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

NOVEL TECHNOLOGIES ENABLING STREAMLINED COMPLETE PROTEOME ANALYSIS

NILS ALEXANDER KULAK

AUS

FRANKFURT AM MAIN, DEUTSCHLAND

2014

ERKLÄRUNG

DIESE DISSERTATION WURDE IM SINNE VON § 7 DER PROMOTIONSORDNUNG VOM 28. NOVEMBER 2011 VON HERRN PROF. MATTHIAS MANN BETREUT.

EIDESSTATTLICHE VERSICHERUNG

DIESE DISSERTATION WURDE EIGENSTÄNDIG UND OHNE UNERLAUBTE HILFE ERARBEITET.

MÜNCHEN, 21.08.2014

....
(UNTERSCHRIFT DES AUTORS / DER AUTORIN)

DISSERTATION EINGEREICHT AM 11.02.2014

1. GUTACHTER: PROF. MATTHIAS MANN

2. GUTACHTER: PROF. THOMAS NEUBERT

MÜNDLICHE PRÜFUNG AM 13.03.2014

SUMMARY

Mass spectrometry (MS)-based proteomics is a powerful technology platform capable to identify and quantify the protein content of highly complex samples. Tremendous improvements of instrumentation over the last decades enabled an extensive coverage of eukaryotic proteomes. However, the capability to analyze near complete cellular proteomes have so far been restricted to specialized mass spectrometry laboratories with dedicated instruments and expertise in sample handling. This is highlighted by the remarkable efforts which were put forth to acquire the first MS measurements of a complete yeast proteome and the very deep coverage of human cell lines. However, technological developments described and discussed in this work may help the technology to be generally applicable and more accessible.

The aim of this thesis was the development of a streamlined and robust LC-MS platform and to benchmark the performance on the yeast model system. We therefore coupled a high-resolution nano-UHPLC setup to a novel bench top quadrupole Orbitrap mass spectrometer. The optimized platform allowed very high identification rates which promised deep proteomic coverage even with single-shot measurements. Testing the platform on Lys-C prepared yeast samples resulted in the identification of nearly 4,000 protein groups in single-shot 4 h measurements which covers nearly the complete proteome of expected 4,500 proteins. The combined depth of six replicates was even higher covering 87% of all verified proteins of the *S.cerevisiae* database and a very high reproducibility of identifications with 92% of the identified protein groups represented in every replicate. We then demonstrated the use of the new platform for system-wide perturbation analysis and performed a heat-shock experiment using the spike-in super-SILAC approach for accurate quantification. The measurements achieved similar proteomic coverage despite the higher sample complexity of SILAC. As expected, significantly up-regulated protein groups belong to the group of heat-shock proteins, down-regulations were dominated by the transcription and translation related processes.

Because of the high performance of the discussed LC-MS platforms, we set forth to also improve the sample preparation for complete proteomics. We therefore challenged every processing step and simplified the overall sample preparation procedure. The entire sample handling workflow from cell material to LC-MS ready peptides was reduced to three processing steps that can be performed in a single enclosed reaction device. Additionally a procedure for SCX and SDB-RPS based fractionation technologies was developed yielding superior performance over classical SAX-based SPE fractionation technologies. Because of the high quantitative reproducibility we further performed protein copy-number estimations and provide very deep proteomic measurements of the model systems *S.cerevisiae*, *S.pombe*, and HeLa to the community (4,570, 4,134, and 9,678 estimated proteins per proteome, respectively).

The developments of the previous projects inspired us to map genome wide protein expression in the model organism *S.cerevisiae*. The project included the deep proteomic measurements of well described nutrient growth conditions, environmental stress conditions, mating types, cell-cycle stages, and native wild-type yeast strains. We furthermore obtained a phosphorylation map of cells grown under normal and heat-shock condition. In the combined analysis of all conditions we identified 5,015 protein groups with 99% certainty and a median sequence coverage of 50% which represents the deepest coverage on protein and peptide level to date. We furthermore acquired the most comprehensive phosphopeptide dataset and observed a very high level of protein phosphorylation covering more than half of all identified protein groups. Acquired mass spectra of more than 130,000 peptides represent a large resource for directed measurements and targeted data analysis.

Quantitative reproducibility and the substantial differences across the conditions still resulted in a large number of stable expressed proteins. We found that 499 proteins did not exceed two-fold regulation across all conditions arguing for a household function. The proteins span the entire abundance range and are involved in various physiological functions such as cellular transport. We furthermore identified two entirely new open reading frames with high statistical accuracy and observed a truncated isoform of a protein believed to be non-existent. The study represents a new level of proteomic coverage of a eukaryotic model organism and represents a paradigm for future mammalian systems.

CONTENT

ABBREVIATIONS

ACN	acetonitrile
ADC	analog-to-digital converter
AIF	all-ion fragmentation, MS^E
AQUA	absolute quantification peptide
CAA	chloro-acetamide
CDS	coding sequences
CID	collision-induced dissociation
DC	direct current
DIA	data-independent acquisition
DMSO	dimethylsulfoxide
DTT	dithiothreithol
ECD	electron capture dissociation
eFT	enhanced Fourier-transformation
emPAI	exponentially modified PAI
ESI	electrospray ionization
ETD	electron-transfer dissociation
FASP	filter-assisted sample preparation
FDR	false-discovery rate
FT	Fourier-transformation
FT-ICR	Fourier-transform ion cyclotron resonance
GdmCl	Guanidinium hydrochloride
h	RF-only hexapole
HCD	higher-energy C-trap dissociation
HPLC	high-performance liquid chromatography
I.D.	inner diameter
IAA	iodo-acetamide
iBAQ	intensity based absolute quantification
ICAT	isotope-coded affinity tag
IP	immuno-precipitation
iTRAQ	isobaric tag for relative and absolute quantitation
LC	liquid-chromatography
LIT	linear ion trap
m/z	mass-to-charge ratio
MALDI	matrix-assisted laser desorption/ionization
MCP	multi-channel plate

Met(O)	methionine sulfoxide
MMA	N-methylmercaptoacetamide
MRM	multiple reaction monitoring
MS	mass spectrometry
MS^1	survey scan, full scan, normal mass spectrum
MS^2	MS/MS scan, tandem scan, fragmentation scan
MudPIT	multidimensional protein identification technology
o	RF-only octapole
OT	Orbitrap
PAI	protein abundance index
PEP	posterior error probability
PrEST	protein epitope signature tag
PTM	post-translation modifications
q	RF-only quadrupole
Q	mass selection quadrupole
QCAT/QconCAT	quantitative concatenated protein
QqQ	triple quadruple
Qq-TOF	quadrupole-TOF
RF	radio frequency
ROS	reactive oxygen species
RP	reversed-phase
SAX	strong anion exchange
SCX	strong cation exchange
SDB-RPS	poly(styrene divinylbenzene) reverse phase sulfonate
SDC	sodium deoxycholate
SDS	sodium dodecylsulfate
SILAC	stable-isotope labeling by amino-acids in cell culture
SIM	selected ion monitoring
SPE	solid-phase extraction
SRM	selected reaction monitoring
StageTip	Stop-and-Go Extraction Tip
TCEP	tris(2-carboxyethyl)phosphine
TGA	thioglycolic acid
TMT	tandem mass tag
TOF	time-of-flight
TPA	total protein approach
UHPLC	ultra-high pressure liquid chromatography
XIC	extracted ion current

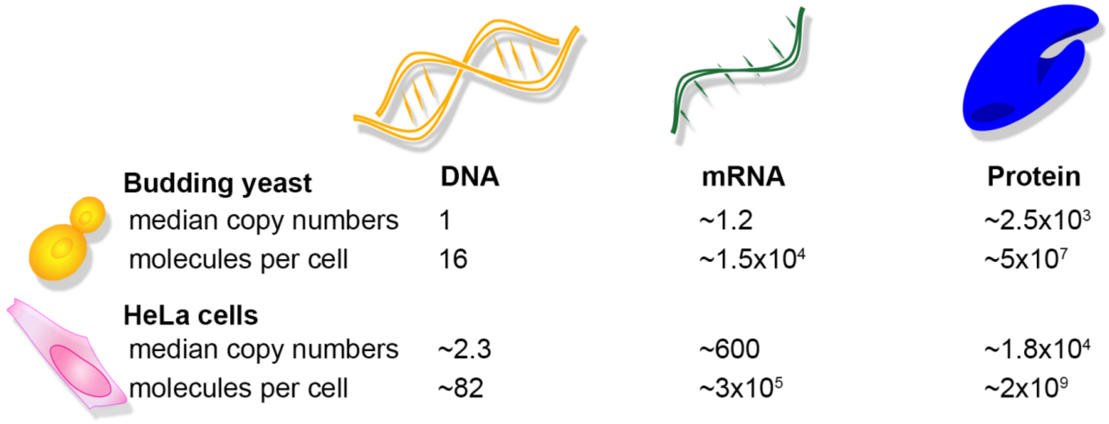
1 INTRODUCTION

TECHNOLOGIES FOR SYSTEMS BIOLOGY

The understanding of complex molecular mechanisms of entire living systems, preferably in a mathematical framework, is termed systems biology. Even though this task appeared to be impossible, and in some respect still does, great strides have been made in defining the ‘parts lists’ of organisms. Demonstrating the feasibility to analyze and sequence a complete eukaryotic genome was achieved in 1996 when the complete DNA sequence of the model organism *Saccharomyces cerevisiae* was reported, containing an estimated 5,885 protein-coding sequences [1]. The human genome project took billions of dollars and years of hard labor, and included many laboratories. A draft version of this genome was announced in 2000 by two rival teams and followed several years later by more complete versions [2, 3]. Despite the much larger size of the human genome the number of protein-coding genes was estimated to be only about 20,000 [4, 5]. In recent years revolutionary next-generation sequencing and chip-based platforms have been developed and these now enable similar coverage in only a few days of measurement.

Transcriptomics – the determination of the identity and amount of RNA species in a biological system – strongly profited from the developments in genomics [6]. Already a year after the publication of the yeast genome the first transcriptome under exponential growth conditions was published and 4,665 expressed transcripts were observed [7]. Transcriptomics delivered increasingly comprehensive data and is now able to determine accurate mRNA copy-number values. However, in most biological systems functional and regulatory processes rely on proteins. Furthermore mRNA and protein expression values were often found to be only weakly correlated, limiting the usefulness of transcriptome measurements as proxies for proteins [8, 9]. Direct analysis of the dynamic proteome is therefore the most desirable approach for global functional analysis and for the understanding a given phenotype in a system-wide manner [10, 11]. However biological complexity and dynamic range increase nearly exponentially from genome to transcriptome to proteome. To analyze an entire proteome is therefore a formidable challenge that had eluted researchers for decades (**Fig. 1**).

The modern field of proteomics has roots in classical microbiological and biochemical tools such as light microscopy, antibody based, and two-hybrid based systems [12]. Even though comprehensive studies applying these technologies have been performed, they were quite laborious and they were always limited by the underlying principle-of-function of the technology employed. In particular, each experiment typically covered only a single protein or small part of the proteome at a time [13-15]. The inventions of matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) allowed gentle ionization and vaporization of intact proteins and peptides. These efficient ionization technologies, combined with increasingly sophisticated mass spectrometric (MS) instrumentation opened new vistas for the study of proteins. Especially ESI became popular for proteomics since analytes are sprayed, volatized, and ionized directly out of solution and could therefore be directly supplied to the mass spectrometer using liquid chromatography systems (LC) [16]. The new field of MS-based proteomics then quickly developed and a wide variety of novel proteomic research areas were born.



	DNA	mRNA	Protein
Budding yeast			
median copy numbers	1	~1.2	~2.5x10^{3}
molecules per cell	16	~1.5x10^{4}	~5x10^{7}
HeLa cells			
median copy numbers	~2.3	~600	~1.8x10^{4}
molecules per cell	~82	~3x10^{5}	~2x10^{9}

FIGURE 1 | Median copy numbers and total molecules per cell of two model systems.
(Data obtained from BioNumber Database (BNIDs) entries: 106198, 100204, 103023, 108248, 104330, 109526, 108425, 109387; Milo et al., 2010 [17])

BOTTOM-UP PROTEOMICS TECHNOLOGIES

In MS-based measurements analytes need to be ionized to enable their observation. These ions enter the mass spectrometer and are focused and guided by electrical fields to the analytical components of the mass spectrometer; the behaviors of the ions in the high vacuum exclusively depend on their mass-to-charge ratio (m/z, unit: Th) and mass spectrometers characterize analytes according to these values. MS-based proteomics can be thought of as consisting of two distinct research fields, the analysis of intact proteins (top-down proteomics) and the analysis of peptides resulting from enzymatic digestion of the proteins (bottom-up proteomics). Even though measurements of highly purified intact proteins, transmembrane proteins, protein complexes and even moderately complex proteomes by MS are possible [18-20], mass-spectrometers are much more sensitive for ions with smaller mass. To generate smaller peptides, which are still unique to the protein from which they originated, dedicated sample preparation workflows were designed that include a proteolytic digestion step. Digestion with sequence specific enzymes has become the most common approach for large-scale proteomic screens. This is because peptides are typically better separated by chromatography, easier to ionize, and they are much more readily fragmented than intact proteins [21]. For the analysis of complex proteomic peptide samples, liquid-chromatography systems are always coupled to mass spectrometers (LC-MS).

SAMPLE PREPARATION FOR BOTTOM-UP PROTEOMICS

Sample preparation is an essential part of the proteomics workflow. The processing steps are typically based on classic biochemical techniques and are commonly adapted to the starting material and the questions addressed in the experiment. Even though dedicated protocols exist there are common steps that are typical for all bottom-up proteomic workflows. The major ones are cell- or tissue-lysis, protein denaturation, reduction of disulfide bonds, alkylation of cysteins, enzymatic digestion of proteins, and peptide clean-up before LC-MS analysis. Interaction studies additionally apply an immuno-precipitation (IP) step before denaturing proteins. If post-translation modifications (PTMs) are of interest, specific affinity-precipitation and enrichment steps targeted to the PTM of interest are added after the proteolytic digestion. In case of deep proteomic measurements fractionation techniques are often applied.

CELL AND TISSUE LYSIS

The very first step of all *in vitro* measurements is the disruption of cellular structures. Generally tissue and cell samples are lysed using one of a small number of established methods. The major choices in lysis are whether denaturing or native protein folding is desired; denaturation for instance may hinder subsequent interaction analysis. Simple boiling of cell material typically denatures the proteins and inactivates the enzymes which can be very useful for those desiring complete proteome coverage and PTM analysis. Bead-milling, grinding, rotor-stator, or blending systems are often used to disrupt rigid structures, but can also extract proteins in their native context if the sample is kept frozen [22, 23]. Ultrasonic homogenization is a softer lysis method in which large nucleic-acid structures are sheared in the process [24]. This may reduce unintentional co-precipitation of DNA- and RNA-binding proteins during centrifugation steps and improve their accessibility to proteolytic enzymes.

In addition to physical cell disruption, most approaches use chemical or enzymatic lysis approaches. As in the case of physical disruption, native or denaturing conditions are chosen. The very strong detergent sodium dodecylsulfate (SDS) has classically been the most common means to solubilize and denature proteins. Weaker detergents, emulsifiers, and surfactants without chaotropic properties such as Triton X-100 are also often used in sample lysis but are mostly applied in interaction proteomics. Surfactants typically help in the breakdown of membrane fractions and to solubilize membrane spanning proteins [25, 26]. If non-denaturing lysis conditions are chosen, native enzymatic activities remain. In case of proteases this can lead to severe loss of intact proteins of interest and PTM-modifying enzymes may alter the extent and quantities of PTMs; therefore inhibitor chemicals are commonly added during the lysis to avoid sample preparation related modifications [27, 28].

Classical workflows include a lysate clarification step after cell disruption. Even though this clarification step is rarely discussed, most workflows include the removal of the insoluble fraction after lysis. Removing crude remnants prior to sensitive chromatography or interaction analysis is necessary to avoid blocking of the column or unspecific retention of proteins in interaction studies. The most common clarification step across sample preparation methods remains high-speed centrifugation and pelleting of insoluble parts; the cleared supernatant is then transferred for further processing [29, 30]. The reactor based sample preparation method developed in the Figeys group applies high-pressure filtration before further processing steps are performed [31]. Special equipment is necessary for the high-pressure filtration method while simple centrifugation can be performed on common laboratory centrifuges. However, centrifugation is less reproducible and sample loss may be more pronounced.

REDUCTION OF DISULFIDE BONDS

Disulfide bonds are naturally occurring inter- and intra-protein crosslinks that stabilize tertiary and quaternary protein structures. These bonds are stable but readily break or form under reductive or mildly oxidizing conditions, respectively. Clearly disulfide bonds need to be broken to efficiently unfold and digest proteins by enzymatic proteolysis. Classically 2-mercaptoethanol was used for this purpose but chemicals with higher reduction efficiency have also been described and employed in sample preparation workflows.

The most common reducing agent for proteomics is dithiothreitol (DTT), which demonstrates very high reducing efficiencies at slightly basic conditions. DTT breaks disulfide bonds by two successive nucleophilic attacks with two thiol-disulfide exchange reactions. After the reaction a kinetically favored ring-shape is formed by DTT; one DTT breaks one disulfide bond and therefore needs to be supplied in excess (**Fig. 2A**) [32]. Because DTT is readily oxidized, it is unstable under aqueous and basic conditions and needs to be freshly prepared before use. A second popular reducing agent is tris(2-carboxyethyl)phosphine (TCEP). It is stable and active across a broader pH range and demonstrates similar disulfide reductions rates [33, 34].

MODIFICATION OF FREE CYSTEINE

Cysteine residues contain a reactive thiol group which readily reacts with various reagents or other free cysteine groups to form disulfide bonds. Because these reactions may interfere with subsequent analysis, cysteine groups are typically irreversibly modified by alkylating reagents. The major requirements of these alkylating reactions are very high specificity while maintaining high efficiencies. Independent of the reagent, free cysteine groups are necessary to efficiently alkylate thiol groups.

The initial reduction of disulfide bonds is an important step before alkylation, but because thiol-based reducing agents are normally used, excessive amounts of the alkylating reagent are necessary to avoid reaction quenching [35]. In contrast to thiol-based reducing agents, TCEP was shown to be compatible with some alkylating reagents and does not necessarily affect the alkylation reaction [36].

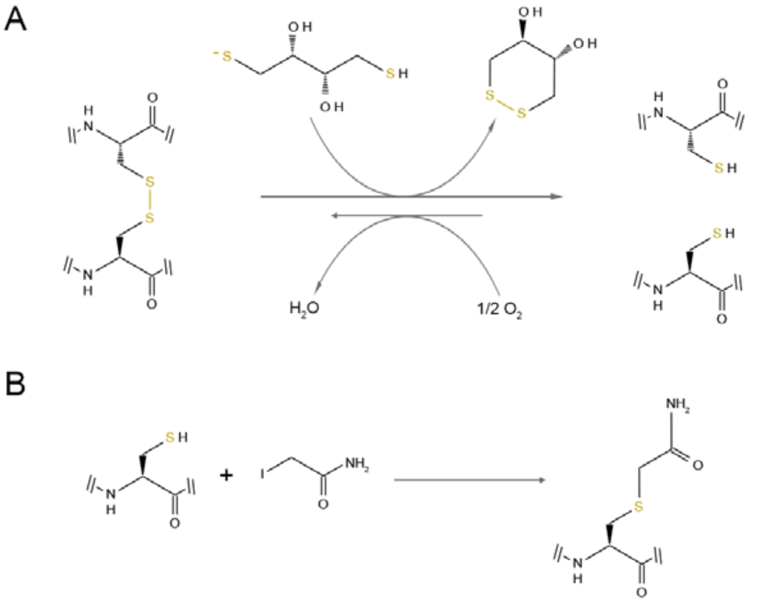


FIGURE 2 | Reactions performed on cysteine side-chains. A Reduction of disulfide bonds by DTT and the spontaneous oxidation forming disulfide bonds. **B** Reaction of iodo-acetamide with free cysteine residues. (Adapted from Cleland et al. and Rombouts et al. [32, 37])

The commonly used alkylating reagents date back to Edman degradation protocols and a wide variety of these chemicals has been described over the last decades. Even though many options exist in principle, only a small subset is applied in proteomic studies. Especially halo-acetamides are highly reactive and selective, leading to near complete alkylation of free cysteine residues. Among proteomic studies iodo-acetamide (IAA) is the most common alkylating reagent yielding stable S-carboxamidomethyl cysteine residues (**Fig. 2B**). Iodide is a better leaving group than other halogens and leads to a fast reaction [35]. Because of the high reactivity it was believed to be a superior choice for alkylating reactions, but IAA has been demonstrated to react with lysine residues resulting in identical composition as the di-glycine remnant tag used to identify the modification in MS-based ubiquitination studies [38]. Such side-reactions can be avoided by early quenching of the alkylating reagent or by using the somewhat less reactive chloro-acetamide (CAA) analog.

ENZYMATIC DIGESTION OF PROTEINS

Enzymatic protein digestion remains the most important and time-consuming step in sample preparation. Even though substrate proteins are more accessible in a denatured state, even minor amounts of strong chaotropes reduce or completely block protease activities; most enzymes are almost entirely incompatible with SDS which is mostly used during cell disruption. Therefore a removal and exchange step is typically employed after protein alkylation. The most common removal methods are protein precipitation (in-solution sample preparation) and filter-assisted removal (FASP). The buffer is then replaced by enzyme compatible surfactants or chaotropic agents [30, 39]. The newly established buffer conditions are typically chosen to be compatible with the applied protease. Classically urea is used as a chaotrope for the enzymatic digestion but recent studies demonstrated that certain other conditions increase protease performance and are better suited for the digestion step [40].

The choice of the protease can have an important role in the outcome of a proteomics experiment. The resulting peptides are analyzed by LC-MS and should therefore be suitable for the surface chemistry of the chromatography material and for the mass spectrometer used. The enzyme should also be highly sequence specific and active. Since most shotgun approaches use the positive ion mode (MS analysis of gas phase cations) the enzyme should generate peptides that are readily chargeable under ESI conditions and the resulting peptides should be of an m/z distribution easily observable by the mass spectrometer. While all these aspects are important, only two proteases have gained broad popularity in proteomic studies: the endoproteases Lys-C and trypsin cleave C-terminal to lysines or lysine and arginine residues, respectively [41].

Lys-C and trypsin are well suited for proteomic studies because nearly all peptide products carry basic lysine or arginine at their C-terminus; under acidic conditions N-termini and the C-terminal amino acids are charged leading to multiple charges. Multiply-charged peptides are highly suited for MS-based proteomics because they yield information-rich fragmentation spectra. They are also easily distinguishable from non-peptide polymers which are typically singly-charged. A combination of both enzymes yields even higher efficiency since Lys-C tolerates higher concentrations of denaturant and the resulting peptides with internal arginines can easily be cleaved by trypsin at lower concentrations [41].

PEPTIDE CLEAN-UP FOR LC-MS ANALYSIS

Most protocols include a final clean-up step before chromatographic separation. Remaining detergents can damage chromatography media and insoluble fragments can block the very sensitive parts of high pressure chromatography set ups; a damaged or clogged column is costly and leads to lost samples and LC-MS downtimes. Salt remnants may cause ionization-suppression in ESI and may form crystals in the MS-components. Some laboratories therefore perform on-line sample clean-up steps using pre-columns and column-guard systems. Risk-free and cheaper solutions are disposable solid-phase extraction (SPE) systems, which can be quickly applied before loading the sample on a LC column. A specialized SPE design for proteomic sample clean-up was introduced by Rappsilber et al. in 2003. So called Stop-and-Go Extraction Tips (StageTip) are constructed of regular pipette tips, which can be equipped with small amounts of Teflon-embedded chromatography material. This simple and inexpensive yet sensitive solution is ideal for the microgram range quantities applied for LC-MS analysis [42].

StageTips can be supplied with various stationary phase chemistries and can be used for one- or multi-dimensional fractionation of the peptide contents [43, 44]. Even though fractionation is possible, StageTips are typically only used for peptide clean-up before LC-MS measurements. Most peptide purification protocols employ reversed-phase C_{18} bead materials comparable to the resin used during the analytical liquid chromatography separation. The peptides are therefore loaded and washed under aqueous conditions removing non-binding salts. The clean peptides are then eluted using volatile organics such as acetonitrile (ACN). The final step of sample preparation is the concentration of the purified peptides and the removal of volatile components using a vacuum concentrator centrifuge.

HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY FOR BOTTOM-UP PROTEOMICS

The very high complexity and dynamic range of digested proteomic samples generally necessitates a molecular separation before mass spectrometric analysis. For this purpose high-performance liquid chromatography (HPLC) systems are commonly coupled to mass spectrometers. The aim of the chromatography is to sufficiently separate peptides in time and upon elution from the column, spray them directly into the mass spectrometer to measure them by MS and MS/MS. The increasing sample complexity and throughput of LC-MS systems also necessitate very high chromatographic reproducibility, chromatographic and electrospray stability, and sensitivity. This can be achieved by nano-flow high-performance liquid chromatography (nano-HPLC) systems and column ovens to improve the reproducibility and chromatographic resolution (**Box 1**) [45, 46]. Nevertheless chromatographic reproducibility remains a major challenge in proteomics.

The resolution of HPLC systems has been studied over the last decades but major improvements became feasible when slurry packing of stationary phases with very small diameter beads became possible. Band broadening and therefore resolving power of a chromatography system is described by the *van Deemter* equation (*HETP = A + B/u + Cu*). The equation takes major commonly occurring physical phenomena in chromatography into account that are major factors effecting peak shape – channeling (*Eddy diffusion, A*), random diffusion (*Diffusion coefficient, B*), mass transfer (*Resistance to mass transfer coefficient within particles, C*), and linear mobile-phase velocity (*u*) (**Box 1, Fig. 3**). A small *HETP* value is desired to achieve high chromatographic resolution and each factor can be optimized by itself [47, 48].

The Eddy diffusion depends mainly on the packing of the column material; large, non-uniform sized or badly packed particles tend to allow channeling, which means that some molecules may travel faster through the packed bed while others are retained and thereby cause peak broadening. The longitudinal diffusion is caused by void volumes and occurs over the separation time, diffusion along the flow-path cause peak broadening, whereas higher mobile-phase velocities can counteract diffusion. Therefore narrow columns with higher flow rates are typically used on HPLC systems. The mass transfer between stationary and liquid phase that is responsible for separation largely results from the porous surface of the packing material. The bead material therefore typically needs to be porous as this increases the surface area of the beads; shallow pores decrease the stagnant mobile phase within the pores and cause faster transfers resulting in less peak broadening. This is typically the case for small beads and therefore the plate height (a measure of separation efficiency) of a column with smaller beads is typically less affected by the linear velocity and the optimal velocity spans a larger range. The mass transfer can further be optimized by increased chromatography temperatures and by causing the stagnant mobile phase to exchange faster.

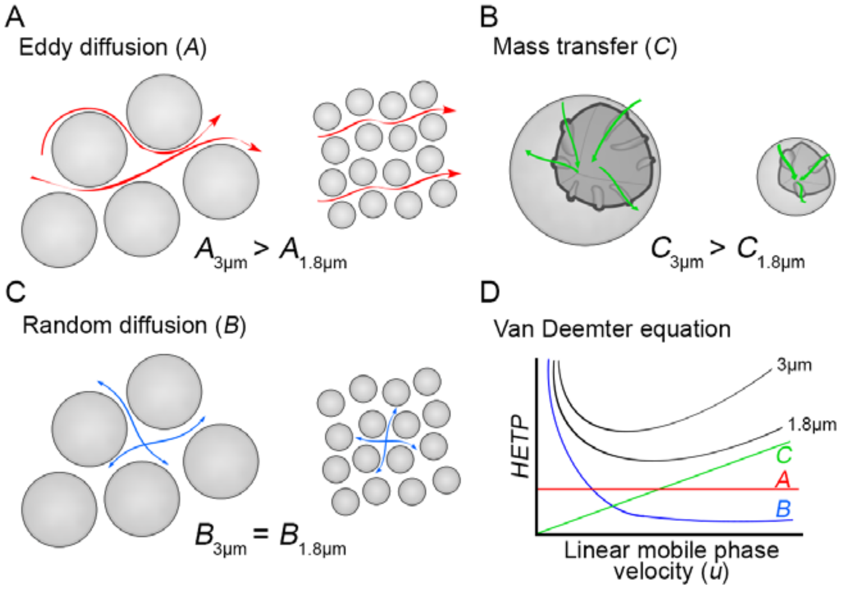

A
Eddy diffusion (A)
$A_{3\mu m} > A_{1.8\mu m}$
B
Mass transfer (C)
$C_{3\mu m} > C_{1.8\mu m}$
C
Random diffusion (B)
$B_{3\mu m} = B_{1.8\mu m}$
D
Van Deemter equation
3µm
1.8µm
C
A
B
HETP
Linear mobile phase velocity (u)

FIGURE 3 | Influence of physical phenomena on chromatographic resolution. A-C: Schematic depiction of factors influenced by particle size. **D:** The effect of flow rate on chromatographic plate height according to the van Deemter equation (adapted from Meyer, 2013 [49]).

These factors are a focus of recent developments and are a reason for the appearance of HPLC-MS systems with long columns (>15cm), narrow inner diameters (I.D., <100 µm), and sub-2µm packing materials running at increased temperatures (50 °C). The major difficulty of these setups is the reproducible and uniform packing of such columns. In particular decreasing bead size and increasing packed-bed lengths cause higher back-pressures of the columns. Because of the higher desired performance of chromatography, ultra-high pressure liquid chromatography (UHPLC) systems are increasingly needed with state of the art systems allowing pressures of 15,000 psi (1,000 bar).

A further aspect of backpressure and peak shape comes from the column outlet, which is typically an emitter tip pointing at the inlet of the mass spectrometer and which should involve as little peak broadening as possible (**Fig. 4**). Because of the direct coupling chromatographic conditions need to be MS compatible. Especially reversed-phase (RP) chromatography media such as C_{18} allow eluting peptides in gradients of volatile liquids such as increasing concentration of acetonitrile (ACN). At this point peptides need to be transferred from the liquid into the gas phase by ESI.

Box 1 – Definitions for chromatography (McNaught and Wilkinson [48])**:**

Mobile phase is liquid running through the stationary phase.

Flow rate **(F_C)** is the volume of mobile phase passing through the column per time (nano-flow typically 100-400 nl/min).

Mobile-phase velocity (u) is the linear velocity at which the mobile phase passes through the average cross-section of the chromatographic bed (typically cm/min). The mobile-phase velocity mainly depends on the inner diameter (I.D.) of the column and the packing material.

Gradient elution is a procedure where the make-up of the mobile phase is changed during the elution process.

Peak width at half height **(w_h)** is the retention time parallel to the baseline at 50% of its maximum.

Peak width at base **(w_b)** is the retention time at the base of a peak.

Hold-up volume **(t_M)** is the retention volume (time) of an unretained compound.

Peak resolution **(in chromatography, R_S)** is defined by the separation of two peaks in terms of their average peak width at base (for adjacent peaks: $t_{R2} > t_{R1}$, $w_{b1} \approx w_{b2}$, $R_S \approx (t_{R2}\text{-}t_{R1})/w_{b2}$).

Height equivalent to a theoretical plate (HETP) is a value describing the theoretical resolving power of a chromatography system.

Plate number (N) is a number to evaluate chromatographic performance assuming a symmetrical Gaussian peak: $N = 5.545(t_R/w_h)^2$ (more plates translates to better chromatography)

Plate height (H_{eff}) is used to evaluate a chromatographic system according to the column length (L) with $H = L/N$.

ELECTRO SPRAY IONIZATION

After chromatographic separation, eluting peptides need to be charged and the liquid evaporated so that they can enter the mass spectrometer's vacuum conditions. ESI is based on the theoretical principles of electrostatic dispersion and the desolvation to gas-phase ions [50, 51]. While theorized already in 1968, the experimental demonstration of electrospray ionization of intact macromolecules was first demonstrated by Fenn et al in 1988 [52]. In electrospray, a spray tip needs to be charged to kilovolt potential (**Fig. 4**). The liquid and the analytes within the liquid are emitted by flow and dispersed into very small, highly charged droplets causing the formation of a spray plume. During the flight, liquid evaporates from these droplets, which leads to higher field densities on the surface of the droplets [53]. When a density threshold is reached smaller droplets are formed (Coulomb explosion) and separate at points of high curvature. The repeated separation of larger to smaller droplets combined with the rapid evaporation finally lead to desolvated ions in the gas phase [54]. In the final step ions enter the transfer capillary of the opposite charge. The capillary is typically heated to assist further evaporation and forms the bridge from atmospheric pressure to vacuum.

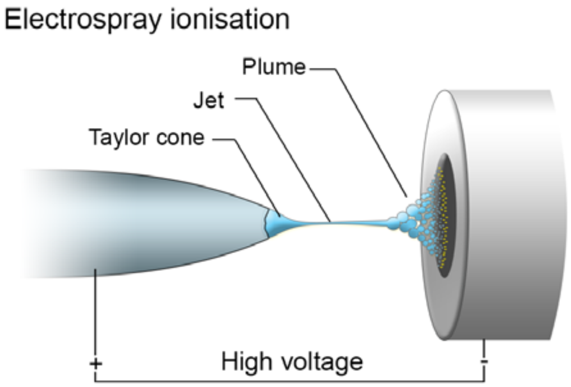


FIGURE 4 | A schematic view of electrospray ionization (ESI). The flow rate of the liquid defines how well the Taylor cone and the desolvation of the electrospray plume perform (adapted from Hahne et al., 2013 [55]).

The efficiency of ESI strongly depends on the formation of small droplets leaving the spray tip. The droplet size mainly depends on the quantity of liquid being sprayed per time (flow rate). A very high flow rate may cause large droplets to fly into the mass spectrometer resulting in disturbed ion formation and intensity variations; such a condition can interfere with quantification and large droplets may carry unionzed analyte as well as contaminants into the mass spectrometer [56]. For this reason drying gas was initially applied to facilitate the evaporation of the liquid; this is not necessary with most of the current nanoflow systems. Because the flow-rate is directly related to the linear velocity by the cross-section area of the column, very narrow columns can be used with low flow-rates while maintaining a high linear flow-velocity across the chromatographic bed. LC-MS setups for proteomics typically use nano-flow columns leading to efficient electrospray, which together achieves high chromatographic resolution and sensitive ionization.

MASS SPECTROMETERS FOR BOTTOM-UP PROTEOMICS

After peptides are sprayed into the mass spectrometer by ESI, they need to be guided to the mass resolving components to be identified and quantified (**Fig. 5, 7, 8**). Peptide ions normally enter the mass spectrometer through an ESI capillary, which forms the bridge between atmospheric pressure and vacuum. Ions leaving the back end of the capillary are often focused to proceed to further components, while non-charged gas molecules are pumped away, improving the sensitivity of the instrument. In some instruments the ion focusing is performed by ring-shaped poles through which the ions travel. Especially the ion-funnel and S-lense technologies demonstrate high sensitivity as the first stage of a mass spectrometer [56, 57]. The mass spectrometric analyzers themselves need to record the masses and intensities of the ions and can precisely determine the peptide sequence by peptide fragmentation. Three technical features are therefore commonly supplied by nearly all mass spectrometers used for standard bottom-up proteomics analysis: mass selection, peptide fragmentation, and mass analysis.

QUADRUPOLES AND MULTIPOLES

Quadrupoles or multipoles are extremely versatile functional components and are therefore components in nearly every mass spectrometer type used in proteomics. They typically consist of four, six, or eight hyperbolic or cylindrical rods which are equally spaced in a pair wise orientation around the center (**Fig. 5A**). The two opposing rods are always set to the same, and the neighboring to opposite potentials. Ions are guided into the center of the rod arrangement and applied voltages lead to attractive and repulsive forces. Typically the voltage on the rods is applied periodically as radio frequency waves, thereby changing attraction and repulsion ($\varphi_0 = U + V\cos(\omega\ t)$; φ_0: total potential, U: direct current (DC) voltage, V: radio frequency (RF) voltage, ω: frequency). This causes ions to follow controlled oscillating trajectories. The radial oscillation distance of a specific m/z is a function of the rod geometry and the applied potentials. Only ions within the physical oscillating distance from the center of the quadrupole to the rods have stable trajectories along the quadrupole ($2r_0$, **Fig. 5A**) [58].

The simplest mode of operation of these multipoles is ion guiding and transmission. The best transmission is achieved by RF-only quadrupoles (q) or 2N-multipoles (hexapoles: h, octapoles: o). Only RF voltages are applied without DC-voltages (U=0) leading to stabilization across a large m/z range (**Fig. 5B**). Higher-order multipoles offer the best guiding and transmission characteristics, and perform best at lower vacuum. They are therefore often used at the front parts of the ion path and in collision cells. Because ions follow the stabilized ion paths, bent RF-only quadrupoles can be used to remove residual neutrals and photons in the early sections of the mass spectrometer or after collision cells [56].

Mass selection quadrupoles (Q) are often employed in MS-based proteomics. As described earlier, specific ion trajectories can be stabilized within a quadrupole but destabilization of all other m/z values is equally important. Qs are found in nearly every mass-spectrometer and act as m/z-filters, with high-performance quadrupoles achieving narrow m/z selection without dramatic loss in transmission. The mass-selection capabilities are also used in conjunction with the detection capabilities of electron multipliers, constituting a mass analyzer (see **Mass Analyzers**). The scanning of the quadrupole is typically performed by linearly increasing the U/V voltages (**Fig. 5B**). Starting from a low to high U/V ratio, smaller to larger m/z values are stabilized by the quadrupole and thereby a specified mass range can be scanned. Because the area under the curve defines the stability region for a given m/z it also reflects the resolving power of quadrupole analyzers; higher resolving powers would result in a steeper scan lines, which would narrow the stable *U*/*V* ratio without ion loss. The resolving power strongly depends on the quality of the quadrupole and the scan speed but it typically only reaches about 1000 along the m/z range which is typically measured in proteomics experiments (**Box 2**) [58, 59].

Trapping quadrupoles are also often used where ion storage or collection is needed. An additional electric field is applied in the front and back of the quadrupole to collect and hold ions within the quadrupole. This is typically achieved by two small additional quadrupoles before and after the main quadrupole or by lenses at the ends (trapping elements) (**Fig. 7A**).

