J. F. Pittet, N. Kaminski, C. Garat, M. A. Matthay, D. B. Rifkin, D. Sheppard, The integrin alpha v beta 6 binds and activates latent TGF beta 1: A mechanism for regulating pulmonary inflammation and fibrosis. *Cell* **96**, 319 (1999); published online EpubFeb 5 (Doi 10.1016/S0092-8674(00)80545-0).
 52. S. T. Eblen, M. P. Fautsch, R. J. Burnette, P. Joshi, E. B. Leof, Cell Cycle-Dependent Inhibition of P34(Cdc2) Synthesis by Transforming Growth-Factor Beta(1) in Cycling Epithelial-Cells. *Cell Growth & Differentiation* **5**, 109 (1994); published online EpubFeb
 53. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, G. O. Consortium, Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25 (2000); published online EpubMay
 54. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, The KEGG databases at GenomeNet. *Nucleic acids research* **30**, 42 (2002); published online EpubJan 1

55. M. Chandra, S. B. Zang, H. Q. Li, L. J. Zimmerman, J. Champer, A. Tsuyada, A. Chow, W. Y. Zhou, Y. Yu, H. Gao, X. B. Ren, R. J. Lin, S. E. Wang, Nuclear Translocation of Type I Transforming Growth Factor beta Receptor Confers a Novel Function in RNA Processing. *Mol Cell Biol* **32**, 2183 (2012); published online EpubJun (Doi 10.1128/Mcb.00320-12).
56. G. Thillainadesan, J. M. Chitilian, M. Iovic, J. N. G. Ablack, J. S. Mymryk, M. Tini, J. Torchia, TGF-beta-Dependent Active Demethylation and Expression of the p15(ink4b) Tumor Suppressor Are Impaired by the ZNF217/CoREST Complex. *Molecular cell* **46**, 636 (2012); published online EpubJun 8 (DOI 10.1016/j.molce.2012.03.027).
57. J. Cox, M. Mann, 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *Bmc Bioinformatics* **13**, (2012); published online EpubNov 5 (Artn S12 Doi 10.1186/1471-2105-13-S16-S12).
58. I. H. Chen, H. H. Wang, Y. S. Hsieh, W. C. Huang, H. I. Yeh, Y. J. Chuang, PRSS23 is essential for the Snail-dependent endothelial-to-mesenchymal transition during valvulogenesis in zebrafish. *Cardiovasc Res* **97**, 443 (2013); published online EpubMar 1 (Doi 10.1093/Cvr/Cvs355).
59. R. Samarakoon, J. M. Overstreet, S. P. Higgins, P. J. Higgins, TGF-beta 1 -> SMAD/p53/USF2 -> PAI-1 transcriptional axis in ureteral obstruction-induced renal fibrosis. *Cell and tissue research* **347**, 117 (2012); published online EpubJan
60. Q. R. Xi, W. He, X. H. F. Zhang, H. V. Le, J. Massague, Genome-wide impact of the BRG1 SWI/SNF chromatin remodeler on the transforming growth factor beta transcriptional program. *Journal of Biological Chemistry* **283**, 1146 (2008); published online EpubJan 11
61. M. Fragiadaki, T. Ikeda, A. Witherden, R. M. Mason, D. Abraham, G. Bou-Gharios, High doses of TGF-beta potently suppress type I collagen via the transcription factor CUX1. *Molecular biology of the cell* **22**, 1836 (2011); published online EpubJun 1
62. G. Mengus, A. Fadloun, D. Kobi, C. Thibault, L. Perletti, I. Michel, I. Davidson, TAF4 inactivation in embryonic fibroblasts activates TGF beta signalling and autocrine growth. *Embo Journal* **24**, 2753 (2005); published online EpubAug 3
63. X. S. Gu, L. F. Zerbini, H. H. Otu, M. Bhasin, Q. L. Yang, M. G. Joseph, F. Grall, T. Onatunde, R. G. Correa, T. A. Libermann, Reduced PDEF expression increases invasion and expression of mesenchymal genes in prostate cancer cells. *Cancer research* **67**, 4219 (2007); published online EpubMay 1
64. B. Wang, L. E. Lindley, V. Fernandez-Vega, M. E. Rieger, A. H. Sims, K. J. Briegel, The T Box Transcription Factor TBX2 Promotes Epithelial-Mesenchymal Transition and Invasion of Normal and Malignant Breast Epithelial Cells. *Plos One* **7**, (2012); published online EpubJul 23
65. N. Ding, R. T. Yu, N. Subramaniam, M. H. Sherman, C. Wilson, R. Rao, M. Leblanc, S. Coulter, M. X. He, C. Scott, S. L. Lau, A. R. Atkins, G. D. Barish, J. E. Gunton, C. Liddle, M. Downes, R. M. Evans, A Vitamin D Receptor/SMAD Genomic Circuit Gates Hepatic Fibrotic Response. *Cell* **153**, 601 (2013); published online EpubApr 25 (DOI 10.1016/j.cell.2013.03.028).
66. S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, M. Mann, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate

- approach to expression proteomics. *Molecular & cellular proteomics : MCP* **1**, 376 (2002); published online EpubMay
67. N. A. Ali, M. P. Molloy, Quantitative phosphoproteomics of transforming growth factor-beta signaling in colon cancer cells. *Proteomics* **11**, 3390 (2011); published online EpubAug (10.1002/pmic.201100036).
68. T. Stasyk, A. Dubrovskaya, M. Lomnytska, I. Yakymovych, C. Wernstedt, C. H. Heldin, U. Hellman, S. Souhelnytskyi, Phosphoproteome profiling of transforming growth factor (TGF)-beta signaling: abrogation of TGFbeta1-dependent phosphorylation of transcription factor-II-I (TFII-I) enhances cooperation of TFII-I and Smad3 in transcription. *Molecular biology of the cell* **16**, 4765 (2005); published online EpubOct (10.1091/mbc.E05-03-0257).
69. P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, M. Sullivan, PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research* **40**, D261 (2012); published online EpubJan (Doi 10.1093/Nar/Gkr1122).
70. M. C. Wilkes, S. J. Murphy, N. Garamszegi, E. B. Leof, Cell-type-specific activation of PAK2 by transforming growth factor beta independent of Smad2 and Smad3. *Mol Cell Biol* **23**, 8878 (2003); published online EpubDec (Doi 10.1128/Mcb.23.23.8878-8889.2003).
71. S. Lamouille, E. Connolly, J. W. Smyth, R. J. Akhurst, R. Derynck, TGF-beta-induced activation of mTOR complex 2 drives epithelial-mesenchymal transition and cell invasion. *J Cell Sci* **125**, 1259 (2012); published online EpubMar 1 (Doi 10.1242/Jcs.095299).
72. J. A. Ubersax, J. E. Ferrell, Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Bio* **8**, 530 (2007); published online EpubJul
73. M. B. Datto, Y. Li, J. F. Panus, D. J. Howe, Y. Xiong, X. F. Wang, Transforming growth factor beta induces the cyclin-dependent kinase inhibitor p21 through a p53-independent mechanism. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 5545 (1995); published online EpubJun 6
74. A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Y. Lin, P. Minguez, P. Bork, C. von Mering, L. J. Jensen, STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**, D808 (2013); published online EpubJan (Doi 10.1093/Nar/Gks1094).
75. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498 (2003); published online EpubNov (Doi 10.1101/Gr.1239303).
76. G. Zhu, Y. Wang, B. Huang, J. Liang, Y. Ding, A. Xu, W. Wu, A Rac1/PAK1 cascade controls beta-catenin activation in colon cancer cells. *Oncogene* **31**, 1001 (2012); published online EpubFeb (Doi 10.1038/Onc.2011.294).
77. M. Mann, N. A. Kulak, N. Nagaraj, J. Cox, The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Molecular cell* **49**, 583 (2013); published online EpubFeb 21 (DOI 10.1016/j.molcel.2013.01.029).

78. Y. Oda, L. Z. Hu, V. Bul, H. Elalieh, J. K. Reddy, D. D. Bikle, Coactivator MED1 Ablation in Keratinocytes Results in Hair-Cycling Defects and Epidermal Alterations. *J Invest Dermatol* **132**, 1075 (2012); published online EpubApr
79. M. Xiong, J. B. Gong, Y. H. Liu, R. Xiang, X. Y. Tan, Loss of vitamin D receptor in chronic kidney disease: a potential mechanism linking inflammation to epithelial-to-mesenchymal transition. *Am J Physiol-Renal* **303**, F1107 (2012); published online EpubOct (DOI 10.1152/ajprenal.00151.2012).
80. S. Sarkar, D. W. Leaman, S. Gupta, P. Sil, D. Young, A. Morehead, D. Mukherjee, N. Ratliff, Y. P. Sun, M. Rayborn, J. Hollyfield, S. Sen, Cardiac overexpression of myotrophin triggers myocardial hypertrophy and heart failure in transgenic mice. *Journal of Biological Chemistry* **279**, 20422 (2004); published online EpubMay 7 (DOI 10.1074/jbc.M308488200).
81. G. M. Di Guglielmo, C. Le Roy, A. F. Goodfellow, J. L. Wrana, Distinct endocytic pathways regulate TGF-beta receptor signalling and turnover. *Nat Cell Biol* **5**, 410 (2003); published online EpubMay (Doi 10.1038/Ncb975).
82. F. Palacios, J. S. Tushir, Y. Fujita, C. D'Souza-Schorey, Lysosomal targeting of E-cadherin: a unique mechanism for the down-regulation of cell-cell adhesion during epithelial to mesenchymal transitions. *Mol Cell Biol* **25**, 389 (2005); published online EpubJan (Doi 10.1128/Mcb.25.1.389-402.2005).
83. M. C. Jones, P. T Caswell, J. C. Norman, Endocytic recycling pathways: emerging regulators of cell migration. *Curr Opin Cell Biol* **18**, 549 (2006); published online EpubOct (
84. X. H. Feng, X. Lin, R. Derynck, Smad2, Smad3 and Smad4 cooperate with Sp1 to induce p15(Ink4B) transcription in response to TGF-beta. *Embo Journal* **19**, 5178 (2000); published online EpubOct 2 (DOI 10.1093/emboj/19.19.5178).
85. A. Iavarone, J. Massague, Repression of the CDK activator Cdc25A and cell-cycle arrest by cytokine TGF-beta in cells lacking the CDK inhibitor p15. *Nature* **387**, 417 (1997); published online EpubMay 22
86. M. Lemaire, C. Froment, R. Boutros, O. Mondesert, A. R. Nebreda, B. Monsarrat, B. Ducommun, CDC25B phosphorylation by p38 and MK-2. *Cell Cycle* **5**, 1649 (2006); published online EpubAug 1 (Doi 10.4161/Cc.5.15.3006).
87. A. Errico, K. Deshmukh, Y. Tanaka, A. Pozniakovsky, T. Hunt, Identification of substrates for cyclin dependent kinases. *Adv Enzyme Regul* **50**, 375 (2010)DOI 10.1016/j.advenzreg.2009.12.001).
88. A. Iavarone, J. Massague, E2F and histone deacetylase mediate transforming growth factor beta repression of cdc25A during keratinocyte cell cycle arrest. *Mol Cell Biol* **19**, 916 (1999); published online EpubJan
89. K. Matsuzaki, M. Date, F. Furukawa, Y. Tahashi, M. Matsushita, K. Sakitani, N. Yamashiki, T. Seki, H. Saito, M. Nishizawa, J. Fujisawa, K. Inoue, Autocrine stimulatory mechanism by transforming growth factor beta in human hepatocellular carcinoma. *Cancer research* **60**, 1394 (2000); published online EpubMar 1
90. M. A. Bjornsti, P. J. Houghton, Lost in translation: Dysregulation of cap-dependent translation and cancer. *Cancer Cell* **5**, 519 (2004); published online EpubJun (DOI 10.1016/j.ccr.2004.05.027).

91. N. Sonenberg, A. Pause, Signal transduction: Protein synthesis and oncogenesis meet again. *Science* **314**, 428 (2006); published online EpubOct 20 (DOI 10.1126/science.1134031).
92. R. Azar, A. Alard, C. Susini, C. Bousquet, S. Pyronnet, 4E-BP1 is a target of Smad4 essential for TGF beta-mediated inhibition of cell proliferation. *Embo Journal* **28**, 3514 (2009); published online EpubNov 18 (DOI 10.1038/emboj.2009.291).
93. J. W. Poulsen, C. T. Madsen, C. Young, F. M. Poulsen, M. L. Nielsen, Using Guanidine-Hydrochloride for Fast and Efficient Protein Digestion and Single-step Affinity-purification Mass Spectrometry. *Journal of proteome research* **12**, 1020 (2013); published online EpubFeb (Doi 10.1021/Pr300883y).
94. N. Nagaraj, R. C. J. D'Souza, J. Cox, J. V. Olsen, M. Mann, Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. *Journal of proteome research* **9**, 6786 (2010); published online EpubDec (Doi 10.1021/Pr100637q).
95. J. R. Wisniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis. *Nature methods* **6**, 359 (2009); published online EpubMay (10.1038/nmeth.1322).
96. S. A. Beausoleil, M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villen, J. Li, M. A. Cohn, L. C. Cantley, S. P. Gygi, Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12130 (2004); published online EpubAug 17 (10.1073/pnas.0404720101).
97. J. Rappsilber, Y. Ishihama, M. Mann, Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical chemistry* **75**, 663 (2003); published online EpubFeb 1
98. J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794 (2011); published online EpubApr 1 (10.1021/pr101065j).
99. D. A. Hosack, G. Dennis, B. T. Sherman, H. C. Lane, R. A. Lempicki, Identifying biological themes within lists of genes with EASE. *Genome Biol* **4**, (2003).
100. L. Zhang, H. Z. Huang, F. F. Zhou, J. Schimmel, C. G. Pardo, T. T. Zhang, T. S. Barakat, K. A. Sheppard, C. Mikanin, J. A. Porter, A. C. O. Vertegaal, H. van Dam, J. Gribnau, C. X. Lu, P. ten Dijke, RNF12 Controls Embryonic Stem Cell Fate and Morphogenesis in Zebrafish Embryos by Targeting Smad7 for Degradation (vol 46, pg 650, 2012). *Molecular cell* **47**, 330 (2012); published online EpubJul 27
101. U. Persson, H. Izumi, S. Souchelnytskyi, S. Itoh, S. Grimsby, U. Engstrom, C. H. Heldin, K. Funo, P. ten Dijke, The L45 loop in type I receptors for TGF-beta family members is a critical determinant in specifying Smad isoform activation. *FEBS letters* **434**, 83 (1998); published online EpubAug 28 (Doi 10.1016/S0014-5793(98)00954-5).

Acknowledgements:

We thank Jürgen Cox, Nadin Neuhauser and Thanatip Viturawong at the MPI Martinsried for helpful discussions. Maarten van Dinther, Long Zhang and Theo van Laar at the Department of Molecular Cell Biology, Leiden assisted with shRNA and imaging work. Michael Wierer, Markus Raeschle, Sean Humphrey, Herbert Schiller and Jeff Liu critically commented on the manuscript. This work was partially supported by European Union 7th Framework project PROSPECTS (Proteomics Specification in Time and Space, grant HEALTH-F4-2008-201645) at the MPI, Martinsried and the Centre for Biomedical Genetics grant at the LUMC, Leiden.

Author contributions

RD, KS, AM, CC, PTD and MM designed and planned the experiments. RD, KS and NN prepared and measures samples by mass spectrometry. RD and KS analyzed the data. RD and AM performed shRNA experiments. RD, KS, AM, PTD and MM interpreted the data and wrote the paper.

Competing interests

The authors declare no competing interests.

Data and materials availability

The mass spectrometry proteomics data are deposited at the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD000496 (for reviewing process the data is accessible with the username: review12691 and password: xsRpTFzK).

Fig 1

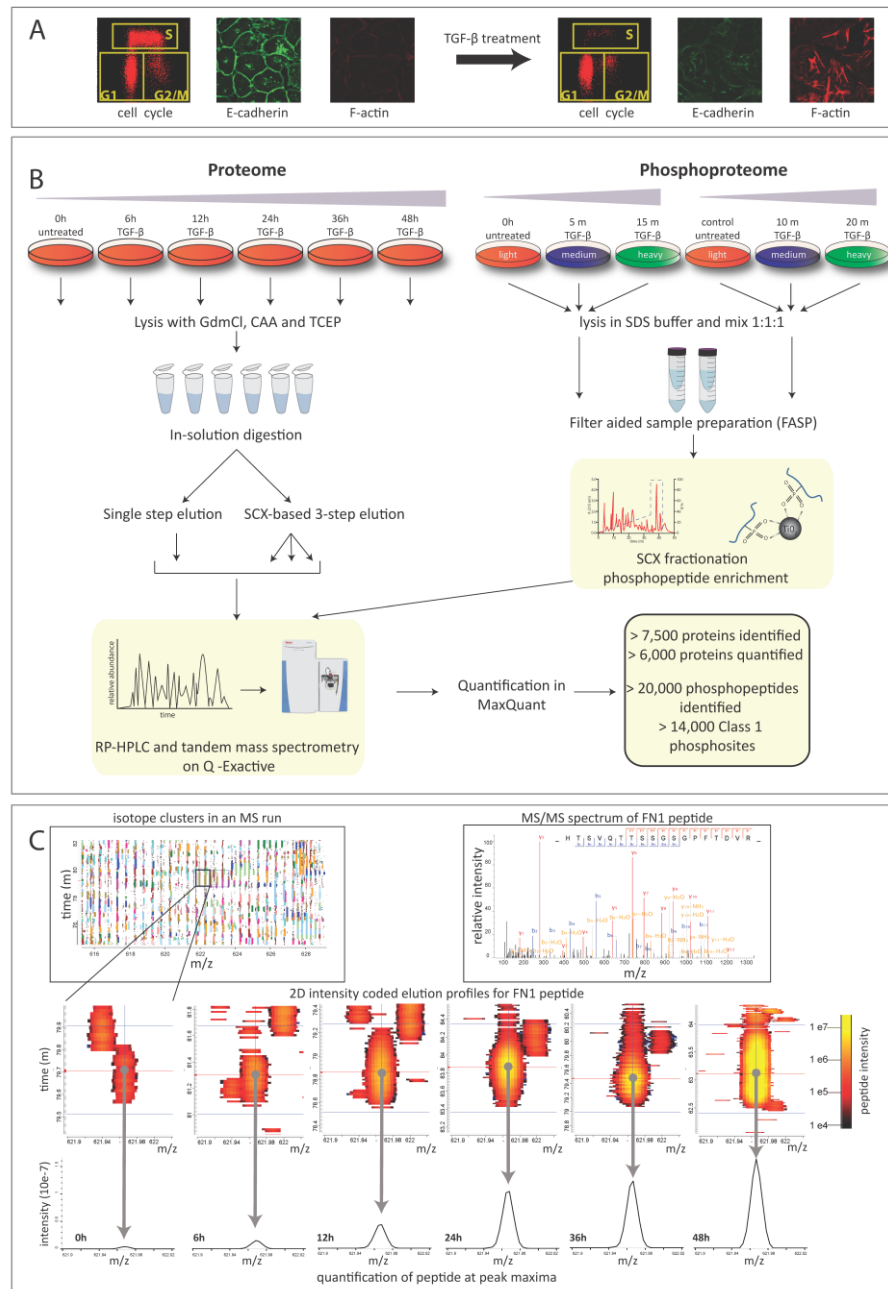


Fig.1: Quantitative proteome and phosphoproteome analysis of TGF- β signaling. (A) HaCaT cells showing induction of G1-cell cycle arrest, loss of E-cadherin and formation of stress fibers upon TGF- β treatment for 40h. (B) For proteome analysis, HaCaT cells are treated with TGF- β for six different time points, processed by in-solution digestion and analyzed either as single-runs or as three SCX fractions. For phosphoproteome analysis, SILAC labeled HaCaT cells are treated with TGF- β using a double-triple SILAC approach with untreated control as common reference point. Equal amount of cell lysates are pooled, digested by the FASP method and phosphopeptides are enriched using SCX chromatography and TiO₂ beads. Both peptides and phosphopeptides are analyzed by LC-MS/MS and raw data is processed in MaxQuant. (B) Identification and quantification of a fibronectin (FN1) peptide: Multiple isotope clusters eluting over time in an LC-MS run with that of FN1 peptide marked (left upper panel) and MS/MS spectrum of the identified peptide (upper right panel). Intensity coded 2D peak of this peptide across different time points of TGF- β treatment in the mass-retention time plane (middle panel) and MS signal intensity corresponding to 3D peak maxima used for quantification (lower panel).