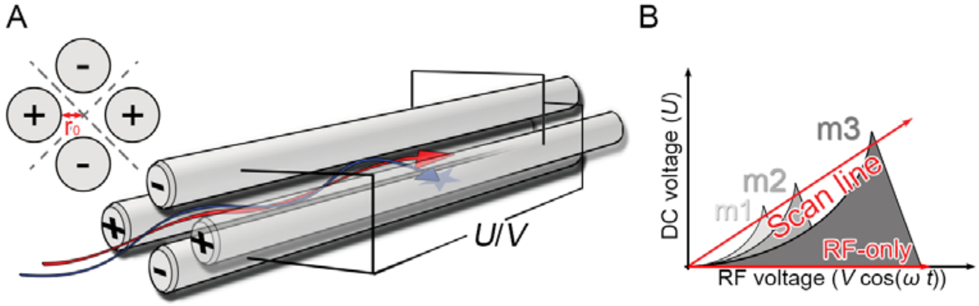


FIGURE 5 | Schematic function of a quadrupole for mass selection or ion guiding. A Conceptual construction of a quadrupole. **B** Diagram of quadrupole scanning and stabilization of ions (adapted from Barner-Kowollik et al., 2012 [58]).

COLLISION CELLS

The second common component of mass spectrometers used in proteomics is the collision cell for peptide fragmentation. Many technologies were developed to fragment ions resulting in different fragment ion distributions. The most prominent fragmentation principles are collision-induced dissociation (CID), higher-energy C-trap dissociation (HCD), and electron-transfer dissociation (ETD), which is a further development of electron capture dissociation (ECD). These fragmentation systems are typically performed in RF-only quadrupoles or multipoles (see **Quadrupoles and multipoles**). This is beneficial because all generated ions are stabilized and the fragmentation can be performed at higher pressures without ion destabilization. In one typical procedure specific peptide precursors are isolated using a mass selection quadrupole and are transferred to a fragmentation cell. The selected peptides are then fragmented by collision (termed collision induced dissociation, CID, higher energy collisional dissociation, HCD) or by transferring electrons to the ions (ETD); resulting fragment ions are then measured by the mass analyzer.

During CID the ion beam is directed into the collision cell. Here peptides of high kinetic energy collide with a neutral gas that is locally leaked into the multipole (typically He, N_2, or Ar). The collisions leads to conversion of kinetic energy into internal energy of the ion (activation); the gained internal energy destabilize the chemical structure and the peptides decompose into smaller peptide fragments. The reaction can be controlled by the gas density and the kinetic energy of the ion and also depends on their m/z. In CID b- and y-ions are both generated but mostly y-ions are observed (**Fig. 6B,C**) [60, 61]. The same working principle of CID is termed HCD when implemented on Orbitrap mass spectrometers. In ion trap CID the ions are excited by an RF field and the precursor can be completely converted to fragments if activated sufficiently long. However, ion trap CID often only fragments to one lower energy state, i.e. water loss or loss of a phospho group, necessitating more complicated fragmentation schemes. HCD has the further advantage over CID to also stabilize lower mass reporter ions after activation; these reporter ions can therefore be used for database search and offer a more complete fragmentation spectrum [62]. Conversely, HCD spectra have a higher proportion of internal ions generated by double fragmentation of peptides [63].

In contrast to collision based fragmentation technologies ETD, as its name implies, activates the peptide ions by transferring electrons. For this purpose special fragmentation cells are used which are filled with radical anions; the radical transfers its electron to the peptide-cation causing the peptide backbone to break. ETD performs relatively better for larger mass ions and a more instantaneous fragmentation method, which makes it more likely to retain weakly bound post-translational modifications [64]. Because it is less sensitive for small peptides and because the technological implementations of ETD is not as well developed as CID based technologies, ETD still remains rarely used but promises high potential for future applications.

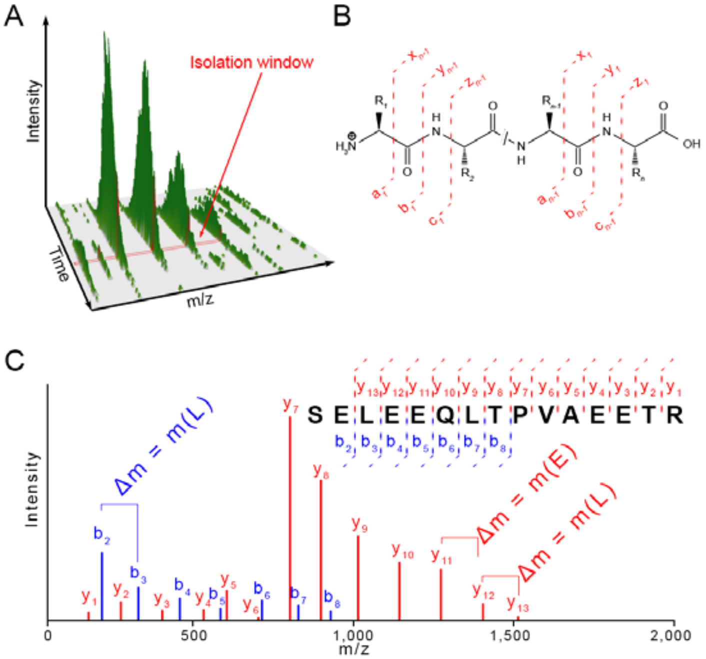

A
Isolation window
Intensity
Time
m/z
B
x_{n-1}
y_{n-1}
z_{n-1}
x_1
y_1
z_1
R_1
R_2
R_{n-1}
R_n
H_3N^+
OH
a_1
b_1
c_1
a_{n-1}
b_{n-1}
c_{n-1}
C
Intensity
y_{13} y_{12} y_{11} y_{10} y_9 y_8 y_7 y_6 y_5 y_4 y_3 y_2 y_1
S E L E E Q L T P V A E E T R
b_2 b_3 b_4 b_5 b_6 b_7 b_8
Δm = m(L)
Δm = m(E)
Δm = m(L)
y_1 b_2 y_2 b_3 y_3 b_4 y_4 b_5 y_5 y_6 b_6 y_7 b_7 y_8 b_8 y_9 y_{10} y_{11} y_{12} y_{13}
0
500
1,000
1,500
2,000
m/z

FIGURE 6 | Peptide mass selection and fragmentation. A 3D elution peak of peptide isotope-patterns. **B** Ion series annotation generated during peptide backbone fragmentation. **C** Fragmentation spectrum and peptide sequencing principle. (Adapted from Steen and Mann, 2004 [61]).

MASS ANALYZER AND DETECTOR

The most important part of a mass spectrometer is the mass analyzer which needs to fulfill certain criteria for defined tasks. The major characteristics of a mass analyzer are the mass resolving-power, mass accuracy, detection limit, and scan speed (**Box 2**). A large variety of mass analyzers exists but four types are mainly used for proteomics because of their cost and performance characteristics. The simplest analyzer is the quadrupole mass filter consisting of a quadrupole and a detector (see **Quadrupoles and multipoles**). The quadrupole is used to scan through the mass range, transmitting one m/z value at a time and an electron multiplier which detects the ions (**Fig. 5B**). Most electron multipliers used in proteomics contain semiconductor channels which are arranged in a way to generate an avalanche of secondary electrons for every incoming ion with sufficient kinetic energy that hits the semiconducting surface (**Fig. 7C**) [56]. An array of channels is called multi-channel plate (MCP) and is often used in mass spectrometers because of the high sensitivity, low production costs, and small footprint. The output current is converted by analog-to-digital converters (ADC); the ADCs mostly define the speed of detection.

A similar working principle is applied in linear ion trap (LIT) analyzers (**Fig. 7A**). LITs consist of hyperbolic quadrupoles with a central selection and two trapping units (see **Quadrupoles and multipoles**). In contrast to the principle of a scanning quadrupole mass analyzer that was described above, a resonance ion ejection is performed. Here ions of each devised m/z in turn are destabilized and ejected through an exit slit in one or two rods of the central quadrupole. The ejected ions are then measured on electron multipliers [65]. Early platforms used these LITs for ion trapping, mass selection, CID fragmentation, and as mass analyzer. While the LIT is capable of all these functions, separating the function into two units is highly beneficial – a first higher-pressure collision unit and a high-vacuum selection and scanning unit. The units operate at the same RF voltages but separate DC supply to transfer ions between the parts [57]. Because of its limited resolution, the LIT is today mainly used in conjunction with a high resolution analyzer.

A different working principle is applied in time-of-flight (TOF) analyzers. Here ions are accelerated, traverse a given distance and the mass of the ions is deduced from the time ions remain in flight, with larger ions traveling slower than smaller ions. Most state of the art TOF instruments use orthogonal acceleration and ion reflectors to precisely time the flight. The ions are filled into the acceleration unit and are pulsed orthogonally along the drift tube. The ions are then reflected in an electrostatic mirror (reflectron) to compensate for the kinetic energy distributions of identical m/z species and to prolong the flight time without increasing the length of the drift tube. Finally the ions are observed on an MCP detector (**Fig. 7B**). Employing long drift tubes and very fast ADCs, resolving powers above 20,000 and a dynamic range of 10^5 can be reached. One challenge of TOF instruments are their large mass drifts over time with mass deviations of 5-50 ppm [58], however, this can be avoided by recalibration of the spectra using internal mass standards or already identified ions.

The highest mass resolving-power is achieved by Fourier-transform ion cyclotron resonance (FT-ICR) analyzers using very high magnetic field. However, these are rarely used because of very high costs. A similar principle, albeit without magnetic fields, is used in Orbitrap mass analyzers (OT), which demonstrate very good performance at lower cost. In a typical scan a defined number of charges are pulsed into the Orbitrap analyzer. Because of the special orientation and shape of the electrodes (outer and inner) ions cycle/orbit the inner axial electrode without any RF fields (**Fig. 7D**) [66]. The centrifugal forces lead to a balanced orbiting while the axial field leads to harmonic oscillation of the ions. A cyclotron frequency inversely proportional to the square root of m/z of the ions is induced and oscillation currents are measured between the two halfs of the outer electrode. The recorded signal is an overlay of those originating from different ion species in the Orbitrap, hence a Fourier-transformation (FT) of the recorded time signal separates waves and amplitudes of the ions giving the m/z and signal intensity of the ions present [67].

The Orbitrap achieves very high mass accuracy and mass-resolving power depending on the number of oscillations recorded (time). The dynamic range of the Orbitrap is limited by the number of charges that can be supplied (dynamic range ~10^4, in proteomics typically <10^3 in one spectrum) [68]. The first generation of the LTQ-Orbitrap family achieved a resolving power of 60.000 at 400 m/z in 1 Hz scans and demonstrated mass accuracies below 2 ppm using internal and 5 ppm with external standards. The Orbitrap can routinely achieve resolving-power above 140.000 at 400 m/z but such resolution settings are rarely used because they increase the scan time. Novel generations of Orbitrap instruments and a smaller diameter Orbitrap (D20) achieve even higher resolving-power with shorter scans [69] (see **Novel MS instrumentation**).

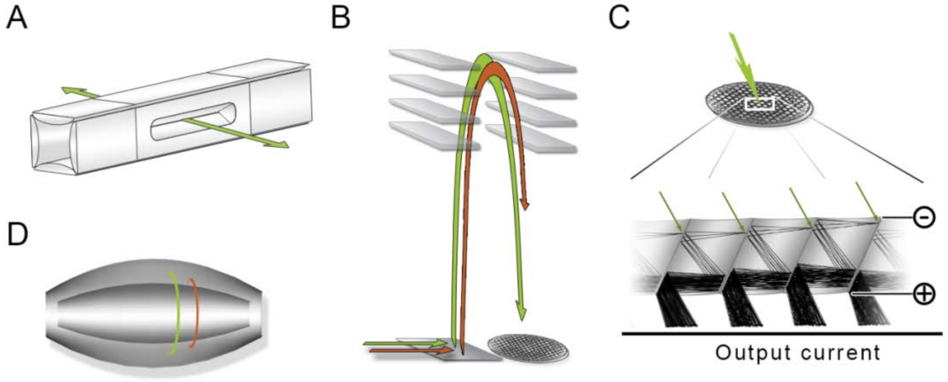


FIGURE 7 | Mass analyzers and detectors used in modern mass spectrometers. A Linear ion-trap quadrupole with scanning capability. **B** Orthogonal time-of-flight (TOF) system with a reflector. **C** Working principle of multi-channel plate (MCP) electron multiplier detector. **D** Schematic of an Orbitrap mass analyzer depicting the central and outer electrodes. (Adapted from Gross and Roepstorff, 2011 [56]).

HYBRID MASS SPECTROMETERS

Nearly all mass spectrometers for bottom-up proteomics offer mass selection, peptide fragmentation, and ion detection but the different instruments are more suited for some tasks than others. The most featured instruments in proteomics are triple quadruple (QqQ) analyzers, Qq-TOFs [16], and Orbitrap instruments [67] (**Fig. 8**). The QqQ instruments consist of three quadrupoles as the name suggests; the first and last quadrupole are mass selection quadrupoles, and the second quadrupole functions as a CID collision cell. The first quadrupole transmits a defined m/z window corresponding to the peptide ion of interest, the second quadrupole is used to fragment the peptide, and the third again filters and scans through the detection m/z window before the ions hit a detector plate [70] (**Fig. 8A**). Qq-TOF instruments have a similar setup but the mass analyzer consists of an orthogonal time-of-flight unit. Because of the lower cost, QqQ and Qq-TOF are often used for targeted analysis (**Fig. 8C**) (see **Experimental designs of bottom-up proteomics**).

Orbitrap instruments contain trap-type mass analyzers and therefore always require external fragmentation devices. The first generations of Orbitraps belonged to the LTQ-Orbitrap family and were equipped with two independent detector types – the linear ion-trap (LTQ part) and the Orbitrap (see **Mass analyzers**). The analyzers are connected by a specialized c-shaped ion trap – the C-trap – where ions can be trapped or re-directed (**Fig. 8B**). The first generation was limited to CID fragmentation in the LIT but later generations were equipped with HCD and optionally ETD fragmentation cells behind the C-trap [57, 67]. Further improvements in ion path, LIT, and the Orbitrap were incorporated over time into the LTQ-Orbitrap family; the last redesigned model – the Orbitrap Fusion Tribrid – was announced 2013 as a Q-OT-qIT instrument featuring a selection quadrupole, C-trap-Orbitrap D20 (denoting the inner diameter of the Orbitrap), and a dual cell linear ion trap analyzer [71].

A different benchtop Orbitrap setup was introduced in 2012 and named Q Exactive (Q-OT); the major hardware differences were the replacement of the LIT mass analyzer by a regular selection quadrupole. Furthermore, ion transmission in this device appears to have improved. The novel hardware construction restricted the instrument to HCD fragmentation only and therefore to an exclusive high-high strategy where all mass measurements are performed in the Orbitrap (high resolution full- and fragmentation-scans). A major operational improvement was the parallel filling option where ions are collected in the C-trap during the preceding scan; this option improves the cycle times up to two-fold because the time for trapping ions is reduced. Additionally an enhanced Fourier-transformation algorithm was introduced (eFT), which doubled the resolving-power per unit time by taking phase information into account [69]. These advances make the Q Exactive a very efficient yet compact and relatively economical instrument for discovery-based proteomics approaches [72].

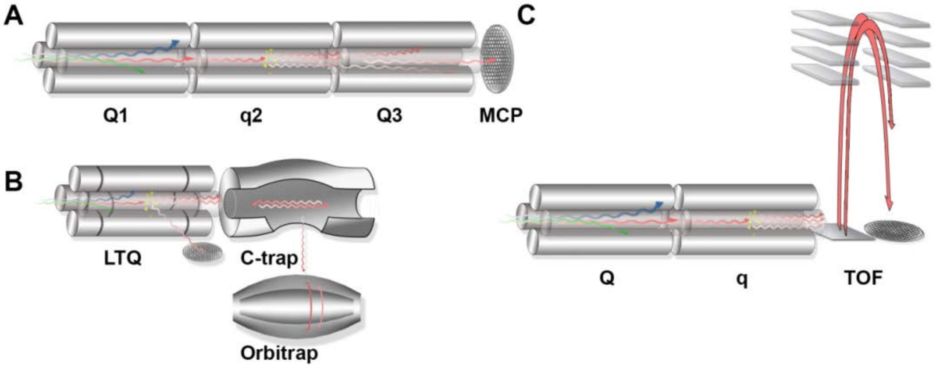


FIGURE 8 | Components of QqQ, LTQ-Orbitrap, and Qq-TOF instruments. **A** Arrangement of quadrupoles of a QqQ instrument, with mass selection quadrupoles (Q1 and Q3), and a fragmentation cell (q2). **B** Dual analyzer arrangement of LTQ-Orbitrap hybrid instruments. **C** Working principle of Qq-TOF instruments with a mass selection (Q), a collision cell (q), and a TOF analyzer.

Box 2 – Definitions for mass spectrometry (McNaught & Wilkinson; Marshall, 2008 [48, 59])**:**

Mass resolution and resolving power (peak width definition) for a single peak made up of singly charged ions at mass m in a mass spectrum is expressed as $m \ / \ \Delta m_{50\%}$ where $\Delta m_{50\%}$ is the width of the peak at 50% of the maximum peak height.

Mass accuracy is the error of the observed mass from the calculated molecular mass. Typically mass spectrometers need regular calibration and mass accuracy may change due to temperature fluctuations.

Dynamic range is defined by the ratio of the most intense to the lowest signal within a single scan.

Signal to noise describes the ratio of the signal caused by the desired ions to chemical or electrical noise.

Detection limit is the smallest amount of material still correctly detectable.

Scan speed is the time needed to acquire a certain signal.

SCAN MODES APPLIED IN MS-BASED PROTEOMICS

State of the art mass-spectrometers allow a variety of scan modes. Especially instruments with trapping capabilities have a large variety of such options. The most simple scan type is the full scan (survey scan, MS^1, or normal mass spectrum); typically no mass selection or fragmentation is applied and the entire m/z range is scanned [16]. In these scans peptide isotope patterns are usually observed; since natural stable-isotope distributions exist for the elements found in peptides so called isotope patterns of all peptide are observed at sufficiently high resolutions. Doubly-charged peptide ions for instance generate isotope patterns with 0.5 Th (m/z = 1/2) and triply-charged peptides 0.33 Th (m/z = 1/3) spacing. Full-scan measurements are typically performed using TOF or Orbitrap analyzers because of their high resolving power [56].

Mass filtering without fragmentation can also be applied – so call selected ion monitoring (SIM) scans. SIM scans are of special interest if high intense ions in another part of the spectrum cause detection issues because of limited dynamic range or space charging [73]. SIM scans can be performed in data dependent or independent manner; in a data dependent scan the window of the SIM scan can even be defined in real-time according to a data independent survey scan.

A rather untypical scan mode is all-ion fragmentation (AIF) or MS^E; here all ions entering the mass spectrometer are fragmented and measured at high resolution. In this and related scan modes, the fragments are related to the precursors via their elution similarity on the chromatographic time axis.

One problem of fragmentation of complex mixtures is the different fragmentation behavior of diverse peptide species. Typically the collision energy is chosen according to the mass and charge (m/z and z) of the precursor ion; since this is not possible in AIF stepped-collision energies can be used instead. Batches of ions are fragmented with stepwise increasing energies; the fragment ions are collected and pooled and are measured together [74]. Other challenges of AIF or MS^E experiments are the high complexity of the AIF scan, the dynamic range, and the interpretation of the measurements. Nevertheless, results rivaling Q Exactive measurements have recently been reported [75].

Scans in which mass selection is applied before fragmentation are typically called MS/MS scans (tandem scan, MS^2). This scan type is commonly used after peptide-fragmentation of a precursor identified in an MS scan (see **Database search**). In so-called topN methods ions are chosen for sequencing according to their intensity; this data-dependent measurement work by the selection of precursor ions according to non-fragmented survey spectra obtained in the same measurement [16]. Similarly single or multiple peptides can be chosen at previous measurements where the chromatographic retention time and exact mass of a selected subset of ions has been determined beforehand. Sequencing events consisting only of measuring a precursor-fragment relationship are then scheduled to take place in the analytical measurements; these measurements are termed selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) when several transitions are monitored [70]. Data-independent scans are generally termed data-independent acquisitions (DIA). A variant of a mass selection and fragmentation DIA mode has recently been reported in which the instrument cycles through a defined number of DIA windows and the resulting fragmentation spectra are compared to a library of peptide fragmentation spectra (SWATH) [76].

Higher order fragmentation scans (MS^n) consist of further fragmentation of the fragments in MS^2 and have been used in specialized applications. One of these is in the field of cross-linking. Here precursors are selected and fragmented with low energy, fragmentation-products are again selected, and fragmented with higher energies. First tests using this scan mode with readily cleavable cross-linkers look promising and the technique could be used beneficially in the field of large-scale structural proteomics [77].

Mass spectrometers with both mass selection and trapping capabilities also introduce controlled ion multiplexing as a scan option. Multiplexed SIM scans, for instance, can trap ions from two or more different mass windows and the collected ions can then be measured together in a single scan. These multiplexing options can be performed in various constellations depending on experimental requirements. In general the conceptual idea is to remove unwanted ions and thereby improve the dynamic range and intensity of the desired ions. The major drawbacks of multiplexing strategies are the prolonged fill times if many different precursors are targeted. This is because of non-parallel ion collection; the windows need to be collected sequentially and while one window is being collected, the ions of the other window(s) are lost.

EXPERIMENTAL DESIGNS OF BOTTOM-UP PROTEOMICS

The workflow of bottom-up proteomics differs strongly across various experimental designs. Three approaches are mostly applied in the proteomic community namely shotgun (discovery-based), directed, or targeted proteomic workflows and acquisition modes [70, 78]. The choice of the approach is often limited by the available hardware because each strategy typically works best on a dedicated MS setup. The major differences of the technologies are the prior knowledge necessary for their application; targeted workflows in particular monitor preselected peptides while directed and discovery-based proteomics work in a more unbiased fashion.

In contrast to the other concepts, targeted proteomics experiments are entirely hypothesis driven; conceptually similar to western-blots, one chooses to monitor only the targeted proteins. But in contrast to anti-body based techniques mass spectrometry directly measures the analyte of interest and it can therefore be more specific and robust. A targeted workflow usually starts by synthesizing, purifying, or purchasing peptides which are expected from the proteolytic digestion of the target proteins of interest and which are unique in their sequence and transitions (‘proteotypic peptides’) [79]. These synthesized peptides are then measured by LC-MS to determine their retention time, m/z-value, and fragmentation spectra (transition coordinates). Alternatively peptide libraries can be used to obtain these values. The actual targeted measurement consists of monitoring the SRM or MRM transitions during the appropriate part of the LC gradient (see **Scan-modes applied in MS-based proteomics**). The much more complex mixtures of biological samples are subjected to the same chromatographic gradient which was used to determine the transition coordinates; the coordinates are then used to mass select and monitor the peptide fragments during their elution. For this purpose precise mass selection would be beneficial at the MS and MS/MS levels but QqQ instruments with relatively modest resolution are usually used for this purpose instead (see **Mass spectrometers for proteomics**) [80]. In contrast to the classical targeted acquisition as described above, the SWATH strategy mentioned above only uses a targeted data analysis. The acquisition of SWATH is independent of prior knowledge while the interpretation of the acquired spectra relies on prior knowledge. After the acquisition the spectra fitting the transition coordinates of desired peptides are searched for the fragment ions from prior knowledge. Because a large number of scans have to be performed in very short time state of the art Qq-TOF instruments must be used [76].

Discovery-based proteomics strategies can be performed without any prior knowledge and are therefore, as the name suggests, primarily used to discover proteins of interest in a biological system. Classically shotgun LC-MS/MS strategies employ topN measurements (see **Scan modes applied in MS-based proteomics**). Because the scans are performed without prior knowledge and because peptide mixtures are typically very complex, pre-fractionation, long measurements, high performance chromatography, and fast scan speeds have been necessary to obtain measurements of low abundant proteins. Because the proteomic coverage of complex samples is strongly hardware dependent, high performance instrumentation is highly beneficial for discovery-based approaches. Increasingly faster and more sensitive LC-MS based systems have been developed over the last years, in particular based on the Orbitrap analyzer, which have enabled a more complete coverage in discovery based proteomic measurements [81].

PEPTIDE SEQUENCING AND IDENTIFICATION

The database search is currently an essential data analysis step as it allows the identification of the peptide species. As described before, peptides are typically mass selected and fragmented in MS^2 scans. The fragment ions are measured and depending on the fragmentation technique mostly y-ions or a mixture of b- and y-ions (or c- and z-ions in case of ETD) are observed. In an ideal scenario every fragment ion of a peptide species would be observed; with the mass differences of the fragment ions corresponding to the mass of each amino-acid lost (**Fig. 6C**). Such a perfect coverage could therefore be used to *de-novo* sequence peptides.

In a real life situation the fragmentation spectra are not perfect; in a typical fragmentation scan peptides are not completely fragmented, many b-ions are lost, internal and side-chains fragments are produced, and co-isolated peptides are co-fragmented making MS^2 spectra difficult to interpret. For these reasons database searches are performed to assign fragment ions to expected sequences. The databases are typically based on deducing possible peptides from protein-coding gene sequences.

Because of ongoing research, sequence databases and their annotations are regularly revised and change over time which can make comparisons to older publications tedious. A major leap for the community was the establishment of UniProt providing unified and well documented databases and annotations for proteomic searches (www.uniprot.org). Typically provided protein sequences are *in silico* digested during a search to generate a set of possible peptide sequences, these sequences are then assigned to spectra. A pre-selection of possible sequences is performed according to the measured mass of the precursor selected for fragmentation [82-84].

The complexity and quality differ widely between different MS^2 spectra even in the same workflow. This requires a quality and statistical control for the search. In the first step of statistical evaluation MS^2 spectra matches are ranked according to their quality and confidence of assignment. This is done by observing the peptide coverage according to mass difference of fragment ions, the correlation of theoretical and experimental spectra, or by the probability that observed peaks could have occurred by chance [85, 86]. The MaxQuant environment for instance calculates a posterior error probability (PEP) for each peptide based on its identification score and length. The length dependent PEP calculation is performed because longer peptides are less likely to occur and may therefore be correct even though their assignment score might be lower. The same algorithms can be applied to assign localization probabilities for post-translational modifications. Classically a simple score cut-off was chosen to guarantee a certain quality but such approaches are statistically insufficient and have been replaced by more sophisticated approaches.

The simplest and most often applied statistical cut-off method is the target-decoy search strategy [87, 88]. For the decoy search a second database of typically reversed peptide sequences is added to the regular search; the decoy search-space is equivalent to the one of the forward database and is assumed to be equally likely to be assigned to a spectrum as an incorrect hit. The knowledge of the reverse hits therefore gives an estimation of false positive assignments. The score in conjunction with the false positive reverse hits can be used to make a false-discovery rate (FDR) cut-off at the peptide level. In a typical experiment a 1% peptide FDR is applied to warrant statistically relevant spectrum assignments meaning that a cut-off is chosen where 1% of the identifications belong to the decoy database [89].

Apart from simple reverse databases, alternatives have been developed. One important aspect is that most peptides have specific amino-acids in the N- or C-terminal position originating from sequence specific enzymatic digestion. These amino-acids are therefore specially treated in the decoy database to avoid peptides with precisely the same mass as peptides in the forward database, which would bias the likelihood of peptide assignment [85].

The third step is a protein-based FDR cut-off. For this purpose peptide hits are assembled into single protein or protein group identifications. If a protein cannot be distinguished from other proteins because the assigned peptides do not contain a unique sequence, they are combined into a protein group. Even though only 1% of all peptide identifications might be false the number of falsely assigned protein groups would be larger if no additional filters were applied. False positive peptides are more likely to result in single protein assignments. A simple assignment to reverse protein groups is not possible and complex calculations are therefore necessary to maintain statistical quality. In the case of the Andromeda search engine the posterior error probability of each protein or protein group is calculated by the multiplication of all sufficiently scoring peptide PEPs. According to these error probabilities, all proteins can be arranged in a list and an FDR cutoff can be performed on a protein level [85, 89].

PROTEIN QUANTIFICATION

Even though the correct identification of a protein is imperative, quantitative information about the identifications is equally indispensable in proteomics. Mass spectrometry, like all biochemical quantification technologies, is affected by the composition of the analytes. As described before, bottom-up MS-based proteomic workflows involve sample preparation, a separation of peptide species by liquid chromatography, the ionization by ESI, and the measurement by mass spectrometry. These steps rely on biochemical differences of the analytes but they also make quantification challenging. Relative quantitative changes of single peptides are a difficult task but the absolute measurements of different peptide or protein species is even more challenging [90].

Every physical step of the proteomics workflow in principle contributes to quantitative variations. The first and most severe cause is the sample preparation; here most deviations originate from pipetting errors, variations in multi-stage processing, enzymatic digestion, and chemical modifications (see **Sample preparation**). The second step of the workflow is the reversed-phase chromatography dimension of the LC-MS setup. Their distinct affinities to the reversed-phase material cause peptides to separate and elute along a gradient. Even minor differences in conditions may affect the chromatographic peak width of a peptide and therefore the number of acquire spectra and the observed peptide intensities. Similarly, electrospray sensitivity and ionization efficiency are subject to variations and individual peptide ionization may depend on the changing chemical composition of the analyte. Furthermore, mass spectrometers and their components may have a bias of transmission according to m/z, which further complicates correct quantification. These problems are the reason for the existence of a whole research field of mass spectrometry based quantification strategies (**Fig. 9**). MS- based quantification technologies can be divided into two strategies: label-free and labeling technologies. These again can be distinguished as relative and absolute quantifications.

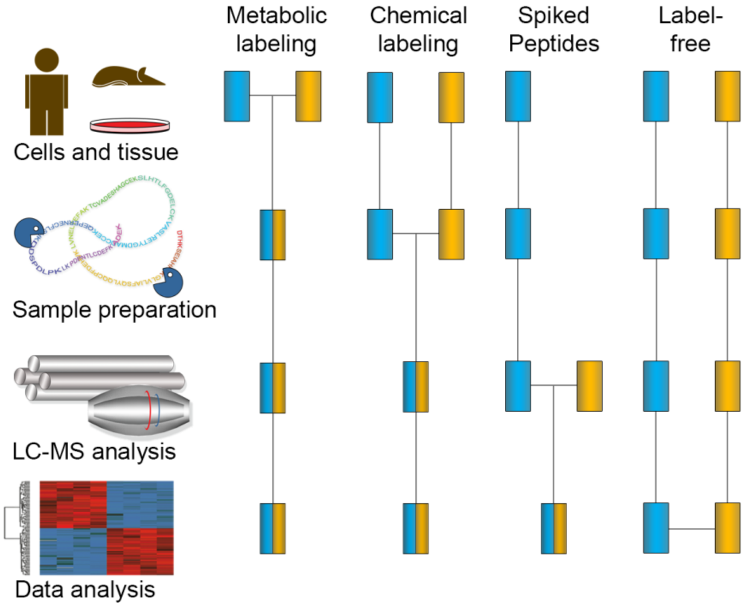


FIGURE 9 | Schemes of the most frequently applied quantification strategies. Metabolic labeling enables combination of samples at the earliest point. Samples in label-free quantification, in contrast, are processed separately and are combined only during bioinformatics analysis. (Adapted from Ong and Mann, 2005; Bantscheff et al. 2007 [90, 91])

METABOLIC STABLE ISOTOPE LABELING

Because of the large variations in sample handling and measurement, internal reference technologies are typically more accurate than label-free methods. Naturally occurring stable isotopes have nearly identical sample handling and ionization behaviors leading to the natural isotope patterns described previously (see **Scan modes applied in MS-based proteomics**). This property can be used to label and distinguish samples at the MS-level. Typically, sample labeling strategies use a combination of stable isotope atoms of hydrogen (D or ^{2}H), carbon (^{13}C), oxygen (^{18}O), or nitrogen (^{15}N) which naturally occur in organic analytes. These labels can be added to the sample or introduced in reference analytes by various approaches and during different stages of the sample handling steps (**Fig. 9**).

The earliest introduction of a stable isotope label is at the level of cell-growth – so called metabolic labeling. The most popular and cost-efficient form is stable-isotope labeling by amino-acids in cell culture (SILAC) [92]. Typically heavy isotope-labeled amino-acids are added to the growth-conditions of a cell culture to be incorporated into the entire proteome. Typically the amino acids that are sequence-specifically recognized during proteolytic digestion are used for labeling to ensure that at least one label is introduced in each peptide. In a quantification experiment heavy- and light- labeled samples are mixed even before sample preparation and are processed and measured together. The ratios between heavy- and light-labelled peptide isoforms reflect the quantitative differences of the peptides. This circumvents all processing artefacts and therefore guarantees the most accurate quantification [91, 93].

A variety of SILAC technologies have evolved and especially interaction and PTM studies with SILAC forward-/reverse-labeling have become popular. In this case an interaction sample and a negative control are labeled by heavy and light amino-acids, respectively, and mixed before sample processing and enrichments. In a second experiment the labeling order can be reverted. Therefore a simple plotting of the forward-/reverse-ratios allows accurate quantification, which distinguishes labelling effects, background-binders, and true interactors/regulators in a simple manner [93]. For large scale studies the so called super-SILAC strategy has become popular [94].

In a simple SILAC experiment only two conditions are compared, making a forward-/reverse-strategy ideal, but in the case of multiple conditions or time-courses a simpler strategy has been described. First a super-SILAC standard is designed that consists of a mixture of heavy-labeled reference proteins. The same super-SILAC standard is added to each of the experimental conditions or patient samples before sample preparation. The ratio-of-ratios in turn yield accurate quantitative information across the conditions. A further strategy is time-dependent labeling (pulsed-SILAC) that can be used to observe production and degradation of protein products [95]. Even though the benefits of SILAC outweigh its limitations in many cases a few caveats remain. One issue with SILAC, at least with current generation instruments, is the loss of dynamic range due to doubling of sample complexity compared to equivalent unlabeled samples. This is typically the reason why SILAC experiments result in fewer peptide and protein identifications. Another restriction is the need to grow cells under labeling conditions. Even though labeling of higher organisms has been achieved, it remains uneconomical in many cases. Additionally very large differences in heavy- and light-labeled samples can lead to a larger number of missing ratios. However, this problem has successfully been addressed by a hybrid label-free/label algorithm. Another limitation is restricted multiplexing of classical amino-acid labels.

CHEMICAL STABLE ISOTOPE LABELING

Chemical labeling strategies have been described for the protein or the peptide level but the latter is much more commonly applied. The classical chemical labeling strategy was termed isotope-coded affinity tag (ICAT). The initial ICAT chemicals consisted of a reactive group (halo-acetamide derivatives), an isotope-labeled linker, and an affinity tag. Only cysteine containing peptides were labeled by alkylation and therefore only these carried the attached affinity handle that was used for enrichment [96]. A more cost-efficient system is dimethyl labeling which can be highly accurate [97]. Typically, regular (CH_2O), median- (CD_2O), and heavy-labeled ($^{13}CD_2O$) formaldehyde is used to methylate primary amines of sample peptides. Lysines, especially, are labeled in a very fast chemical reaction. Labeled peptides are then combined and measured by LC-MS leading to peptide ratios similar to SILAC ratios. The sample complexity is also increased leading to less identification [98]. Minor drawbacks are slight retention-time shifts between deuterated and hydrogen isotope labels. These shifts can be computationally compensated to some degree.

The most elegant but also expensive chemical labeling is isobaric labeling. Especially two systems became popular for bottom-up proteomics: tandem mass tag (TMT) and isobaric tag for relative and absolute quantitation (iTRAQ) [99-102]. Isobaric labels consist of a reactive group (typically labeling primary- and secondary amines) and two isotope-labeled regions connected by a CID-cleavable linker. One region acts as reporter region while the other is used to counterbalance the mass so that various reporter/balance combinations lead to the same mass for the labeled precursor. Upon fragmentation, the linker region breaks and only the differently heavy reporter-region is seen in the low mass part of the spectrum. This tandem mass spectrum-based quantification in principle enables accurate quantification ratios in single scans; several fragment ions can carry a reporter region and therefore several ratios can be measured in a multiplexed way. Apart from their high costs, a major drawback of the reporter ion based strategies is ‘ratio compression’ due to the fact that co-eluting labeled peptides with similar mass are co-fragmented and contribute to the observed ratios.

LABEL-FREE QUANTIFICATION

While label-free quantification is in principle the most straightforward approach, isotope-labeling technologies are normally more accurate since they use internal standards to circumvent variations during sample processing. State of the art label-free approaches combined with high resolution measurements can correct for many of the sample processing-dependent differences with bioinformatic algorithms. Nevertheless, higher reproducibility of sample handling would be a major step towards accurate label-free quantification.

A very simple label-free quantification approach is the counting of MS2 spectra acquired for each protein. This spectral counting approach only provides a very indirect measure and strongly depends on the acquisition modes applied. A more direct approach is the use of raw signal intensities observed of each peptide belonging to the protein to be quantified. The area under the curve for MS spectra and along the elution profile can be integrated yielding more accurate mass and intensity read-outs, these calculated values can then be compared to the signal intensities of other measurements of the same peptide [103, 104]. The ion intensity-based approach gains from high mass resolution and from high density of MS scans because the profiles of the eluting peptides are then captured more accurately [105, 106]. High mass resolution is also necessary to successfully distinguish the extracted ion currents (XICs) of peptides, a prerequisite for accurate quantification. Innovative and complex algorithms have been developed to achieve higher accuracy in label-free quantifications. Most of these algorithms normalize the total intensities to achieve better comparisons between samples and runs.

ABSOLUTE QUANTIFICATION STRATEGIES

Absolute quantification of proteins is the most difficult but often also most desirable level of quantification. All approaches to determine absolute protein abundance are based on labeling or label-free quantification technologies. In case of spike-in methods using labeled standards, quantification proceeds as it does in relative quantification (ratios to the standard are determined). Therefore, the accuracy of the absolute determination is directly related to the underlying relative quantification. The earliest introduction of an absolute quantified standard is at the level of proteins. This is the case in a highly accurate approach that is based on protein epitope signature tag (PrEST) proteins. This SILAC-PrEST strategy makes use of a pre-existing expression library created for the immunization and generation of anti-bodies against the entire human proteome. These PrEST constructs contain of an affinity and solubility tag, which are used for purification, absolute quantification of the construct, respectively. Furthermore there is a 100 to 150 aa region that is sequence identical to a part of the target protein but with no homology to any other protein region and this region allows the absolute quantification of the endogenous target protein [107]. Other label-based strategies make use of artificial constructs such as QCAT/QconCAT (quantitative concatenated protein) [108, 109] or AQUA (absolute quantification peptides) in which labeled, synthetic peptides constitute the standard [110].