Fig 2

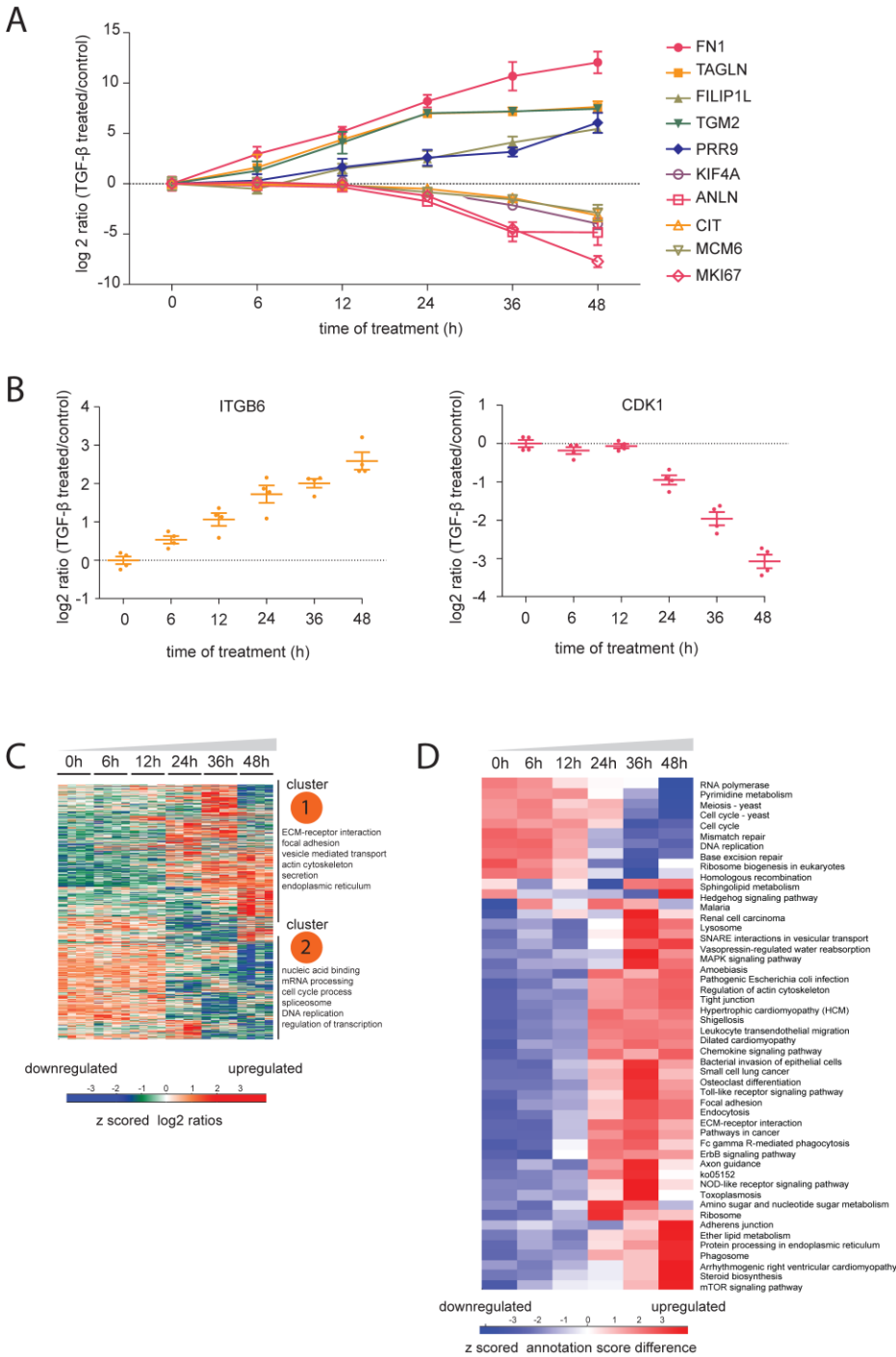


Fig. 2: The TGF-β responsive proteome. (A) Quantitative profiles (mean ±SEM of quadruplicates) of the top five most up-regulated and down-regulated proteins. (B) Quantification of individual replicates is shown as points with mean ±SEM for the known TGF-β targets CDK1 and integrin β6. (C) Heatmap of the significantly regulated proteins by unsupervised hierarchical clustering and enriched annotation categories. (D) Annotation matrix of KEGG pathways enriched at different time points shown as a heatmap after clustering.

Fig 3

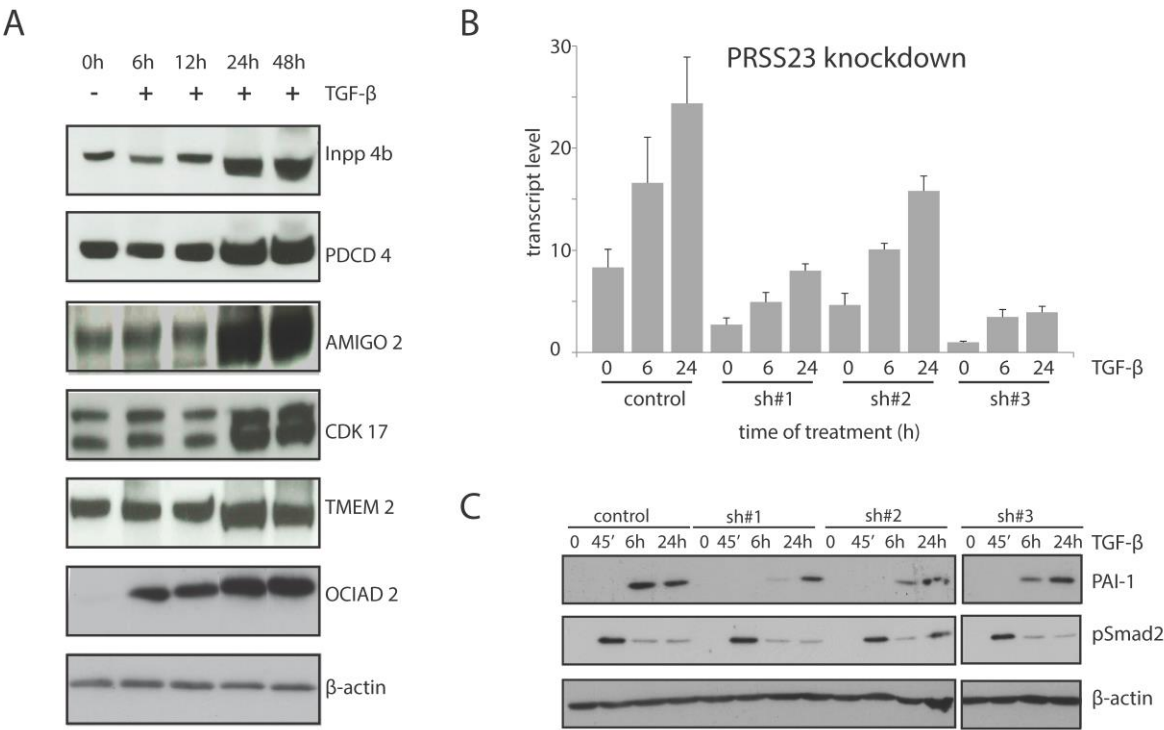


Fig. 3: Independent assessment of novel proteins induced by TGF- β . (A) Immunoblot analysis of INPP4b, PDCD4, AMIGO2, CDK17, TMEM2, OCIAD2 and β -actin after TGF- β treatment. (B) qRT-PCR based transcript level analysis of PRSS3 induction by TGF- β at the indicated time points and its knockdown using 3 different shRNAs (sh#1, sh#2 and sh#3). The columns and error bars represent mean \pm SD of triplicates. (C) Effect of PRSS23 knockdown on PAI-1 expression and SMAD2 phosphorylation.

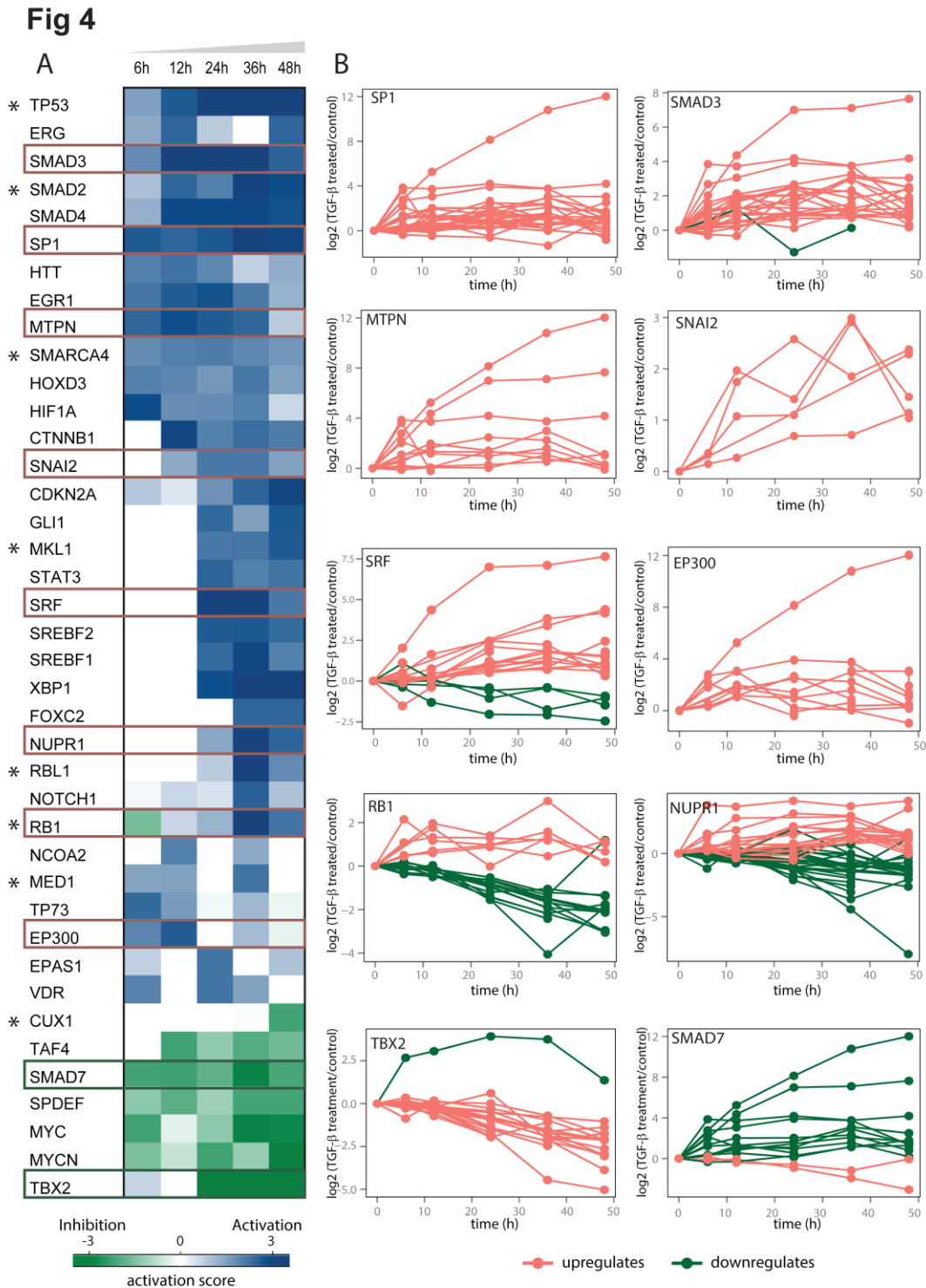


Fig. 4: Upstream transcription regulator analysis of the TGF-β-regulated proteome (A) A heatmap of transcription factors predicted to be activated (blue) or inhibited (green) by the TGF-β treatment in at least one of the 5 time points. Transcription factors with regulated phosphosites are marked with asterisk. (B) Expression profiles for target proteins of a few transcription factors (marked with boxes in A) used to calculate the activation score.