Label-free approaches to estimate protein copy numbers mostly rely on normalization algorithms that correlate a mass spectrometry quantity to the amount of the protein. A first attempt to normalize protein abundance was the protein abundance index (PAI) [111]. The PAI is calculated by dividing the number of observed peptides per protein by the number of theoretically observable peptides. This approach was refined into the exponentially modified PAI (emPAI) derived by calculating 10^{PAI}-1, which correlates roughly with absolute protein amount [112]. A more complex algorithm normalizes the average MS signal of the top three intense peptides of a protein by a signal response factor (Top3 method) [113]. Another variation to these themes was introduced in 2011. It consists of dividing the sum of intensities of observed peptides of a protein by the number of theoretically observable peptides (intensity based absolute quantification, iBAQ) resulting in a proxy of protein abundance [114]. A straightforward estimation for protein copy number estimations was introduced by Wisniewski et al and termed total protein approach (TPA) [115]. Here the percentage of protein quantity is first calculated by dividing the label-free intensity of each protein by the summed intensities of all proteins. In the next step the absolute quantity is calculated by dividing the percent quantity by the molecular weight (g/Mol) of the protein, the Avogadro constant (Mol) and by the total protein content per cell. This approach is only applicable to deep proteomic datasets but has shown promise for the estimation of copy numbers in a variety of analytical situations.

COMPLETE PROTEOMICS

The greatest challenge of MS-based proteomics is the quantitative and comprehensive analysis of the entire protein content of tissues, cells, or even entire organisms. While some groups refer to complete proteomes as the measurement of every protein coding gene or every isoform of every protein coding gene, we mean by this the measurement of at least one form of an expressed protein coding gene (**Article 2**). Typically cells and organisms express thousands of proteins in a given state with a very wide range of protein copy numbers; this generates extremely complex samples with a very large dynamic range.

The comparatively low-complex eukaryote *S. cerevisiae* expresses around 4,500 different proteins at a given time while the human cancer cell-line HeLa appears to contain approximately 12,000 different proteins [116]. The complexity alone makes the analytical task seem overwhelming, but when the dynamic range challenge is added it appears nearly impossible. Protein copy numbers range from approximately 10 to $1x10^6$ and about ~10 to $5x10^7$ in *S. cerevisiae* and HeLa, respectively (**Article 3**). The estimated dynamic range of the proteins of blood plasma is even larger and spans more than ten orders of magnitude [117]. A sample prepared for LC-MS analysis is more complex still since proteins are digested into multiple peptides; the peptides may carry various charges, be *in vivo* or *in vitro* modified, and ionize with different efficiencies further increasing complexity and dynamic range. These factors typically lead to hundreds of thousands of peptide species in a complete proteomic sample.

The task of complete proteomic coverage necessitates the best performance of every area of MS-based proteomics: proper sample preparation can reduce the overall sample complexity, high resolution liquid chromatography can help with the complexity and dynamic range during the measurement, and faster, more sensitive, and accurate mass spectrometers can improve the identification rates and the total number of peptide identifications. Recent advances in all these research fields have now resulted in impressive proteomic coverage. Likewise the time required for a comprehensive measurement has been reduced tremendously over the last few years (**Article 1**).

PRE-FRACTIONATION TECHNIQUES FOR COMPLETE PROTEOMICS

The aim of complete proteomic sample preparation is to provide peptides of all proteins for the LC-MS/MS analysis. The peptides should be provided without any suppression of protein classes or properties. For this reason classical protocols employ protein solubilization and membrane spanning proteins extraction by strong detergents such as SDS [30, 46, 118]. The SDS is removed before digestion and the resulting peptides represent a picture of the complete proteome. But this also leads to very complex mixtures which span a high dynamic range. To reach deeper proteomic coverage extensive sample fractionation techniques were employed before LC-MS/MS analyses which intend to reduce the complexity of the sample.

Typically complex samples are pre-fractionated at the protein or peptide level, preferably in an orthogonal dimension to the final LC measurement; peptide fractionation techniques are usually better suited because even insoluble proteins may provide soluble peptides. For protein based fractionation, strong detergents are required and SDS-PAGE with in-gel digestion remains the best option for complete proteomics [29]. During the in-gel protocol single bands from the SDS-gel are cut out and proteins are extracted in a lengthy procedure. Some problems of this method are sample loss of proteins during the extraction which leads to uncontrolled quantification errors and the general workload for the procedure [119]. It can be said that SDS-gel based techniques can be useful for samples with few highly abundant proteins such as muscle fibers where the abundant protein can be removed entirely from the analysis.

A much simpler technique is the OFFGEL fractionation where either proteins but preferably peptides are fractionated in an electric field on top of an immobilized pH gel [29, 120]. The solubilized peptides settle in the well corresponding their pI and can be readily further processed, either by protein digestion or peptide clean-up. The method has previously demonstrated to be highly efficient in terms of fractionation efficiency but large sample quantities have to be used due to severe losses. A single OFFGEL procedure typically takes around two days and large quantities of ampholytes need to be removed prior LC-MS/MS analysis which can be difficult depending on the nature of the sample. However, the first study to accomplish a complete coverage of the yeast proteome successfully applied the OFFGEL method [29].

Alternatively, affinity-based chromatographic pre-fractionation techniques demonstrate a high potential for peptide fractionation but often reach lower fractionation efficiencies. Generally cation-exchange, anion-exchange, and reversed-phase fractionations are applied before the LC-MS/MS analysis [44, 121-123]. These dimensions can be coupled directly to the LC-MS/MS or can be performed independently before the measurements. Methods such as the multidimensional protein identification technology (MudPIT) are highly efficient and nearly lossless and have demonstrated high throughput and very good coverage [124]. However, on-line fractionation approaches necessitate dedicated LC set-ups with additional high-pressure pumps and special buffer systems. Furthermore the applied buffers for the first dimension need to be compatible with the final dimension or the samples need to be desalted on-line, increasing the complexity of the LC systems.

Off-line fractionation techniques are typically simpler and solid-phase extraction (SPE) based methods are easily applicable without dedicated machinery. Cost-efficient solutions are StageTip-based fractionation techniques [43, 44]. StageTips are typically used for sample-cleanup and desalting after digestion. Wisniewski et al demonstrated very deep proteomic coverage using a FASP-SAX (strong anion exchange) combination [121]: lysis is performed in SDS and the samples prepared on a molecular-weight cut-off membrane (FASP), the resulting peptides are fractionated by SAX-based StageTips and the eluted peptides are desalted on C_{18}-based StageTips. The procedure proved to be very cost efficient and is compatible with few micrograms starting material.

NANO-UHPLC WITH SUB-2-MICRON PARTICLES FOR DEEP COVERAGE

Sample pre-fractionation techniques can be used as a first dimension of separation but the final liquid chromatography step almost always has the highest resolution power and is essential to reach deep coverage. Liquid chromatography is performed on line with the mass spectrometry measurements and therefore has to be optimized in conjunction with the mass spectrometer. The chromatographic resolution and MS scan speed need to be matched to each other; a low resolution chromatography system with a very high acquisition rates typically lead to low coverage even though the mass spectrometer capable of better performance. It is therefore necessary to couple a chromatographic system with high resolution to mass spectrometers with high scan speeds to reach deep proteomic coverage.

The resolution of a chromatographic system can be described by its plate number (**Box 1**). The plate number is directly related to the length of the packed bed and the plate height which in turn, is directly related to the particle size of the chromatographic resin. For high resolution chromatography very long columns with small particle size have become more mainstream than they were before. These developments also result in higher backpressures and ultra-high pressure liquid-chromatography (UHPLC) systems have now become available [45, 125, 126]. Recent publications demonstrate that novel UHPLC systems together with high sequencing speeds are now capable of identifying the near complete yeast proteomes in single measurements without pre-fractionation (**Article 1**; **Article 3**) [71]. The concept of single-shot measurements is a tremendous advance towards higher reproducibility, better quantification, and higher throughput.

STATE-OF-THE-ART MASS SPECTROMETERS

The sequencing of thousands of peptide species in a short time necessitates high acquisition rates yielding high quality spectra. The introduction of the bench-top Q Exactive instruments was a hallmark for very high acquisition rates and the most recent Orbitrap Fusion Tribrid further increased the scan rates. Higher scan rates typically necessitate higher sensitivity or need to come with better identification rates. This became very obvious with the high-high strategy employed on the new Orbitrap instruments; even though fewer scans can be performed per time on the Orbitrap in contrast to a LIT analyzer, the overall identification rates are much higher leading to larger numbers of peptide identification per time [57]. High resolution scans with efficient fragmentation can lead to high identification rates but sufficient ion quantities need to be available if higher scan rates are intended.

The sensitivity of the detector is as important as the sensitivity of the overall ion path; with the simplified and shorted ion path of the Q Exactive instruments a higher sensitivity was achieved but the time necessary to collect sufficient numbers remains a limiting factor [127]. Recently a new member of the Q Exactive family – the Q Exactive Plus – was announced featuring an upgraded bent flatapole and a better selection quadrupole compared to the predecessor (http://planetorbitrap.com/q-exactive-plus). These improvements will improve the general throughput and quality of MS^2 spectra.

Contributing to the overall sensitivity is the efficiency of the electrospray ionization. A recent publication from the Kuster laboratory demonstrated improved ESI sensitivity when dimethylsulfoxide (DMSO) is added to the LC buffer conditions [55]. While the exact working principle remains unclear it was theorized that DMSO reduces the surface tension and increases the Rayleigh limit of the sprayed ion droplets; this again leads to a faster and more efficient ion formation.

S. CEREVISIAE PROTEOMICS

The behavior and phenotype of eukaryotic cells is often represented by the classic model organism *S. cerevisiae*. Similar to the phenotype are proteomic measurements of *S. cerevisiae* a predictor for more complex eukaryotes; the advances in the analysis of the yeast proteomes can be directly applied and translated to the measurements of higher eukaryotes. Like genome- and transcriptome-wide sequencing approaches, recent yeast proteome measurements are a great resource for the community of basic biological research. Because proteomics is the observation of actual protein products, actual expression of proteins and their function can be studied. Amongst the many identifications of characterized proteins are large numbers of uncharacterized proteins. These may offer new opportunities to explore protein functions and may still hold a vast amount of unknown cellular mechanisms.

One rather special protein group of the *S. cerevisiae* UniProt database are proteins annotated to be “dubious”; these proteins are uncertain to be expressed. Even though the exact number varies over time approximately 10% of the S. cerevisiae databases (657 of 6,630 entries of release 12.2013) are dubious coding sequences (CDS). This number gives an additional estimation of the actual protein FDR within a measured dataset [29]. Measurements of 4,000 protein identifications with 1% FDR cutoff should therefore contain no more than 4 protein entries annotated as dubious CDS. While these values cannot be used to absolutely quantify the FDR, they give a hint of the quality of acquired data. Especially early complete proteomic studies based on tagging approaches demonstrated high rates of false positives in contrast to LC-MS/MS based measurements (TAP-tagging: 26 of 4,251; GFP-tagging: 23 of 4,154; LC-MS/MS: 3 of 4,399 protein identifications) [13, 29].

According to LC-MS/MS based strategies, targeted as well as discovery-based methods have been applied to demonstrate the applicable proteomic coverage. For targeted proteomics the Aebersold group synthesized 28,000 peptides for a database representing the entire yeast proteome and observed 97% of these synthetic peptides. In the measurements of the biological samples only 2,509 proteins could be observed [128]. In contrast to the targeted approaches similar efforts were performed by the Mann lab and with elaborate fractionation techniques and many measurements 4,399 protein groups could be observed in discovery-based experiments [29]. Even deeper coverage of wild-type yeast was achieved with a similar effort and employing a large variety of proteases by the Heck group reporting 4,401 protein identifications [123]. These approaches were highly time consuming and not feasible in routine application. Hence there was a high interest and effort to simplify the technology to achieve similar proteomic depth with less effort.

Thakur et al demonstrated the feasibility to achieve deep proteomic coverage in single-shot measurements of 8-h reaching 2,990 proteins [46]. With a new generation of LC systems, novel high-resolution bench-top mass spectrometers with very high sequencing speeds we could achieve the remarkable depth of more than 4,000 protein groups in single 4-h measurements reaching nearly the complete yeast proteome (approximately 90% of the expected expressed yeast proteome) (**Article 1**). A further beneficial alteration to achieve such a proteomic depth was the use of the endopeptidase Lys-C instead of trypsin, reducing the complexity and thereby facilitating proteomic coverage. Recently a near complete proteomic coverage of median 3,977 protein identifications was reported applying 70 min measurements on the novel Orbitrap Fusion Tribrid instrument (total 4,395 protein identifications; 16 dubious identifications) [71]. These new technological breakthroughs make proteomics a competitive technology for well-established genomics and transcriptomics systems in terms of coverage. Quantification of proteins is a further aspect of technological advances.

2 PUBLICATIONS

ARTICLE 1: A NOVEL PLATFORM FOR SINGLE-SHOT LC-MS MEASUREMENTS

SYSTEM-WIDE PERTURBATION ANALYSIS WITH NEARLY COMPLETE COVERAGE OF THE YEAST PROTEOME BY SINGLE-SHOT ULTRA HPLC RUNS ON A BENCH TOP ORBITRAP

AUTHORS: NAGARJUNA NAGARAJ[‡§], NILS A. KULAK[‡§], JUERGEN COX[‡], NADIN NEUHAUSER[‡], KORBINIAN MAYR[‡], OLE HOERNING[¶], OLE VORM[¶], AND MATTHIAS MANN[‡]

§ *Equal contribution*
‡ *Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry*
¶ *Thermo Fisher Scientific, Edisonsvej 4, DK-5000 Odense C, Denmark*

PROLOGUE:

Complete proteomic measurement of eukaryotic systems is a highly challenging task. The complexity and dynamic range of the proteome span many orders of magnitude. This is especially pronounced when analyzing peptides resulting from proteolytic digestion of complex protein samples. Initial attempts to observe peptide species of every expressed protein have relied on extensive pre-fractionation technologies that are supposed to reduce sample complexity during LC-MS measurements. A first successful complete eukaryotic proteome – of *S. cerevisiae* – was achieved in 2008 applying various fraction technologies. Work by Thakur et al. in 2011 demonstrated that even highly complex mixtures could potentially be analyzed in single LC MS measurements while retaining adequate proteomic depth. Subsequently, the introduction of a high-performance bench top Orbitrap mass spectrometer (Q Exactive) promised even higher identification rates per time.

We reasoned that an optimized chromatographic setup would capture the potential of the newly introduced MS instrument, but that a better LC system was also needed. A recently developed nano-UHPLC system (Proxeon Easy nLC-1000), capable of delivering a pressure up to 1000 Bar, was therefore supplied with a column setup to realize very high chromatography resolution along extended gradient lengths. Optimized self-packed 50cm long and 75µm I.D. pulled emitter tip columns with 1.8µm C_{18} bead material promised to be a highly efficient combination with the Q Exactive. Additionally we observed that in yeast higher proteomic coverage could be obtained when Lys-C digestion (rather than trypsin digestion) was combination with FASP. Integrating all these improvements into a single platform proved highly successful. The tip based sample preparation, the high-resolution chromatography system, and the high-performance mass spectrometer turns out to be a very efficient combination and enabled a near complete coverage of the yeast proteome with 4h measurements.

We demonstrated the power of the system using a biological perturbation of regular growth in yeast. For accurate quantification we employed the spike-in super-SILAC approach in the context of heat stress. This allowed system-wide proteome quantification, which was rigorously statistically evaluated for stress response related changes. Most of the significant changes were expected, providing a positive control, however, we also observed down-regulation of protein expression pathways on various regulatory levels, which provided interesting insights into yeast biology.

⌘ Author's Choice

System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap*Ⓢ

Nagarjuna Nagaraj‡§, Nils Alexander Kulak‡§, Juergen Cox‡, Nadin Neuhauser‡, Korbinian Mayr‡, Ole Hoerning¶, Ole Vorm¶, and Matthias Mann‡‖

Yeast remains an important model for systems biology and for evaluating proteomics strategies. In-depth shotgun proteomics studies have reached nearly comprehensive coverage, and rapid, targeted approaches have been developed for this organism. Recently, we demonstrated that single LC-MS/MS analysis using long columns and gradients coupled to a linear ion trap Orbitrap instrument had an unexpectedly large dynamic range of protein identification (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, 10.1074/mcp.M110.003699). Here we couple an ultra high pressure liquid chromatography system to a novel bench top Orbitrap mass spectrometer (Q Exactive) with the goal of nearly complete, rapid, and robust analysis of the yeast proteome. Single runs of filter-aided sample preparation (FASP)-prepared and LysC-digested yeast cell lysates identified an average of 3923 proteins. Combined analysis of six single runs improved these values to more than 4000 identified proteins/run, close to the total number of proteins expressed under standard conditions, with median sequence coverage of 23%. Because of the absence of fractionation steps, only minuscule amounts of sample are required. Thus the yeast model proteome can now largely be covered within a few hours of measurement time and at high sensitivity. Median coverage of proteins in Kyoto Encyclopedia of Genes and Genomes pathways with at least 10 members was 88%, and pathways not covered were not expected to be active under the conditions used. To study perturbations of the yeast proteome, we developed an external, heavy lysine-labeled SILAC yeast standard representing different proteome states. This spike-in standard was employed to measure the heat shock response of the yeast proteome. Bioinformatic analysis of the heat shock response revealed that translation-related functions were down-regulated prominently, including nucleolar processes. Conversely, stress-related pathways were up-regulated. The proteomic technology described here is straightforward, rapid, and robust, potentially enabling widespread use in the yeast and other biological research communities. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.013722, 1–11, 2012.***

Yeast is one of the most well established model systems in molecular biology. It is used to study a large range of conserved cellular processes, including the cell cycle, metabolism, and stress responses. Yeast was the first organism whose genome was sequenced completely (1), and many other systems-wide biology screens were first carried out in the yeast model (2–6). Large scale proteomics has also been pioneered in yeast, identifying first hundreds and then thousands of proteins (7–13). Using three different analytical strategies, including one with subcellular fractionation and two involving peptide separation into 24 fractions, our group has reported a substantially complete proteome of yeast as judged against genome-wide tagging experiments (14). However, the expertise and analysis times associated with in-depth proteome measurements have so far precluded the widespread adoption of in-depth proteomics in the yeast research community. Targeted proteomics, in the form of multiple reaction monitoring, offers a possible solution to this problem and has recently been used to detect proteins throughout the dynamic range of the yeast proteome, as well as to quantify changes in key proteins after metabolic shift (15). However, targeted proteomics aims at the characterization of relatively few key proteins across many conditions, and it is therefore less well suited to the discovery of biological responses on a global scale.

Both the multiple reaction monitoring experiments and analyses of the total features detectable in the MS retention time contour plots suggest that a very large number of peptides are

From the ‡Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany and ¶Thermo Fisher Scientific, Edisonsvej 4, DK-5000 Odense C, Denmark

⌘ *Author's Choice*—Final version full access.
Received August 19, 2011, and in revised form, October 19, 2011
Published, MCP Papers in Press, October 20, 2011, DOI 10.1074/mcp.M111.013722

present in LC-MS runs of total proteome digests (16, 17). We recently investigated the dynamic range of single LC-MS/MS runs and found that even very low-abundance proteins could be detected in this mode (18). Furthermore, direct analysis without prefractionation implies high sensitivity because only a few micrograms of peptides are required to load the column to capacity. However, our previous study was performed with a dedicated chromatographic setup and would not be straightforward to adopt for nonspecialized groups.

A novel mass spectrometer, the Q Exactive, couples a mass selective quadrupole to the Orbitrap analyzer (19). In this bench top instrument, precursor ions are selected by the quadrupole, fragmented by higher energy collisional dissociation (20), and measured at high resolution and mass accuracy in the Orbitrap analyzer. Cycle times for a top10 method (survey scan followed by up to 10 MS/MS scans) are ~1 s, more than twice as fast as with previous instruments of the Orbitrap family. Thus the Q Exactive offers the potential to analyze many more peptides in a given time, with very high MS/MS data quality. We wanted to combine these benefits with ultra HPLC (UHPLC),[1] which was not available to us in the previous single-run analyses. Taking advantage of a newly developed compact UHPLC system termed the EASY-nLC 1000, we achieved higher chromatographic performance with relatively long columns and small particle diameters. Here, we describe this simple but powerful bench top platform and evaluate its capability to characterize the yeast proteome in high throughput but also in-depth fashion.

To quantify proteome states in yeast, SILAC labeling can be employed in the standard format, which requires labeling both the control and the experimental conditions (21). To enable even more streamlined systems analysis of perturbations of the yeast proteome, we further wanted to decouple the SILAC metabolic labeling step from the actual experiments by using a "spike-in" SILAC strategy (22). Here we developed such a standard, taking into account several proteome states of yeast. We then used this standard to quantify yeast proteome changes upon heat shock, an important perturbation frequently encountered with temperature-sensitive mutant strains and synchronization experiments (23, 24).

EXPERIMENTAL PROCEDURES

Yeast Culture and Lysis—The yeast strain W303 MATα was grown in YPD medium until early- to mid-log phase and was harvested by centrifugation at 4000 × *g* for 5 min at 4 °C. The cell pellet was resuspended in 100 mM Tris, pH 7.6, containing 100 mM dithiothreitol and 5% SDS. The lysates were heated to 95 °C for 5 min followed by sonication using a Bioruptor Sonicator (20 kHz, 320 W, 60 s cycles) for 15 min at the maximum power to achieve complete lysis. The lysate was centrifuged at 16,000 × *g* for 5 min to clarify the protein extract.

Yeast Spike-in Standard—The W303 MATα strain for heavy lysine labeling was constructed by deletion of the Lys2 gene using the pYM-*natNT2* plasmid according to Janke *et al.* (25). The cells were labeled only with heavy lysine, and not heavy arginine, to reduce sample complexity and avoid arginine to proline conversion. The spike-in standard was used to compare expression levels across different conditions. We cultured 250 ml to log phase (A_{600} = 0.9) in SCD medium containing [$^{13}C_6/^{15}N_2$]L-lysine. To represent further biological conditions in the spike-in mix, we also cultured cells with 2% ethanol as the carbon source as well as at higher temperature (37 °C for 30 min after previous culture at 24 °C). These three conditions were mixed in equal proportions to produce the spike-in mix. This quantity of cultured cells would be sufficient for thousands of spike-in experiments in single-shot measurements (at a few μg/analysis) and hundreds of experiments with an up front pipette-based strong anion exchange fractionation step (26).

Yeast Heat Shock Treatment—Yeast was cultured to mid-log phase to obtain an A_{600} of 2.5 for cells at 24 °C in the YPD medium and was subsequently shifted to 37 °C via water bath incubation to achieve uniform and efficient heat transfer. Samples were collected at t = 0 and 30 min after incubation at 37 °C to analyze the proteome changes upon heat shock. The samples were lysed as described above.

Protein Digestion—Proteins were digested using the FASP method (27). Briefly, 140 μg of protein was loaded on the filter, and SDS was completely replaced by washing two to three times with buffer containing 8 M urea. The proteins were then alkylated using iodoacetamide, and the excess reagent was washed through the filters. The reduced and alkylated proteins were digested using endoproteinase LysC, which cleaves at the C terminus of lysine residues, with an enzyme to protein ratio of 1:50. Peptides obtained by FASP were desalted using C_{18} StageTips (28).

Ultra High Pressure Easy LC—The Thermo Scientific EASY-nLC 1000 (Thermo Fisher Scientific, Odense, Denmark) is a split-free, nano-flow LC designed to operate at ultra high pressures up to 1000 bars (15,000 p.s.i.). The system employs two direct-drive syringe pumps to generate binary gradients with minimum stable flow down to ~50 nL/min. Flow and pressure sensors (one set for each mobile phase) are placed immediately upstream from the high pressure mixing Tee such that sensor output can accurately control the gradient. The LC system is preconfigured, requiring only two liquid connections by which the user connects the column(s) to the eluent flow line and a waste/venting line. This simplicity facilitates daily use, and further ease-of use is obtained by a finger-tight fitting, named NanoViper (Thermo Fisher Scientific), that ensures zero dead volume seals up to 1200 bars. This compact LC instrument, with its maximum pressure limit of 1000 bars, enables the use of long columns with linear velocity of 250 nl/min in the temperature range of 35 °C, rather than the relatively high temperatures of up to 60 °C required in our previous setup without ultra high pressure (18).

LC-MS/MS—Peptides were loaded on a 50-cm column with 75-μm inner diameter, packed in-house with 1.8-μm C_{18} particles (Dr Maisch GmbH, Germany). Reversed phase chromatography was performed using the Thermo EASY-nLC 1000 with a binary buffer system consisting of 0.5% acetic acid (buffer A) and 80% acetonitrile in 0.5% acetic acid (buffer B). The peptides were separated by a linear gradient of buffer B up to 40% in 240 min for a 4-h gradient run with a flow rate of 250 nl/min in the EASY-nLC 1000 system. The column was operated at a constant temperature of 35 °C regulated by an in-house designed oven with a Peltier element (18). The LC was coupled to a Q Exactive mass spectrometer (19) (Thermo Fisher Scientific) via the nanoelectrospray source (Proxeon Biosystems, now Thermo Fisher Scientific). The Q Exactive was operated in the data-dependent mode with survey scans acquired at a resolution of 50,000

[1] The abbreviations used are: UHPLC, ultra HPLC; SILAC, stable isotope labeling of amino acids in cell culture; FASP, filter-aided sample preparation; GO, gene ontology.

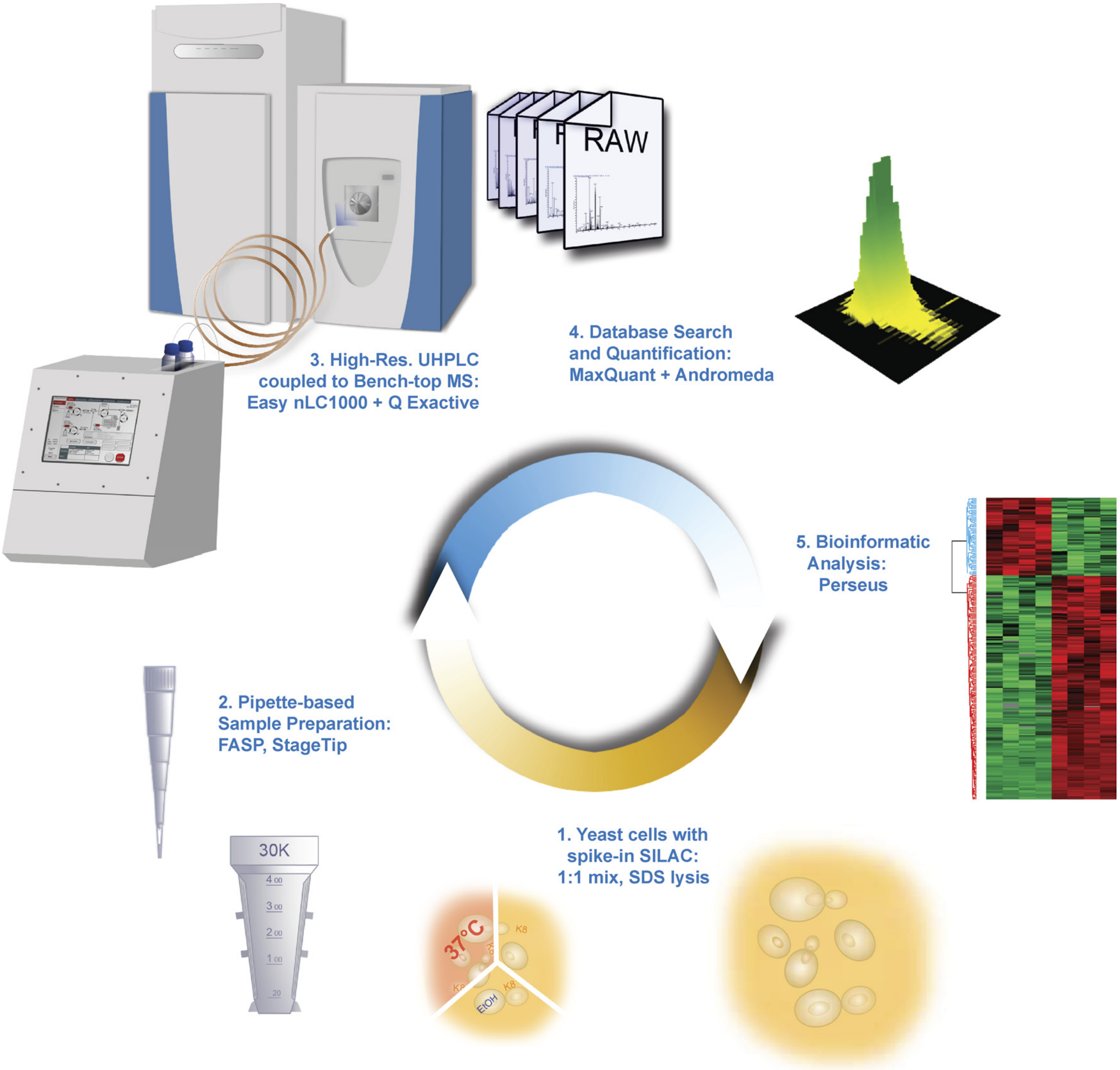

RAW
3. High-Res. UHPLC coupled to Bench-top MS: Easy nLC1000 + Q Exactive
4. Database Search and Quantification: MaxQuant + Andromeda
5. Bioinformatic Analysis: Perseus
2. Pipette-based Sample Preparation: FASP, StageTip
30K
1. Yeast cells with spike-in SILAC: 1:1 mix, SDS lysis
37°C
K8
EtOH

FIG. 1. **Minimalistic proteomics setup.** Yeast samples were lysed and prepared by the FASP method. Peptides were purified on StageTips and placed in an autosampler, which loads them directly on to a relatively long column (50 cm). The binary gradient system is provided by an UHPLC system (EASY nLC 1000) system coupled to a bench top quadrupole Orbitrap mass spectrometer (Q Exactive) via a nanoelectrospray source. The data obtained were analyzed in the MaxQuant computational proteomics platform, and bioinformatics analyses were performed using the Perseus tool.

at *m/z* 400 (transient time = 256 ms). Up to the top 10 most abundant isotope patterns with charge ≥2 from the survey scan were selected with an isolation window of 1.6 Thomsons and fragmented by higher energy collisional dissociation (20) with normalized collision energies of 25. The maximum ion injection times for the survey scan and the MS/MS scans were 20 and 60 ms, respectively, and the ion target value for both scan modes were set to 1E6. Repeat sequencing of peptides was kept to a minimum by dynamic exclusion of the sequenced peptides for 40 s.

Data Analysis—The raw files were processed using the MaxQuant computational proteomics platform (29) version 1.2.0.34. The fragmentation spectra were searched against the yeast ORF database (release date of February 3, 2011; 6752 entries) using the Andromeda search engine (30) with the initial precursor and fragment mass tolerances set to 7 and 20 ppm, respectively, and with up to two missed cleavages. Carabamidomethlyation of cysteine was set as a fixed modification, and oxidation of methionine and protein N-terminal acetylation were chosen as variable modifications for database searching. Both peptide and protein identifications were filtered at 1% false discovery rate and thus were not dependent on the peptide score. Bioinformatics analysis was performed using the Perseus tools available in the MaxQuant environment. All enrichment analysis and analysis of variance tests were performed with Benjamini-Hochberg correction at a false discovery rate of 2%. The raw data are available from the Tranche proteome repository with the following access code: Bz9hlKJ5EaEq/rgoVH0+fHehRgTSaCcD2 + 879Q1JnJm3d9sFaCpNgFnPPZT9WFu5K5mXKz8o1B9qaK7WBFxdFPu2ThkAAAAAAAAPmA = =.

RESULTS AND DISCUSSION

The Single-shot LC-MS/MS System—We aimed to devise a shotgun proteomics workflow with the lowest possible number of processing and analysis steps and consequently high

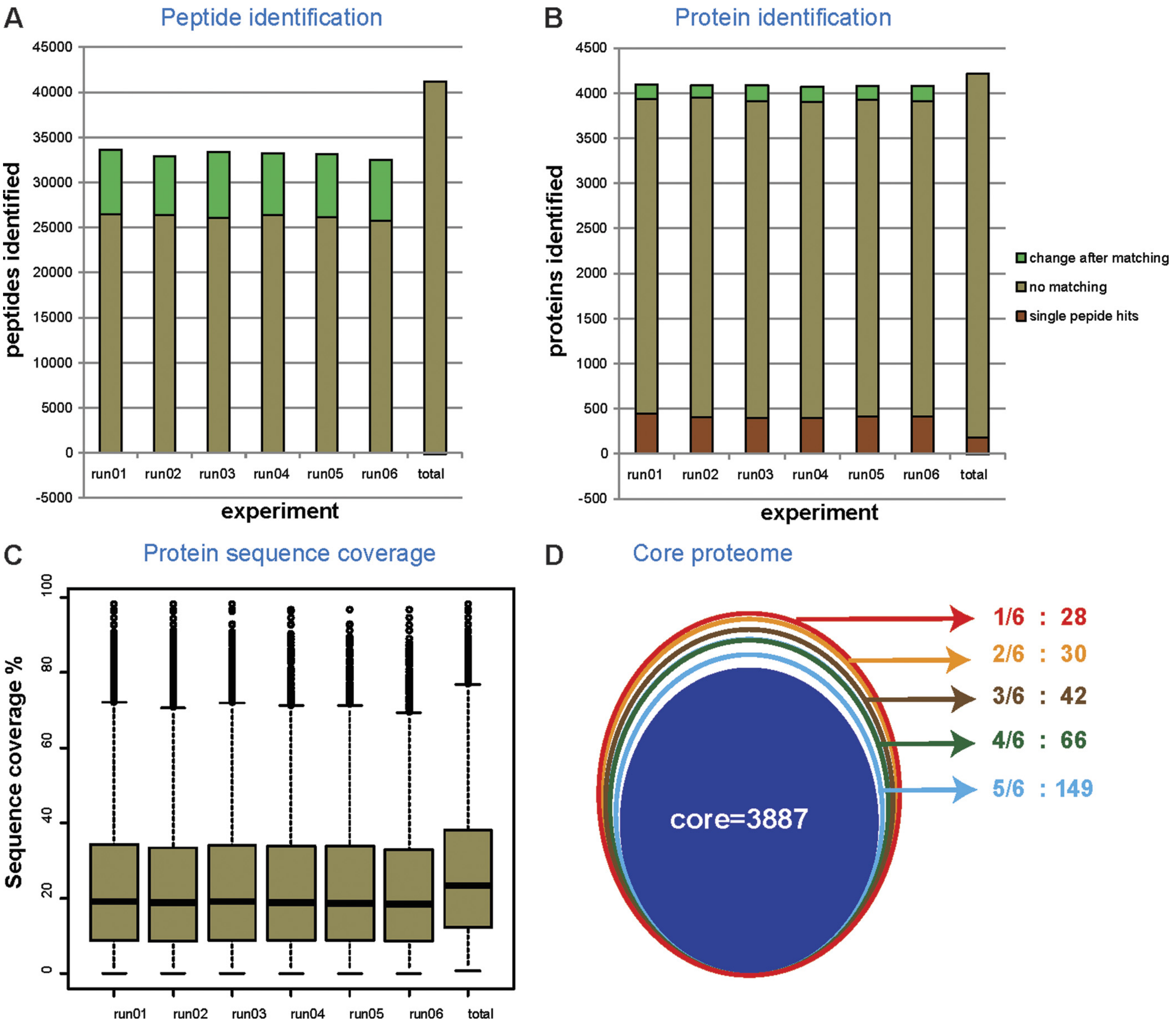


FIG. 2. **In-depth coverage of the yeast proteome.** *A*, number of peptides identified in individual runs with and without matching between the runs. Peptides identified by matching are indicated in *green*. *B*, proteins identified in individual runs with the gain from matching between the runs indicated in *green*. Proteins identified with single peptide hits are shown in *red*. *C*, the median sequence coverage of individual runs after matching was ~17%. The median sequence coverage from the combined run for 4099 proteins was 22.9% as shown. *D*, the *conjoint circles* represent the frequency of identification of proteins in the six runs. Proteins identified in all six runs were designated as core proteome in the innermost circle.

robustness (Fig. 1). Yeast cells were lysed in the presence of SDS, ensuring efficient denaturation and solubilization of all protein classes ("Experimental Procedures"). The proteins were reduced to peptides by LysC digestion using the FASP method (27), and the resulting peptides were purified on StageTips (28). These procedures only involve pipette-based operations, and they can be performed in several hours and in parallel for several conditions. Peptide mixtures were then loaded onto the autosampler of the UHPLC system (EASY-nLC 1000) and analyzed in an automated manner by LC-MS/MS on the bench top quadrupole Orbitrap mass spectrometer (Q Exactive) (19). The LC setup does not use precolumns or flow splitting, avoiding sample loss and reducing solvent consumption. The UHPLC system itself is designed for compactness and simplicity ("Experimental Procedures").

To facilitate deep sampling of the proteome, we employed relatively long columns and small particle sizes (50 cm, 1.8 μm). This was readily accommodated by the UHPLC pump, which produced a stable flow of 250 nL/min at 500 bars. Another advantage of the UHPLC system is its ability to load samples at a higher flow rate and to equilibrate columns more

TABLE I

Coverage of Saccharomyces Genome Database annotations and GO biological process terms that are de-enriched

	Yeast ORFs	Proteins identified	Proteins identified − core proteome
Total	6717	4206 (63%)	3887 (58%)
Saccharomyces Genome Database			
Verified	4941	3856 (87%)	3587 (73%)
Uncharacterized	857	335 (39%)	287 (33%)
Dubious	809	2 (0%)	2 (0%)
Transposable Elements	89	17 (19%)	13 (15%)
Pseudogenes	21	1 (5%)	1 (5%)
Silenced	4	1 (25%)	1 (5%)
GO biological process			
Maltose metabolic process	11	1 (9%)	0 (0%)
Synapsis	10	2 (18%)	2 (18%)
Multidrug transport	11	0 (0%)	0 (0%)

quickly, leading to a shortening of overhead times. We found the combination of a 50-cm column and 4-h gradients to be a good combination for standard use.

Depth of Analysis of the Yeast Proteome—Having established the single-shot workflow, we next measured six yeast cell lysates, which simulates an experiment with triplicate control and triplicate perturbation. Approximately 4 μg of peptide material was loaded onto the 50-cm column and separated with the 4-h gradients. Joint analysis of the six LC-MS/MS files in MaxQuant resulted in an average of 26,173 ± 286 peptide identifications with unique amino acid sequence for the single runs. Transferring identifications between the runs based on their mass precision and retention time ("match between runs" feature in MaxQuant) led to 33,122 ± 405 sequence-unique peptide identifications per single run (Fig. 2*A*). Together, 41,035 peptides were identified from this experiment, which took ~24 h of total measurement time. Even though LysC peptides are on average larger than tryptic peptides and therefore more difficult to identify, the identification rates for runs were above 51%. This is presumably due to the high mass accuracy enabled by the high resolution higher energy collisional dissociation spectra.

When matching between the runs, 4084 ± 8 proteins were identified per run. In the combined data set, 4206 proteins were identified (not counting contaminants such as keratins), and only 180 of these had a single peptide (Fig. 2*B* and supplemental Tables I and II and other supplemental materials containing the spectra of all the proteins identified with single peptides). We repeated the database search with an arbitrary Andromeda peptide score threshold of 60, which is high for a database with the size of the yeast proteome, and still identified 4137 proteins. This further demonstrates that our data do not rely on low scoring peptides. Our previous study using 8-h gradients, a custom LC setup, and the previous generation Orbitrap instrument identified just under 3000 proteins in a triplicate experiment (18). Here we achieved dramatically increased performance—close to the complete expressed proteome (see below)—with a very streamlined and minimalistic proteomic system.

Median sequence coverage of identified proteins was 23.4% with a median of seven peptide sequences (Fig. 2*C*). Many more peptides can be detected in LC MS plots than are sequenced and identified by tandem mass spectrometry. In our data set, the median intensity of the fragmented isotope patterns was ~10-fold higher than that of the nonfragmented isotope patterns (supplemental Fig. 1). This suggests that many more yeast peptides are present in the single-runs than are fragmented and identified, although they may not be accessible to data-driven LC-MS/MS (19).

A key challenge in shotgun proteomics is the "missing value" problem, which refers to the absence of data on particular proteins or peptides in some of the measurements of a series and which is caused by the semi-random nature of peak selection for fragmentation. Remarkably, when comparing identifications in different subsets of the single-shot analyses, we found that a full 3887 of the 4206 proteins (92%) were identified in all six runs (termed *core* in Fig. 2*D*), and 96% were identified in at least five of the six data sets. This indicates that for the vast majority of the proteins, there is no or very little "missing value problem." At the peptide level, naturally, overlap is not as high, but 75% of the peptides are still identified in at least five of the six runs (supplemental Fig. 2). High reproducibility between the single runs is presumably a consequence of the very high sequencing speed of the Q Exactive, combined with the efficient matching of peptides between runs by MaxQuant.

To assess the completeness of our data set, we compared it against our previous in-depth study (14). Despite differences in the yeast background (W303 *versus* S288C), somewhat different conditions and slight reannotation of the yeast genome in the past 4 years, 95% of the 4206 genes found here were contained in our previous data set. Of the 217 proteins not reported there, 133 were identified in six of six runs (*core* in Fig. 2*D*). Yeast has 809 ORFs that are classified as "dubious" by the Saccharomyces Genome Database, and these ORFs are thought not to encode a corresponding protein (Table I). As described before (14), this set of genes provides a useful independent test of false positive identification rates.

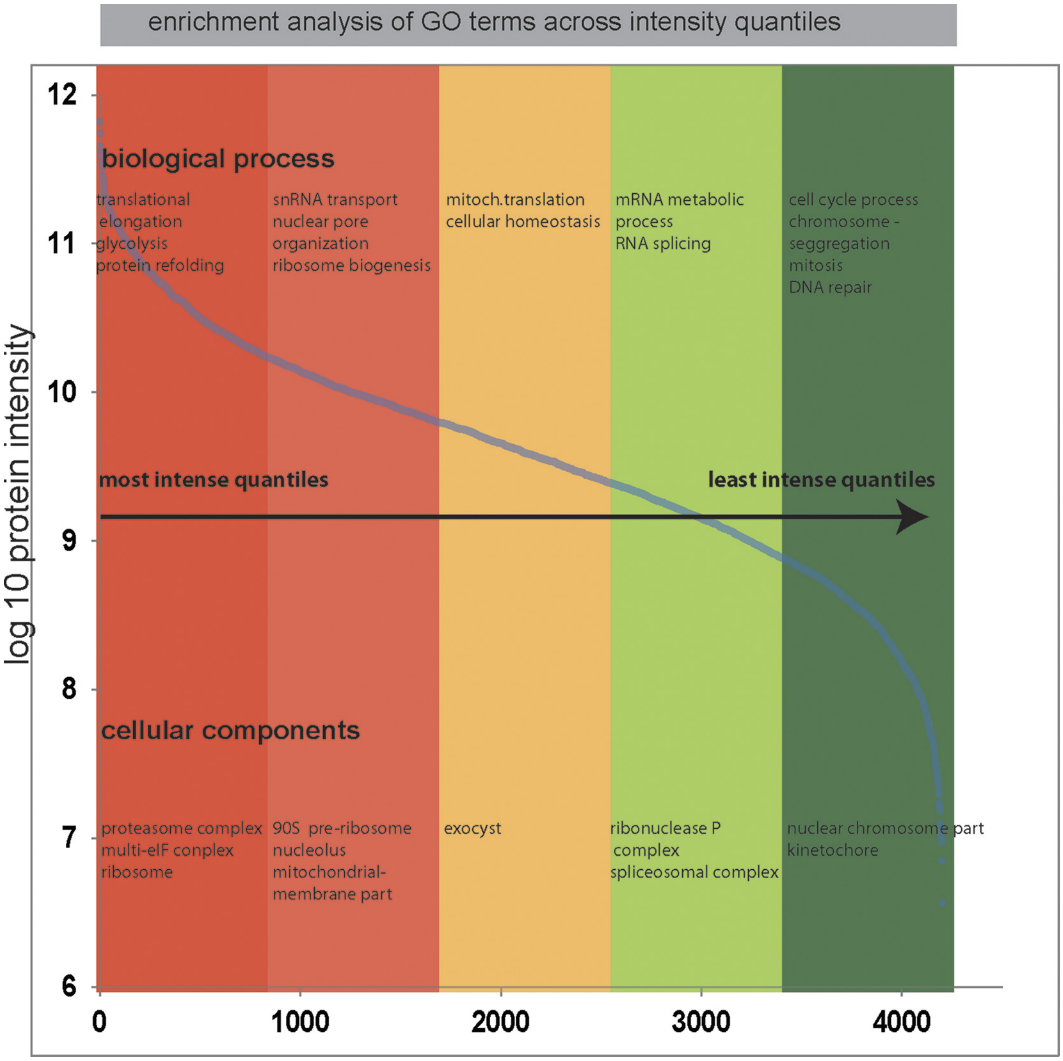


FIG. 3. **Dynamic range of the identified proteome.** Expression levels of identified proteins were roughly estimated using their summed peptide intensities. The proteins were ranked into five quantiles based on their abundance. A Fisher exact test extracted enriched GO terms in each quantile (false discovery rate < 0.02 after Benjamini-Hochberg correction).

The combined single-shot data set only identified two dubious ORFs ("majority" protein column in supplemental Table II), whereas on the basis of a 1% false positive rate, we would have expected five false positives hits in this subset (1% of 809 dubious ORFs given our coverage of the 6717 yeast ORFs; Table I). Furthermore, one of the two dubious ORF hits was also found in our previous study because one of only four hits in this subset (YBR126W-A), which suggests that it may not in fact be a false positive. These data provide independent evidence that our false positive rate is below 1%.

Pathway Analysis of the Detected Proteome—Table I indicates that the six single-shot runs together identified 78% of the ORFs verified as genuine gene products by the Saccharomyces Genome Database; therefore at least this number is expressed as proteins in laboratory yeast. Many pathways and functions are not needed under laboratory conditions, and the corresponding proteins may not be expressed. At 88%, coverage of the proteins in the Kyoto Encyclopedia of Genes and Genomes database was very high in the single shot yeast proteome, as was the coverage of the three gene ontology (GO) categories (GOCC, 85%; GOMF, 82%; and GOBP, 85% CC-cell component, MF-molecular function, BP-biological process). (Because some pathways consist of only a few proteins, we restricted the analysis to pathways with 10 proteins or more; coverage would be even higher without this filter.) Interestingly, the pathways with most missing proteins belong to sugar metabolism and meiosis (Table I), functions that are not expected to be active in haploid yeast growing in glucose media.

Dynamic Range of the Single-shot Proteome—Given the number of identified proteins, we expected the single-shot proteome to have a large dynamic range of protein expression. Indeed, the integrated peptide signals for all the identified proteins spanned approximately 5 orders of magnitude in

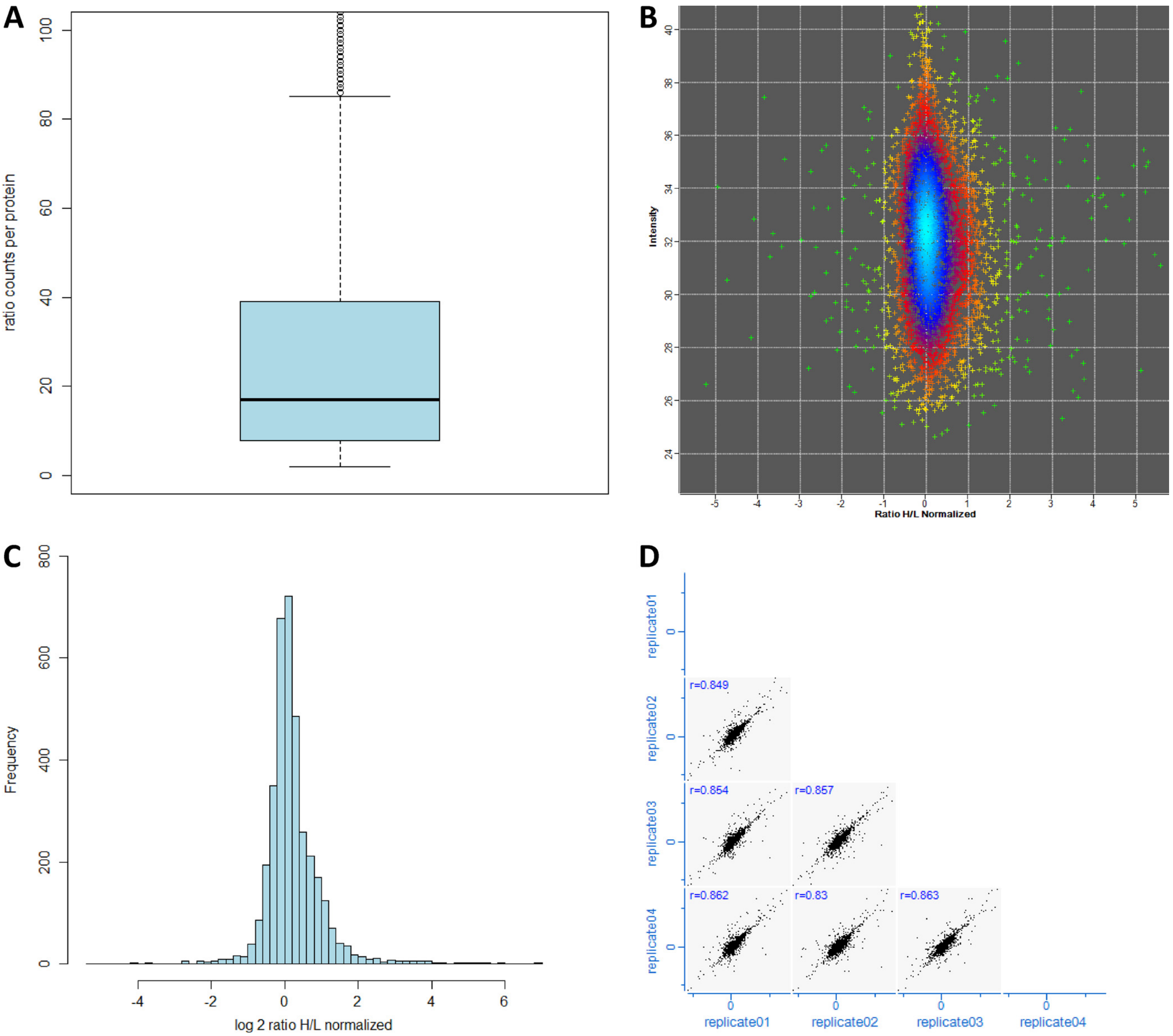


FIG. 4. **Quantification of the yeast proteome using spike-in SILAC labeling.** From four single-shot runs, more than 3200 proteins were quantified with respect to the spike-in SILAC mix. *A*, box plot of the number of ratio counts contributing to quantification of each protein. *B* and *C*, distribution of protein ratios to the spike-in SILAC standard. *D*, the reproducibility of spike-in SILAC quantification of the biological replicates as illustrated by the protein ratio correlations as shown here.

the single-shot measurements (Fig. 3). A recent multiple reaction monitoring study examined the detectability of 127 proteins chosen to represent the full range of the yeast protein expression from most abundant to least abundant protein classes (15). Our single shot proteome included 121 of these proteins, and the six missing proteins were all in the lowest abundance classes. All of the proteins in the category “less than 50 copies/cell” were identified, but they may have been misclassified (18). Together, these results indicate that our data set covered a remarkably large dynamic range.

Bioinformatic enrichment analysis of GO terms in the most abundant quantiles of the distribution, as expected, placed the cell cytoskeleton and biogenesis-related functions among the functions carried out by the most abundant proteins. Cell cycle-related functions are diluted down in nonsynchronized cells and accordingly were enriched in the lowest quantile.

Performance of a Spike-in SILAC Standard—Although SILAC has become a standard and highly accurate quantification method in many systems, the requirement for metabolic labeling prevents some researchers from adopting this technology. Furthermore, in some systems the requirement for media free of external amino acids may impose restrictions on the intended experiments. These issues are addressed by a spike-in SILAC approach (22). In that strategy, a standard

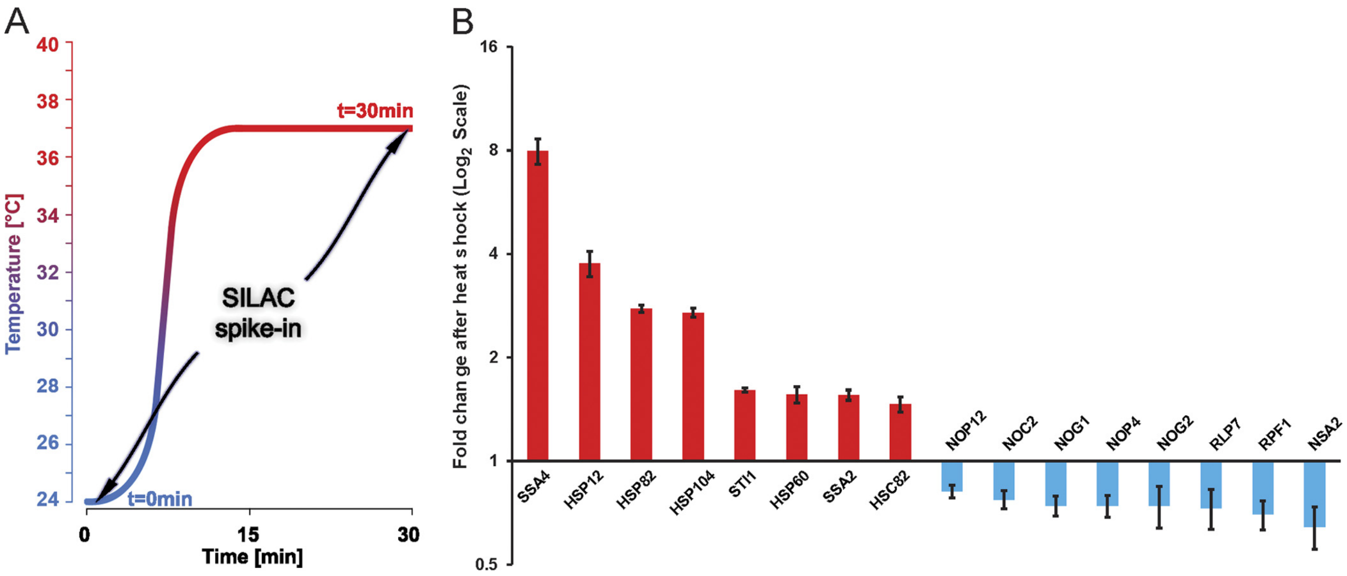


FIG. 5. **Quantitation of heat shock response by spike in SILAC strategy.** *A*, schematic representation of the heat shock experiment. Samples at $t = 0$ and 30 min after incubation at 37 °C were mixed 1:1 with spike-in SILAC standard grown at 30 °C. *B*, fold change represented in $\log_2$ ratios is shown for selected proteins.

representing the proteome of interest is heavy lysine-labeled and serves as a reference across diverse experiments. Biological experiments can be performed as usual, and the spike-in standard is mixed in before sample preparation.

To enable a spike-in strategy for yeast, we SILAC-labeled the W303 MATα strain in which the Lys2 gene was knocked out by homologous recombination. A relatively small amount of standard is sufficient for a large number of experiments ("Experimental Procedures"). It is advantageous to choose the standard so that it represents diverse conditions. Therefore we also cultured yeast under a different growth condition (2% ethanol) and a temperature stress condition. The spike-in mix was prepared by combining all three conditions in equal amounts. To test quantification with the spike-in SILAC standard in single-run conditions, we mixed it into yeast growing under normal laboratory conditions in rich media. Quadruplicate single-run analyses together identified 3794 yeast proteins (supplemental Table III). This number is somewhat lower than in the above "label-free" experiments because SILAC doubles the complexity of the peptide mixtures and because the number of runs was lower. Of these proteins, 3656 and 3553 were quantified with two and three "ratio counts," respectively, which designates valid SILAC quantification ratios in the MaxQuant analysis. The median number of ratio counts/protein was 16 (Fig. 4*A*). Despite using a spike-in SILAC standard including several conditions, the distribution of the ratios in these single-run experiments was very narrow, with 89% of the protein ratios within a 2-fold change (Fig. 4, *B* and *C*). Furthermore, correlation analysis between all of the individual replicates resulted in *R* values of at least 0.83 (Fig. 4*D*). Remarkably, inclusion of the ethanol growth condition in the mix now enabled complete identification of the glycolysis and gluconeogenesis pathways, TCA cycle, and glyoxylate cycle (45 of 45 proteins) as targeted in the recent multiple reaction monitoring study (15). These results demonstrate that the yeast spike-in SILAC adequately represents the yeast proteome and that it performs well in single-run quantification analysis.

Systems-wide Response to Heat Shock—To test the single-run workflow in a systems biology context, we chose to investigate the heat shock response. This is a much studied stress response in yeast. Despite many microarray studies (31, 32), no in-depth proteomic study of this process has been reported. In addition, heat shock is an inevitable component of experiments involving temperature-sensitive mutants, and it would therefore be interesting to know how heat shock modulates the proteome.

The heat shock experiment was performed by shifting the yeast cultures from 24 to 37 °C, taking time points at 0 min and after 30 min at 37 °C (Fig. 5*A*). The samples were combined with the spike-in standard and analyzed by 4-h single runs in quadruplicates. After MaxQuant analysis with the "matching between run" feature, we identified 4072 proteins. The heat shock data set had an overlap of 3708 proteins with the core proteome depicted in Fig. 2*D*. We filtered for proteins that had at least been quantified twice at both time points and obtained 3152 yeast proteins (supplemental Table IV).

Fig. 5*B* shows the fold change of proteins with significant change upon heat shock on a $\log_2$ scale. For every protein, these fold changes were calculated as "ratios of ratios" by dividing the ratios of the unlabeled samples to spike-in SILAC standard (light to heavy ratio) for control ($t = 0$) and heat shock ($t = 30$ min). One of the proteins with the highest fold change (close to 4-fold induction) was HSP12 (heat shock

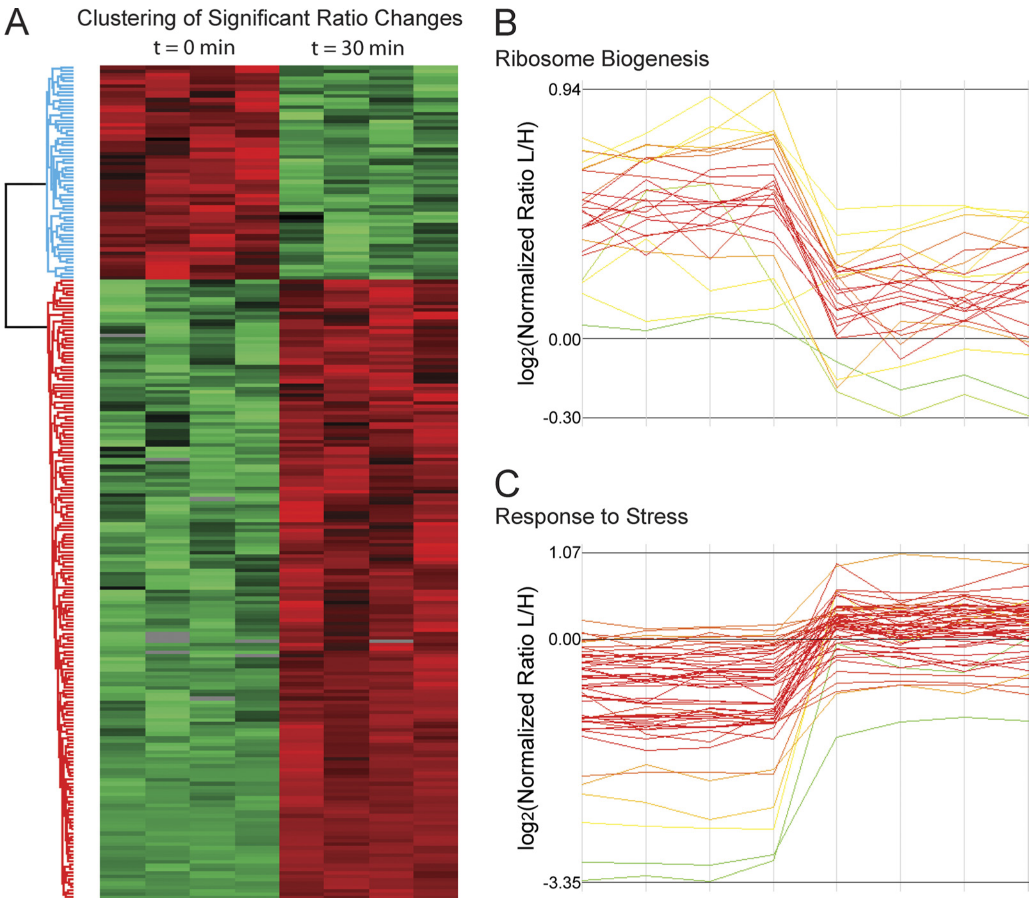


FIG. 6. **Hierarchical clustering of significantly changing proteins.** *A*, clustering of significantly up- and down-regulated proteins upon heat shock. Significance was determined by analysis of variance with correction for multiple hypothesis testing. *B* and *C*, expression patterns for clusters enriched for ribosome biogenesis (*B*) and response to stress (*C*) show the two major trends of protein regulation.

protein 12), which is known to be highly induced by heat shock as well as other stress factors (33). Other heat shock proteins were also up-regulated, including SSA4, SSA2, HSP104, HSP82, and HSP60 (Fig. 5*B*), and this group displayed the highest fold changes overall. Among the down-regulated factors, we noticed a prominent group of proteins involved in ribosomal biogenesis. For example, NSA2, NOG1, RPF1, NOP4, and NOP12 were all down-regulated significantly. The fold changes of these proteins were between 0.6 and 8.0, which was still reliably quantified by MaxQuant (see *error bars* in Fig. 5*B*).

Next we explored the global proteomics response using the Perseus bioinformatics environment that is part of MaxQuant. We performed one-way analysis of variance between the quadruplicates at $t = 0$ and $t = 30$ min and Benjamini-Hochberg correction for multiple hypothesis testing with a cutoff false discovery rate value of 0.02. This yielded 234 proteins that were significantly changing in expression (supplemental Table V). More than half of these proteins were up-regulated (Fig. 6*A*). Enrichment analysis of either set revealed the GO terms "nucleolus" and "ribosome biogenesis" as highly significantly down-regulated ($p < 10^{-16}$). Among the up-regulated proteins, the GO categories "response to stress" and "catabolic process" were most dominant. The profiles of the proteins responsible for these effects are plotted in Fig. 6 (*B* and *C*). As a control, we inspected the profiles in the category "transport," which is not significantly changing upon heat shock. These profiles do not display a coherent trend upon heat shock.

Closer inspection of the down-regulated processes highlighted additional categories related to the regulation of translation. For example, proteins belonging to "tRNA metabolic processes", which are needed for translation initiation and elongation, are all significantly down-regulated during heat shock ($p < 10^{-5}$). By the same token, rRNA transcription, maturation, and ribosome assembly would be expected to be down-regulated, and this is indeed what our bioinformatics analysis shows. The nucleolus itself is the site for many of these processes and is independently known to be a key sensor of cellular stress (34). Our analysis now pinpoints proteins responsible for this interesting connection.

Conclusion and Outlook—Here we have devised a minimalistic proteomic workflow consisting only of pipette-based preparation of digested yeast cell lysate, spike-in SILAC as the quantification technology, single UHPLC-runs on a bench top mass spectrometer and data analysis by the freely available MaxQuant framework. Despite its simplicity, this technology reaches very large coverage of the yeast proteome and readily allows system-wide analysis of a perturbation such as stress response.

Attractive features of our workflow include its sensitivity and rapid analysis times. Because there are no requirements for labeling, experiments can be performed according to standard protocols, and standard yeast strains can be employed. We believe that the single-shot system is indeed a valid third approach between in-depth shotgun proteomics employing fractionation and targeted approaches. That said, there are many applications of proteomics where the single shot technology as described here would not be the ideal approach. For example, very large sequence coverage of the proteome, as needed to distinguish all isoforms, cannot be expected of this strategy. Likewise, analysis of post-translational modifications usually requires enrichment and fractionation steps. However, almost all the improvements made to enable nearly complete coverage of the yeast proteome would carry over to the analysis of fractions in a standard shotgun proteomics approach.

Here we have applied the single-shot technology to the yeast model system. The human proteome is much more complex than the yeast proteome, but with further advances in technology, it is possible that much of that proteome will also be analyzable by single-shot approaches.

* This work was supported by European Commission's 7th Framework Programme Grant Agreement HEALTH-F4-2008-201648/PROSPECTS. 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.

S This article contains Tables I–V and Figs. 1 and 2.

§ These authors contributed equally to this work.

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

REFERENCES

1. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996) Life with 6000 genes. *Science* **274,** 546, 563–567
2. Bader, G. D., Heilbut, A., Andrews, B., Tyers, M., Hughes, T., and Boone, C. (2003) Functional genomics and proteomics: Charting a multidimensional map of the yeast cell. *Trends Cell Biol.* **13,** 344–356
3. Jorgensen, P., Breitkreutz, B. J., Breitkreutz, K., Stark, C., Liu, G., Cook, M., Sharom, J., Nishikawa, J. L., Ketela, T., Bellows, D., Breitkreutz, A., Rupes, I., Boucher, L., Dewar, D., Vo, M., Angeli, M., Reguly, T., Tong, A., Andrews, B., Boone, C., and Tyers, M. (2003) Harvesting the genome's bounty: Integrative genomics. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 431–443
4. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003) Global analysis of protein expression in yeast. *Nature* **425,** 737–741
5. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691
6. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Ménard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A. M., Shapiro, J., Sheikh, B., Suter, B., Wong, S. L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H., and Boone, C. (2004) Global mapping of the yeast genetic interaction network. *Science* **303,** 808–813
7. Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., and Mann, M. (1996) Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 14440–14445
8. Figeys, D., Ducret, A., Yates, J. R., 3rd, and Aebersold, R. (1996) Protein identification by solid phase microextraction-capillary zone electrophoresis-microelectrospray-tandem mass spectrometry. *Nat. Biotechnol.* **14,** 1579–1583
9. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
10. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247
11. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
12. de Godoy, L. M., Olsen, J. V., de Souza, G. A., Li, G., Mortensen, P., and Mann, M. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **7,** R50
13. Swaney, D. L., Wenger, C. D., and Coon, J. J. (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9,** 1323–1329
14. de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–1254
15. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138,** 795–806
16. Köcher, T., Swart, R., and Mechtler, K. (2011) Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Anal. Chem.* **83,** 2699–2704
17. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793
18. 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,** 10.1074/mcp.M110.003699
19. 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,** 10.1074/mcp.M111.011015
20. 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
21. Walther, T. C., Olsen, J. V., and Mann, M. (2010) Yeast expression proteomics by high-resolution mass spectrometry. *Methods Enzymol.* **470,** 259–280
22. 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. Protocols* **6,** 147–157
23. Futcher, B. (1999) Cell cycle synchronization. *Methods Cell Sci.* **21,** 79–86
24. Walker, G. M. (1999) Synchronization of yeast cell populations. *Methods*

Cell Sci. **21,** 87–93

25. Janke, C., Magiera, M. M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., and Knop, M. (2004) A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. *Yeast* **21,** 947–962
26. 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
27. Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
28. 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
29. 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
30. 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
31. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257
32. Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12,** 323–337
33. Praekelt, U. M., and Meacock, P. A. (1990) HSP12, a new small heat shock gene of *Saccharomyces cerevisiae*: Analysis of structure, regulation and function. *Mol. Gen. Genet.* **223,** 97–106
34. Boulon, S., Westman, B. J., Hutten, S., Boisvert, F. M., and Lamond, A. I. (2010) The nucleolus under stress. *Mol. Cell* **40,** 216–227

ARTICLE 2: A PERSPECTIVE ON COMPLETE PROTEOMIC MEASUREMENTS

THE COMING AGE OF COMPLETE, ACCURATE, AND UBIQUITOUS PROTEOMES

AUTHORS: MATTHIAS MANN[1,2,*], NILS A. KULAK[1], NAGARJUNA NAGARAJ[1], AND JUERGEN COX[1]

1 Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry
2 Novo Nordisk Foundation Center for Protein Research, University of Copenhagen
** Correspondence*

PROLOGUE:

Recent developments of LC-MS platforms allow increasing coverage of proteome measurements. The achievement of near complete proteome coverage of a eukaryotic model system inspired us to investigate the state of the art of comprehensive shotgun proteomics measurements. Focus of this perspective was the recent advances and technological breakthroughs that now allow deep measurements of yeast and mammalian cells. We discuss the necessary developments to achieve deep proteomic coverage in more routine settings with limited acquisition times and their potential use for system-wide studies.

Molecular Cell
Perspective

The Coming Age of Complete, Accurate, and Ubiquitous Proteomes

Matthias Mann,[1,2,*] Nils A. Kulak,[1] Nagarjuna Nagaraj,[1] and Jürgen Cox[1]
[1]Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
*Correspondence: mmann@biochem.mpg.de

http://dx.doi.org/10.1016/j.molcel.2013.01.029

High-resolution mass spectrometry (MS)-based proteomics has progressed tremendously over the years. For model organisms like yeast, we can now quantify complete proteomes in just a few hours. Developments discussed in this Perspective will soon enable complete proteome analysis of mammalian cells, as well, with profound impact on biology and biomedicine.

Introduction

Proteins are the biochemical actors in all cellular processes, and diseases almost always manifest at the level of proteins. Accordingly, analysis of specific proteins of interest is ubiquitous in biological research. Usually, this relies on long-established techniques such as staining of gel-separated proteins or antibody-based methods. In an age of whole-genome analysis and systems biology, however, it would be desirable to determine how the entirety of all expressed proteins, the proteome, changes in the process of interest. In effect, such a capability could transform the traditional protein by protein approach in biomedical research into an investigation of the entire cellular system. The appeal of unbiased and large-scale analysis of the total protein complement of a biological system in a given state has in fact always been evident. Indeed, attempts to map proteomes date back even further than those to map genomes (O'Farrell, 1975). However, accurate large-scale protein characterization had to await the development of analysis methods equal in power to those that can be brought to bear on oligonucleotides. For proteins, this technology is mass spectrometry (MS), a nearly universal detection method that has no principal limits to its specificity and sensitivity down to the level of single molecules. Biomolecules have been made amenable to MS by the development of the MALDI and electrospray soft ionization methods in 1988 for which the Nobel Prize in Chemistry was awarded in 2002. Initially, MS-based characterization of proteins took the form of peptide mapping of isolated proteins, meaning the mass measurement of peptides derived from enzymatic digestion. The masses and fragmentation spectra of the peptides were assigned to the known protein sequence. This basic principle was then extended by searching for peptide mass and peptide fragment information in an amino acid sequence database, which enabled the identification of previously unknown proteins with unprecedented sensitivity (Wilm et al., 1996). Even complex peptide mixtures generated from unseparated protein complexes proved to be amenable to liquid chromatography combined with online electrospray analysis by MS and tandem MS (MS/MS) (Link et al., 1999). By the early 2000s, the basic pieces of the "shotgun proteomics" pipeline were in place and a steady march toward more and more complex protein samples began (Aebersold and Mann, 2003). However, until very recently, it was still thought that daunting technological challenges would keep complete proteome analysis impractical for the foreseeable future (Malmström et al., 2007).

In addition to complex peptide mixture analysis, MS-based technologies have been developed for different applications. For instance, MALDI ionization is used for "imaging" of the surface molecules of tissues (Schwamborn and Caprioli, 2010), and "top-down" electrospray analysis of intact proteins can reveal important information about the combination of modifications on a mature protein or about the structure of protein complexes (Tran et al., 2011; Zhou et al., 2011). Since these methods do not aim at complete proteome characterization, we will not discuss them here. Likewise "targeted proteomics" employs basically the same workflow as shotgun proteomics, but like traditional protein chemistry methods, it aims at rapid and simplified analysis of already known candidate proteins (Picotti et al., 2009; Wolf-Yadlin et al., 2007).

In this Perspective, we explore the idea that MS-based shotgun proteomics is now becoming sufficiently powerful to tackle complete proteomes in a sensitive, accurate, and streamlined manner. We show that this is already the case for the yeast proteome and highlight the evidences that suggest it will soon be true of mammalian proteomes as well. Proteomics will ideally complement and perhaps extend genomic methods such as microarrays and next-generation sequencing, which are extremely powerful but which by their nature cannot directly interrogate the proteome.

What Is a Complete Proteome?

Sequencing of the first genomes, and especially of the human genome, was the signature accomplishment of the new "large-scale biology." While requiring great ingenuity and large international efforts, achievement of a complete reference genome was relatively easy to define. This is not the case in proteomics, because in contrast to the linear array of genomic information, a proteome has a high dynamic range and varies with time and space. If one takes into account posttranslational modifications and protein isoforms, even the definition of what constitutes

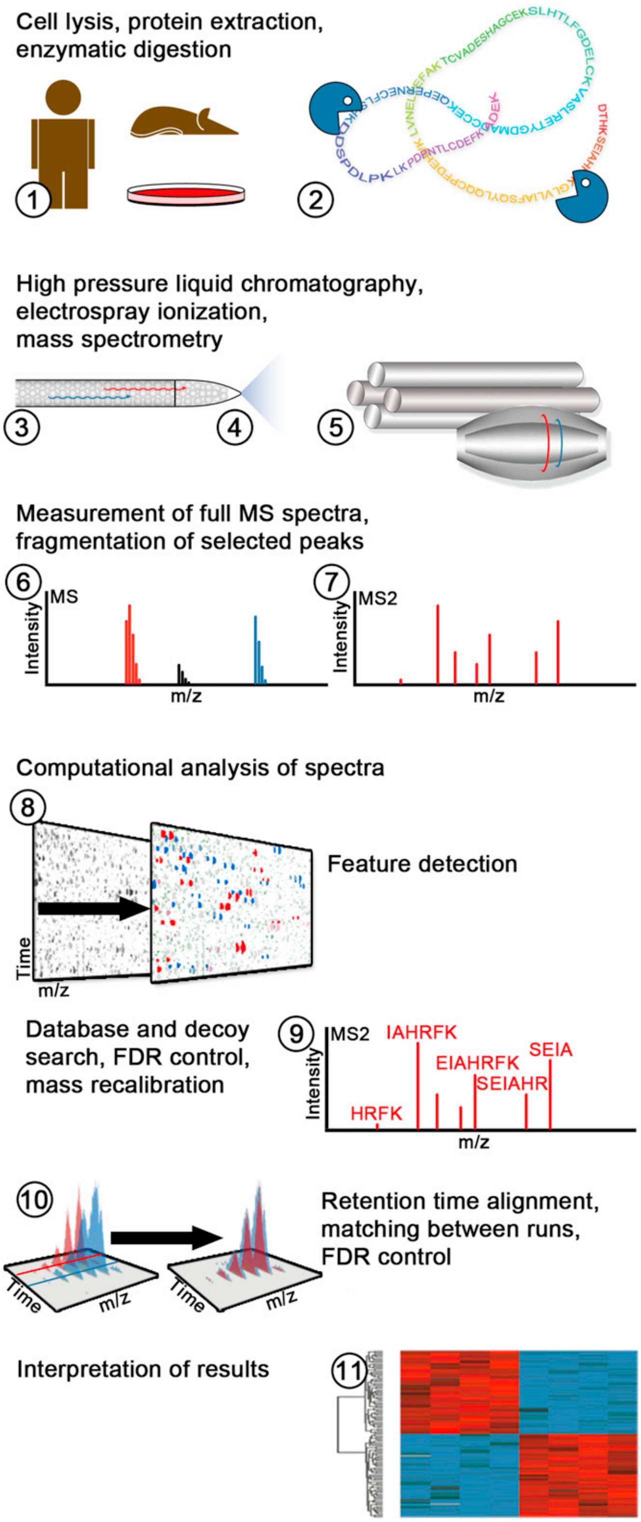

Cell lysis, protein extraction, enzymatic digestion
1
2
High pressure liquid chromatography, electrospray ionization, mass spectrometry
3
4
5
Measurement of full MS spectra, fragmentation of selected peaks
6
MS
Intensity
m/z
7
MS2
Intensity
m/z
Computational analysis of spectra
8
Feature detection
Time
m/z
Database and decoy search, FDR control, mass recalibration
9
MS2
Intensity
IAHRFK
SEIA
EIAHRFK
SEIAHR
HRFK
m/z
10
Retention time alignment, matching between runs, FDR control
Time
m/z
Time
m/z
Interpretation of results
11

Figure 1. Shotgun Proteomics Workflow for Complete Proteome Measurements

The first three steps pertain to sample preparation and on-line peptide separation. Steps 4–7 represent electrospray ionization of peptides, analysis in the mass spectrometer (quadrupole-Orbitrap analyzer), resulting in survey scans of the eluting peptides as well as high-resolution fragment spectra. Computational proteomics makes up the remaining steps of the pipeline, including bioinformatic and systems biological interpretation of the results (shown for the MaxQuant framework).

a different protein, and therefore the number of proteins in a given biological system, is subject to diverse interpretations (Cox and Mann, 2011). From an analytical perspective, a comprehensive proteome could mean all of the proteins identifiable by a state of art mass spectrometric methodology, but this is of course a limited view (Beck et al., 2011a). Alternatively, a comprehensive proteome could mean identification of one or few protein representatives of all of the protein coding genes in an organism. The Chromosome-Centric Human Proteome Project has broken down this task chromosome by chromosome and assigned them to different countries (Paik et al., 2012). A complete proteome could also consist of characterization of all the possible isoforms and modification states of all expressed proteins. This may be impossible to achieve, regardless of advances in technology, because a single coding region can give rise to many chemically distinct species, and, when multiplied by looking at all coding regions, the number of possible species is astronomical. From a biological perspective, identification and quantification of at least one protein from every genomic locus that is expressed in a given biological system would already deliver almost all of the benefits. This pragmatic definition of a complete proteome, which we adopt here, also provides a clear goalpost for technology development. Furthermore, by ensuring at least minimal information for the lowest abundance proteins, peptide sequence coverage of all other proteins would inevitably be substantial—because more-abundant proteins are identified with more of their peptides. This would provide a rich source of information about protein isoforms and other protein variants. Clearly, the achievement of a complete proteome defined in this way is not the endpoint for proteomics technology. Instead, it opens up for further developments aimed at maximizing information about functional protein variations within and between proteomes—somewhat akin to the way reference genomes were followed by the current era of studying individual genome differences.