Fig 5

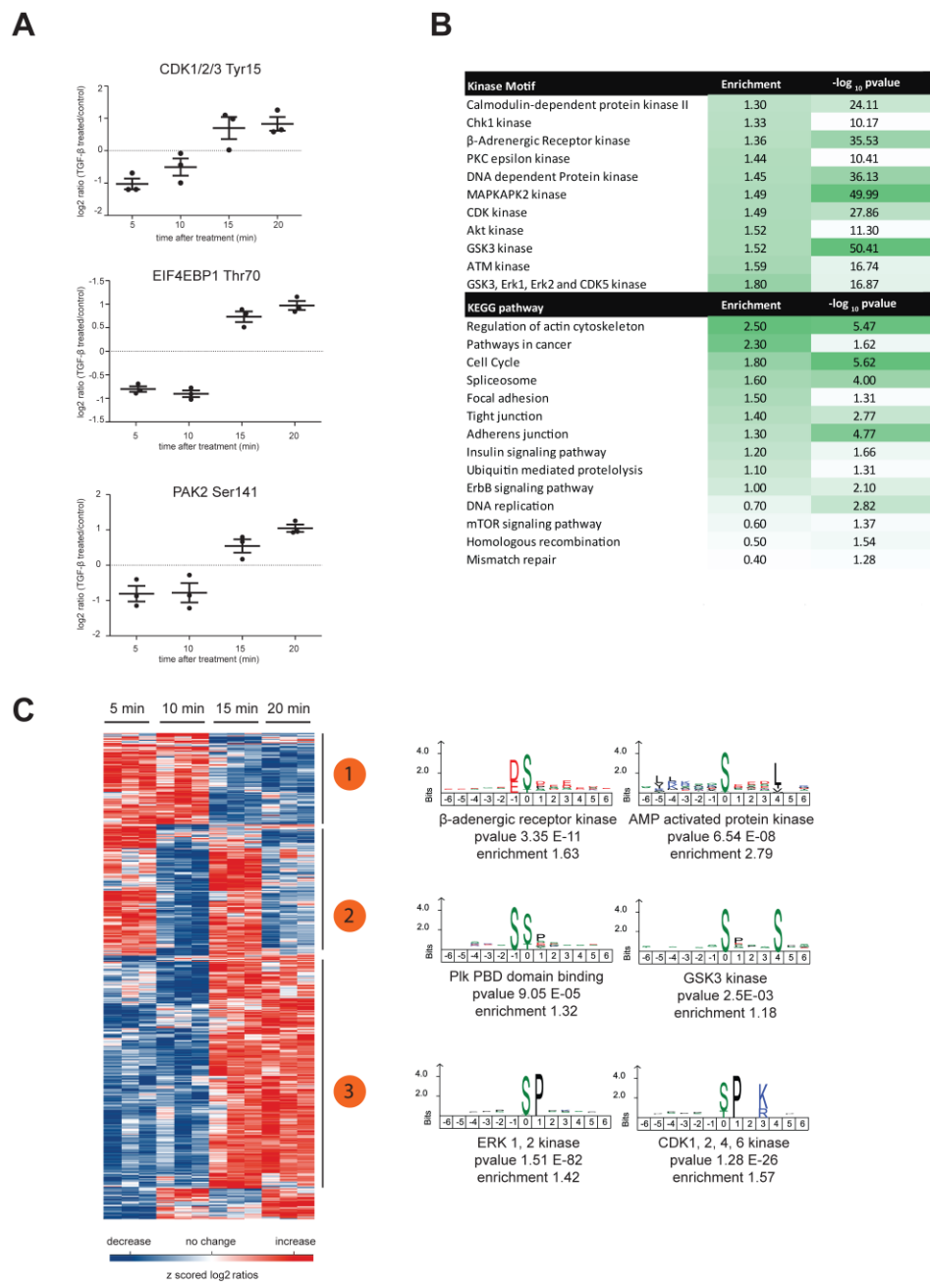


Fig. 5: The TGF-β responsive phosphoproteome (A) Quantification of individual replicates is shown for Tyr15 on CDK1/2/3, Thr70 on EIF4EBP1 and Ser141 on PAK2 after TGF-β treatment. (B) Protein kinase substrate motifs and KEGG pathways found to be enriched in the regulated dataset with corresponding p-values and enrichment factors (color-coded by values). (C) Hierarchical clustering of regulated phosphorylation sites in three distinct clusters marked as clusters 1, 2 and 3 and kinase substrate motifs enriched in individual clusters with their p-values (motifs generated using iomics.ugent.be/icelogoserver/main.html).

Fig 6

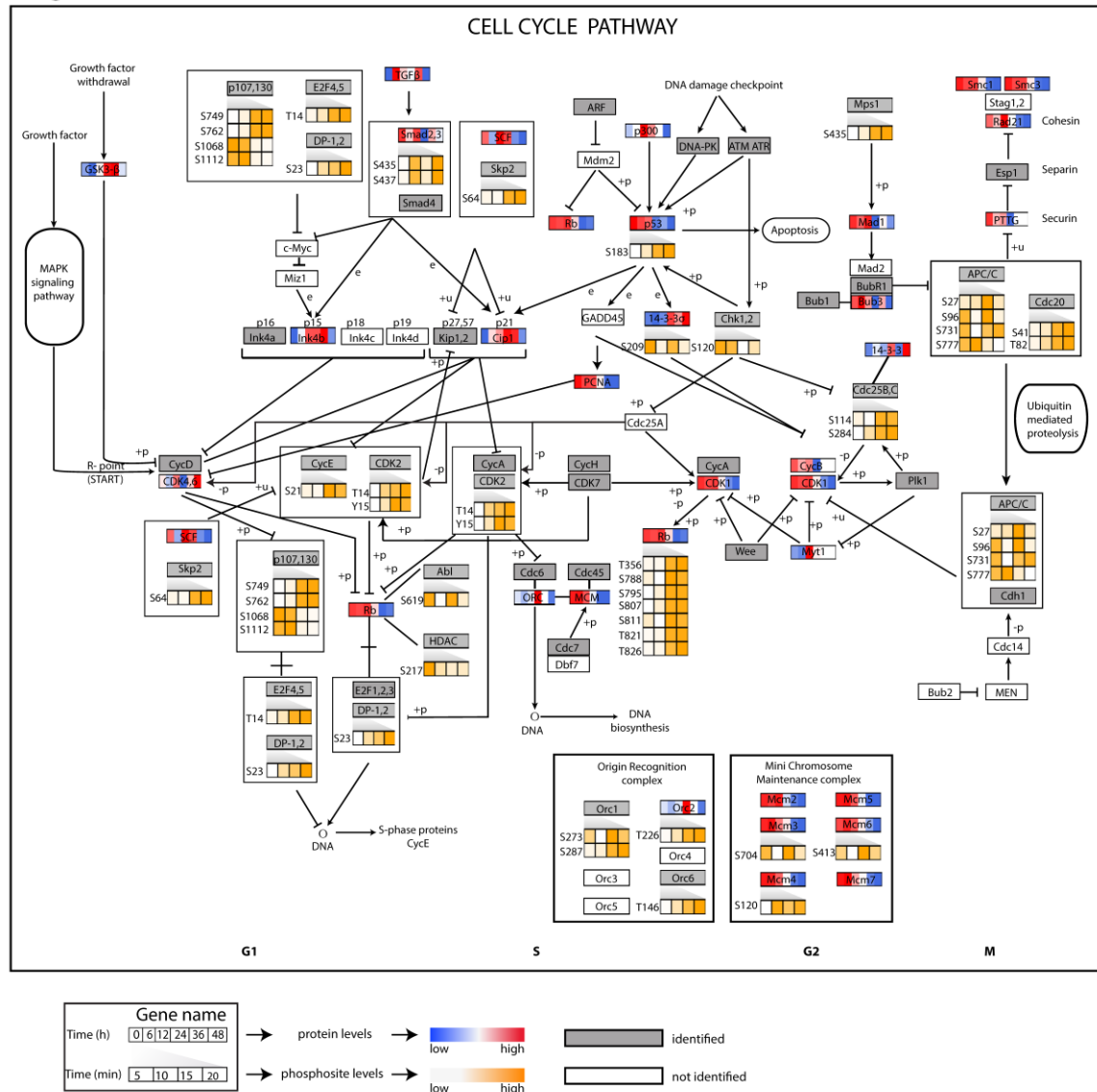


Fig. 6: Coverage of the cell cycle pathway in our proteomic and phosphoproteomic analysis in a KEGG based schematic. Temporal profiles of proteins (top) and phosphosites (bottom) are depicted as color coded bars divided into time points (color-code for protein regulation is from red to blue and for individual sites from orange to white; see bar). Proteins identified but not significantly regulated are colored grey and any protein not identified by MS is white.

Fig7

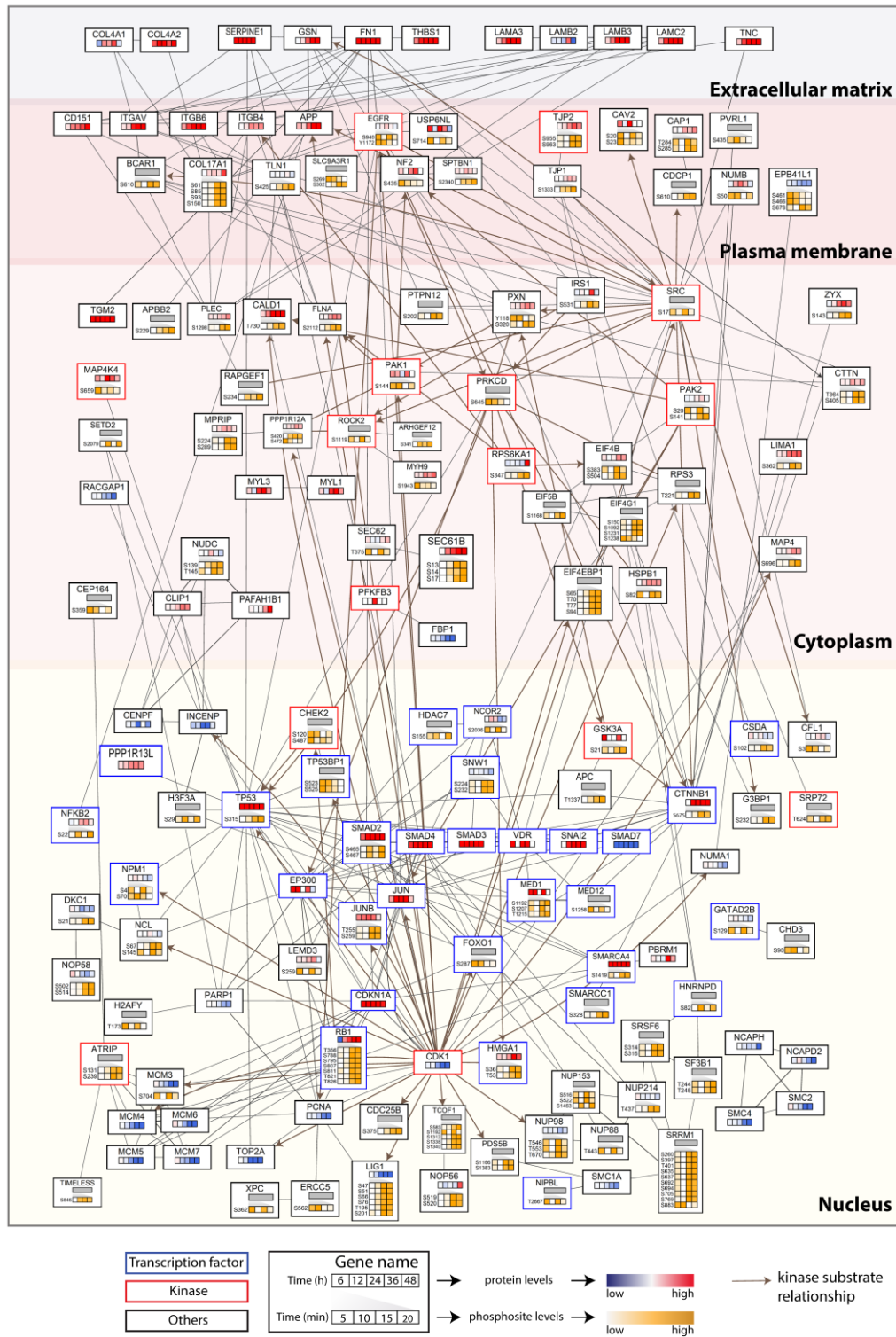


Fig. 7: Spatio-temporal view of TGF- β induced cellular effects depicted as a high confidence protein-protein interaction network, whose members are significantly regulated at expression and/or phosphorylation levels. Protein kinase-substrate relationships are shown by brown arrows and regulated protein expression and/or their early phosphorylation are depicted by keys explained in the legend to Fig. 6. Transcription factors are shown with blue margins and kinases with red.

Fig: S1A

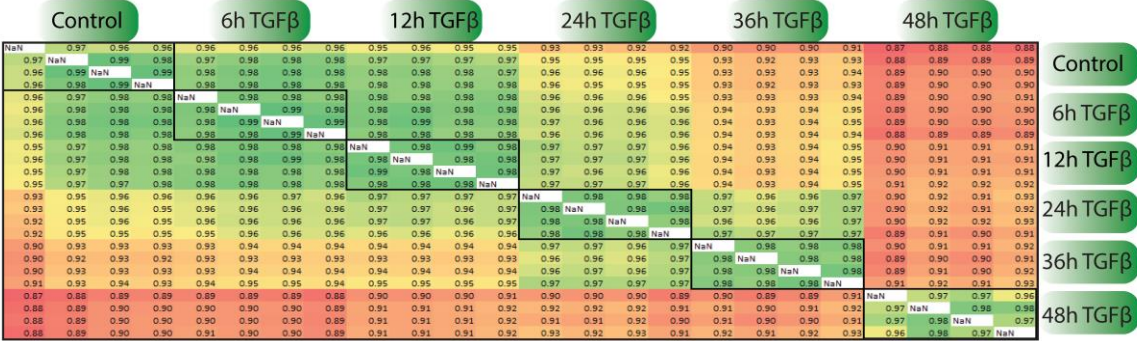


Fig. S1: A color-coded Pearson correlation matrix for protein quantification across biological quadruplicates measured for six different time points after TGF-β treatment.

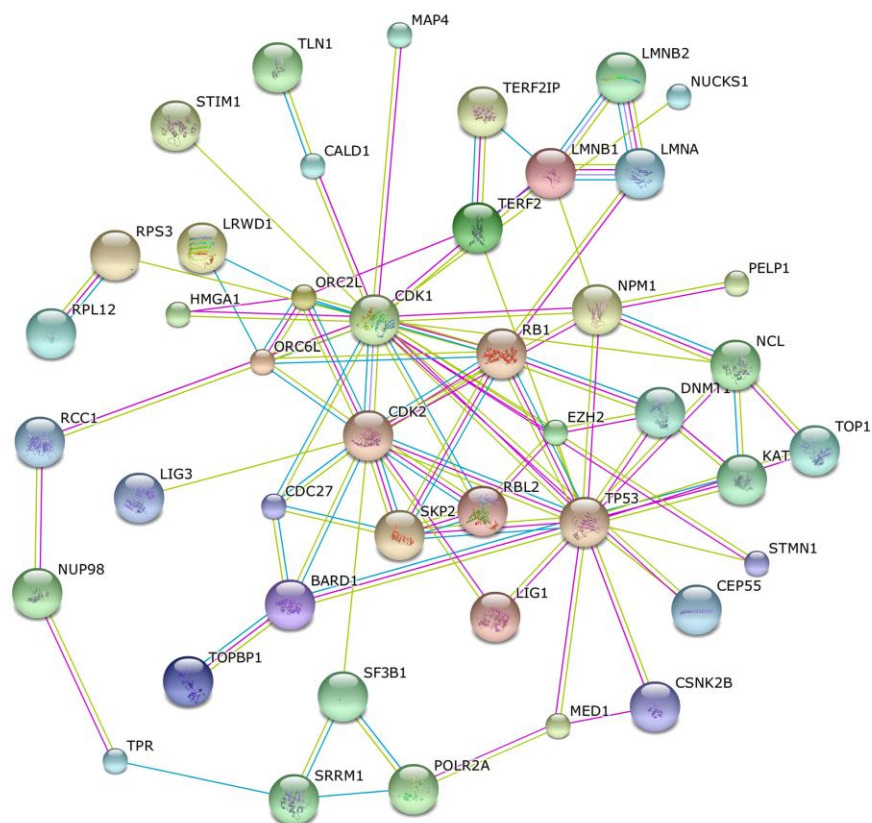
FIG S1 CDK1/2 substrates found regulated upon TGF- β treatment

Fig. S2: Interaction map (generated using STRING) showing high confidence interaction of CDK1/2 with various substrates that were regulated at CDK1 specific phosphorylation sites at early time points of TGF- β treatment.

Conclusions and perspectives

Work in this thesis, developed the analysis of large scale phospho-proteomics by HCD fragmentation. With this technology in hand, we investigated a human cancer cell line, acquiring the largest phospho-proteome to date. I then applied our technology in combination with a sophisticated experimental design to analyze the early phosphoproteome and downstream protein changes to uncover a network driving downstream signal response to TGF- β . This was made possible due to advancement in technology, sample preparation and data analysis algorithms leading to near complete coverage of the proteome in a few hours of measurement time. Very soon, it will also be possible to obtain a deep coverage of PTMs in similar single runs. This would be a very desirable next goal for the PTM field. Such rapid analysis of elaborate time courses can further enable detailed study of spatio-temporal signaling responses to intra and extracellular cues. With such rapid measurements, analysis of many samples in a short span of time is becoming a reality thus narrowing the gap of throughput between proteomics and genomics technologies.

Another major goal in the future will be to move from the exhaustive list of quantified phosphorylation sites to characterizing biologically relevant phosphorylation events.

In combination, bioinformatics and computational biology when applied to integrate several layers of omics technologies can very well provide useful information on kinase-regulated phosphorylation events, crosstalk between pathways and the spatiotemporal regulation of PTMs to sketch the complex circuitry of signaling at a systems level.

The recent advance in MS throughput also opens up possibilities to apply MS-based proteomics to the clinic. Any clinical study typically involves measurement of a few hundred samples in a given patient cohort to account for clinical heterogeneity. This possibility is even more realistic after the introduction of bench top mass spectrometers, such as the Q Exactive, which have already helped to spread proteomics to many laboratories. Recent reports demonstrate that the proteome and its PTMs are intact in frozen samples and FFPE samples, opening up for investigation of large cohorts.

The field of MS-based proteomics still faces challenges especially at the level of data acquisition and analysis. The developments detailed in this thesis have addressed a number of these

challenges already. Improvements in the instrumental capabilities including in sensitivity, sequencing speed, and cost effectiveness combined with further development in analysis tools and data modeling will very soon enable this technology to become even more powerful and to be truly on par with genomics technologies.