Progress in the Proteomics Workflow

Before complete proteomes could become a realistic prospect, a large number of discoveries and developments related to the basic shotgun proteomics workflow had to happen first (Figure 1). Although often overlooked, technologies aimed at improving and simplifying the sample preparation prior to the actual MS measurement are crucial because they determine if the entire proteome is accessible for exploration. As a result of progressive improvements, proteome measurements can now be unbiased in the protein classes covered, including low-abundance proteins such as transcription factors or difficult-to-extract membrane proteins (Wiśniewski et al., 2009). This is in marked contrast to the early days of proteomics, when these proteins were often missing entirely from published data sets.

There has been a trend away from using extensive cellular fractionation to improve the depth of proteome coverage. While this approach is informative, especially in an organellar proteomics context, it has diminishing returns for full-proteome measurements. This is because even rigorous fractionation only alters the relative abundance of proteins by a factor of ten to 100, meaning that de-enriched proteins are still easily detected by modern mass spectrometers. Furthermore, cellular fractionation multiplies the number of samples to be analyzed, increasing measurement time per proteome and inherently limiting overall sensitivity. Instead, two other areas have turned out to be crucial for in-depth proteome characterization: the chromatography setup preceding online peptide analysis and the mass spectrometers themselves. Recent workflows now tend to push peptide separation to its limits, using high-pressure high-performance liquid chromatography pumps, very small bead particles as column material, and relatively long columns and gradients (Köcher et al., 2011; Thakur et al., 2011). Together, these factors yield high peptide-separation capacity and maximize the number of eluting peptides that the mass spectrometer can isolate and fragment. Increasingly, the combined power of the LC system and the mass spectrometer make it possible to dispense with any upfront protein or peptide separation (“single-shot” or “single-run” analysis, see below), but usually such a step is still included.

The largest hardware improvements have been in the mass spectrometers themselves. Notably, they have become extremely fast. In one second, they can execute the basic measurement cycle, which involves acquiring a survey mass spectrum then fragmenting ten of the eluting peptides (known as the “top 10” method) (Andrews et al., 2011; Michalski et al., 2011). Furthermore, they are now more sensitive and they have a greater dynamic range. Most importantly, instruments with low mass resolution have almost universally been replaced by instruments with resolution in the tens of thousands (MS resolution is defined as peak width divided by mass; a dimensionless quantity). This makes it possible to routinely achieve mass accuracies in the ppm or even sub-ppm range with obvious benefits for the certainty of peptide identification (Cox and Mann, 2008). With high mass resolution, coeluting peptides of similar mass are readily distinguished. This is a precondition for their accurate quantification, which is now a feature of more and more proteomics projects and which has been reviewed in depth elsewhere (Bantscheff et al., 2012; Bantscheff et al., 2007). In brief, the most accurate methods for relative quantification of two or more proteomes are still based on the metabolic incorporation of heavy or light stable isotopes into the entire proteomes to be compared, followed by combined mass spectrometric analysis. Protein turnover can be investigated for every protein in the proteome by dynamic or pulsed versions of these techniques (Hinkson and Elias, 2011). Chemical labeling with stable isotope reagents is also very widely employed. In principle, all proteomic samples are amenable to chemical labeling but care must be taken that they are processed in the same way. In “label-free quantification,” the mass spectrometric signals of the peptides are directly compared between different proteome measurements, meaning that this form of quantification can be performed on any high resolution MS data, provided that measurement conditions remain identical between separate measurements. Metabolic labeling can determine protein changes within a precision as high as a few percent, whereas abundance changes generally need to be at least 2-fold to confidently detect them in a label-free format. Proteomics is also capable of absolute quantification, either with the added peptide signal of a protein (a label-free method) or with isotope-labeled standards, and their accuracies vary accordingly.

The sophisticated algorithms needed for accurate label-free quantification are one example of the astounding improvements made in “computational proteomics,” which has become a research field of its own in recent years. Previously, interpretation of proteomics data was at best semimanual, and data analysis times frequently stretched to several months for a single project. In contrast, the much larger data sets generated in current projects can be processed in a completely automated manner and with quantitative and statistical rigor on par with any other field of large-scale biology. Together, the above developments have laid the groundwork for the proteome analyses described below.

The Yeast Proteome

Apart from being an experimental system of choice for studying many basic biological functions, yeast has long been a testing ground for large-scale biology. Budding yeast has about 6,600 open reading frames, and tagging experiments had shown that more than 4,000 of these are expressed in normal growth conditions (Ghaemmaghami et al., 2003; Huh et al., 2003). As a demonstration that MS-based proteomics can identify and quantify an entire proteome, haploid and diploid yeast were grown in media containing light or heavy SILAC-labeled amino acids, respectively (de Godoy et al., 2008). Combined lysates were either subjected to cellular fractionation (which proved to be comparatively ineffective) or analyzed after one step of peptide separation before LC MS/MS (which proved to be highly efficient). Analysis of the results from these extensive measurements in the MaxQuant framework (Cox and Mann, 2008) identified 4,399 yeast proteins with a confidence of 99%. In each abundance range of the yeast proteome, essentially equal numbers of proteins were detected by MS than had been shown to be expressed by the tagging experiments mentioned above (Figure 2A, from de Godoy et al. [2008]). Due to SILAC quantification, the general absence of the pheromone pathway in diploid yeast was immediately apparent, but the data also pinpointed members of the pathway that must have functions in diploid yeast unrelated to mating (since they were expressed in both cell types). Detailed analysis of transcriptome changes against proteome changes revealed a number of processes that were controlled mainly at the transcript or mainly at the proteome level (Cox and Mann, 2012; de Godoy et al., 2008).

While these results demonstrated that shotgun proteomics can indeed acquire a complete proteome, the entire project involved measurement times of several months, clearly impractical for routine applications. Recently, the yeast proteome was revisited with the latest technology, with the aim of radically simplifying the workflow (Nagaraj et al., 2012). In this minimalistic approach, sample preparation is reduced to a single step, digestion of yeast cells, and there was no fractionation before mass spectrometric analysis (Figure 2B). A benchtop quadrupole Orbitrap mass spectrometer was employed in which peptides are analyzed very rapidly and fragments are always recorded with

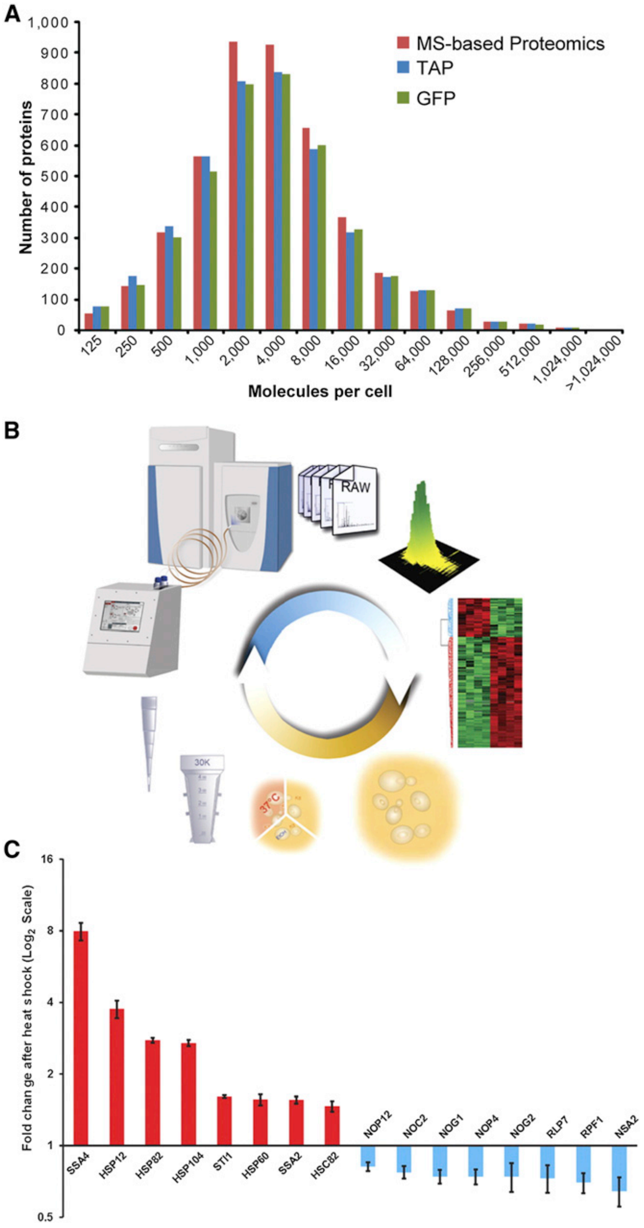


Figure 2. The Complete Yeast Proteome

(A) Comparison of shotgun proteomics results against tandem affinity tagging (TAP) or green fluorescent protein (GFP) tagging experiments. (Reprinted from de Godoy et al. [2008].)

(B) Streamlined system for rapid analysis of nearly the entire yeast proteome. (Adapted from Nagaraj et al. [2012], American Society for Biochemistry and Molecular Biology.)

(C) Median fold changes for heat shock proteins (red) and nucleolar proteins (blue). Error bars represent the SD from quantification with a yeast spike-in SILAC standard. (Reprinted from Nagaraj et al. [2012].)

very high mass accuracy (Michalski et al., 2011). In single 4 hr measurements, this streamlined setup was able to identify more than 4,000 yeast proteins, nearly the complete proteome in this growth state. Sample requirements in single-run analysis are inherently reduced, and the entire analysis could be performed with a few micrograms of peptide material.

This technology was then employed to investigate how the expression levels of each yeast protein changes during heat shock. Answering such a straightforward question by standard western blotting experiments would require a separate experiment for each open reading frame. Instead, the proteome of yeast was grown at normal temperatures and at high temperatures, followed by quadruplicate MS measurements for accurate statistics in little more than a day. Among the significantly changing proteins, heat shock proteins were upregulated (serving as positive controls), and members of the ribosomal biosynthesis pathway were downregulated moderately but highly reproducibly. While temperature stress is a simple perturbation, the employed single-run workflow is entirely generic and could be applied to any biological situation that ultimately involves protein regulation. At least for the yeast proteome, single-run analysis now provides an alternative to targeted proteomics because it shares its beneficial features of rapid and sensitive measurements and still retains the advantages inherent in systems-wide approaches.

Mammalian Proteomes

Attempts to map mammalian proteomes started with 2D gel electrophoresis decades ago (O'Farrell, 1975). Although there sometimes were thousands of spots on these gels, recent analysis of such patterns by MS revealed that they only represented a few hundred different protein coding regions. In contrast, MS-based proteomics in mammalian systems started with small protein complexes and gradually worked its way up to complex organelles (Yates et al., 2005). Reports with more than a few thousand proteins in mammalian systems have only appeared in the last few years (Wiśniewski et al., 2009).

Two recent papers have taken a first serious stab at characterizing human proteomes in comprehensive depth (Beck et al., 2011b; Nagaraj et al., 2011). Both investigated human cancer cell lines (HeLa or U2OS), which are widespread models in cell biology and also have the advantage of being relatively homogeneous and reproducible biological systems. Furthermore, both used generally similar shotgun proteomics strategies, including the measurement of very large numbers of fractions. These were the first proteome measurements to identify more than 10,000 different human proteins in a single experimental system, providing a lower limit to the complexity of mammalian cell line proteomes. To ask how close to completion these proteomes actually were, Beck et al. used modeling tools showing that their mass-spectrometric measurements had gone to saturation, i.e., that addition of further replicates would not materially change the depth of the detected proteome. Nagaraj et al. compared their proteome with deep sequencing of the transcriptome of the same cells. The RNA sequencing data contained 16,500 transcripts from protein-coding genes. Their histogram showed a bimodal distribution whose lower abundance part is probably not functional (Hebenstreit et al., 2011). Commonly used filtering criteria for RNA-seq data dropped this number to less than 12,000 genes. Comparison of the filtered transcriptome and the proteome of HeLa cells suggested that perhaps 10,000 to 12,000 different protein coding loci are expressed in this cancer cell line and that the measured proteome was not very far from completion.

Both studies incorporated special measurements to estimate the copy numbers of the measured proteins. Beck et al. used heavy labeled peptides and extrapolation from measured peptide intensities (Malmström et al., 2009), whereas Nagaraj et al. verified their absolute expression estimates with the PrEST SILAC method (Zeiler et al., 2012). MS-derived signals indicated a dynamic range of protein expression that extends over more than six orders of magnitude. At the same time, 90% of the HeLa proteome is contained within a range of less than 60-fold above and below the median expression level of 18,000 copies per cell. Such data now allows messenger RNA (mRNA) and protein abundance estimation not only for individual proteins but for protein complexes, pathways, and compartments. For instance, protein tyrosine kinases turned out to have a large copy number range from very low expression levels up to the top quarter of expression values. Such information has previously not been available at all on a large scale and is extremely useful when building qualitative or quantitative models of cellular processes or disease mechanisms.

Using a more streamlined approach, Geiger et al. extended the HeLa proteome measurement to 11 common human cell lines and identified more than 10,000 proteins in each (Geiger et al., 2012). With this data, researchers working with these or similar cell lines can now check whether their proteins of interest are expressed and, if so, at what levels and how these levels vary across common cell lines. Such information can reveal that a protein is constantly expressed, i.e., part of the "household proteome," or if it is coexpressed with other proteins in specific cell types (Schaab et al., 2012). Interestingly, this study underlined conclusions from the ProteinAtlas project (Uhlen et al., 2010; Lundberg et al., 2010) that proteins tend to be expressed ubiquitously, with the character of the tissue being shaped more by the level of expression of the proteins rather than by their presence and absence.

With a somewhat less extensive depth of proteome coverage, several recent reports have investigated overall proteome properties in specific biological contexts, for example, stem cell proteomics. Slight but significant changes in embryonic stem cells (ESCs) versus induced pluripotent stem cells (iPSCs) were found at the proteome and phosphoproteome level. Notably, it was possible to detect the remaining imprints of the cell line used to derive the iPSCs (Phanstiel et al., 2011). Studying the differentiation of human stem cells, Blagoev and coworkers found dynamic changes in much of the proteome and at least half of more than 20,000 phosphorylation sites. These results indicated that DNA methyltransferases (DNMTs) are regulated by phosphorylation in this process. Furthermore, the authors observed an interesting association between DNMTs and PAF1, providing a possible molecular link for the silencing of OCT4 and NANOG during differentiation (Rigbolt et al., 2011). Similarly, Heck and coworkers quantified proteome differences between human ESCs, iPSCs, and a fibroblast cell line, reporting more than 10,000 proteins in the combined cell types and confirming high similarity between ESCs and iPSCs as opposed to the fibroblasts (Munoz et al., 2011).

The above studies were all done in cell lines, which can be much easier to analyze than tissue samples. Accordingly, there are only a few in-depth studies of tissue proteomes. A total of 12,000 proteins were reported in a study of phosphorylation sites in nine different mouse tissues (Huttlin et al., 2010). In a very recent colon cancer study, more than 7,500 proteins were quantified across patient matched normal mucosa, primary carcinoma, and nodal metastases (Wiśniewski et al., 2012). Unexpectedly in light of previous microarray studies, the authors observed a very large change in the proteomes of normal tissue and primary carcinoma, whereas there were hardly any statistically significant changes between primary tumor and nodal metastasis.

All the above investigations required significant resources, which makes routine applications impractical. To determine what part of the proteome could be obtained in a few hours with the amounts of material comparable to standard western blotting, the approach described above for single-run analysis of the yeast proteome (Nagaraj et al., 2012) was extended to analyze lysates from 11 cell lines in single 4 hr gradients (N. Nagaraj, et al., 2012, ASMS conference). Remarkably, this led to a detected cellular proteome of around 8,000 different proteins in each of these systems (Figure 3A). The first 184 proteins already accounted for half the proteome mass, while the last 5,600 added up to less than 5%. Protein dynamic range in these single-run measurements exceeded six orders of magnitude, and different abundance ranges were enriched for different biological functions (Figures 3B and 3C). Thus, a very large percentage of the mammalian proteome can be captured in a short time.

Impact of Complete Proteome Analysis

The developments discussed above open up for a future in which complete proteome measurements are not only possible but also streamlined and easily applicable. This vision has already been put into practice for yeast. It is now in principle possible to quantify the entire yeast proteome on a benchtop mass spectrometer, with similar sample amounts and measurement times as those used for western blotting, but probing for all expressed proteins simultaneously. Processing of the data is completely automated and straightforward. As we have shown here, it is only a matter of time until the same approach will deliver essentially complete human cell line proteomes. Tissue analysis and especially the analysis of body fluids are more challenging still and the latter may need entirely novel modes of analysis.

Although the analysis of complete proteomes could soon be feasible in a straightforward manner, we strongly emphasize that the required technology is by no means widespread or readily available. This indeed remains one of the Achilles' heels of modern proteomics: the capabilities of leading laboratories have taken long to translate into general accessibility for the entire research community. Illustrating this point is a recent study in which only seven out of 27 laboratories correctly identify the protein constituents of an equimolar mixture of just 20 proteins (Bell et al., 2009). Given the impressive capabilities of modern-day proteomics, it is crucial that more resources are invested into making the technology available to the broad scientific community. There are no intrinsic reasons why proteomics should be any less affordable, sensitive, or streamlined than current deep sequencing methods.

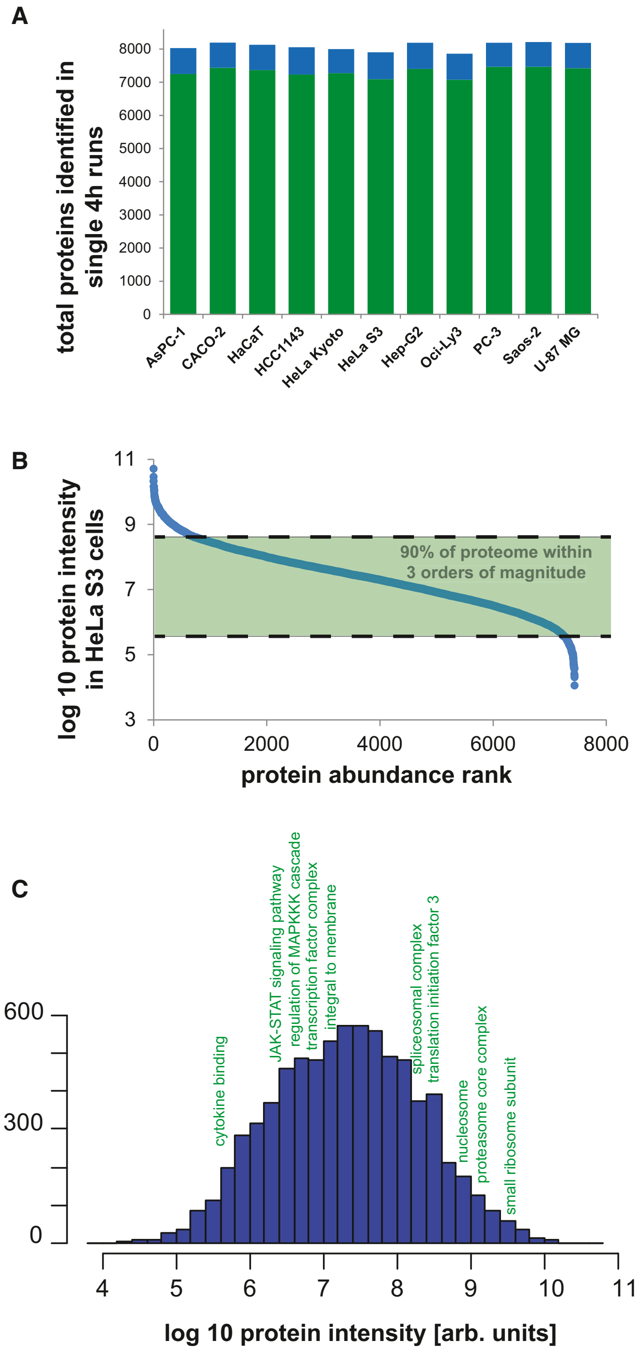


Figure 3. Single-Run Analysis of the Human Proteome

(A) Eleven different cell lines were measured by single 4 hr LC MS/MS runs. Numbers of identified proteins are indicated with (blue) or without (green) "matching between runs" in the MaxQuant environment.

(B) Dynamic range of the single-run proteome spans more than six orders of magnitude, but 90% of the MS signals of the identified proteins are within three orders of magnitude.

(C) Binned histogram of estimated protein copy numbers. Significantly enriched protein categories compared to the entire proteome abundance distribution are annotated (calculated by 1D enrichment [Cox and Mann, 2012]).

Tools are emerging to determine the specific points of regulation in the gene expression cascade. For instance, a pioneering study comparing the turnover of the mRNAs and proteins found that differences in translation rates have a major impact on protein levels relative to mRNA, explaining why mRNA and protein levels often correlate only modestly (Schwanhäusser et al., 2011). In the future, it will be very interesting to investigate whether or not the astonishing complexity of the genome and transcriptome, with its myriad forms of different RNA molecules, is transmitted to the level of the proteome.

Comprehensive expression proteomics is powerful and multidimensional. It can be used to study many aspects of protein function and regulation, including turnover, localization, and protein-biomolecule interactions, all with essentially the same shotgun proteomics pipeline. Improved pipelines for deep expression proteomics can now be used for more exhaustive analysis of PTMs, including phosphorylation, glycosylation, ubiquitination, SUMOylation, and many other modifications. Illustrating how biochemical and cell biological experiments can be reimagined with quantitative proteomics, Lamond and coworkers performed subcellular fractionation of differently SILAC-labeled cells into compartments. This allowed them to take an unbiased look at how the proteins move from one cellular compartment to another in response to perturbations. Importantly, modern proteomics technology in principle allows researchers to resolve protein isoforms and modification states, opening up entirely new perspectives on cell biology (Ahmad et al., 2012; Boisvert et al., 2012). It is also possible to learn about changes in the interaction landscape of the cell over time. For example, Kristensen et al. observed changes in small protein complexes in response to signaling events by combining protein correlation profiling with classical size exclusion chromatography (Kristensen et al., 2012).

In conclusion, this Perspective has shown that proteomics can be as comprehensive as other "omics" approaches. The ability to characterize all expressed proteins at once could transform any cell biological experiment into a systems biology study. Furthermore, it is clear that deep proteome characterization could have many applications in the clinic. Because proteins are often closer to biological functions than either DNA or RNA, we believe that this is an especially important area to explore, given the potentially far-reaching benefits for understanding disease processes, monitoring drug efficacy, and classifying patients.

ACKNOWLEDGMENTS

We thank members of the Proteomics and Signal Transduction group for fruitful comments and especially Marco Hein for help with the figures. Work in M.M.'s laboratory is supported by the Max Planck Society for the Advancement of Science and by PROSPECT, a Seventh Framework Programme of the European Union (grant agreement HEALTH-F4-2008-201648).

REFERENCES

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Ahmad, Y., Boisvert, F.M., Lundberg, E., Uhlen, M., and Lamond, A.I. (2012). Systematic analysis of protein pools, isoforms, and modifications affecting turnover and subcellular localization. Mol. Cell. Proteomics *11*, M111.013680.

Andrews, G.L., Simons, B.L., Young, J.B., Hawkridge, A.M., and Muddiman, D.C. (2011). Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). Anal. Chem. *83*, 5442–5446.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. *389*, 1017–1031.

Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal. Bioanal. Chem. *404*, 939–965.

Beck, M., Claassen, M., and Aebersold, R. (2011a). Comprehensive proteomics. Curr. Opin. Biotechnol. *22*, 3–8.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011b). The quantitative proteome of a human cell line. Mol. Syst. Biol. *7*, 549.

Bell, A.W., Deutsch, E.W., Au, C.E., Kearney, R.E., Beavis, R., Sechi, S., Nilsson, T., and Bergeron, J.J.; HUPO Test Sample Working Group. (2009). A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat. Methods *6*, 423–430.

Boisvert, F.M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., Barton, G., and Lamond, A.I. (2012). A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol. Cell. Proteomics *11*, M111.011429.

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. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. Annu. Rev. Biochem. *80*, 273–299.

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.

de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell. Proteomics *11*, M111.014050.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S.A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol. Syst. Biol. *7*, 497.

Hinkson, I.V., and Elias, J.E. (2011). The dynamic state of protein turnover: It's about time. Trends Cell Biol. *21*, 293–303.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.

Köcher, T., Swart, R., and Mechtler, K. (2011). Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. Anal. Chem. *83*, 2699–2704.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. Nat. Methods *9*, 907–909.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., and Yates, J.R., 3rd. (1999). Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. *17*, 676–682.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol. *6*, 450.

Malmström, J., Lee, H., and Aebersold, R. (2007). Advances in proteomic workflows for systems biology. Curr. Opin. Biotechnol. *18*, 378–384.

Malmström, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E.W., and Aebersold, R. (2009). Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. Nature *460*, 762–765.

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*, M111, 011015.

Munoz, J., Low, T.Y., Kok, Y.J., Chin, A., Frese, C.K., Ding, V., Choo, A., and Heck, A.J. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol. Syst. Biol. *7*, 550.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol. Cell. Proteomics *11*, M111.013722.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. *250*, 4007–4021.

Paik, Y.K., Jeong, S.K., Omenn, G.S., Uhlen, M., Hanash, S., Cho, S.Y., Lee, H.J., Na, K., Choi, E.Y., Yan, F., et al. (2012). The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat. Biotechnol. *30*, 221–223.

Phanstiel, D.H., Brumbaugh, J., Wenger, C.D., Tian, S., Probasco, M.D., Bailey, D.J., Swaney, D.L., Tervo, M.A., Bolin, J.M., Ruotti, V., et al. (2011). Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat. Methods *8*, 821–827.

Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B., and Aebersold, R. (2009). Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. Cell *138*, 795–806.

Rigbolt, K.T., Prokhorova, T.A., Akimov, V., Henningsen, J., Johansen, P.T., Kratchmarova, I., Kassem, M., Mann, M., Olsen, J.V., and Blagoev, B. (2011). System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. Sci. Signal. *4*, rs3.

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*, M111.014068.

Schwamborn, K., and Caprioli, R.M. (2010). Molecular imaging by mass spectrometry—looking beyond classical histology. Nat. Rev. Cancer *10*, 639–646.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Thakur, S.S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, 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.

Tran, J.C., Zamdborg, L., Ahlf, D.R., Lee, J.E., Catherman, A.D., Durbin, K.R., Tipton, J.D., Vellaichamy, A., Kellie, J.F., Li, M., et al. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature *480*, 254–258.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature *379*, 466–469.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

Wiśniewski, J.R., Ostasiewicz, P., Duś, K., Zielińska, D.F., Gnad, F., and Mann, M. (2012). Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Mol. Syst. Biol. *8*, 611.

Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D.A., and White, F.M. (2007). Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc. Natl. Acad. Sci. USA *104*, 5860–5865.

Yates, J.R., 3rd, Gilchrist, A., Howell, K.E., and Bergeron, J.J. (2005). Proteomics of organelles and large cellular structures. Nat. Rev. Mol. Cell Biol. *6*, 702–714.

Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M., and Mann, M. (2012). A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. Mol. Cell. Proteomics *11*, O111.009613.

Zhou, M., Morgner, N., Barrera, N.P., Politis, A., Isaacson, S.C., Matak-Vinković, D., Murata, T., Bernal, R.A., Stock, D., and Robinson, C.V. (2011). Mass spectrometry of intact V-type ATPases reveals bound lipids and the effects of nucleotide binding. Science *334*, 380–385.

ARTICLE 3: THE MINIMALISTIC SAMPLE PREPARATION

MINIMAL, ENCAPSULATED PROTEOMIC-SAMPLE PROCESSING APPLIED TO COPY-NUMBER ESTIMATION IN EUKARYOTIC CELLS

AUTHORS: NILS A. KULAK[1], GARWIN PICHLER[1], IGOR PARON[1], NAGARJUNA NAGARAJ[1], AND MATTHIAS MANN[1]

1 *Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry*

PROLOGUE:

The above described development of a robust and powerful LC-MS platform was a major breakthrough for easy and complete measurements of proteomes. Since every component of the platform was developed with a view towards simplified handling, a more robust sample preparation procedure was the logical next step for better overall applicability. The initial aim was to streamline the pre-existing filter based FASP workflow and to achieve better reproducibility and the capability to multiplex the sample handling. We therefore started out to challenge every processing step in the complex sample preparation workflow. For this we used the SILAC technology to quantify the effect of each of the changes made to the established FASP protocol.

The first major breakthrough was the combination of lysis, reduction, and alkylation, removing two independent and highly time consuming procedures. With this adaptation to the protocol, approximately 1 hour of lab-work is saved. The second essential and unconventional alteration to classical protocols was to perform the proteolytic digestion in the same buffer that was used for tissue- or cell-lysis. This in term reduced the processing pipeline from cell to peptides to a two stage protocol. Depending on the workflow applied such a change reduces the entire procedure by 3 hours or even an entire day. The final breakthrough was to perform the entire procedure within a single enclosed container. For this we built on the StageTips, which are routinely used for peptide clean-up. Together this resulted in a minimalistic three-step protocol without changing reaction vessels.

The new sample processing procedure entailed a wide range of positive effects. Excellent quantitative reproducibility and sensitivity was observed for all tested samples. The simplified procedure also allows multiplexed sample processing and specially designed 96-well processing blocks in principle allow processing of hundreds of samples in parallel. Because the sample preparation is performed in StageTips, fractionation techniques can be readily added. We therefore further developed the aspect of simple peptide fractionation and achieved excellent results using novel SPE fractionation technologies. We here obtained the deepest proteome coverage of exponentially growing *S. cerevisiae* and *S. pombe* and one of the most comprehensive quantitative datasets of the human cancel cell line HeLa reported so far. This is even more remarkable considering that it was done with only 24 hour gradient times.

Because of the very high quantitative reproducibility and close to comprehensive proteomic coverage, we used the data set for protein copy number estimates. These should provide a useful resource for basic research using these model systems. We further observed evolutionary variations and similarities across the model systems that argue for very high functional conservation. With its excellent coverage and data quality the workflow and resulting data sets define a benchmark for proteomic studies.

Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells

Nils A Kulak, Garwin Pichler, Igor Paron, Nagarjuna Nagaraj & Matthias Mann

Mass spectrometry (MS)-based proteomics typically employs multistep sample-preparation workflows that are subject to sample contamination and loss. We report an in-StageTip method for performing sample processing, from cell lysis through elution of purified peptides, in a single, enclosed volume. This robust and scalable method largely eliminates contamination or loss. Peptides can be eluted in several fractions or in one step for single-run proteome analysis. In one day, we obtained the largest proteome coverage to date for budding and fission yeast, and found that protein copy numbers in these cells were highly correlated (R^2 = 0.78). Applying the in-StageTip method to quadruplicate measurements of a human cell line, we obtained copy-number estimates for 9,667 human proteins and observed excellent quantitative reproducibility between replicates (R^2 = 0.97). The in-StageTip method is straightforward and generally applicable in biological or clinical applications.

Bottom-up MS-based proteomics involves the separation of peptides by liquid chromatography (LC), coupled to electrospray ionization and peptide analysis in the mass spectrometer. Sample preparation in proteomics is an important part of the workflow because it determines the overall sensitivity, accuracy and robustness of the entire analysis[1]. It consists of a multistep procedure that begins with the extraction and solubilization of the protein material, and is followed by denaturation, reduction and alkylation of cysteines, and enzymatic digestion. Peptide mixtures that result from digestion need to be cleaned up for LC-MS/MS. Sample preparation can also include additional separation steps such as fractionation at the organelle, protein or peptide levels.

By combining advances in both analytical and computational proteomics workflows, and using extensive sample preparation and fractionation strategies, we have previously reported the identification and quantitation of essentially the entire proteome of exponentially growing yeast[2]. Subsequently, we achieved nearly equivalent proteome coverage with a single-run shotgun proteomics approach, direct LC-MS/MS analysis without prefractionation[3–6], using a benchtop quadrupole Orbitrap mass analyzer[7]. In this work we set out to radically simplify proteome-sample preparation by eliminating or combining steps and performing all processing steps in a single, enclosed volume. We applied the method to determine copy numbers of the proteomes of budding yeast (*Saccharomyces cerevisiae*), fission yeast (*Schizosaccharomyces pombe*) and a human cancer cell line (HeLa cells).

RESULTS

Development and validation of the method

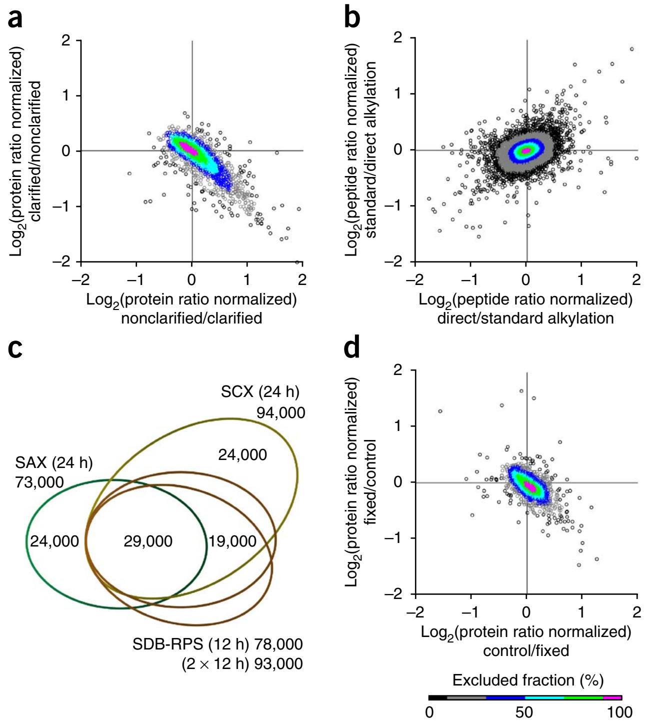
Protein characterization typically involves cell lysis, clarification of the lysate, optional enrichment of the protein or the protein class of interest, protein separation on acrylamide gels followed by detection with antibodies or MS. Strong detergents such as SDS ensure protein solubilization before clarification, but they need to be removed for enzymatic digestion of in-gel[8], in-solution[9,10] or protein reactor–based[11–14] approaches. Although such approaches are very robust, they necessarily involve several drastic milieu changes of the proteomic samples, with attendant losses, biases and possible introduction of contaminants. Furthermore, they are time-consuming and laborious.

To radically simplify the proteomic sample preparation workflow, we aimed to retain a filter-aided sample preparation (FASP)-like reactor-based method[12] but to avoid the use of strong detergents that are incompatible with proteolytic digestion and LC-MS/MS analysis as such detergents necessitate the use of molecular weight cut-off filters. As the reactor, we used stop-and-go extraction tips (StageTips)[15], which consist of a pipette tip with an inserted C_{18} disc that is usually used for final peptide cleanup before LC-MS/MS. Starting from the established FASP protocol[12], we first processed digests of light and heavy isotope–containing yeast and HeLa cells (encoded by stable isotope labeling by amino acids in cell culture (SILAC)[16]) with or without clarification. We used MaxQuant[17] for quantitative analysis. Most proteins were equally abundant in clarified and unclarified lysate, but for yeast, an outlier population was enriched for Gene Ontology (GO) categories 'intrinsic to membrane' and 'nucleus' in the upper right quadrant (two-sided, false discovery rate–based two-dimensional annotation enrichment[18]; $P = 5.6 \times 10^{-11}$, $P = 8.9 \times 10^{-8}$, respectively), indicating that clarification preferentially depleted these protein classes (**Fig. 1a** and **Supplementary Fig. 1a**). When we

Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. Correspondence should be addressed to M.M. (mmann@biochem.mpg.de).

RECEIVED 24 MAY 2013; ACCEPTED 20 DECEMBER 2013; PUBLISHED ONLINE 2 FEBRUARY 2014; DOI:10.1038/NMETH.2834

Figure 1 | Validation of improvements incorporated in the iST method. (**a**) SILAC-based ratios (normalized to median log ratio of zero as is normally done by MaxQuant) of clarified over nonclarified *S. cerevisiae* lysates before proteolytic digestion. The excluded fraction shows the content in a discrete manner in terms of distance from the density center (see color bar in **d**; e.g., black data points represent 5% and the outermost population. Outlier points in the bottom-right quadrant originate from proteins enriched in nonclarified lysate (**Supplementary Note**). (**b**) Comparison of stepwise over simultaneous lysis and alkylation using SILAC-labeled HeLa cells. Here peptide ratios are plotted unlike in the other panels. (**c**) Peptide identifications from 20 µg starting material fractionated by SAX, SCX and SDB-RPS StageTips. Numbers of unique peptides identified with each approach are shown. (**d**) Comparison of the iST method applied to formalin-fixed HeLa cells and nonfixed cells.



quantified SDS-solubilized (FASP protocol) and urea-solubilized nonclarified HeLa cell lysates, we observed no differences (**Supplementary Fig. 1b**). These experiments showed that SDS can indeed be omitted if there was no prior clarification step, even for difficult-to-lyse samples such as yeast.