References

1. Nagaraj, N., et al., *Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation*. Journal of Proteome Research, 2010. **9**(12): p. 6786-6794.
2. Deeb, S.J., et al., *Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles*. Molecular & Cellular Proteomics, 2012. **11**(5): p. 77-89.
3. Venter, J.C., et al., *The sequence of the human genome*. Science, 2001. **291**(5507): p. 1304-+.
4. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
5. Chen, R., et al., *Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes*. Cell, 2012. **148**(6): p. 1293-1307.
6. Altelaar, A.F.M., J. Munoz, and A.J.R. Heck, *Next-generation proteomics: towards an integrative view of proteome dynamics*. Nature Reviews Genetics, 2013. **14**(1): p. 35-48.
7. Abreu, R.D., et al., *Global signatures of protein and mRNA expression levels*. Molecular Biosystems, 2009. **5**(12): p. 1512-1526.
8. Maier, T., M. Guell, and L. Serrano, *Correlation of mRNA and protein in complex biological samples*. Febs Letters, 2009. **583**(24): p. 3966-3973.
9. Vogel, C. and E.M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses*. Nature Reviews Genetics, 2012. **13**(4): p. 227-232.
10. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics*. Nature, 2003. **422**(6928): p. 198-207.
11. Domon, B. and R. Aebersold, *Review - Mass spectrometry and protein analysis*. Science, 2006. **312**(5771): p. 212-217.
12. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line*. Mol Syst Biol, 2011. **7**: p. 548.
13. Cox, J. and M. Mann, *Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology*. Annual Review of Biochemistry, Vol 80, 2011. **80**: p. 273-299.
14. Bensimon, A., A.J.R. Heck, and R. Aebersold, *Mass Spectrometry-Based Proteomics and Network Biology*. Annual Review of Biochemistry, Vol 81, 2012. **81**: p. 379-405.
15. Gibson, T.J., *Cell regulation: determined to signal discrete cooperation*. Trends Biochem Sci, 2009. **34**(10): p. 471-82.
16. Yamashita, M. and J.B. Fenn, *Electrospray Ion-Source - Another Variation on the Free-Jet Theme*. Journal of Physical Chemistry, 1984. **88**(20): p. 4451-4459.
17. Fenn, J.B., et al., *Electrospray Ionization for Mass-Spectrometry of Large Biomolecules*. Science, 1989. **246**(4926): p. 64-71.
18. Karas, M. and F. Hillenkamp, *Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons*. Analytical Chemistry, 1988. **60**(20): p. 2299-2301.
19. Eberl, H.C., et al., *A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics*. Molecular Cell, 2013. **49**(2): p. 368-378.
20. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing*. Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.
21. Anderson, N.L. and N.G. Anderson, *The human plasma proteome - History, character, and diagnostic prospects*. Molecular & Cellular Proteomics, 2002. **1**(11): p. 845-867.
22. Michalski, A., et al., *Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer*. Molecular & Cellular Proteomics, 2011. **10**(9).
23. Wang, Y., et al., *Mass-selective ion accumulation and fragmentation in a linear octopole ion trap external to a Fourier transform ion cyclotron resonance mass spectrometer*. International Journal of Mass Spectrometry, 2000. **198**(1-2): p. 113-120.

24. Schwartz, J.C.S.J., CA), Senko, Michael W. (Sunnyvale, CA), *Two-dimensional quadrupole ion trap operated as a mass spectrometer*. US Patent Number: 6797950, 2004.
25. Colinge, J. and K.L. Bennett, *Introduction to computational proteomics*. Plos Computational Biology, 2007. **3**(7): p. 1151-1160.
26. Kingdon, K.H., *A method for the neutralization of electron space charge by positive ionization at very low gas pressures*. Physical Review, 1923. **21**(4): p. 408-418.
27. Knight, R.D., *Storage of Ions from Laser-Produced Plasmas*. Applied Physics Letters, 1981. **38**(4): p. 221-223.
28. Makarov, A., *Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis*. Analytical Chemistry, 2000. **72**(6): p. 1156-1162.
29. Scigelova, M. and A. Makarov, *Orbitrap mass analyzer - Overview and applications in proteomics*. Proteomics, 2006: p. 16-21.
30. Makarov, A., et al., *Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer*. Analytical Chemistry, 2006. **78**(7): p. 2113-2120.
31. Michalski, A., et al., *Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes*. Molecular & Cellular Proteomics, 2012. **11**(3).
32. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis*. Nat Methods, 2007. **4**(9): p. 709-12.
33. Michalski, A., et al., *A Systematic Investigation into the Nature of Tryptic HCD Spectra*. Journal of Proteome Research, 2012. **11**(11): p. 5479-5491.
34. Steen, H., et al., *Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway*. Journal of Biological Chemistry, 2002. **277**(2): p. 1031-1039.
35. Hung, C.W., et al., *Collision-induced reporter fragmentations for identification of covalently modified peptides*. Analytical and Bioanalytical Chemistry, 2007. **389**(4): p. 1003-1016.
36. Olsen, J.V., et al., *Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap*. Molecular & Cellular Proteomics, 2005. **4**(12): p. 2010-2021.
37. Olsen, J.V., et al., *A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed*. Mol Cell Proteomics, 2009. **8**(12): p. 2759-69.
38. Geiger, T., J. Cox, and M. Mann, *Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation*. Molecular & Cellular Proteomics, 2010. **9**(10): p. 2252-2261.
39. Nagaraj, N., et al., *System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap*. Molecular & Cellular Proteomics, 2012. **11**(3).
40. Bogdanov, B. and R.D. Smith, *Proteomics by FTICR mass spectrometry: Top down and bottom up*. Mass Spectrometry Reviews, 2005. **24**(2): p. 168-200.
41. Chait, B.T., *Mass spectrometry: Bottom-up or top-down?* Science, 2006. **314**(5796): p. 65-66.
42. Yates, J.R., C.I. Ruse, and A. Nakorchevsky, *Proteomics by Mass Spectrometry: Approaches, Advances, and Applications*. Annual Review of Biomedical Engineering, 2009. **11**: p. 49-79.
43. Zhang, H. and Y. Ge, *Comprehensive Analysis of Protein Modifications by Top-Down Mass Spectrometry*. Circulation-Cardiovascular Genetics, 2011. **4**(6): p. 711-+.
44. Kelleher, N.L., et al., *Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry*. Journal of the American Chemical Society, 1999. **121**(4): p. 806-812.
45. Zabrouskov, V., et al., *Stepwise deamidation of ribonuclease A at five sites determined by top down mass spectrometry*. Biochemistry, 2006. **45**(3): p. 987-992.
46. Siuti, N. and N.L. Kelleher, *Decoding protein modifications using top-down mass spectrometry*. Nature Methods, 2007. **4**(10): p. 817-821.

47. Heck, A.J.R., *Native mass spectrometry: a bridge between interactomics and structural biology*. Nature Methods, 2008. **5**(11): p. 927-933.
48. Marsh, J.A., et al., *Protein Complexes Are under Evolutionary Selection to Assemble via Ordered Pathways*. Cell, 2013. **153**(2): p. 461-470.
49. Morgner, N. and C.V. Robinson, *Linking structural change with functional regulation-insights from mass spectrometry*. Curr Opin Struct Biol, 2012. **22**(1): p. 44-51.
50. Dancik, V., et al., *De novo peptide sequencing via tandem mass spectrometry*. Journal of Computational Biology, 1999. **6**(3-4): p. 327-342.
51. Taylor, J.A. and R.S. Johnson, *Sequence database searches via de novo peptide sequencing by tandem mass spectrometry*. Rapid Communications in Mass Spectrometry, 1997. **11**(9): p. 1067-1075.
52. Shevchenko, A., et al., *Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels*. Analytical Chemistry, 1996. **68**(5): p. 850-858.
53. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis*. Nat Methods, 2009. **6**(5): p. 359-62.
54. Olsen, J.V., S.E. Ong, and M. Mann, *Trypsin cleaves exclusively C-terminal to arginine and lysine residues*. Molecular & Cellular Proteomics, 2004. **3**(6): p. 608-614.
55. Wisniewski, J.R. and M. Mann, *Consecutive Proteolytic Digestion in an Enzyme Reactor Increases Depth of Proteomic and Phosphoproteomic Analysis*. Analytical Chemistry, 2012. **84**(6): p. 2631-2637.
56. Blagoev, B., et al., *Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics*. Nature Biotechnology, 2004. **22**(9): p. 1139-1145.
57. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics*. Anal Chem, 2003. **75**(3): p. 663-70.
58. Ishihama, Y., et al., *Microcolumns with self-assembled particle frits for proteomics*. Journal of Chromatography A, 2002. **979**(1-2): p. 233-239.
59. Marco Y. Hein, K.S., Jürgen Cox, Matthias Mann, *Proteomic Analysis of Cellular Systems*. Handbook of Systems Biology, 2013: p. 3-25.
60. Thakur, S.S., et al., *Deep and Highly Sensitive Proteome Coverage by LC-MS/MS Without Prefractionation*. Molecular & Cellular Proteomics, 2011. **10**(8).
61. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data*. Electrophoresis, 1999. **20**(18): p. 3551-3567.
62. Cox, J., et al., *Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment*. Journal of Proteome Research, 2011. **10**(4): p. 1794-1805.
63. Picotti, P. and R. Aebersold, *Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions*. Nature Methods, 2012. **9**(6): p. 555-566.
64. Zhao, Y.M. and O.N. Jensen, *Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques*. Proteomics, 2009. **9**(20): p. 4632-4641.
65. Xu, G.Q., J.S. Paige, and S.R. Jaffrey, *Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling*. Nature Biotechnology, 2010. **28**(8): p. 868-U154.
66. Wagner, S.A., et al., *A Proteome-wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles*. Molecular & Cellular Proteomics, 2011. **10**(10).
67. Kim, W., et al., *Systematic and quantitative assessment of the ubiquitin-modified proteome*. Mol Cell, 2011. **44**(2): p. 325-40.
68. Choudhary, C., et al., *Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions*. Science, 2009. **325**(5942): p. 834-840.

69. Bremang, M., et al., *Mass spectrometry-based identification and characterisation of lysine and arginine methylation in the human proteome*. Molecular Biosystems, 2013. **9**(9): p. 2231-2247.
70. Kettenbach, A.N. and S.A. Gerber, *Rapid and Reproducible Single-Stage Phosphopeptide Enrichment of Complex Peptide Mixtures: Application to General and Phosphotyrosine-Specific Phosphoproteomics Experiments*. Analytical Chemistry, 2011. **83**(20): p. 7635-7644.
71. Zielinska, D.F., et al., *Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints*. Cell, 2010. **141**(5): p. 897-907.
72. Hirabayashi, J. and K. Kasai, *Separation technologies for glycomics*. J Chromatogr B Analyt Technol Biomed Life Sci, 2002. **771**(1-2): p. 67-87.
73. Olsen, J.V., et al., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks*. Cell, 2006. **127**(3): p. 635-48.
74. Ficarro, S.B., et al., *Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae*. Nat Biotechnol, 2002. **20**(3): p. 301-5.
75. Gruhler, A., et al., *Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway*. Molecular & Cellular Proteomics, 2005. **4**(3): p. 310-327.
76. Andersson, L. and J. Porath, *Isolation of Phosphoproteins by Immobilized Metal (Fe-3+) Affinity-Chromatography*. Analytical Biochemistry, 1986. **154**(1): p. 250-254.
77. Villen, J., et al., *Large-scale phosphorylation analysis of mouse liver*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(5): p. 1488-1493.
78. Pinkse, M.W.H., et al., *Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-nanoLC-ESI-MS/MS and titanium oxide precolumns*. Analytical Chemistry, 2004. **76**(14): p. 3935-3943.
79. Macek, B., M. Mann, and J.V. Olsen, *Global and Site-Specific Quantitative Phosphoproteomics: Principles and Applications*. Annual Review of Pharmacology and Toxicology, 2009. **49**: p. 199-221.
80. Larsen, M.R., et al., *Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns*. Mol Cell Proteomics, 2005. **4**(7): p. 873-86.
81. Zhou, H.J., et al., *Enhancing the Identification of Phosphopeptides from Putative Basophilic Kinase Substrates Using Ti (IV) Based IMAC Enrichment*. Molecular & Cellular Proteomics, 2011. **10**(10).
82. Boersema, P.J., S. Mohammed, and A.J. Heck, *Phosphopeptide fragmentation and analysis by mass spectrometry*. J Mass Spectrom, 2009. **44**(6): p. 861-78.
83. Palumbo, A.M., J.J. Tepe, and G.E. Reid, *Mechanistic insights into the multistage gas-phase fragmentation behavior of phosphoserine- and phosphothreonine-containing peptides*. Journal of Proteome Research, 2008. **7**(2): p. 771-779.
84. Villen, J., S.A. Beausoleil, and S.P. Gygi, *Evaluation of the utility of neutral-loss-dependent MS3 strategies in large-scale phosphorylation analysis*. Proteomics, 2008. **8**(21): p. 4444-4452.
85. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review*. Analytical and Bioanalytical Chemistry, 2007. **389**(4): p. 1017-1031.
86. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Molecular & Cellular Proteomics, 2002. **1**(5): p. 376-386.
87. Hebert, A.S., et al., *Neutron-encoded mass signatures for multiplexed proteome quantification*. Nat Methods, 2013. **10**(4): p. 332-4.
88. Kruger, M., et al., *SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function*. Cell, 2008. **134**(2): p. 353-364.
89. Soufi, B., et al., *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) Applied to Quantitative Proteomics of Bacillus subtilis*. Journal of Proteome Research, 2010. **9**(7): p. 3638-3646.

90. Larance, M., et al., *Stable-isotope labeling with amino acids in nematodes*. Nature Methods, 2011. **8**(10): p. 849-U114.
91. Sury, M.D., J.X. Chen, and M. Selbach, *The SILAC Fly Allows for Accurate Protein Quantification in Vivo*. Molecular & Cellular Proteomics, 2010. **9**(10): p. 2173-2183.
92. Ong, S.E. and M. Mann, *A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)*. Nature Protocols, 2006. **1**(6): p. 2650-2660.
93. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue*. Nature Methods, 2010. **7**(5): p. 383-U64.
94. Schwanhaussner, B., et al., *Global analysis of cellular protein translation by pulsed SILAC*. Proteomics, 2009. **9**(1): p. 205-209.
95. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nature Biotechnology, 1999. **17**(10): p. 994-999.
96. Hsu, J.L., et al., *Stable-isotope dimethyl labeling for quantitative proteomics*. Analytical Chemistry, 2003. **75**(24): p. 6843-6852.
97. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Molecular & Cellular Proteomics, 2004. **3**(12): p. 1154-1169.
98. Thompson, A., et al., *Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS*. Analytical Chemistry, 2003. **75**(8): p. 1895-1904.
99. Hanke, S., et al., *Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level*. Journal of Proteome Research, 2008. **7**(3): p. 1118-1130.
100. Zeiler, M., et al., *A Protein Epitope Signature Tag (PREST) Library Allows SILAC-based Absolute Quantification and Multiplexed Determination of Protein Copy Numbers in Cell Lines*. Molecular & Cellular Proteomics, 2012. **11**(3).
101. Kirkpatrick, D.S., S.A. Gerber, and S.P. Gygi, *The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications*. Methods, 2005. **35**(3): p. 265-273.
102. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein*. Molecular & Cellular Proteomics, 2005. **4**(9): p. 1265-1272.
103. Schwanhaussner, B., et al., *Global quantification of mammalian gene expression control*. Nature, 2011. **473**(7347): p. 337-342.
104. Lubner, C.A., et al., *Quantitative Proteomics Reveals Subset-Specific Viral Recognition in Dendritic Cells*. Immunity, 2010. **32**(2): p. 279-289.
105. Hubner, N.C., et al., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions*. Journal of Cell Biology, 2010. **189**(4): p. 739-754.
106. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*. Nature Biotechnology, 2008. **26**(12): p. 1367-1372.
107. Zubarev, R. and M. Mann, *On the proper use of mass accuracy in proteomics*. Molecular & Cellular Proteomics, 2007. **6**(3): p. 377-381.
108. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. Nature Methods, 2007. **4**(3): p. 207-214.
109. Marx, H., et al., *A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics*. Nature Biotechnology, 2013. **31**(6): p. 557-+.
110. Liotta, L.A., M. Ferrari, and E. Petricoin, *Written in blood*. Nature, 2003. **425**(6961): p. 905-905.
111. Araujo, R.P., E.F. Petricoin, and L.A. Liotta, *Critical dependence of blood-borne biomarker concentrations on the half-lives of their carrier proteins*. Journal of Theoretical Biology, 2008. **253**(3): p. 616-622.

112. Ioannidis, J.P.A., *A roadmap for successful applications of clinical proteomics*. Proteomics Clinical Applications, 2011. **5**(5-6): p. 241-247.
113. Apweiler, R., et al., *Approaching clinical proteomics: current state and future fields of application in fluid proteomics*. Clinical Chemistry and Laboratory Medicine, 2009. **47**(6): p. 724-744.
114. Apweiler, R., et al., *Approaching Clinical Proteomics: Current State and Future Fields of Application in Cellular Proteomics*. Cytometry Part A, 2009. **75A**(10): p. 816-832.
115. Parker, C.E., et al., *Mass-spectrometry-based clinical proteomics--a review and prospective*. Analyst, 2010. **135**(8): p. 1830-8.
116. Boyd, R.S., M.J. Dyer, and K. Cain, *Proteomic analysis of B-cell malignancies*. J Proteomics, 2010. **73**(10): p. 1804-22.
117. Quackenbush, J., *Microarray analysis and tumor classification - Reply*. New England Journal of Medicine, 2006. **355**(9): p. 960-960.
118. McDermott, U., J.R. Downing, and M.R. Stratton, *GENOMIC MEDICINE Genomics and the Continuum of Cancer Care*. New England Journal of Medicine, 2011. **364**(4): p. 340-350.
119. Johnson, L.N. and D. Barford, *The effects of phosphorylation on the structure and function of proteins*. Annu Rev Biophys Biomol Struct, 1993. **22**: p. 199-232.
120. Hunter, T., *Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling*. Cell, 1995. **80**(2): p. 225-36.
121. Manning, G., et al., *The protein kinase complement of the human genome*. Science, 2002. **298**(5600): p. 1912-34.
122. Forrest, A.R., et al., *Phosphoregulators: protein kinases and protein phosphatases of mouse*. Genome Res, 2003. **13**(6B): p. 1443-54.
123. Cohen, P.T., *Protein phosphatase 1--targeted in many directions*. J Cell Sci, 2002. **115**(Pt 2): p. 241-56.
124. Matthews, H.R., *Protein kinases and phosphatases that act on histidine, lysine, or arginine residues in eukaryotic proteins: a possible regulator of the mitogen-activated protein kinase cascade*. Pharmacol Ther, 1995. **67**(3): p. 323-50.
125. Flores-Riveros, J.R., et al., *Substrate phosphorylation catalyzed by the insulin receptor tyrosine kinase. Kinetic correlation to autophosphorylation of specific sites in the beta subunit*. J Biol Chem, 1989. **264**(36): p. 21557-72.
126. Hunter, T., *The age of crosstalk: phosphorylation, ubiquitination, and beyond*. Mol Cell, 2007. **28**(5): p. 730-8.
127. Zolnierowicz, S. and M. Bollen, *Protein phosphorylation and protein phosphatases*. De Panne, Belgium, September 19-24, 1999. EMBO J, 2000. **19**(4): p. 483-8.
128. Olsen, J.V., et al., *Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis*. Sci Signal, 2010. **3**(104): p. ra3.
129. Ullrich, A. and J. Schlessinger, *Signal transduction by receptors with tyrosine kinase activity*. Cell, 1990. **61**(2): p. 203-12.
130. Pawson, T., *Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems*. Cell, 2004. **116**(2): p. 191-203.
131. Smith, M.J., et al., *Screening for PTB domain binding partners and ligand specificity using proteome-derived NPXY peptide arrays*. Mol Cell Biol, 2006. **26**(22): p. 8461-74.
132. Nurse, P., *A long twentieth century of the cell cycle and beyond*. Cell, 2000. **100**(1): p. 71-8.
133. http://www.cellsignal.com/reference/kinase_disease.html.
134. Nowell, P.C. and D.A. Hungerford, *Chromosome studies on normal and leukemic human leukocytes*. J Natl Cancer Inst, 1960. **25**: p. 85-109.
135. Konopka, J.B., S.M. Watanabe, and O.N. Witte, *An alteration of the human c-abl protein in K562 leukemia cells unmask associated tyrosine kinase activity*. Cell, 1984. **37**(3): p. 1035-42.

136. Carroll, M., et al., *CGP 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing BCR-ABL, TEL-ABL, and TEL-PDGFR fusion proteins*. *Blood*, 1997. **90**(12): p. 4947-52.
137. Rask-Andersen, M., M.S. Almen, and H.B. Schiøth, *Trends in the exploitation of novel drug targets*. *Nat Rev Drug Discov*, 2011. **10**(8): p. 579-90.
138. Garay, J.P. and J.W. Gray, *Omics and therapy - a basis for precision medicine*. *Mol Oncol*, 2012. **6**(2): p. 128-39.
139. Fabbro, D., et al., *Targeting cancer with small-molecular-weight kinase inhibitors*. *Methods Mol Biol*, 2012. **795**: p. 1-34.
140. Fedorov, O., S. Muller, and S. Knapp, *The (un)targeted cancer kinome*. *Nat Chem Biol*, 2010. **6**(3): p. 166-169.
141. Jordan, J.D., E.M. Landau, and R. Iyengar, *Signaling networks: the origins of cellular multitasking*. *Cell*, 2000. **103**(2): p. 193-200.
142. Pan, C., et al., *Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics*. *Mol Cell Proteomics*, 2009. **8**(12): p. 2796-808.
143. Beausoleil, S.A., et al., *Large-scale characterization of HeLa cell nuclear phosphoproteins*. *Proc Natl Acad Sci U S A*, 2004. **101**(33): p. 12130-5.
144. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse*. *Nucleic Acids Research*, 2012. **40**(D1): p. D261-D270.
145. Diella, F., et al., *Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins*. *BMC Bioinformatics*, 2004. **5**: p. 79.
146. Gnad, F., et al., *PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites*. *Genome Biology*, 2007. **8**(11).
147. Moses, H.L., et al., *Transforming growth factor production by chemically transformed cells*. *Cancer Res*, 1981. **41**(7): p. 2842-8.
148. Moustakas, A., et al., *Mechanisms of TGF-beta signaling in regulation of cell growth and differentiation*. *Immunol Lett*, 2002. **82**(1-2): p. 85-91.
149. Tsukazaki, T., et al., *SARA, a FYVE domain protein that recruits Smad2 to the TGFbeta receptor*. *Cell*, 1998. **95**(6): p. 779-91.
150. Feng, X.H. and R. Derynck, *Specificity and versatility in tgfbeta signaling through Smads*. *Annu Rev Cell Dev Biol*, 2005. **21**: p. 659-93.
151. Kang, Y., C.R. Chen, and J. Massague, *A self-enabling TGFbeta response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells*. *Mol Cell*, 2003. **11**(4): p. 915-26.
152. Hocevar, B.A., T.L. Brown, and P.H. Howe, *TGF-beta induces fibronectin synthesis through a c-Jun N-terminal kinase-dependent, Smad4-independent pathway*. *EMBO J*, 1999. **18**(5): p. 1345-56.
153. Mu, Y., S.K. Gudey, and M. Landstrom, *Non-Smad signaling pathways*. *Cell Tissue Res*, 2012. **347**(1): p. 11-20.
154. Massague, J., *TGF beta signalling in context*. *Nature Reviews Molecular Cell Biology*, 2012. **13**(10): p. 616-630.
155. Bierie, B. and H.L. Moses, *Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer*. *Nat Rev Cancer*, 2006. **6**(7): p. 506-20.
156. Akhurst, R.J. and R. Derynck, *TGF-beta signaling in cancer--a double-edged sword*. *Trends Cell Biol*, 2001. **11**(11): p. S44-51.
157. Heldin, C.H., M. Vanlandewijck, and A. Moustakas, *Regulation of EMT by TGFbeta in cancer*. *FEBS Lett*, 2012. **586**(14): p. 1959-70.
158. Tsuji, T., S. Ibaragi, and G.F. Hu, *Epithelial-mesenchymal transition and cell cooperativity in metastasis*. *Cancer Res*, 2009. **69**(18): p. 7135-9.