In standard proteomic workflows, lysis is performed before the reduction of disulfide bridges with dithiothreitol, which is followed by alkylation of free cysteines. These steps cannot be combined because iodoacetamide reacts with dithiothreitol. We found that the reducing agent tris(2-carboxyethyl)phosphine (TCEP) is compatible with the alkylating agent chloroacetamide, which allows these chemicals to be incorporated directly into the lysis buffer and eliminates the need to perform reduction and alkylation as separate steps. The simplified alkylation procedure was as efficient as the previous multistep reaction (as, for instance, applied in the FASP protocol[12]) and did not bias the results at the peptide or protein level (**Fig. 1b**).

Having combined lysis, reduction and alkylation into a single step, we reasoned that intact HeLa cells could be lysed in a single chemical reactor without interfering with downstream analysis. This dramatically reduced opportunities for contamination, sample loss and sample preparation–related modifications (**Supplementary Fig. 1c**,**d**).

We initially tried separating complex peptide mixtures into six fractions with StageTips containing strong anion exchange (SAX) material (referred to as 'SAX StageTips')[15], which required desalting in C_{18}-containing StageTips before MS analysis. We found that the use of strong cation exchange (SCX) resins and volatile elution buffers allowed us to perform the peptide separation and clean-up in one device; this eliminated the need to use an organic solvent for the activation step, which is otherwise needed to prepare the bead material for peptide binding (**Supplementary Figs. 2–4**). We found that six-fraction SCX outperformed six-fraction SAX and even a three-fraction poly(styrenedivinylbenzene) reverse phase sulfonate (SDB-RPS) approach resulted in higher peptide numbers than the six-fraction SAX approach (**Fig. 1c** and **Supplementary Fig. 5**).

To test whether our in-StageTip (iST) processing method was compatible with fixed samples such as those encountered in biobanks, we fixed HeLa cells with formalin for 15 min before sample processing. We obtained clean peptides with comparable yield to that for nonfixed HeLa cells (**Fig. 1d**). There was little if any protein class–specific change in protein abundance, and total peptide signal after a 15-min fixation was ~63% of the signal without fixation and 33% of the signal after 16 h of fixation.

in-StageTip processing protocol, automation and scaling

The entire iST sample-preparation method is thus performed in a single device, which consists of an enclosed reaction chamber whose bottom part serves as a barrier during processing, and as a filtration and separation medium during elution (**Fig. 2a**,**b**). Cells or other protein material are pipetted into the reaction chamber from above and, depending on their nature, are disintegrated by boiling, sonication or bead milling in a lysis buffer that already incorporates the reduction and alkylation reagent. Guanidinium hydrochloride and sodium deoxycholate have advantages over urea as the lysis agent because of their temperature stability and chemical inertness[19,20]. Excluding digestion, the five manual steps of the final protocol (Online Methods and **Supplementary Video 1**) can be performed in less than 30 min. Starting materials of submicrogram to ~20 µg protein content can readily be processed; for larger amounts, we perform lysis and digestion in a separate tube and add the resulting peptide material to the device.

Samples can be eluted into autosampler vials of the LC-MS/MS system in one step for single-run proteomic analysis or in several elution steps if peptide fractionation is desired. We found the combination of our standard 4 h gradient durations with three-fraction elution (3 × 4 h gradient duration) or six-fraction elution (6 × 4 h gradient duration) to be particularly efficient (data not shown).

We also developed a version of the protocol for use with an in-house–made 96-well device (**Fig. 2c**). Performance with that protocol was indistinguishable from performance using the protocol for the single device (data not shown), and total processing time per sample became negligible. Because of the ease of handling these 96-well devices, we routinely use them for multiple-sample processing.

Copy numbers in *S. cerevisiae* and *S. pombe*

Protein copy numbers are of great interest to the biological and systems biological communities, and we reasoned that a streamlined,

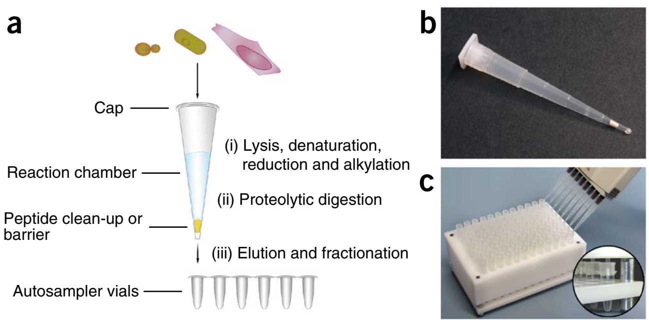

a
Cap
Reaction chamber
Peptide clean-up or barrier
Autosampler vials
(i) Lysis, denaturation, reduction and alkylation
(ii) Proteolytic digestion
(iii) Elution and fractionation
b
c

Figure 2 | Minimal sample-processing protocol performed in an enclosed volume is amenable to automation and scaling. (**a**) Outline of the iST sample-processing method. Cells or other protein material are directly transferred into a StageTip and are processed in three steps. (**b**) Enclosed iST reactor. (**c**) 96-well iST device for multiple-sample processing. Inset, shows StageTips reaching into PCR tubes.

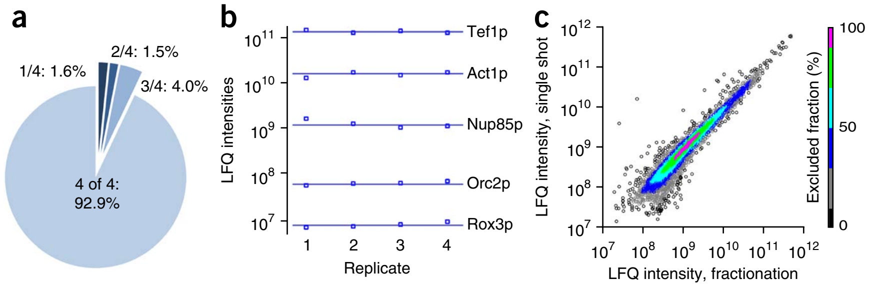
minimal sample-processing method such as the iST method could provide unbiased values. We grew *S. cerevisiae* in four biological replicates and processed them in parallel in the 96-well format (100-μl cultures, each at optical density at 600 nm (OD_{600}) of 0.8). In four single-run analyses together we identified 4,270 protein groups, and we detected 97% of them in at least three of the four replicates with high quantitative reproducibility (34.4% total median sequence coverage and, 46,125 total unique-sequence peptide identifications; **Fig. 3a**,**b** and **Supplementary Table 1**). In our recent yeast proteome analysis in single-run mode, in which we had used the same downstream LC-MS/MS setup[5], the mean identification in each individual run was 4,084 protein groups (33,122 ± 405 (±s.d.) unique-sequence peptide identifications and 23.4% median sequence coverage), whereas using the iST method we identified an even greater number (4,144 protein groups, 37,880 ± 1,771 (±s.d.) unique-sequence peptides and 27.2% median sequence coverage).

SCX fractionation of a yeast sample directly from the reaction device into six autosampler vials, followed by essentially the same LC-MS/MS approach as the single-run analysis, resulted in the identification of 4,577 protein groups, which, to our knowledge, is the largest expressed yeast proteome reported to date. We did not identify any of the 656 dubious open reading frames, which are thought not to represent expressed messages or proteins (Online Methods). Excellent correlation of label-free intensity values with those of a single-run analysis ($R^2 = 0.91$) showed that iST fractionation did not introduce biases, even in the very-low-intensity region (**Fig. 3c**).

The deepest previously reported proteome of exponentially growing *S. cerevisiae* had been measured using five different proteolytic enzymes as well as extensive, column-based SCX fractionation of peptides[21]. We identified 94.9% of the proteins identified in the previous study in our six-fraction data set and also identified an additional 400 proteins, among which intrinsic membrane proteins were significantly enriched ($P = 9.4 \times 10^{-6}$, one-dimensional annotation enrichment[18]; **Fig. 4a**). We next used the label-free MS signal for each protein as a fraction of the total MS signal of the proteome[22] to estimate copy numbers for 4,570 yeast proteins (Online Methods and **Supplementary Table 1**). Copy numbers have previously been established for 21 yeast proteins using synthetic peptide standards[23], and our values agree well within the expected uncertainties ($R^2 = 0.82$; **Fig. 4b**). The most abundant yeast protein we identified, at 1.6×10^6 copies per cell, was the glycolytic enzyme Tdh3p, which is encoded in three genomic loci. The yeast protein with the median abundance value had ~800 copies per cell, and copy number values for 90% of the proteins were within a range of 2,000. The six origin-recognition complex members had a median copy number of ~300 ± 150 (±s.d.), a value that is interesting to compare to the estimated 500 origins of replication in *S. cerevisiae*[24]. We found that more than 763 yeast proteins had less than 100 copies per cell (**Fig. 4c**), a much larger proportion than reported in a classical study of yeast protein copy numbers[25]. This population was significantly enriched for the GO terms 'cell cycle process' and 'DNA repair' ($P < 9 \times 10^{-10}$ and $P < 1.8 \times 10^{-4}$, respectively). For very-low-abundance proteins, a weak MS signal may introduce uncertainties; nevertheless, we measured largely consistent copy numbers for members of the anaphase promoting complex (APC), ~30 APCs per cell. Proteins only present in certain cellular states were often found with very low apparent copy numbers, for example, the cyclin CLB2 (G2/M phase) at 100 copies or the kinase inhibitor FAR1 (G1 phase) at ~50 copies. This illustrates that our data set includes contributions from several different proteomic states.

S. pombe diverged from *S. cerevisiae* more than 400 million years ago, which makes for an interesting comparison. The deepest proteomic study of *S. pombe* used several growth conditions and very extensive, orthogonal fractionation to identify 3,542 proteins[26]. Using the six-fraction iST approach on exponentially growing cells only, we identified 4,087 proteins by searching against the same database as that used in ref. 26 (**Supplementary Table 1**). This represents 80% of *S. pombe* open reading frames and covers 96.5% of the previous proteome as well as 670 additional, generally low-abundance proteins (**Fig. 4d** and **Supplementary Fig. 6**). In comparison to another deep *S. pombe* proteome[27], our *S. pombe* copy numbers agreed very well with data reported for 34 proteins for which isotope-labeled standards had been synthesized

Figure 3 | Quantitative reproducibility of in-depth analysis of *S. cerevisiae* proteome and copy-number estimation. (**a**) Frequency of protein identification of one, two, three or all four biological replicates. (**b**) MS signals (label-free quantification (LFQ) intensities from the MaxQuant output) of five proteins representing the entire dynamic range of measured protein expression in four biological replicates. (**c**) Comparison of LFQ intensities determined in single-shot analysis to LFQ intensities determined in six-fraction analysis.



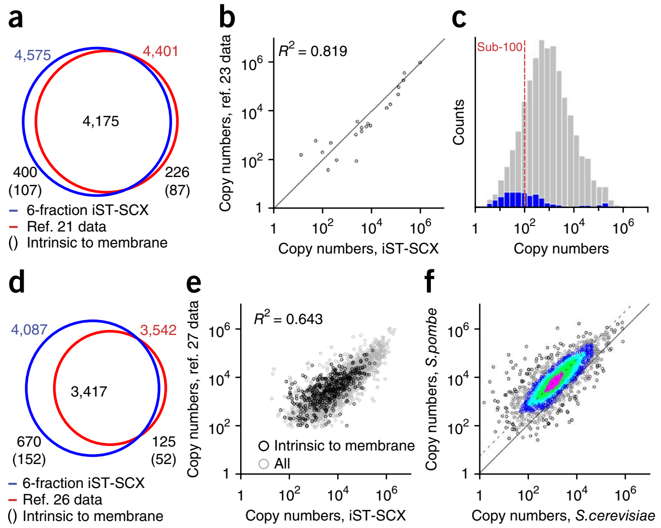


Figure 4 | In-depth analysis of yeast proteomes and estimation of yeast copy numbers. (**a**) Comparison of identified proteins using six-fraction iST-SCX analysis to the previous deepest-coverage experimental *S. cerevisiae* proteome[21]. Number of identified proteins are indicated. (**b**) Correlation of estimated copy numbers using six-fraction iST-SCX analysis of the *S. cerevisiae* data set to previously reported copy numbers obtained using 21 synthetic peptide standards[23]. Diagonal line indicates perfect correlation. (**c**) Distribution of estimated copy numbers for the *S. cerevisiae* data set. Dashed line indicates 100 copies per cell. Blue bars represent proteins identified uniquely in the six-fraction iST-SCX analysis. (**d**) Comparison of identified proteins using six-fraction iST-SCX analysis to proteins reported in a recent study presenting the previous deepest proteome of *S. pombe*[26]. (**e**) Correlation of estimated copy numbers using six-fraction iST-SCX analysis to copy numbers reported in another recent in-depth analysis of *S. pombe*[27]. (**f**) Correlation of estimated copy numbers between *S. pombe* and *S. cerevisiae* orthologs. The excluded fraction is color-coded as in **Figure 3c**. Solid and dashed lines indicate perfect correlation and correlation offset owing to greater protein content per cell, respectively.

($R^2 = 0.89$; **Supplementary Fig. 7**) and there was no apparent bias against any protein class, including intrinsic membrane proteins (**Fig. 4e**). The most abundant proteins had around 10^6 copies per cell, similar to *S. cerevisiae*, but the proportion of proteins with fewer than 100 copies per cell was much lower (17% versus 3%). Median copy number was 5,137, about sixfold higher than in *S. cerevisiae*. The lowest-expressed 5% of the proteome was significantly enriched for replication fork processing–related and DNA repair–related proteins ($P < 1.1 \times 10^{-6}$ and $P < 1.2 \times 10^{-6}$, respectively, one-dimensional annotation enrichment). This fraction of the proteome contains many so-far uncharacterized *S. pombe* open reading frames (59 of 207 proteins; $P < 3.9 \times 10^{-5}$, one-dimensional annotation enrichment).

Despite the compressed dynamic number range of *S. pombe* proteins (**Supplementary Fig. 8**), there was a high rank correlation ($r_s = 0.78$) of the copy numbers of the 1,773 eggNOG orthologs[28] between the yeast species (**Fig. 4f** and **Supplementary Table 2**). Enrichment analysis of the yeast proteomes against each other[18] suggests that almost all functional categories are highly conserved between the species (**Supplementary Fig. 9**). The few outliers, such as copy-number values for vacuolar, cell wall and Golgi apparatus components were significantly higher for *S. cerevisiae* proteins ($P < 10^{-3}$ for all) and they relate to well-known morphological differences between the species.

Copy numbers in the human proteome

Applying the six-fraction workflow to HeLa cells in quadruplicate yielded highly reproducible label-free quantification values ($R^2 > 0.96$, **Supplementary Fig. 10**). We obtained copy numbers for 9,667 human proteins (Online Methods, **Fig. 5a** and **Supplementary Table 1**), which we validated by previous measurements of SILAC-labeled protein fragments[29] ($R^2 = 0.81$; **Fig. 5b**). Median protein expression was 21,000 copies per cell, close to the 18,000 copies that we determined in a previous in-depth HeLa cell proteome study[30] and comparable to the 10,000 copies per cell that had been reported for the U2OS cell line[31]. Our calculations resulted in an estimate of 2.98×10^9 protein molecules per HeLa cell. We observed 581 proteins with more than 10^6 copies, which included histone components, members of the proteosomal core complex, ribosomal proteins, metabolic enzymes and proteins associated with folding, as observed previously[32]. Ranking HeLa cell proteins by copy number revealed that the vast majority (92.6%) were expressed with copy numbers between 100 and 10^6 (**Fig. 5c**). For any two proteins, there was a

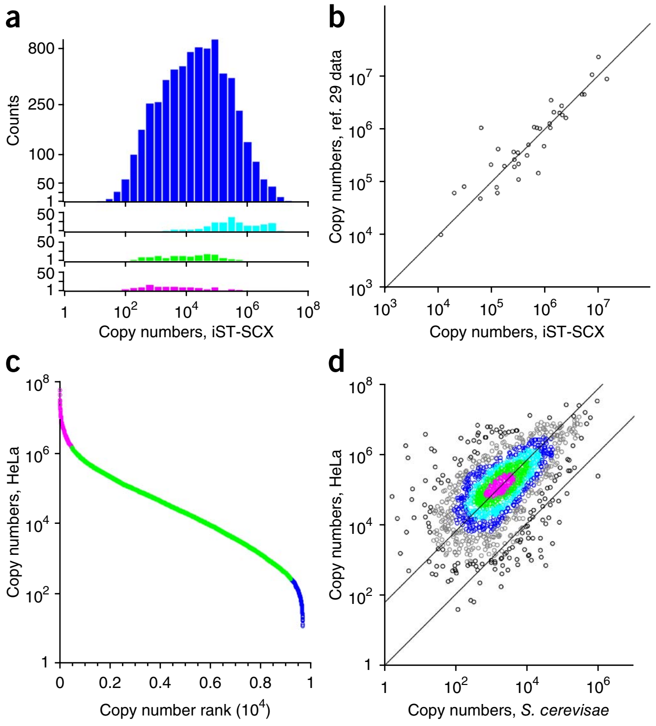


Figure 5 | In-depth analysis of the human proteome and estimation of copy numbers using the iST method. (**a**) Distribution of estimated copy numbers. Turquoise, green and pink bins indicate 'ribonucleoprotein complex' (GO term cellular component), 'protein kinase activity' (GO term molecular function) and 'receptor activity' (GO term molecular function), respectively. (**b**) Correlation of estimated median copy numbers using technical quadruplicates of six-fraction iST-SCX analysis to copy numbers reported in a previous study using SILAC-labeled protein fragments[29]. Line indicates perfect correlation. (**c**) Ranking of HeLa cell proteins by median copy number. Pink, most abundant 5%; green, median 90%; blue, least abundant 5%. (**d**) Correlation of estimated copy numbers between human HeLa cell and *S. cerevisiae* orthologs. Lines indicate offset in copy numbers, which is due to greater protein content per cell in HeLa cells.

64% chance that their copy numbers were within a factor 100 of each other and a 40% chance that they did not differ more than tenfold (**Supplementary Fig. 11**). As expected, protein classes involved in biogenesis such as ribonucleoproteins generally had high copy numbers. Regulatory protein classes, such as those with kinase activity or receptor activity did not have members with very high copy numbers but were otherwise evenly distributed in abundance (**Fig. 5a**). *S. cerevisiae* and human are evolutionary separated by ~10^9 years, and human cells are much larger than yeast cells, but the rank-order correlation of the copy numbers was still high ($r_s = 0.59$, **Fig. 5d**).

At the level of complexes, pathways and individual proteins, the catalytic and scaffold subunits of the phosphatase PP2A have a median copy number of ~80,000, whereas the regulatory subunits that drive specificity were about half as abundant (copy number of 50,000). (See **Supplementary Table 1** for copy numbers mentioned here and below.) The PP2A alpha isoform was more abundant than the beta isoform (80,000 versus 20,000 copies per cell, respectively), confirming previous reports based on northern blots[33]. A main marker of autophagy is MAP1LC3, a protein present in three isoforms, A, B and C. We found that the B isoform was expressed in greater quantity than the A isoform (~400,000 versus ~40) as observed before[34], in accordance with the high basal autophagy levels in this system[35].

Factors involved in DNA replication such as DNA polymerases and helicases were also present in high copy numbers and expressed at comparable levels. Copy numbers for core members of the eukaryotic replicative helicase complex MCM2-7 were ~670,000, much higher than the number of origins of replication. A similar observation already has been made in yeast[36]. In contrast, MCM8, which is part of the MCM8-9 complex and is required for DNA-damage tolerance was expressed at only 1,400 copies. There was also concordance of stoichiometry-adjusted copy numbers among members of the DNA-repair protein complex Mre11-Rad50-Nbs1 (ref. 37) (~87,000, ~82,000 and ~65,000, respectively). In contrast, there were large differences in copy numbers between DNA-repair complexes within the same repair pathway, for instance, Fanconi anemia (FA) associated proteins FANCD2 and FANCI were present in high copy numbers (85,000 and 220,000), whereas members of the FA core complex were much less abundant (median copy number, 2,000). We similarly found an uneven distribution of copy numbers for pathway members involved in repair of DNA double-strand breaks (for example, Rad50-Mre11 and CtIP (RBBP8); ~170,000 and ~1,700, respectively) providing useful insights into pathway architecture.

DISCUSSION

Our minimal iST protocol allows proteome-sample preparation in five pipetting steps and can readily be performed in a 96-well format. We believe that this proteomic method is now even simpler, more robust and faster than ubiquitously used standard procedures such as western blots. Advantages of our workflow are especially apparent in comparison to the sophisticated preparation protocols used to obtain accurate transcriptome results with RNA sequencing (RNA-seq)[38]. The simplicity of this iST method also reduces opportunities for contamination, sample loss and workflow-induced post-translational modifications.

The estimation of absolute protein abundance in a cell type can yield important insight into its biology[2,23,25,29]. Combining the iST processing method with high-resolution, high-mass-accuracy LC-MS/MS, we obtained for *S. pombe* and *S. cerevisiae* the deepest proteomes reported to date and the largest copy number resource for a human cell line, to our knowledge. Total analysis time per proteome was a little more than a day, compared to weeks typically required using more complex workflows[2,21,26,27]. Our analysis revealed that the proteome of exponentially growing *S. cerevisiae* already has contributions from cell-specific states such as phases of the cell cycle or DNA-damage response in a subset of the cells. Copy number estimates from our data in three different species compared very well with accurate absolute quantification using isotope-labeled standards, which showed that a straightforward workflow using no or minimal fractionation and a single protease can comprehensively cover the proteome. This is also supported by a relatively high rank-order correlation ($r_s = 0.64$) to a previous MS-independent data set in which all yeast genes had been tagged and quantified separately[25].

Although cellular proteomes are known to have an extremely large dynamic range, we found that most of the yeast and human proteomes are expressed within a relatively small factor of the median expression number. Finally, we found that copy numbers of proteins are generally highly conserved across vast evolutionary distance, indicating that the functional proteome imposes constraints in addition to the more familiar sequence conservation of individual orthologs.

The *S. cerevisiae*, *S. pombe* and HeLa cell line proteomes obtained with the iST workflow have been uploaded to the MaxQB database[39], where they can easily be visualized and analyzed. We believe the iST method will be useful for quickly and comprehensively producing copy number estimates for a large variety of biological systems, which will be a useful reference for understanding biological processes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The raw mass spectrometric data used in this study and the MaxQuant analysis files are available via proteomeXchange: PXD000269. (raw data and MaxQuant analysis files). MaxQuant database: P004 (protein copy-number estimations).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Braun and P. Georgescu (Ludwig Maximilian University of Munich) for providing us with fission yeast pellets, F. Sacco, M. Steger, D. Walther and M. Räschle for help concerning biological interpretations. M. Hein, A. Dalfovo, C. Schaab and J. Cox helped with figures, videos and bioinformatics analysis of our data sets. Work in M.M.'s laboratory is supported by the Max Planck Society for the Advancement of Science and PROSPECT, a 7th framework program of the European Union (grant agreement HEALTH-F4-2008-201648).

AUTHOR CONTRIBUTIONS

N.A.K. and M.M. developed and invented the method; G.P. and N.N. contributed in the developments; N.A.K., G.P., I.P. and N.N. performed the experiments; and N.A.K., G.P., N.N. and M.M. designed and interpreted the experiments, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Altelaar, A.M. & Heck, A.J. Trends in ultrasensitive proteomics. *Curr. Opin. Chem. Biol.* **16**, 206–213 (2012).
2. de Godoy, L.M. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).
3. Thakur, S.S. *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10**, M110 003699 (2011).
4. Kocher, T., Swart, R. & Mechtler, K. Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Anal. Chem.* **83**, 2699–2704 (2011).
5. 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. *Mol. Cell. Proteomics* **11**, M111 013722 (2012).
6. Yamana, R. *et al.* Rapid and deep profiling of human induced pluripotent stem cell proteome by one-shot NanoLC-MS/MS analysis with meter-scale monolithic silica columns. *J. Proteome Res.* **12**, 214–221 (2013).
7. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, M111 011015 (2011).
8. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).
9. Chen, E.I., McClatchy, D., Park, S.K. & Yates, J.R. III. Comparisons of mass spectrometry compatible surfactants for global analysis of the mammalian brain proteome. *Anal. Chem.* **80**, 8694–8701 (2008).
10. 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.* **7**, 5028–5032 (2008).
11. Manza, L.L., Stamer, S.L., Ham, A.J., Codreanu, S.G. & Liebler, D.C. Sample preparation and digestion for proteomic analyses using spin filters. *Proteomics* **5**, 1742–1745 (2005).
12. Wisniewski, J.R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
13. Ethier, M., Hou, W., Duewel, H.S. & Figeys, D. The proteomic reactor: a microfluidic device for processing minute amounts of protein prior to mass spectrometry analysis. *J. Proteome Res.* **5**, 2754–2759 (2006).
14. Zhou, H., Ning, Z., Wang, F., Seebun, D. & Figeys, D. Proteomic reactors and their applications in biology. *FEBS J.* **278**, 3796–3806 (2011).
15. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
16. Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
17. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
18. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13**, S12 (2012).
19. Poulsen, J.W., Madsen, C.T., Young, C., Poulsen, F.M. & Nielsen, M.L. Using guanidine-hydrochloride for fast and efficient protein digestion and single-step affinity-purification mass spectrometry. *J. Proteome Res.* **12**, 1020–1030 (2013).
20. Leon, I.R., Schwamme, V., Jensen, O.N. & Sprenger, R.R. Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol. Cell. Proteomics* **12**, 2992–3005 (2013).
21. Peng, M. *et al.* Protease bias in absolute protein quantitation. *Nat. Methods* **9**, 524–525 (2012).
22. Wisniewski, J.R. *et al.* Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **8**, 611 (2012).
23. Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B. & Aebersold, R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell* **138**, 795–806 (2009).
24. Nieduszynski, C.A., Hiraga, S., Ak, P., Benham, C.J. & Donaldson, A.D. OriDB: a DNA replication origin database. *Nucleic Acids Res.* **35**, D40–D46 (2007).
25. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
26. Gunaratne, J. *et al.* Extensive mass spectrometry-based analysis of the fission yeast proteome: The *S. pombe* PeptideAtlas. *Mol. Cell. Proteomics* **12**, 1741–1751 (2013).
27. Marguerat, S. *et al.* Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151**, 671–683 (2012).
28. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–D289 (2012).
29. Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M. & Mann, M. A protein epitope signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol. Cell. Proteomics* **11**, O111.009613 (2012).
30. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548 (2011).
31. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549 (2011).
32. Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
33. Cristobal, I. *et al.* PP2A impaired activity is a common event in acute myeloid leukemia and its activation by forskolin has a potent anti-leukemic effect. *Leukemia* **25**, 606–614 (2011).
34. Kar, R., Singha, P.K., Venkatachalam, M.A. & Saikumar, P. A novel role for MAP1 LC3 in nonautophagic cytoplasmic vacuolation death of cancer cells. *Oncogene* **28**, 2556–2568 (2009).
35. Tanida, I., Minematsu-Ikeguchi, N., Ueno, T. & Kominami, E. Lysosomal turnover, but not a cellular level, of endogenous LC3 is a marker for autophagy. *Autophagy* **1**, 84–91 (2005).
36. Forsburg, S.L. Eukaryotic MCM proteins: beyond replication initiation. *Microbiol. Mol. Biol. Rev.* **68**, 109–131 (2004).
37. Williams, R.S., Williams, J.S. & Tainer, J.A. Mre11-Rad50-Nbs1 is a keystone complex connecting DNA repair machinery, double-strand break signaling, and the chromatin template. *Biochem. Cell Biol.* **85**, 509–520 (2007).
38. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
39. Schaab, C., Geiger, T., Stoehr, G., Cox, J. & Mann, M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol. Cell. Proteomics* **11**, M111 014068 (2012).

ONLINE METHODS

HeLa cell culture. Human epithelial carcinoma cells of the line HeLa (ATCC, S3 subclone) were cultured in SILAC DMEM where applicable (PAA Laboratories, E15-086), supplemented with 10% dialyzed FBS (PAA Laboratories, A15-107), 20 mM glutamine (PAA Laboratories, M11-006), 1% penicillin-streptomycin (PAA Laboratories, P11-010), 42 mg/l L-arginine (Sigma-Aldrich, A6969) and 62 mg/l L-lysine (Sigma-Aldrich, L8662). Cells were tested for mycoplasma contamination. For preparation of heavy isotope–labeled peptides, medium contained 42 mg/l [$^{13}C_6{}^{15}N_4$]arginine (Arg^{10}, Cambridge Isotope Laboratories, CNLM-539) and 61 mg/l [$^{13}C_6{}^{15}N_2$]lysine (Lys^8, Cambridge Isotope Laboratories, CNLM-291) instead of the natural amino acids. Cells were cultured for six passages until they were fully labeled. The cells were collected by centrifugation at 200*g* for 10 min, washed once with cold PBS and resuspended in cold PBS. Cell viability and number counts were performed according to the manufacturer using a Countess Automated Cell Counter (Life technologies, C10227).

Yeast cell culture. Budding yeast (*S. cerevisiae*) strains BY4741, YBR115C (*lys2* deletion strain) and fission yeast (*S. pombe*) strain SP286 were acquired from EUROSCARF, Thermo Scientific or Bioneer, respectively. The wild-type strain BY4741 was grown in YPD medium (20 g/l Bacto peptone (BD, 211677), 10 g/l yeast extract (Fisher Scientific, BP1422-2)) supplemented with 2% w/v glucose (Sigma-Aldrich, G7021). SILAC labeling of YBR115C was achieved by growing the cells for at least eight generations in SC medium supplemented with L-$^{13}C_6{}^{15}N_2$-lysine (Cambridge Isotope Laboratories, CNLM-291) and 2% w/v glucose. Fission yeast was grown in YES medium (5 g/l Bacto yeast extract supplemented with 3% w/v glucose and 250 mg/l of each adenine (Sigma-Aldrich, A2786), L-histidine (Sigma-Aldrich, H6034), L-leucine (Sigma-Aldrich, L8912), uracil (Sigma-Aldrich, U1128) and L-lysine (Sigma-Aldrich, L862)). Cells were grown at 30 °C to an OD_{600} of 0.6, harvested by centrifugation at 500*g* for 5 min at 4 °C, washed once with water and stored at −80 °C.

Tryptophan fluorescence emission assay for protein quantification. Protein concentrations were determined by tryptophan fluorescence emission at 350 nm using an excitation wavelength of 295 nm. Briefly, 1 µl of sample was solubilized in 200 µl of 8 M urea, and tryptophan at a concentration of 0.1 µg/µl was used to build a standard calibration curve (0.25–1.5 µl). Protein concentration in samples was estimated considering the emission of 0.1 µg/µl tryptophan equivalent to the emission of 7 µg/µl of human protein extract, assuming that tryptophan accounts for 1.3% on the human protein amino acid composition, on average.

in-StageTip lysis, reduction and alkylation. Quantities of up to 20 µg protein material were loaded directly onto the enclosed StageTips (Eppendorf epT.I.P.S., 0030073266); larger quantities were lysed and digested in a separate vial before loading a StageTip. To avoid clogging, typically 14-gauge StageTip plugs were used. Unless otherwise stated, approximately 10 µg or 20 µg protein starting material was used for single-shot or fractionation sample preparations, respectively (**Fig. 2, Supplementary Video 1** and **Supplementary Table 3**). Cells were lysed in lysis buffer (**Supplementary Table 3**) at a ratio of 1–5 µg protein per 1 µl lysis buffer (*S. cerevisiae*, *S. pombe* and HeLa cells contain approximately 3 pg/cell, 9 pg/cell and 200 pg/cell of protein, respectively). To simplify calculation, yeast cells corresponding to 1 ml culture at OD_{600} = 1 should be lysed in 60 µl lysis buffer, and 10^6 HeLa cells should be lysed in 300 µl lysis buffer. The lysates were boiled for 5 min and then sonicated to denature proteins, shear DNA and enhance cell disruption using a water-bath sonicator for enclosed StageTips (Bioruptor, model UCD-200, Diagenode) for 15 min at level 5, or a Sonifier for large volumes (>500 µl) (model 250, Branson Ultrasonics) for 1 min at duty cycle 20% and output control 3. If bead-milling is desired, an adaptor for a bead-milling system (MP Biomedicals, FastPrep-24) can be constructed by drilling a centered 2 mm diameter hole at the bottom of a 2 ml screw-cap micro tube (Sarstedt AG & Co., 72.694.006). The enclosed StageTip filled with ~100 µl beads (Lysing Matrix Y, MP Biomedicals) can be placed inside the bead-milling adaptor.

Lysates were diluted for digestion using a dilution buffer (**Supplementary Table 3**). The dilution buffer should contain respective amounts of proteolytic enzyme to ensure a ratio of 1:50 (micrograms of enzyme to micrograms of protein). Digestion was performed at 37 °C overnight. Peptides were acidified for C_{18}, SDB-XC, SDB-RPS and SCX materials and basified for SAX material (**Supplementary Table 3**). The StageTips were centrifuged using an in-house-made StageTip centrifuge (identical specifications to the Sonation StageTip centrifuge) for up to 2,000*g*; for higher centrifugation speed, Eppendorf tube adaptors (STH01, Sonation) were used. The StageTip was washed 1–3 times using 100 µl washing buffer depending on the number of plugs (**Supplementary Table 3**). Elutions were performed using 60 µl elution buffer depending on the StageTip material and whether fractionation was intended (**Supplementary Table 3**). All eluted materials were collected in autosampler vials and dried using a SpeedVac centrifuge at room temperature (Eppendorf, Concentrator plus, 5305 000.304). If remnants were visible after drying, the pellet was resuspended in double-distilled water followed by a second drying step. Only SAX elutions needed additional C_{18} desalting. Peptides were resuspended in buffer A* (2% acetonitrile (ACN) and 0.1% trifluoroacetic acid (TFA)) and were briefly sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510).

96-well processing device. A 96-well StageTip holder (length × width × height: 127 mm × 85 mm × 32 mm) was designed by drilling conical holes with the measures of the pipette tips (4 mm diameter) in a spacing corresponding to a 96-well PCR plate (9 mm). The material of the block was polyoxymethylene (POM). A second holder for PCR tubes was designed using a plate of the same material and equal area but 5 mm height. In the second holder, holes were drilled with equal spacing to hold PCR tubes (5 mm diameter). Spacers of 11 mm height above and 6 mm below the PCR holder were placed in the corners to of the two plates to maintain the distance between the StageTip holder and the PCR tube holder and to guarantee a correct alignment of the StageTips and the PCR tubes.

Phase transfer surfactant-aided in-solution sample preparation. In-solution sample preparation was performed as previously described[40] with some adjustments for a more comprehensive comparison of the method. In brief, 1,000, 10,000 and 100,000

doi:10.1038/nmeth.2834

HeLa cells were resuspended in 5 µl 1% (w/v) sodium deoxycholate, 10 mM TCEP, 40 mM 2-chloroacetamide (CAA), 100 mM Tris, pH 8.5, and subsequently lysed by 5 min boiling at 95 °C and sonication (Bioruptor, model UCD-200, Diagenode) for 15 min at level 5. Cell debris were pelleted by centrifugation at 13,200 r.p.m. for 5 min and the clarified lysate was transferred into a new vial. The lysate was diluted 1:10 for LysC-trypsin digestion (0.4 µg of each enzyme in double distilled water), and the digestion was performed overnight at 37 °C. The digest was acidified with 50 µl 2% TFA and sodium deoxycholate was extracted using 50 µl ethyl acetate and vigorous shaking. The organic phase was removed after centrifugation at 13,200 r.p.m. for 5 min. Finally, the peptides were desalted on C_{18} StageTips (**Supplementary Table 3**). The LC-MS set up used for in-solution experiments was the same as described below.

Liquid chromatography and mass spectrometry. Approximately 1 µg or 2 µg of peptides were loaded for 2 h or 4 h gradients, respectively. Peptides were separated on a 50-cm 75-µm inner diameter column packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH). Reverse-phase chromatography was performed with an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific), which was coupled to the Q Exactive mass spectrometer (Thermo Fisher Scientific) via a nanoelectrospray source (Thermo Fisher Scientific). Peptides were loaded in buffer A (0.1% (v/v) formic acid) and eluted with a non-linear 120-min or 240-min gradient of 5–60% buffer B (0.1% (v/v) formic acid, 80% (v/v) acetonitrile) at a flow rate of 250 nl/min. After each gradient, the column was washed with 95% buffer B for 3 min and reequilibrated with buffer A for 3 min. Column temperature was kept at 50 °C by an in-house–designed oven with a Peltier element and operational parameters were monitored in real time by the SprayQc software[41]. MS data were acquired with an automatic switch between a full scan and up to five or ten data-dependent MS/MS scans (topN method). Target value for the full scan MS spectra was 3×10^6 charges in the 300–1,700 *m/z* range with a maximum injection time of 20 ms and a resolution of 70,000 at *m/z* 400. Isolation of precursors was performed with a 1.6 *m/z* window and a fixed first mass of 100.0 *m/z*. Precursors were fragmented by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 25 eV. MS/MS scans were acquired at a resolution of 17,500 at *m/z* 400 with an ion target value of 1×10^6 and a maximum injection time of 60 ms. Repeat sequencing of peptides was minimized by excluding the selected peptide candidates for 45 s.

Data analysis. MS raw files were analyzed by MaxQuant software (version 1.3.10.12) and peak lists were searched either against the human Uniprot FASTA database version of 25 February 2012 (81213 entries), against the *S. cerevisiae* Uniprot FASTA database version of 25 February 2012 (6,649 entries) or the *Saccharomyces* genome database–based *S. cerevisiae* FASTA database orf_trans.20100105 (5,904 entries) or against the *S. pombe* Uniprot FASTA database version of 2 April 2013 (5,096 entries) or *S. pombe* FASTA database version of 2 April 2013 (5,031 entries) and a common contaminants database (247 entries) by Andromeda search engine[42] with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. False discovery rate was set to 0.01 for proteins and peptides (minimum length of 7 amino acids) and was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine, and a maximum of two missed cleavages were allowed in the database search. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 p.p.m. and an allowed fragment mass deviation 20 p.p.m. Quantification of SILAC pairs was carried out by MaxQuant with standard settings and without the requantification option.