References

159. Scheel, C. and R.A. Weinberg, *Cancer stem cells and epithelial-mesenchymal transition: Concepts and molecular links*. Seminars in Cancer Biology, 2012. **22**(5-6): p. 396-403.
160. Akhurst, R.J. and A. Hata, *Targeting the TGFbeta signalling pathway in disease*. Nat Rev Drug Discov, 2012. **11**(10): p. 790-811.

Acknowledgements

Let us rise up and be thankful, for if we didn't learn a lot today, at least we learned a little, and if we didn't learn a little, at least we didn't get sick, and if we got sick, at least we didn't die; so, let us all be thankful.

-The Buddha

I firstly would like to thank Prof. Matthias Mann for giving me an opportunity to work in his laboratory. Thanks Matthias, for all your support and help during my tenure at the MPI. Thanks to your tutelage I have learnt many things in these 5 years which will be helpful in the long run. Thanks for the possibility to attend many conferences and meetings and for the great working atmosphere!

I am extremely thankful to Nagarjuna Nagaraj, Alexandre Zougman aka Sasha and Jacek Jacek Wisniewski, my first office-mates for their help especially during the first few months in the lab. Naga, Tami and Chuna for literally holding my hands and teaching me to walk and talk in the world of mass spectrometry. I have learnt a lot from you guys.

Thanks to Korbinian Mayr, Igor Paron and Peter Bandilla for all their help with fixing mass spectrometers and answering my questions about the machine.

Thanks to Johannes Graumann, for the late evening sessions of discussing a range of things right from science to how to fix my safety helmet for biking.

I am also grateful to Susi, Heike, Claudia, Stefan, Ivo, Alfred, Peter and Anna with whom I had the opportunity to collaborate over scientific projects and from whom I learnt a lot about many areas in science. Peter and Anna, you especially took good care of me during my stay in Leiden and have been of a lot of help over the years.

The cooking group, Tami, Sara, Mara, Naga, Helen, and Kirti for the good times shared over our many meals every afternoon. Mario and Ivan for the many movies and fun times shared together in my earlier years here. Thanks to Werner who was also part of the 'Orbi-5' maintenance team who taught me many things. Damla, Sally, Nina, Maxi, Marlis, Jüergen, Stefka, Richard, Chris and Anne for many good moments shared that I will cherish.

Thanks also goes to my second (and last) office, the 'girls office' aka 'Blümchen Büro' with whom I have shared moments of happiness and sadness, with whom I have talked about a lot of things on this planet. Kirti, Michal, Annette, Nadin and Scarlet, you are the best office mates to have.

Acknowledgements

Thanks to Tar, who introduced me to two things: ESME and R programing. Thanks to Sean, Michael and Jeff. Though I did not spend much time with you guys, you have been of great help and support in the last months.

Thank goes to Alison and Theresa for handling all administrative issues and being there whenever I needed help.

Kirti Sharma, a great friend and companion who cannot be thanked enough. You have inspired me and taught me many lessons for life that I will never forget. Ranga, Juhi, Anoop, Kirti, Siju, Venky, Naga, Archana, Arathi and Krishna with whom I have spent many evenings over dinner, weekends partying and much more. Your support and understanding made life away from home easier.

Lastly, I thank my parents and sister for all their love and understanding over these years. Your support and encouragement over these years has been instrumental. Mike, thanks for your love and understanding especially in the last few months of stress and deadlines.

Résumé

Rochelle. C. J. D'Souza

MSc (Biotechnology)

Department of Proteomics and Signal Transduction

Max Planck Institute of Biochemistry

Am Klopferspitz 18

82152 Martinsried, Germany

Ph: +49-89-8578-2098, +49-176-62514299

ACADEMIC EDUCATION

Nov 2008-Present	PhD with Prof Matthias Mann at the Max Planck Institute for Biochemistry, Germany
July 2003-May 2005	Master of Science (M.Sc.) in Biotechnology , University of Mysore, India First Class with Distinction (72%)
June 2000 - Apr 2003	Bachelor of Science (B.Sc.) in Microbiology, Zoology and Chemistry , University of Mangalore, India First Class with Distinction (86%)

TEACHING/ RESEARCH EXPERIENCE

Sep 2007-Oct 2008	Research associate at Physiology Institute, Justus Liebig University, Germany
Aug 2006- Aug 2007	Guest Faculty in Biotechnology at the Department of Biosciences, Mangalore University, India
Jan 2006 - July 2006	Industrial research at Suven Life Sciences, Hyderabad, India for 6 month training

RESEARCH EXPERIENCE

- **Present Position**

Project Studying deregulated cancer cell signaling using MS- based proteomics

Place Max Planck Institute of Biochemistry, Germany

Guide Prof. Dr. Matthias Mann

Duration Since November 2008

- **Research associate at Justus Liebig University**

Project Regulation of the expression and function of axon/angiogenesis guidance molecules in embryonic stem cells by mechanical strain

Place Physiology Institute, Justus Liebig University, Germany

Guide Prof. Dr. Heinrich Sauer

Duration September 2007-October 2008

- **M.Sc Project**

Project: Construction of TH-Puro^r/ SV40- Neo^r Vector for the Enrichment of Dopaminergic Neurons in Differentiating Mouse ES Cells.

Place: National Centre for Biological Sciences (NCBS), Bangalore, India

Guide: Prof. Dr. Mitradas M. Panicker

Duration: December 2004- March 2005

- **Summer Research Fellowship (B.Sc)**

Project: Cloning, Expression and Purification of Protein kinase H (PknH) in *Mycobacterium tuberculosis*.

Place: Institute for Genomics and Integrative Biology (IGIB), Delhi, India

Guide: Prof. Dr.Yogendra Singh

Duration: April 2002- May2002

AWARDS

- Awarded Council of Scientific and Industrial Research (CSIR) and Qualified National Eligibility Test for Lectureship conducted Govt. of India in December 2004
- Selected for the Biotech Consortium of India (BCIL) by Department of biotechnology (DBT), Ministry of science and technology, Govt. of India in 2005
- Selected for summer research fellowship by Jawaharlal Nehru centre for advanced scientific research (JNCASR), Bangalore, India in April 2002
- Awarded gold medal at the university level for best student in Microbiology by Mangalore University, India in 2004.

WORK RELATED RESPONSIBILITIES

- Served as organiser of soft-skill courses for the graduate programme, Max Planck institute of biochemistry from 2009-2012
- Vice president of student college council 2002-2003

PUBLICATIONS

- Deeb S, D'souza R.C., Cox J, Schmidt-Supprian M and Mann M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Molecular and cellular proteomics, 11: 77-89 (2012)
- Nagaraj N*, D'souza R.C.* Cox J, Olsen J.V.O. and Mann M. Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. Journal of proteome research, 9: 6786-6794 (2010)
- Sauer H, Ruhe C, Müller JP, Schmelter M, D'Souza R.C. and Wartenberg M. Reactive oxygen species and upregulation of NADPH oxidases in mechanotransduction of embryonic stem cells. Methods in molecular biology, 477: 397-418 (2008)
- Sharma K, Chandra H, Gupta P. K., Pathak M, Narayan A, Meena L. S, D'Souza R. C., Chopra P, Ramachandran S and Singh Y. PknH, a transmembrane Hank's type serine/threonine kinase from Mycobacterium tuberculosis is differentially expressed under stress conditions. FEMS Microbiol Lett, 233:107-13 (2004)

MANUSCRIPTS SUBMITTED/PREPARATION

- **D'souza R.C***, Sharma K*, Korhonen A, Nagaraj N, Choudhary C, ten Djike P and Mann M. Studying TGF- β signaling using a high resolution, quantitative mass spectrometric approach.
- **D'souza R.C.**, Deeb S, Badura S, Pfeiffer H, Ottmann O, Serve H, Cox J and Mann M. Deciphering the mechanisms of resistance in acute lymphoblastic leukemia using phospho-proteomic analysis
- Sharma K, **Dsouza R.C.**, Tyanova S., Wiśniewski, J. R., Cox J and Mann M. A very deep and high-accuracy phosphoproteome reveals fundamental differences between tyrosine and serine/threonine phosphorylation events

ORAL/POSTER PRESENTATIONS

- Oral presentations
 - Keystone symposium- Omics meets cell biology held at Alpbach, Austria in May 2011
 - GRK 534 "Biological Basis of Vascular Medicine" Annual Retreat at Rauischolzhausen, Germany
 - Poster presentations
 - 13th Hunter meeting held in the Hunter Valley, Australia in March 2013
 - 22nd Biennial Congress of the European Association for Cancer Research held in Barcelona, Spain in August 2012
 - 60th ASMS Conference on Mass Spectrometry and Allied Topics held in Vancouver, Canada in May 2012
 - HUPO 2011 10th world congress held in Geneva, Switzerland in September 2011
 - Gordon Research Conference on Phosphorylation & G-Protein Mediated Signaling Networks held in Biddeford, USA in 2010
-