Bioinformatics analysis. Data analysis was performed with Perseus software in the MaxQuant computational platform and by R statistical computing environment. All enrichment analysis and analysis of variance tests were performed with Benjamini-Hochberg correction at a false discovery rate of 0.02.

Absolute quantification of protein abundances (copy numbers) were computed using peptide label-free quantification values, sequence length and molecular weight as described before[22] based on a total protein per cell value of 3 pg, 10 pg or 200 pg for *S. cerevisiae*, *S. pombe* or HeLa cells, respectively.

To assign protein orthologs between *S. cerevisiae*, *S. pombe* and HeLa cells, the Uniprot identifier was annotated with its corresponding eggNOG identifier[28]. In case of the same eggNOG identifier for multiple Uniprot identifiers, the median copy number for the corresponding protein groups was calculated. The resulting data set contained information about the UniProt identifier of the identified protein groups, the protein and gene name, the copy number as well as the eggNOG identifier, indicating orthologs between *S. cerevisiae*, *S. pombe* and HeLa cells (**Supplementary Table 2**).

To analyze GO-term differences between orthologs, we used the 2D Annotation Enrichment technique[18] that employs a two-dimensional generalization of the nonparametric two-sample test and uses the Benjamini-Hochberg false discovery rate to correct for multiple-hypotheses testing.

40. Masuda, T., Tomita, M. & Ishihama, Y. Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J. Proteome Res.* **7**, 731–740 (2008).
41. Scheltema, R.A. & Mann, M. SprayQc: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* (11 May 2012).
42. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

doi:10.1038/nmeth.2834

Genome-wide protein expression map accross environmental and evolutionary states in yeast reveals specific adaptations and conservations.

GARWIN PICHLER*, NILS A. KULAK*, SEAN HUMPHREY, MARCO Y. HEIN, IGOR PARON, JÜRGEN COX, NAGARJUNA NAGARAJ AND MATTHIAS MANN

DEPARTMENT OF PROTEOMICS AND SIGNAL TRANSDUCTION, MAX PLANCK INSTITUTE OF BIOCHEMISTRY, AM KLOPFERSPITZ 18, 82152 MARTINSRIED, GERMANY

RUNNING TITLE:

DYNAMIC GENOME-WIDE REGULATION OF PROTEIN EXPRESSION IN YEAST

**Equal contribution*

#Correspondence to M.M. (mmann@biochem.mpg.de)

SUMMARY

Absolute copy numbers of all expressed proteins in changing environmental and evolutionary states would help understanding cellular functions in a system-wide manner. We use a recently developed proteomics platform to quantify the protein-coding genome of budding yeast under diverse environmental conditions and evolutionary states. We identified 91% of all verified proteins, discover novel ones and achieved median peptide sequence coverage of more than 55% of all yeast proteins. This nearly complete protein expression map covers essentially all biological pathways and their variation across metabolic and evolutionary states. Moreover, more than half of the proteome can be phosphorylated. Expression levels of 499 proteins are highly stable under changing physiological conditions and during evolution providing a likely household proteome. Key members of ergosterol and sphingolipid biosynthesis in champagne yeast are significantly up- or down-regulated under various conditions, pinpointing a change of plasma membrane composition during evolutionary adaption. This system-wide study provides a new perspective on biological regulation and provides a paradigm towards comprehensive quantification for complex eukaryotic protein-coding genomes.

HIGHLIGHTS

- Quantification of the complete protein-coding yeast proteome
- Dynamic regulation under environmental stress and evolution
- More than half of the protein-coding proteome can be phosphorylated
- Proteomics pinpoints evolutionary pathway optimization

INTRODUCTION

Yeast is one of the best described eukaryotic model organism for studying highly conserved biological pathways and functions. The complete genome sequence of the budding yeast *Saccharomyces cerevisiae* in combination with many different system-wide biological screens provides insights into the complex regulation and conservation of cellular function (Bader et al., 2003; Ghaemmaghami et al., 2003; Goffeau et al., 1996; Huh et al., 2003; Jorgensen et al., 2003; Tong et al., 2004). In particular, the precise quantification of differences in protein expression is poised to become a key to understand biological phenomena (Mallick and Kuster, 2010; Walther and Mann, 2010).

Recent technology developments have enabled the mapping and quantification of the yeast transcriptome using RNA sequencing (Nagalakshmi et al., 2008), single-molecule sequencing (Lipson et al., 2009) and the genome-wide monitoring of translation via ribosome profiling (Ingolia et al., 2009). Transcriptomic analyses are valued for their low cost, high speed and general accessibility (DeRisi et al., 1997; Schena et al., 1995), however quantified mRNA is only a genetic intermediate and does not provide insights into regulatory processes such as post-translational modifications (Grimsrud et al., 2010; Gygi et al., 1999). In parallel to RNA quantification techniques, large scale proteomics has improved tremendously over the last 20 years, identifying first hundreds and then thousands of proteins (de Godoy et al., 2008; Figeys et al., 1996; Link et al., 1999; Peng et al., 2003; Shevchenko et al., 1996; Washburn et al., 2001). Very recent developments of the proteomics platform applying new generation LC systems and novel high-resolution bench-top mass spectrometers with very high sequencing speeds have enabled the identification of more than 4,000 proteins in single-shot measurements (Hebert et al., 2014; Nagaraj et al., 2012). The single-shot concept is a significant improvement towards higher reproducibility, better quantification, and higher throughput and is a valid third approach between in depth shotgun proteomics employing pre-fractionation and targeted approaches (Picotti et al., 2013) . The largest proteome coverage to date for exponentially growing budding yeast identified 4,575 proteins with 99% certainty, comprising approximately 69% of the complete annotated yeast genome (Kulak et al., 2014).

Uncovering the mechanism underlying the robustness of a system under extreme environments and the evolutionary adaption to changing environments provides valuable insights into the dynamic regulation of cellular functions. Initial research on cellular adaptations was focused on the transcriptional response of yeast populations to environmental stress conditions over time (Causton et al., 2001; Gasch et al., 2000; Hughes et al., 2000; Nagalakshmi et al., 2008). Proteins, however, are the key components maintaining cellular functions and are consequently highly regulated in response to their surrounding milieu and even minor changes in protein abundance can lead to drastic changes in cellular physiology. We here set out to investigate the dynamics of the complete protein-coding genome of *S. cerevisiae* in response to environmental stress conditions and discovered evolutionary adaptions between laboratory and non-laboratory yeast strains. This system-wide study provides new perspectives into system biology and provides an outlook towards comprehensive quantification of all proteins in complex eukaryotic genomes.

RESULTS

Workflow for in-depth quantification of the protein-coding genome of yeast

Recent developments of the MS-based proteomics workflow now enable the accurate quantification of the entire protein content of exponentially growing yeast within a few hours of measurement time (Kulak et al., 2014; Nagaraj et al., 2012). This inspired us to try to identify and accurately quantify the complete protein-coding genome of *S. cerevisiae*. To this end we prepared haploid budding yeast in different physiological cell states such as those found under stress (heat, oxidative, osmotic, DNA damage), metabolic changes (with ethanol and galactose as the sugar source), different cell-cycle phases (cell-cycle arrest in G2/M and G1) and starvation. Additionally, we cultured yeast of different mating types and native, non-laboratory champagne as well as baker's yeast (**Figure 1A**). Harvested yeast cells were processed and proteolytically digested using the minimalistic, encapsulated proteomic sample preparation method recently developed in our group (Kulak et al., 2014). We acquired the proteomes of biological duplicates of all different yeast states in single-run mode (**Figure 1**). To additionally capture peptide sequences covering as much as possible of the entire yeast genome, we also included simple pre-fractionation steps, used additional proteolytic enzymes and enriched phosphorylated peptides from yeast growing under normal and heat stress.

All samples were measured using a high-resolution UHPLC setup coupled to the quadrupole Orbitrap mass spectrometer (Q Exactive) essentially as previously described (Nagaraj et al., 2012). Each single run consisted of a 4h gradient during which more than 25,000 MS and 100,000 MS/MS spectra were acquired (**Figure 1B**). When each duplicate was analyzed separately in MaxQuant at a false discovery rate (FDR) of 1%, we identified 5,199 unique protein groups covering 93% of all verified proteins of the yeast genome (**Figure 1B** and **Table 1**). Although we identified more than 4,000 proteins in each condition, 147 proteins only appeared in a single cell state (**Figure 1C**). Among these cell state specific proteins, the mating factor alpha-1 is known to be expressed and secreted only by cells of mating type alpha, and was only observed in BY4742 (MATα) cells. These and other cell type specific proteins with known function validate the stringency of our MS-based proteomic workflow.

Next we searched our dataset against an *in silico* six-frame translation database of the yeast genome. Although the genomic annotation of *S. cerevisiae* has been studied and redefined over nearly two decades, we identified 12 entirely novel proteins with more than two unique peptide identifications and extremely low posterior-error probabilities (Protein PEP < 3 x 10^{-5}). Two of these novel proteins are of special interest because of their entirely separate genetic coding regions and their very high sequence coverage (80.7% and 48.8%, respectively **Figure 1D**). Moreover, we found evidence for a special protein isoform of Abp140p, which is known for its translational frame-shift and its actin-binding and methyltransferase activity (Noma et al., 2011). The truncated isoform which results from regular translation of the mRNA code was previously believed to be non-existing, yet our mass spectrometric data proves the existence of the smaller truncated isoform (Supplementary Data).

VERY LARGE PEPTIDE SEQUENCE COVERAGE OF THE PROTEIN-CODING YEAST GENOME

In order to quantitatively compare the dynamics of the protein-coding genome, we jointly analyzed all 304 LC-MS/MS files in the MaxQuant environment (Cox and Mann, 2008). This analysis resulted in an average of 38,317±7,249 unique peptide identifications for the single runs. On average 39,778±5,471 peptides were identified in samples using Lys-C as proteolytic enzyme, while samples digested with trypsin, and especially Asp-N, and Glu-C resulted in lower identification rates. In total 132,053 peptides were identified (**Figure 2A**). The median sequence coverage of identified proteins was 53.9%, which is very high for shotgun proteomic experiments. Starvation and sporulation conditions as well as non-laboratory yeast strains resulted in lower sequence coverage compared to the other conditions (**Figure 2B**). To assess the completeness of our data, we compared our identifications with all theoretical peptides of Lys-C with a minimal length of 7 amino acids (see Materials and Methods). In our combined data set, we identified 60,589 out of the 124,506 theoretical peptides. This represents the most extensive, high-confidence *S. cerevisiae* peptide library to date and complements the previously synthesized yeast library comprising 28.000 tryptic peptides (Picotti et al., 2013). We identified more than half of these synthesized peptides, even though that library consists of tryptic and our measurements mainly of Lys-C peptides. The total set of identified peptides showed no detectable bias due to their physiochemical properties such as protein hydrophobicity (**Figure 2D**) and even analytically difficult peptides with very basic isoelectric point values were evenly covered (**Figure 2E**).

COVERAGE OF THE PROTEIN-CODING YEAST GENOME

Using the “match between runs” feature in MaxQuant, we identified 4,268±243 protein groups per run and only a very low number of them (567±123) were identified by protein unique peptides (**Figure 3A**). In the combined dataset, 5,015 proteins were identified with 99% certainty, excluding common contaminants, covering around 90% of the protein-coding genome (annotated as “verified” and “uncharacterized” in SGD) (**Figure 3B**). Next, we compared our complete protein data set against our previous in-depth study (Kulak et al., 2014). This showed that we here extended the previously reported protein identifications by 475 proteins enriched for “ammonium transmembrane transporter activity” (P = 5.7 x 10^{-6}), “nitrogen utilization” (P = 9.9 x 10^{-5}) and “reproductive processes” (P = 1.4 x 10^{-5}) and observed a remarkable overlap of 4,527 protein groups. The missing 37 proteins were reported as low abundant with an estimated median copy number of only 13 (**Figure 3C**).

The saccharomyces genome database (SGD) classifies 5,056 and 753 protein coding open reading frames (ORFs) as “verified” and “uncharacterized” respectively (Cherry et al., 2012). Here, we identified 4,605 (91%) “verified” and 367 (49%) “uncharacterized” ORFs and covered 76% of all ORFs in the *S. cerevisiae* database (**Table 1**). Additionally, a subset are classified as “dubious” (691 ORFs), “transposable elements” (90 ORFs), “pseudogenes” (18 ORFs) and “silenced” (4 ORFs), resulting in a total number of 6,612 annotated ORFs in the *S. cerevisiae* database. Of special interest are the "dubious" annotations, which are described as unlikely to encode an expressed protein. They therefore provide an independent means to estimate false positive protein identifications (de Godoy et al., 2008). The combined data set, as described above only identified one dubious ORF, whereas based on a 1% false discovery rate (FDR) 5 identifications would have been acceptable. Remarkably, applying a highly stringent 0.1% FDR cut-off did not remove the single dubious ORF identified.

Notably, around 91% and 97% of all measured "verified" or "uncharacterized" ORFs were identified by at least two peptide sequences. GO enrichment analysis of 451 proteins, which are classified as "verified" and which were not identified in our dataset, revealed terms such as "ascospore wall assembly" ($P = 5.3 \times 10^{-26}$), "asparagine catabolic process" ($P = 6.1 \times 10^{-5}$) and "condensed chromosome" ($P = 3.0 \times 10^{-8}$) (**Figure 3D**). These are pathways that are expected to be active under very specific conditions. As described above, we identified the Mating factor alpha-1 exclusively in BY4742 cells in a combined dataset of separately processed LC-MS/MS files while the identification did not pass the FDR cut-off of the combined analysis (**Figure 1B**).

Our workflow using yeast grown under all relevant physiological conditions and the addition of native champagne and baker's yeast identified 91% of all "verified" ORFs, covering almost all biological pathways and functions annotated in the Kyoto Encyclopedia of Genes and Genomes (KEGG: 95% coverage) database and Gene Ontology (GO cellular-component: 85%; GO molecular-function: 92%; and GO biological process: 91%) categories with more than 10 members (**Supplementary Table 1; KEGG, GO**). The complete dataset completely covers a remarkable 57 (50%) KEGG, 282 (34%) GOCC, 121 (30%) GOMF and 454 (34%) GOBP annotated pathways or functions (**Figure 3E**). Interestingly, at least 80% of the proteins annotated in one specific pathways or function are identified in 89% of all pathways or functions and no pathways were covered with less than 25% unless it was entirely absent (**Figure 3F**). This data demonstrates that at the pathway level, our yeast proteome achieved near complete coverage of all known biological pathways and functions. The pathways with most missing proteins belong to processes involved in "ascospore wall" or "non-homologous end-joining"**,** which again reflect very specialized developmental and stress programs. Note that we restricted the analysis to pathways consisting of at least 10 proteins, as described previously (Nagaraj et al., 2012).

We next used the label-free MS signal for each protein as a fraction of the total MS signal of the proteome to estimate copy numbers according to the total protein abundance (TPA) calculation (Wisniewski et al., 2012). Comparison of our estimated copy numbers to previous published transcriptome datasets using single-molecule sequencing (Lipson et al., 2009) shows a higher dynamic-range and abundance of protein copy numbers compared to transcripts. Estimated copy numbers were found in the range from 1 to 1.000.000 copies per cell. In contrast, the dynamic range of mRNA transcripts spans 4 orders of magnitude. On average, proteins were around 16 times more abundant than their respective mRNAs similar as in a study for *S. pombe* (Marguerat et al., 2012). The highest and lowest difference between mRNA and protein levels were detected for Swh1p, involved in lipid transport, and Met17p, involved in amino-acid biosynthesis, respectively.

DYNAMIC REGULATION OF THE PROTEIN-CODING GENOME

Next we investigated the system-wide proteomic response to environmental stress and during evolution. Despite many microarray studies of yeast under different environmental stress conditions (Causton et al., 2001; Gasch et al., 2000), no comprehensive in-depth proteomic study of different physiological conditions has been reported. Here, we have quantified the proteomes in response to different stress conditions, in different mating types, in cell-cycle stages, during metabolic changes, under starvation and in champagne and baker's yeast. To compare the label-free proteomes of all different measurements to each other, we performed a principal component analysis (PCA) as described previously (Deeb et al., 2012). This revealed common expression differences between conditions belonging to either stress, mating type, cell-cycle, metabolism, starvation and native yeast (**Figure 4A**). Baker's yeast for instance, which was directly processed from a refrigerated dry-yeast block, clustered with the starvation condition, and is distant to other native yeast measurements.

We then filtered for proteins that are significantly changing determined by analysis of variance with correction for multiple hypothesis testing. Hierarchical clustering was done with the median LFQ intensity between two biological replicates which was normalized by Z-score. Similar to the PCA analysis, conditions belonging to either stress, mating type, cell-cycle, metabolism, starvation and native yeast showed the highest similarity (**Figure 4B**). Cluster enrichment analysis revealed protein changes for terms such as “galactose metabolism” and “reproductive process” in yeast growing in galactose or arrested in G1 phase of the cell-cycle, respectively (**Supplementary Figure 1; Profile Plots**). Comparing the dynamic range of proteins from ORFs annotated as “verified” or “uncharacterized” showed a smaller dynamic range, spanning around 2 orders of magnitude, for “uncharacterized” ORFs. Moreover “uncharacterized” ORFs were mostly found in the range of 100 to 10.000 copies per cell, whereas “verified” ORFs were found between 1 to 1.000.000 copies per cell (**Figure 4C**). Interestingly, Fisher’s exact test in each protein abundance quartile of BY4741 revealed enrichment of low abundant protein groups involved in “response to nutrients” ($P = 1.5 \times 10^{-5}$). In contrast, low abundant proteins identified in champagne yeast were enriched for “mitotic chromosome condensation” ($P = 7.7 \times 10^{-6}$) and “ubiquitin-dependent endocytosis” ($P = 4.5 \times 10^{-5}$) (**Figure 4D**).

In addition to protein identifications, we used an optimized enrichment strategy for phosphorylated peptides to identify an in-depth yeast phosphoproteome. The deepest previously reported data set of yeast phosphosites was constructed by consolidating twelve publicly available phosphoproteomes (Amoutzias et al., 2012). We identified 80% of the reported phosphorylated proteins and identified an additional 499 new ones (**Figure 4E**). The majority of the identified 13,262 class I phosphorylation sites in our comprehensive data sets are on serine (84.7%) followed by threonine (14.9%). Interestingly, we did find a small but confidently identified subset of tyrosine phosphorylated peptides (0.4%), confirming the extremely low extent of tyrosine phosphorylation (0.027%) reported in a previous study (Chi et al., 2007) (**Figure 4F**). Proteins with tyrosine phosphorylation sites are significantly enriched for “kinase activity” ($P = 5.0 \times 10^{-7}$). At least half of the proteome can be phosphorylated (2,408 phosphoproteins) and the abundances of the phosphoproteins identified similarly span the entire range of measured protein abundances (**Figure 4G**).

THE HOUSEHOLD PROTEOME

Cells have to adapt to changing environmental conditions and can alter the protein expression program to maintain cellular functions. This environmental stress response program includes genes whose expression is stereotypically altered during stressful environmental changes (Gasch et al., 2000). In contrast, housekeeping genes are typically constitutive genes that are required for maintaining basic cellular functions and are expressed at relatively constant levels under different physiological conditions. Based on the differential expression of proteins during environmental stress or due to evolutionary changes we defined different classes in the context of the household proteome. The maximum fold change of protein expression in comparison to normal growth of the laboratory wild-type strain (BY4741) for proteins identified in all measurements were plotted against the negative log p-value calculated using a multiple sample ANOVA test (**Figure 5A**). Class I household proteins were regulated up to 2-fold with a negative log p value up to 2 and are considered as tightly regulated. 499 Class I household proteins were enriched for pathways such as "GTPase activator activity" (P = 4.6 x 10^{-5}), "establishment of protein localization" (P = 1.8 x 10^{-6}) and "structural constituent of ribosome" (P = 3.1 x 10^{-6}) (**Figure 5B**). In contrast, Class IV proteins, which were identified as highly up- or down-regulated, were enriched for biological pathways such as "cytochrome-c oxidase activity" (P = 2.8 x 10^{-7}), "respiratory chain complex IV" and (P = 3.5 x 10^{-6}) "TCA cycle" (P = 3.3 x 10^{-6}). Notably, Class I household proteins were identified across the entire dynamic range of estimated copy numbers (**Figure 5C**). No proteins below 100 copies per cell were determined as Class I household proteins. The Saccharomyces Genome Deletion Project reported 1,156 genes as essential for growth on rich glucose media (Giaever et al., 2002; Winzeler et al., 1999). We identified 684 of these essential genes in all conditions and 171 (25%) are found to be Class I, 315 (46.05%) Class II, 178 (26.02%) Class III and a small subset of 20 (2.92%) Class IV.

In general, around 50% more proteins are up- than down-regulated during environmental stress or during evolution in comparison to standard laboratory conditions (**Figure 5D**). Enrichment analyses revealed pathways such as "aerobic respiration" (P = 3.3 x 10^{-5}) and "establishment of localization" (P = 6.3 x 10^{-5}) or "ribosome" (P = 2.3 x 10^{-5}) and "glycosyltransferase" (P = 1.4 x 10^{-4}) as up- or down-regulated, respectively.

A PROTEOMIC VIEW OF METABOLIC BIOSYNTHESIS PATHWAYS

Our deep quantitative dataset and the resulting nearly complete coverage of all annotated KEGG and GO terms provide a detailed insight into the regulation of cellular functions in response to environmental stress conditions. Based on the sum of the 'relative abundance' of individual enzymes associated with specific metabolic pathways, we defined a certain metabolic pathway to be up- or down-regulated. The yeast plasma membrane regulate the selective uptake and/or secretion of solutes and maintains the structure and rigidity of the cell (van der Rest et al., 1995). The yeast plasma membrane consists of phospholipid, sphingolipid and ergosterol and the biosynthesis of these membrane components is highly conserved between eukaryotes. Interestingly, in champagne yeast proteins involved in the biosynthesis of sphingolipid and ergosterol are significantly down- and up-regulated, respectively (**Figure 6A, B, C**). Moreover, a direct comparison of all proteins involved in ergosterol biosynthesis between champagne yeast and BY4741 showed a significant up-regulation of nearly all members. Ergosterol, the yeast homolog of cholesterol, is an essential component of yeast cells, maintains the membrane integrity and was investigated as an important factor for ethanol tolerance of yeast cells (Swan and Watson, 1998). Sphingolipids serve as components of membrane rafts and regulate numerous key cell functions. A recent publication demonstrated that enzyme activities involved in sphingolipid biosynthesis decreased upon heat stress (Chen et al., 2013).

In response to nitrogen starvation in the presence of a poor carbon source, diploid yeast cells produce haploid cells through the development program of sporulation, involving meiosis and spore morphogenesis (Chu et al., 1998). Interestingly, we found that enzymes involved in the tricarboxylic acid cycle (TCA cycle) are up-regulated during sporulation with a peak after 14 hours (Figure 6F). The TCA cycle combines catabolic and anabolic functions and generates energy through the oxidation of acetate derived from carbohydrates, fats and proteins into carbon oxide and ATP. A direct comparison of all proteins involved in the TCA cycle between yeast cells grown under normal condition and during sporulation, revealed an up-regulation of every factor (Figure 6G and Figure 6H), demonstrating an unexpected link between the TCA cycle and sporulation in yeast. Notably, a *Bacillus subtilis* mutant with a deletion in the gene encoding isocitrate dehydrogenase had greatly reduced ability to form the polar division septum (Matsuno et al., 1999).

DISCUSSION

MS-BASED PROTEOMICS FOR THE STUDY OF GENOME-WIDE EXPRESSION

Here, we nearly comprehensively quantified the entire proteome of yeast in fundamental cellular states as well as in champagne and baker's yeast. We applied very recently developed MS-based technologies to identify peptides covering around 50% of those encoded in the genome which resulted in the identification of approximately 90% of all protein-coding genome sequences. We further identified two entirely new yeast genes, and also observed an expressed, truncated isoform of a protein previously believed to be non-existent. Our data constitute a proteomic library, representing by far the deepest accurately quantified *S. cerevisiae* proteome, providing a valuable resource for follow-ups studies and comprehensive insights into the cellular regulation and evolutionary adaption to environmental stress conditions. Moreover, because of its very large coverage, this library provides a resource for targeted as well as non-targeted proteomics. The complete data set has been deposited in the publicly available MaxQB database, which provides a user-friendly interface to directly access the whole dataset (Schaab et al., 2012).

S. cerevisiae is the best studied eukaryotic model organism and has been used for a large number of system-wide studies, which identified almost all primary metabolites, enzymes and metabolic pathways. In addition, the high conservation of genes and regulatory mechanism between *S. cerevisiae* and more complex eukaryotes has made yeast a main model organism for mathematical modelling of complex biological pathways. However, research on cellular regulation of gene expression in yeast has so far focused on the transcriptional response of yeast populations to environmental stress conditions over time (Causton et al., 2001; Gasch et al., 2000; Hughes et al., 2000; Nagalakshmi et al., 2008). Here we added a comprehensive description of the dynamic proteome in many different metabolic and other cellular states. We found that yeast expresses more than 4,000 proteins in each of these states, therefore the proteomes of different cellular states are characterized by quantitative rather than qualitative differences. This is similar to what has been observed in protein expression in mammalian cell and tissue types (Geiger et al., 2012; Lundberg et al., 2010). Despite the substantial quantitative differences across the conditions there still were a large number of stably expressed proteins. We found that 499 proteins did not exceed two-fold regulation across all conditions, arguing for a household function. Interestingly, these proteins span the entire abundance range and are involved in diverse physiological functions such as "GTPase activator activity" and "establishment of protein localization". Another interesting subset of 201 proteins is dynamically regulated and correspond to orthologs of uncharacterized human proteins. This set could provide insights into their human counterpart. For example, the voltage-gated potassium channel subunits beta-1 and

-3 (KCNAB and KCNAB3) are orthologous to a putative pyridoxal reductase (Uniprot ID: Q06494), which is strongly down-regulated during G1 cell-cycle arrest, suggesting a link to the cell cycle. In addition, 147 proteins are exclusively expressed in one specific cellular state.

Besides the identification of a large number of proteins in parallel in a single experiment, our analyses open up the possibility of investigating amino acid substitutions resulting from single nucleotide polymorphism (SNPs) and alternative splicing of mRNA transcripts. SNPs are the most abundant forms of genomic sequence variation among populations of individuals (Altshuler et al., 2005; Gibbs et al., 2003; Sachidanandam et al., 2001). Our MS-base proteomic workflow identified peptide sequences that cover more than 50% of the yeast genome. This suggests that the same technology could realistically identify many SNPs and their regulation at the protein level in mammalian cells or tissues. Moreover, the deep coverage allows the study of alternative splicing events that produce multiple protein isoforms from individual genes, generating complex proteomes (Matlin et al., 2005).

In conclusion, our combination of a minimalistic, encapsulated pipette-based preparation of digested yeast cell lysates, the label-free quantification technology and single-shot proteomics enables streamlined and precise system-wide analysis. Our study presents a new level of proteomic coverage of a eukaryotic model organism and demonstrates a paradigm for future mammalian systems, including clinical applications in human diseases.

EXPERIMENTAL PROCEDURES

YEAST CELL CULTURE

Unless otherwise noted, cells were grown at 30°C in YPD media (20 g/L Bacto^{TM} peptone (BD, 211677), 10 g/L yeast extract (Fisher Scientific, BP1422-2) supplemented with 2% w/v glucose (Sigma-Aldrich, G7021). Budding yeast (*Saccharomyces cerevisiae*) strains BY4741, BY4742 and BY4743 were acquired from EUROSCARF, Germany. Commercially available champagne (Arauner, Germany; Art.Nr. 0015) and baker’s yeast (Wieninger Hefe; Germany) were used. Champagne yeast was grown at 27°C in YPD media supplemented with 2 or 20% w/v glucose. Baker’s yeast was either grown under normal conditions or directly processed after purchase. For cell cycle arrest in G1-phase or G2/M-phase, BY4741 cells were grown to early log phase and incubated with either 10 µg/mL alpha-factor for 3 hours or with 15 µg/mL nocodazole for 3h, respectively. BY4741 cells were grown for 72 hours to reach stationary phase. For heat stress, BY4741 cells were grown to early log phase at 30°C, pelleted, resuspended in YPD at 37°C warm and incubate for 45 minutes at 37°C. To test different sugar sources, YP media was supplemented with either 2% ethanol or galactose . For oxidative stress, BY4741 cells were grown to early log phase and incubated with 1 mM menadione for 1 hour. For sporulation, BY4743 cells were grown to late log phase, pelleted and incubated in sporulation media (1 g/L yeast extract (Fisher Scientific, BP1422-2), 10 g/L KCl supplemented with 0.5% w/v glucose (Sigma-Aldrich, G7021)). Cells were harvested after 2, 7, 14 and 72 hours of incubation. For nitrogen starvation, BY4741 cells were grown to late log phase, pelleted and incubated in sporulation media (6.8 g/L Bacto^{TM} yeast nitrogen base (BD, 291940)).

20 mg/L uracil, 250 mg/L ammonium sulfate supplemented with 20% w/v glucose (Sigma-Aldrich, G7021) and harvested after 3 hours. In addition, synthetic SCD media was used instead of nitrogen starvation media. For DNA damage stress, BY4741 cells were grown to early log phase and incubated with 1 mM cisplatin for 1 hour. For osmotic stress, BY4741 cells were grown to early log phase, 0.4M NaCl was added to the media and cells were harvested after 5 or 20 minutes. In general, cells were grown until an OD600 of 0.6 was reached, harvested by centrifugation at 500 g for 5 minutes at 4°C, washed once with water and stored at -80°C. Protein concentrations were determined by tryptophan fluorescence emission at 350 nm using an excitation wavelength of 295 nm.

iST LYSIS, REDUCTION AND ALKYLATION

Sample preparation was done as described in (Kulak et al., 2014). Briefly, cells were lysed in GdmCl lysis buffer (6 M GdmCl, 10 mM TCEP, 40 mM CAA, 100 mM Tris pH 8.5), sonicated and diluted 1:3 or 1:10 with dilution buffer (10 % ACN 25 mM Tris pH 8.5) for LysC or trypsin digestion, respectively. For GluC or AspN proteolytic digestion, the lysate was diluted 1:10 with 5 % ACN 25 mM Tris pH 7.8 or 10 % ACN 25 mM Tris pH 8.5, respectively. The dilution buffer contained appropriate amounts of proteolytic enzyme to ensure a ratio of 1:50 (µg enzyme : µg protein). Digestion was performed overnight at 37°C or 25°C for GluC. Peptides were acidified, loaded on either SCX or SDB-RPS material and eluted or fractionated. All fractions were collected in autosampler vials and dried using a SpeedVac centrifuge (Eppendorf, Concentrator plus, 5305 000.304). Peptides were resuspended in buffer A* (2 % ACN, 0.1 % TFA) and were briefly sonicated (Branson Ultrasonics, Ultrasonic Cleaner Model 2510).

PHOSPHOPEPTIDE ENRICHMENT

Cells were lysed in GdmCl buffer and proteolytically digested using LysC and trypsin (1:70). We acidified 5 mg of peptides per biological replicate with 0.5% TFA, centrifuged and loaded on equilibrated 100 mg (3cc) SepPak C18 cartridges. After washing with 1% TFA, peptides were eluted in 2 mL SepPak elution buffer (75% ACN, 0.1% TFA). Phosphorylated peptides were incubated with a 10-fold excess (10 x peptide quantity) of TiO2 beads, resuspended in loading buffer (80% ACN, 6% TFA) in a Bioruptor for 5 minutes at 4°C. Beads were pelleted, resuspended in 400 µL wash buffer (60% ACN, 1% TFA) and washed additionally 3 times with 200 µL wash buffer. Beads were resuspended in 100 µL transfer buffer (80% ACN, 0.5% acetic acid) and moved to C8 StageTips. Peptides were eluted 3 times with 20 µL elution buffer (40% ACN, 15% NH4OH), concentrated to 20 µL using a SpeedVac and acidified with 1 µL 100% TFA. Peptides were loaded on equilibrated SDB-RPS StageTips, washed with 100 µL 0.1% TFA and eluted with 60 µL SDB-RPS elution buffer (80% ACN, 5% NH4OH). Samples were concentrated to 2 µL using a SpeedVac centrifuge and briefly sonicated after adding 6 µL buffer A*.

LIQUID CHROMATOGRAPHY AND MS

Liquid chromatography and MS were performed as described (Kulak et al., 2014). Briefly, approximately 2 µg pf peptides were loaded for 4h gradients separated on 50-cm columns. Reverse phase chromatography was performed with an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific), coupled to the Q Exactive mass spectrometer (Thermo Fisher Scientific) via a nanoelectrospray source (Thermo Fisher Scientific). Peptides were loaded in buffer A (0.1% (v/v) formic acid) and eluted with a nonlinear 240-min gradient or a 120-min gradient for phosphopeptides. Operational parameters were real-monitored by the SprayQC software (Scheltema and Mann, 2012).

DATA ANALYSIS

MS raw files were analyzed by MaxQuant software (version 1.4.1.6) (Cox and Mann, 2008) and peak lists were searched against the *S. cerevisiae* Uniprot FASTA database version 2/25/2012 (6649 entries) and a common contaminants database (247 entries) by the Andromeda search engine (Cox et al., 2011) with cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. MS raw files for phosphoproteome analyses were additionally searched with phosphor (STY) as variable modification. False discovery rate (FDR) was usually set to 0.01 for proteins and peptides (minimum length of 7 amino acids) and was determined by searching a reverse database. Enzyme specificity was set as C-terminal to Arg and Lys, and a maximum of 2 allowed missed cleavages. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation 20 ppm. Analyzed RAW files will be deposited at PRIDE (http://www.proteomeexchange.org).

BIOINFORMATICS ANALYSIS

Data analysis was performed with the Perseus software in the MaxQuant computational platform and in the R statistical computing environment. All enrichment analyses and analyses of variance tests were performed with Benjamini-Hochberg correction at a false discovery rate of 0.02. Categorical annotation was supplied in form of Gene Ontology (GO) biological process (BP), molecular function (MF) and cellular component (CC), as well as participation in a KEGG pathway. All annotations were extracted from the UniProt database. Hierarchical clustering and 2D annotation enrichment were based on label free intensities (LFQ) of the samples (Luber et al., 2010). Data was imputed by creating a Gaussian distribution of random numbers with a standard deviation of 30% in comparison to the standard deviation of measured values, and one standard deviation down-shift of the mean to simulate the distribution of low signal values. Two sample t-tests were performed with FDR=0.05. Hierarchical clustering of significantly different proteins was performed after z-score normalization. Absolute quantification of protein abundances (copy numbers) were computed using peptide label free quantification values, sequence length and molecular weight as described before (Wisniewski et al., 2012) based on a normalization using measured intensities of histone peptides. Relative protein abundance (mass) was calculated as described before (Wisniewski et al., 2012) (Wisniewski et al., 2012). Briefly the intensity of individual protein was divided by summed intensity of the all proteins.

AUTHOR CONTRIBUTIONS

G.P., N.A.K. and M.M. conceived and designed the experiments. G.P., N.A.K. and I.P. performed the experiments, G.P. and S.H. measured and analyzed the phosphoproteome and G.P., N.A.K., N.N. and M.M. interpreted the experiments. M.Y.H. and J.C. provided tools for bioinformatical analyses. G.P., N.A.K. and M.M. wrote the manuscript.

ACKNOWLEDGEMENT

Alpha-factor was kindly provided by Dr. Zuzana Storchova (MPI, Martinsried). We thank the whole Mann laboratory for helpful discussions. We are thankful to Korbinian Mayr for technical assistance. Work in M.M.'s laboratory is supported by the Max Planck Society for the Advancement of Science and by the European Union 7th Framework project PROSPECTS (Proteomics Specification in Time and Space, grant HEALTH-F4-2008-201645) at the MPI, Martinsried.

REFERENCES

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., Donnelly, P., Gibbs, R.A., Belmont, J.W., Boudreau, A., Leal, S.M.*, et al.* (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Amoutzias, G.D., He, Y., Lilley, K.S., Van de Peer, Y., and Oliver, S.G. (2012). Evaluation and properties of the budding yeast phosphoproteome. Molecular & cellular proteomics : MCP *11*, M111 009555.

Bader, G.D., Heilbut, A., Andrews, B., Tyers, M., Hughes, T., and Boone, C. (2003). Functional genomics and proteomics: charting a multidimensional map of the yeast cell. Trends in cell biology *13*, 344-356.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S., and Young, R.A. (2001). Remodeling of yeast genome expression in response to environmental changes. Molecular biology of the cell *12*, 323-337.

Chen, P.W., Fonseca, L.L., Hannun, Y.A., and Voit, E.O. (2013). Coordination of Rapid Sphingolipid Responses to Heat Stress in Yeast. Plos Comput Biol *9*.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R.*, et al.* (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res *40*, D700-D705.

Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O.G., and Hunt, D.F. (2007). Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America *104*, 2193-2198.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology *26*, 1367-1372.

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., 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*, 1251-1254.

Deeb, S.J., D'Souza, R.C., Cox, J., Schmidt-Supprian, M., and Mann, M. (2012). Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Molecular & cellular proteomics : MCP *11*, 77-89.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

Figeys, D., Ducret, A., Yates, J.R., 3rd, and Aebersold, R. (1996). Protein identification by solid phase microextraction-capillary zone electrophoresis-microelectrospray-tandem mass spectrometry. Nature biotechnology *14*, 1579-1583.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Molecular biology of the cell *11*, 4241-4257.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Molecular & cellular proteomics : MCP.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-741.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B.*, et al.* (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387-391.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H.M., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y.*, et al.* (2003). The International HapMap Project. Nature *426*, 789-796.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M.*, et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Grimsrud, P.A., Swaney, D.L., Wenger, C.D., Beauchene, N.A., and Coon, J.J. (2010). Phosphoproteomics for the Masses. Acs Chem Biol *5*, 105-119.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol Cell Biol *19*, 1720-1730.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2014). The one hour yeast proteome. Molecular & cellular proteomics : MCP *13*, 339-347.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H.Y., He, Y.D.D.*, et al.* (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109-126.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686-691.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218-223.

Jorgensen, P., Breitkreutz, B.J., Breitkreutz, K., Stark, C., Liu, G., Cook, M., Sharom, J., Nishikawa, J.L., Ketela, T., Bellows, D.*, et al.* (2003). Harvesting the genome's bounty: integrative genomics. Cold Spring Harbor symposia on quantitative biology *68*, 431-443.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nature methods.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., and Yates, J.R., 3rd (1999). Direct analysis of protein complexes using mass spectrometry. Nature biotechnology *17*, 676-682.

Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P., and Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. Nature biotechnology *27*, 652-658.

Luber, C.A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M.*, et al.* (2010). Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. Immunity *32*, 279-289.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenas, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol *6*, 450.

Mallick, P., and Kuster, B. (2010). Proteomics: a pragmatic perspective. Nature biotechnology *28*, 695-709.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bahler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell *151*, 671-683.

Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Bio *6*, 386-398.

Matsuno, K., Blais, T., Serio, A.W., Conway, T., Henkin, T.M., and Sonenshein, A.L. (1999). Metabolic imbalance and sporulation in an isocitrate dehydrogenase mutant of Bacillus subtilis. J Bacteriol *181*, 3382-3391.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344-1349.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). 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.

Noma, A., Yi, S., Katoh, T., Takai, Y., Suzuki, T., and Suzuki, T. (2011). Actin-binding protein ABP140 is a methyltransferase for 3-methylcytidine at position 32 of tRNAs in Saccharomyces cerevisiae. Rna *17*, 1111-1119.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. Journal of proteome research *2*, 43-50.

Picotti, P., Clement-Ziza, M., Lam, H., Campbell, D.S., Schmidt, A., Deutsch, E.W., Rost, H., Sun, Z., Rinner, O., Reiter, L.*, et al.* (2013). A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature *494*, 266-270.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L.*, et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-933.

Schaab, C., Geiger, T., Stoehr, G., Cox, J., and Mann, M. (2012). Analysis of high accuracy, quantitative proteomics data in the MaxQB database. Molecular & cellular proteomics : MCP *11*, M111 014068.

Scheltema, R.A., and Mann, M. (2012). SprayQc: A Real-Time LC-MS/MS Quality Monitoring System To Maximize Uptime Using Off the Shelf Components. Journal of proteome research.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. Science *270*, 467-470.

Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M. (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proceedings of the National Academy of Sciences of the United States of America *93*, 14440-14445.

Swan, T.M., and Watson, K. (1998). Stress tolerance in a yeast sterol auxotroph: role of ergosterol, heat shock proteins and trehalose. Fems Microbiol Lett *169*, 191-197.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M.*, et al.* (2004). Global mapping of the yeast genetic interaction network. Science *303*, 808-813.

van der Rest, M.E., Kamminga, A.H., Nakano, A., Anraku, Y., Poolman, B., and Konings, W.N. (1995). The plasma membrane of Saccharomyces cerevisiae: structure, function, and biogenesis. Microbiological reviews *59*, 304-322.

Walther, T.C., and Mann, M. (2010). Mass spectrometry-based proteomics in cell biology. The Journal of cell biology *190*, 491-500.

Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature biotechnology *19*, 242-247.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H.*, et al.* (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science *285*, 901-906.

Wisniewski, J.R., Ostasiewicz, P., Dus, K., Zielinska, D.F., Gnad, F., and Mann, M. (2012). Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Mol Syst Biol *8*, 611.

FIGURE AND TABLE LEGENDS

Figure 1

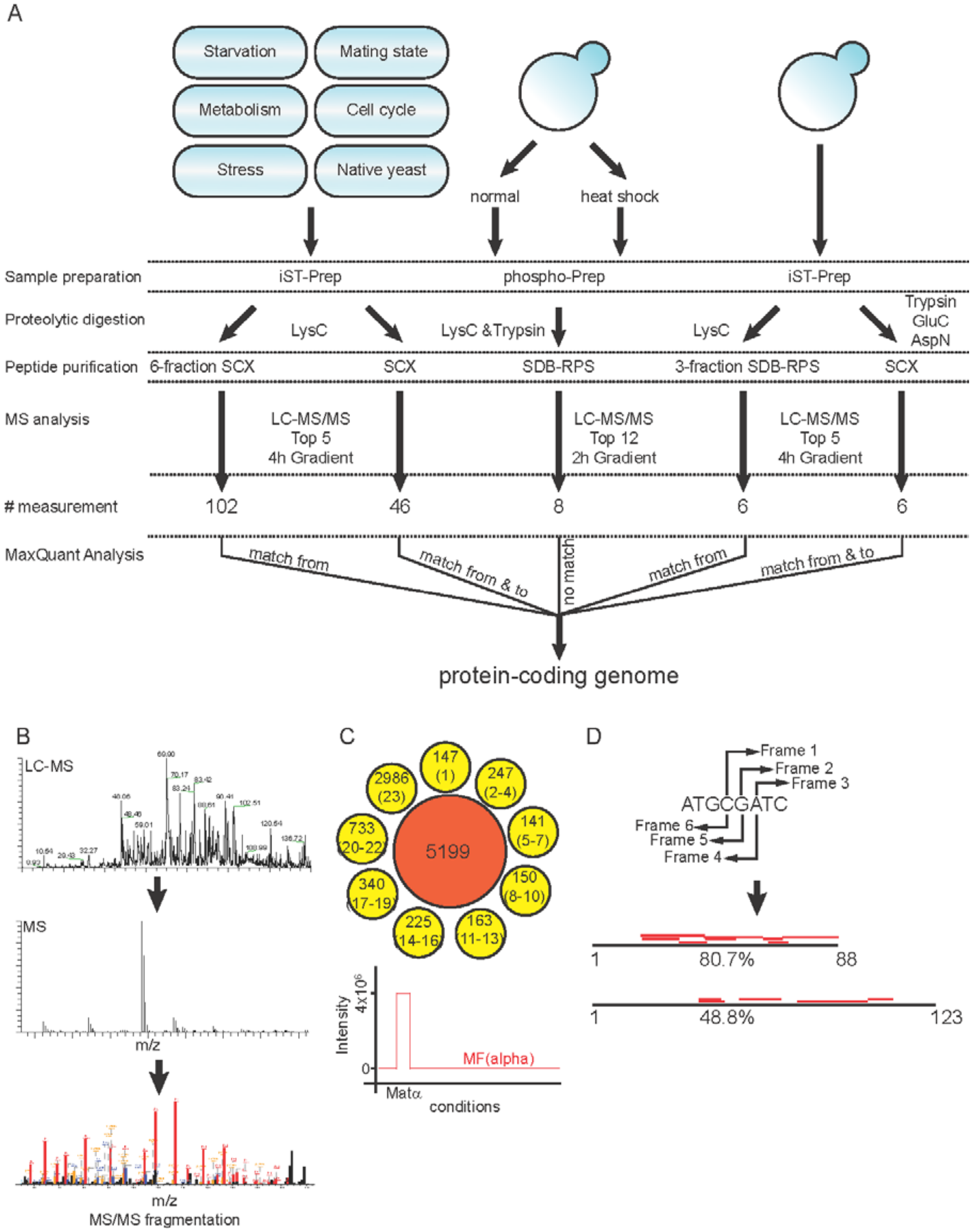


FIGURE 1. WORKFLOW FOR IN-DEPTH QUANTIFICATION OF THE COMPLETE PROTEIN-CODING GENOME OF YEAST

(A) Haploid budding yeast was grown in different physiological cell states such as stress (heat, oxidative, osmotic, DNA damage), metabolic changes (ethanol and galactose as sugar source), different cell-cycle phases (cell-cycle arrest in G2/M and G1) and starvation. Additionally, yeasts of different mating types as well as native, non-laboratory champagne and baker's yeast were cultured. Yeast samples were processed as described before (Kulak et al., 2014). We employed different pre-fractionation techniques, used different proteolytic enzymes and enriched phosphorylated peptides in this study. (B) All samples were measured using a high-resolution UHPLC setup coupled to a quadrupole Orbitrap mass spectrometer (Q Exactive, Thermo Fischer Scientific) as previously described (Nagaraj et al., 2012). (C) Distribution of protein groups exclusively identified in a certain set of growth conditions (number in brackets). Measured raw intensities of mating factor alpha (Matα) over all conditions. (D) Identification of two novel proteins in the yeast genome. Red lines indicate identified peptides, their length and position within the genome sequence.

Figure 2

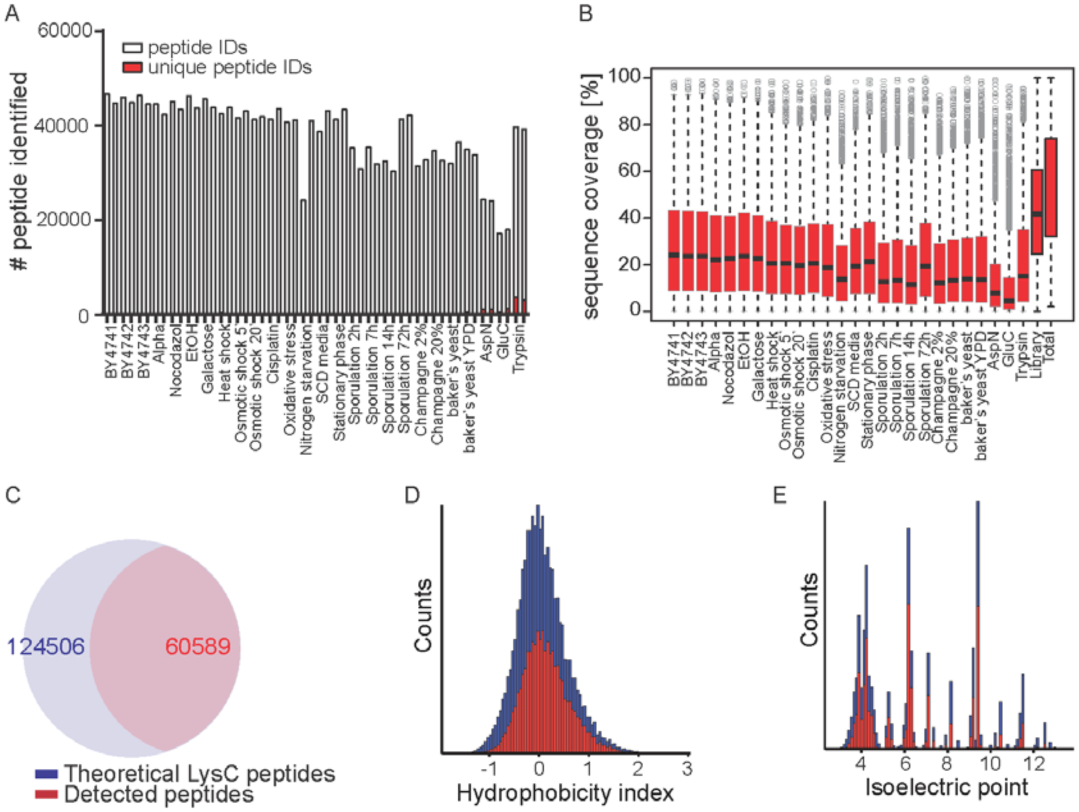

A
60000
40000
20000
0
# peptide identified
peptide IDs
unique peptide IDs
BY4741
BY4742
BY4743
Alpha
Nocodazol
EtOH
Galactose
Heat shock
Osmotic shock 5'
Osmotic shock 20'
Cisplatin
Oxidative stress
Nitrogen starvation
SCD media
Stationary phase
Sporulation 2h
Sporulation 7h
Sporulation 14h
Sporulation 72h
Champagne 2%
Champagne 20%
baker's yeast
baker's yeast YPD
AspN
GluC
Trypsin
B
100
80
60
40
20
0
sequence coverage [%]
BY4741
BY4742
BY4743
Alpha
Nocodazol
EtOH
Galactose
Heat shock
Osmotic shock 5'
Osmotic shock 20'
Cisplatin
Oxidative stress
Nitrogen starvation
SCD media
Stationary phase
Sporulation 2h
Sporulation 7h
Sporulation 14h
Sporulation 72h
Champagne 2%
Champagne 20%
baker's yeast
baker's yeast YPD
AspN
GluC
Trypsin
Library
Total
C
124506
60589
Theoretical LysC peptides
Detected peptides
D
Counts
-1
0
1
2
3
Hydrophobicity index
E
Counts
4
6
8
10
12
Isoelectric point

FIGURE 2. SUMMARY OF PEPTIDE SEQUENCE COVERAGE OF THE PROTEIN-CODING YEAST GENOME

(A) Number of peptides identified in individual runs of two biological replicates for each condition. Peptides unique for individual runs are indicated in red. In total 132,053 sequence distinct yeast peptides were identified. (B) The median sequence coverage of identified proteins was 53.9%. (C) Comparison of all possible Lys-C peptides to actually detected Lys-C peptides. (D) Hydrophobicity distribution for all theoretical (blue) and for all detected LysC peptides (red). (E) Isoelectric point distribution for all theoretical (blue) and for all detected LysC peptides (red).

Figure 3

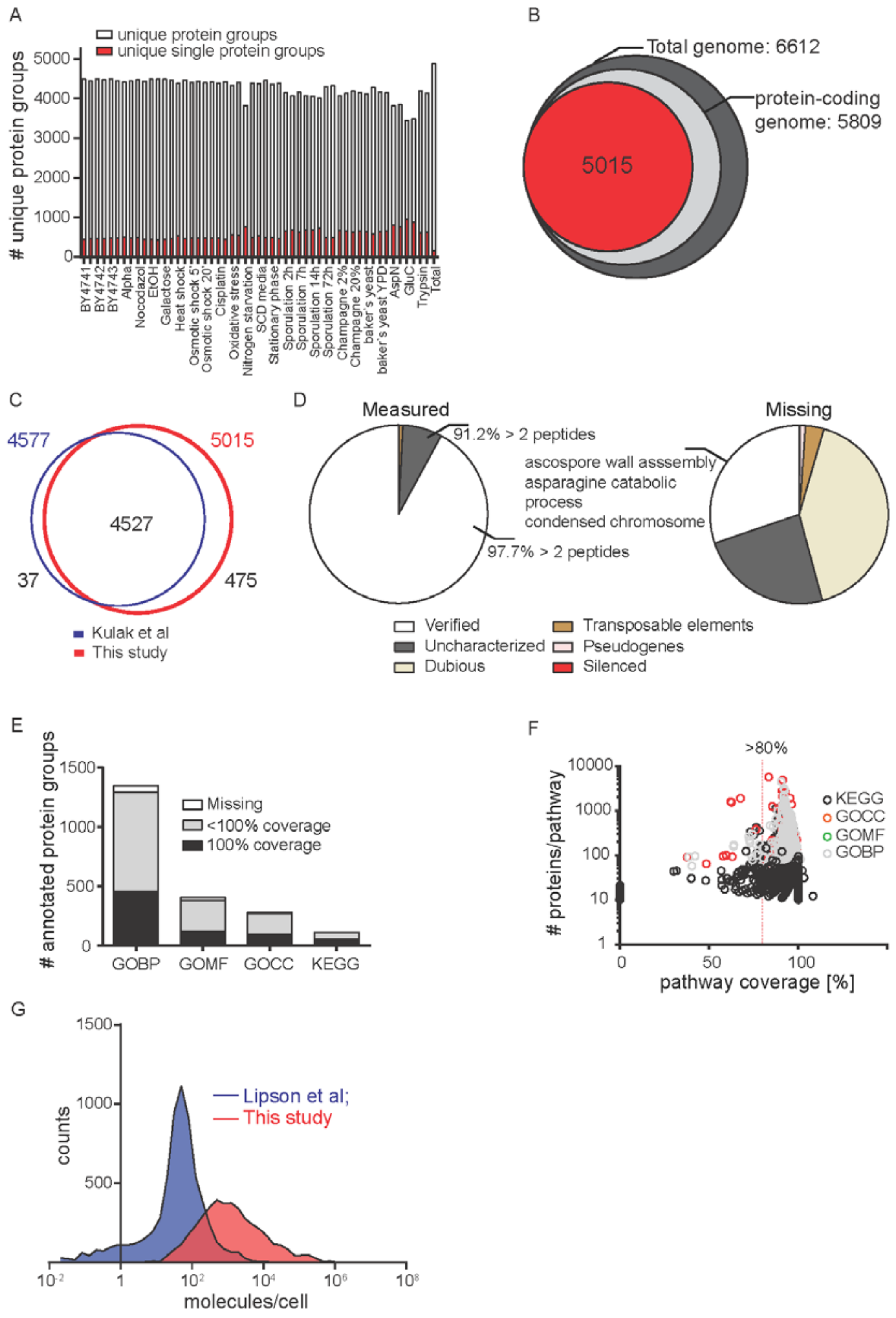

A
unique protein groups
unique single protein groups
# unique protein groups
5000
4000
3000
2000
1000
0
BY4741
BY4742
BY4743
Alpha
Nocodazol
EtOH
Galactose
Heat shock
Osmotic shock 5'
Osmotic shock 20'
Cisplatin
Oxidative stress
Nitrogen starvation
SCD media
Stationary phase
Sporulation 2h
Sporulation 7h
Sporulation 14h
Sporulation 72h
Champagne 2%
Champagne 20%
baker's yeast
baker's yeast YPD
AspN
GluC
Trypsin
Total
B
Total genome: 6612
protein-coding genome: 5809
5015
C
4577
5015
4527
37
475
Kulak et al
This study
D
Measured
91.2% > 2 peptides
97.7% > 2 peptides
Missing
ascospore wall asssembly
asparagine catabolic process
condensed chromosome
Verified
Uncharacterized
Dubious
Transposable elements
Pseudogenes
Silenced
E
# annotated protein groups
1500
1000
500
0
Missing
<100% coverage
100% coverage
GOBP
GOMF
GOCC
KEGG
F
>80%
# proteins/pathway
10000
1000
100
10
1
0
50
100
pathway coverage [%]
KEGG
GOCC
GOMF
GOBP
G
1500
1000
500
counts
Lipson et al;
This study
10-2
1
102
104
106
108
molecules/cell

FIGURE 3. COMPREHENSIVE QUANTIFICATION OF THE PROTEIN-CODING YEAST GENOME

(A) Unique protein groups identified in single runs of two biological replicates for each condition. Proteins identified with unique single peptide hits are shown in red. (B) Schematic comparison of all annotated yeast ORFs (6612; dark grey) to all confirmed protein-coding ORFs (5809; light grey) and the total number of identified protein groups 5015; red). (C) Comparison of identified protein groups to the previous deepest coverage of an experimental S. cerevisiae proteome (Kulak et al., 2014). (D) Distribution of identified and missing protein groups classified by the presence type, as annotated in SGD. The percentage of measured protein groups identified by more than 2 peptides is indicated for verified and uncharacterized proteins. Enriched GO categories (Fisher exact test) of non-identified proteins in the presence type verified. (E) Coverage of biological pathways classified in the Kyoto Encyclopedia of Genes and Genome (KEGG) and Gene Ontology (GO) database. Pathways covered 100%, below 100%, and which are completely missing are indicated. (F) Number of annotated pathway members for each KEGG or GO category. The red dotted line indicates pathway coverage of at least 80%. (G) Abundance distribution for mRNAs by single-molecule sequencing (blue) (Lipson et al., 2009) and proteins (red).

Figure 4

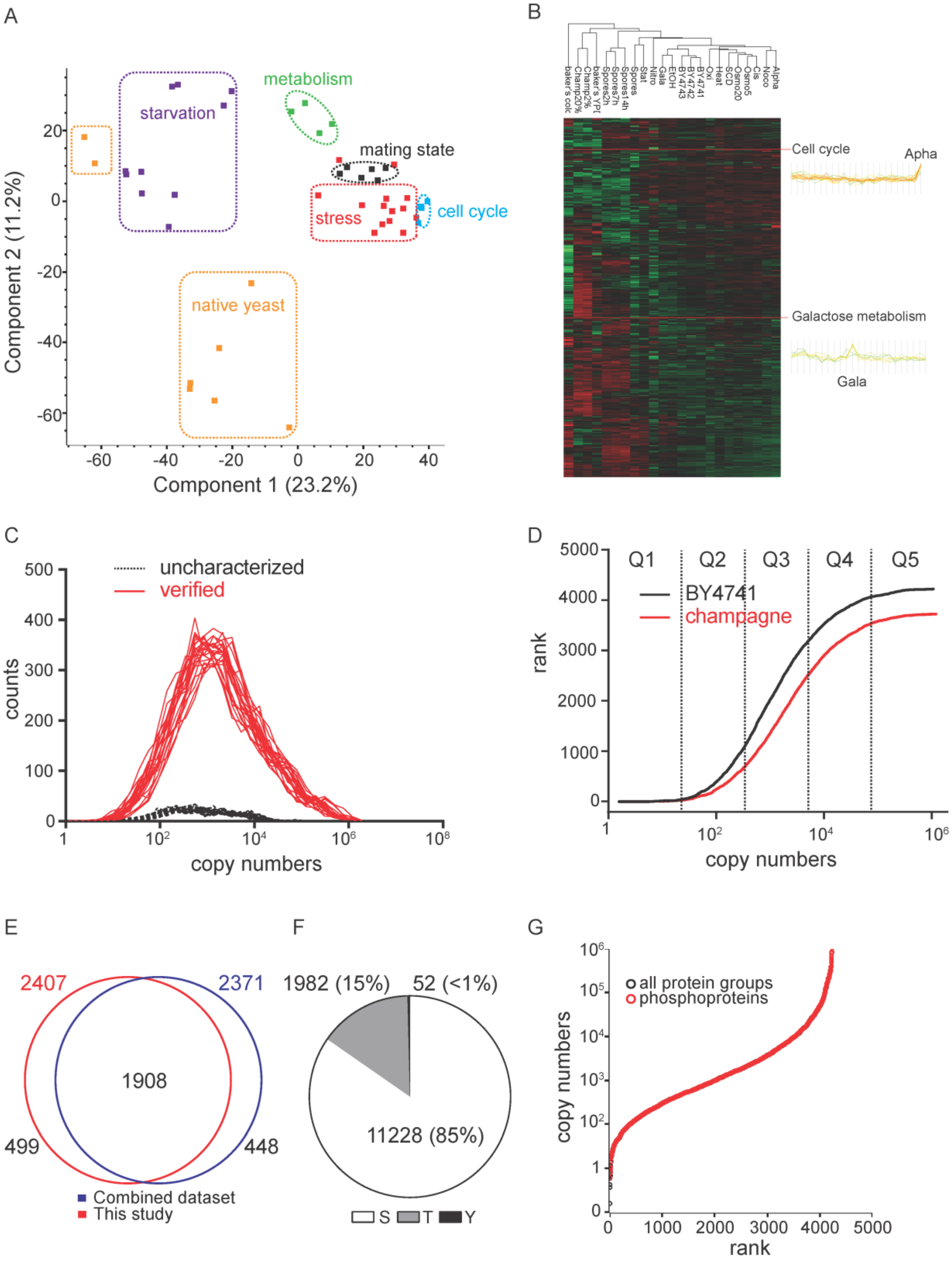

A
starvation
metabolism
mating state
stress
cell cycle
native yeast
Component 2 (11.2%)
Component 1 (23.2%)
20
0
-20
-40
-60
-60
-40
-20
0
20
40
B
baker's colc
Champ20%
Champ2%
baker's YPI
Spores2h
Spores7h
Spores14h
Spores
Stat
Nitro
Gala
EtOH
BY4743
BY4742
BY4741
Oxi
Heat
SCD
Osmo20
Osmo5
Cis
Noco
Alpha
Cell cycle
Apha
Galactose metabolism
Gala
C
uncharacterized
verified
counts
copy numbers
500
400
300
200
100
0
1
10^2
10^4
10^6
10^8
D
Q1
Q2
Q3
Q4
Q5
BY4741
champagne
rank
copy numbers
5000
4000
3000
2000
1000
0
1
10^2
10^4
10^6
E
2407
2371
1908
499
448
Combined dataset
This study
F
1982 (15%)
52 (<1%)
11228 (85%)
S
T
Y
G
all protein groups
phosphoproteins
copy numbers
rank
10^6
10^5
10^4
10^3
10^2
1
0
0
1000
2000
3000
4000
5000

FIGURE 4. DYNAMIC REGULATION OF THE PROTEIN-CODING GENOME

(A) Principal component analysis (PCA) based on protein intensity (LFQ) values for all different conditions and for native champagne and baker's yeast. The following conditions are grouped and color-coded together: metabolism (green; EtOH, galactose); mating state (black; BY4741, BY4742, BY4743); cell-cycle (light blue; G1 and G2/M arrest); stress (red; heat, osmotic, oxidative and DNA damage stress); starvation (purple; SCD media, nitrogen starvation, sporulation); native yeast (champagne yeast grown in 2 and 20% glucose, baker's yeast directly or cultivated). (B) Hierarchical clustering of significantly changing proteins. Significance was determined by ANOVA with correction for multiple hypothesis testing. Average LFQ intensities were normalized by Z-score. (C) Copy numbers of proteins by presence type, as annotated in SGD, and as an overlay of all different measurements. (D) Ranking of estimated copy numbers for BY4741 (black) and champagne yeast (red). Proteins were ranked into five quantiles based on their abundance. (E) Comparison of identified phosphoproteins to a combination of twelve publicly available data sets (Amoutzias et al., 2012). (F) Relative distribution of serine, threonine and tyrosine phosphorylation sites. (G) Distribution of protein abundance of the phosphoproteome.

Figure 5

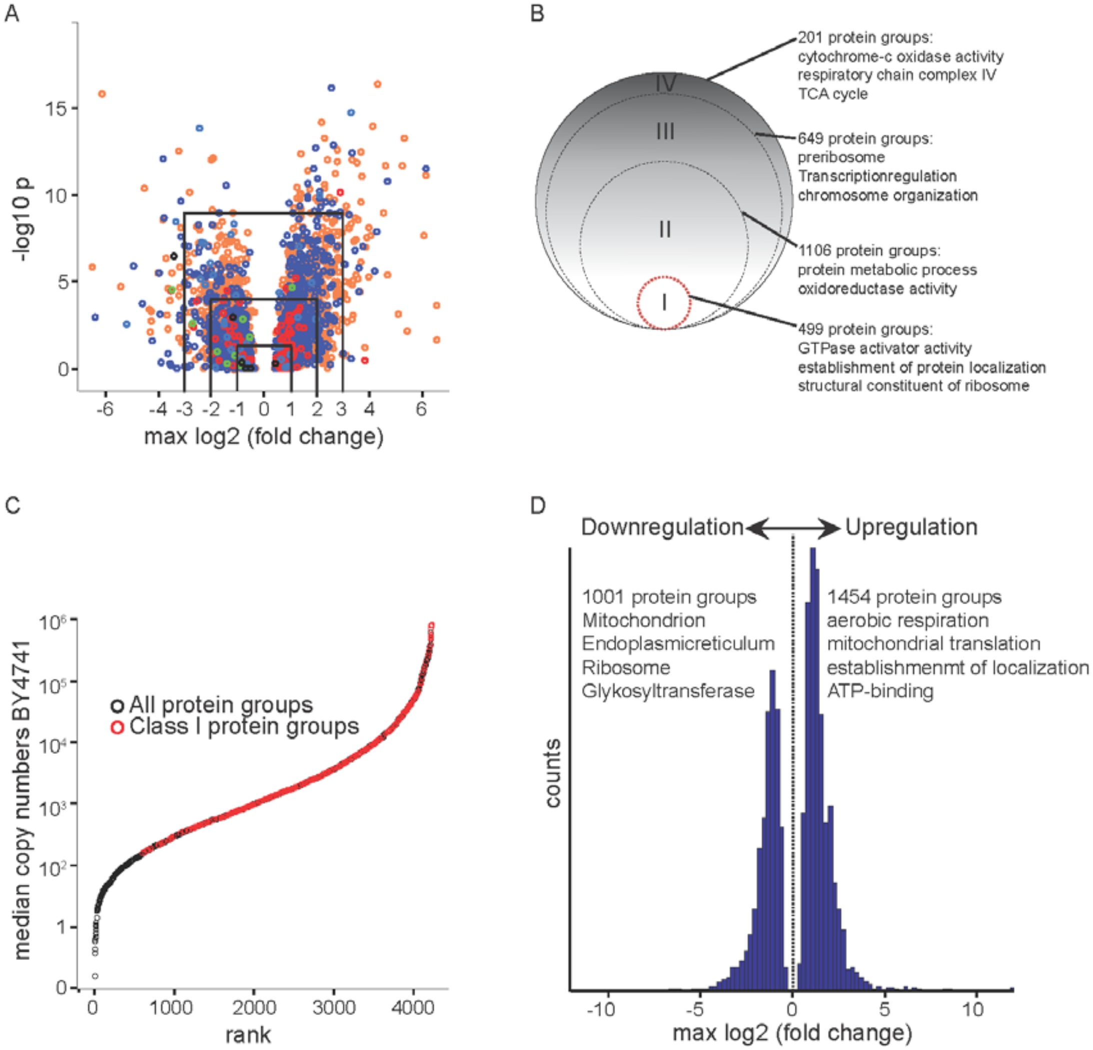


FIGURE 5. HOUSEHOLD PROTEOME

(A) Different classes of protein groups are assigned based on the maximal fold change between all measurements and on the negative logarithmic p-value for a t-test of differences between samples. Maximal fold changes represent the maximal up-or down-regulation of a protein measured in all conditions. The fold change is the average protein intensity value (LFQ) of two biological replicates divided by the equivalent value in BY4741. The following values are assigned to the different classes: Class I (Max. fold change=2; -log10 p=2); Class II (Max. fold change=4; -

log10 p=4); Class III (Max. fold change=8; -log10 p=8); Class IV no threshold. (B) Enriched GO terms extracted by Fisher exact test. (C) Ranking of estimated copy numbers for BY4741. Red dots indicate Class I household proteins. (D) Distribution of the maximal fold change for all identified proteins. Enriched pathways and functions for down- or up-regulated proteins extracted by Fisher exact test.

Figure 6

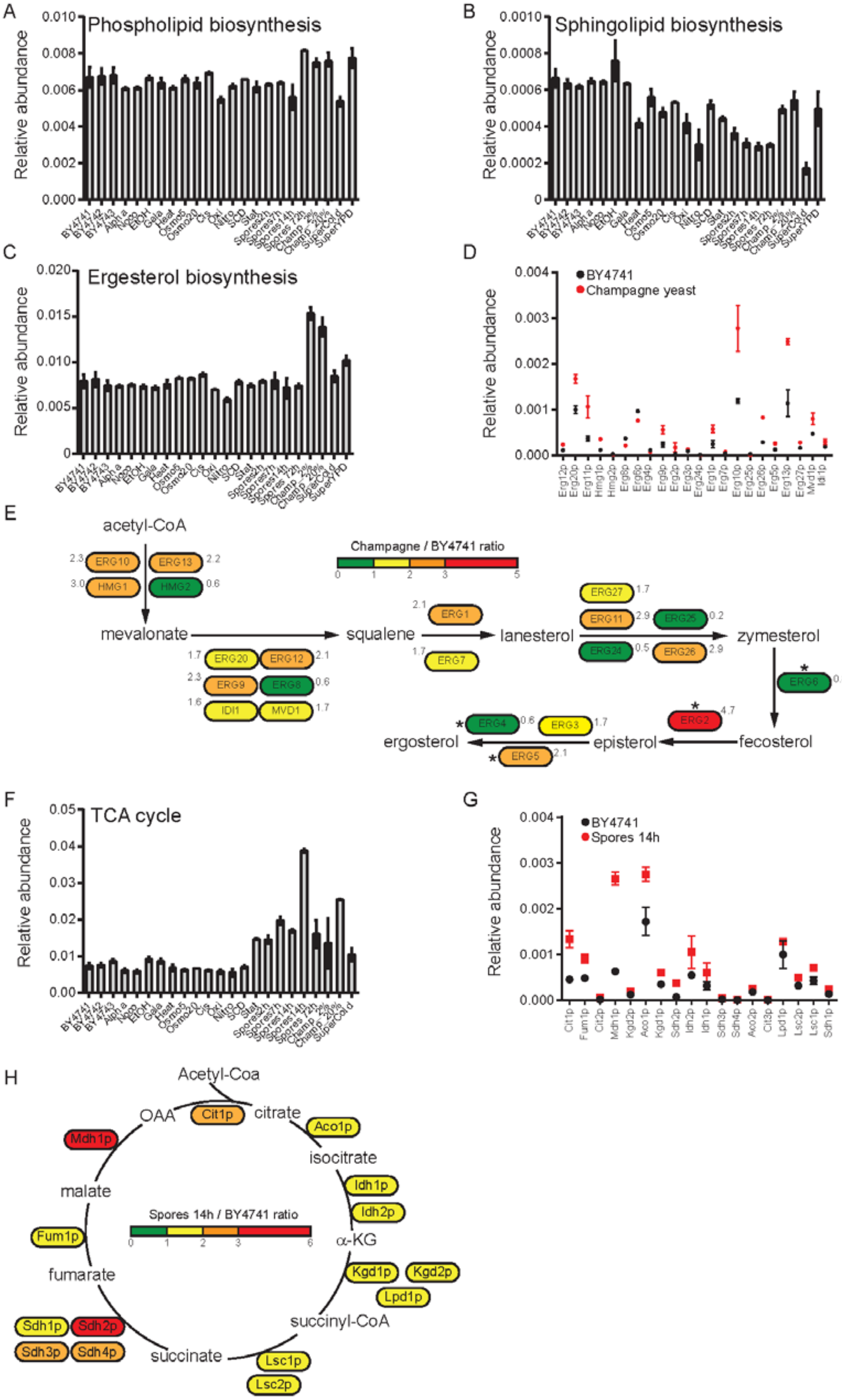
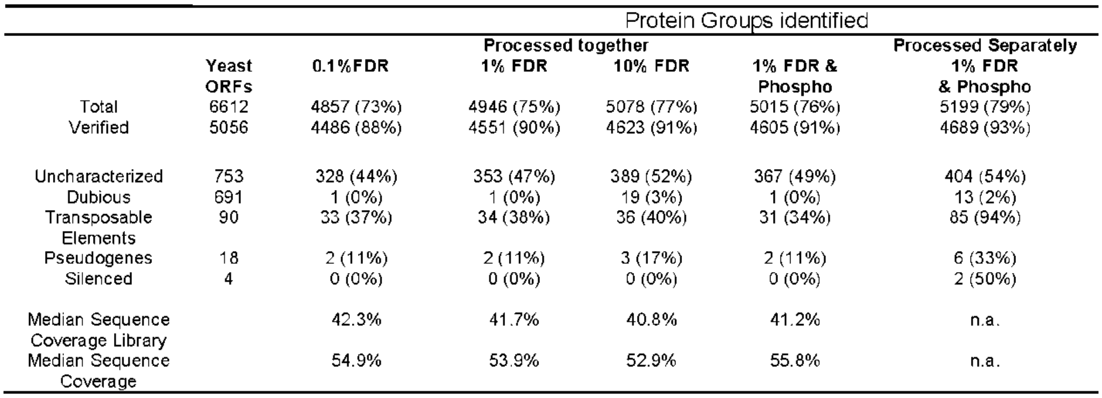

A
Phospholipid biosynthesis
Relative abundance
0.010
0.008
0.006
0.004
0.002
0.000
BY4741
BY4742
BY4743
Alpha
NoO2
EtOH
Gala
Heat
Osmo5
Osmo20
Cts
Oxi
Nitro
SCD
Stat
Spores2h
Spores7h
Spores14h
Spores 72h
Champ 2%
Champ 20%
SuperCd
SuperYPD
B
Sphingolipid biosynthesis
Relative abundance
0.0010
0.0008
0.0006
0.0004
0.0002
0
C
Ergesterol biosynthesis
Relative abundance
0.020
0.015
0.010
0.005
0
D
BY4741
Champagne yeast
Relative abundance
0.004
0.003
0.002
0.001
0.000
Erg12p
Erg20p
Erg11p
Hmg1p
Hmg2p
Erg8p
Erg6p
Erg4p
Erg9p
Erg2p
Erg3p
Erg24p
Erg1p
Erg7p
Erg10p
Erg25p
Erg26p
Erg5p
Erg13p
Erg27p
Mvd1p
Idi1p
E
acetyl-CoA
Champagne / BY4741 ratio
0
1
2
3
5
2.3 ERG10
ERG13 2.2
3.0 HMG1
HMG2 0.6
mevalonate
squalene
lanesterol
zymesterol
1.7 ERG20
ERG12 2.1
2.3 ERG9
ERG8 0.6
1.6 IDI1
MVD1 1.7
2.1 ERG1
1.7 ERG7
ERG27 1.7
ERG11 2.9
ERG25 0.2
ERG24 0.5
ERG26 2.9
*
ERG6 0.8
*
ERG2 4.7
fecosterol
episterol
ergosterol
* ERG4 0.6
ERG3 1.7
* ERG5 2.1
F
TCA cycle
Relative abundance
0.05
0.04
0.03
0.02
0.01
0.00
G
BY4741
Spores 14h
Relative abundance
0.004
0.003
0.002
0.001
0.000
Cit1p
Fum1p
Cit2p
Mdh1p
Kgd2p
Aco1p
Kgd1p
Sdh2p
Idh2p
Idh1p
Sdh3p
Sdh4p
Aco2p
Cit3p
Lpd1p
Lsc2p
Lsc1p
Sdh1p
H
Acetyl-Coa
OAA
Cit1p
citrate
Aco1p
isocitrate
Idh1p
Idh2p
α-KG
Kgd1p
Kgd2p
Lpd1p
succinyl-CoA
Lsc1p
Lsc2p
succinate
Sdh1p
Sdh2p
Sdh3p
Sdh4p
fumarate
Fum1p
malate
Mdh1p
Spores 14h / BY4741 ratio
0
1
2
3
6

FIGURE 6. EVOLUTIONARY CONSERVATION AND REGULATION OF BIOLOGICAL PATHWAYS UNDER DIFFERENT CONDITIONS

(A) Relative abundance of proteins involved in phospholipid, (B) sphingolipid or (C) ergosterol biosynthesis. (D) The ratio of the relative abundance of proteins involved in ergosterol biosynthesis between champagne yeast and BY4741. (E) Proteins involved in ergosterol biosynthesis are color-coded according to median fold change between champagne yeast and BY4741. The ratio of relative abundance is indicated for each protein. (F) Relative abundance of proteins involved in the tricarboxylic acid cycle (TCA cycle). (G) The ratio between BY4743 cultured in sporulation media for 14h and BY4741 of the relative abundance between proteins involved in the TCA cycle. (H) Proteins involved in the TCA cycle are color-coded according to median fold change between BY4743 cultured in sporulation media for 14h and BY4741. The ratio of relative abundance is indicated for each protein.

All bars represent average values ± st dev for two biological replicates.



		Protein Groups identified				
		Processed together				Processed Separately
	Yeast ORFs	0.1%FDR	1% FDR	10% FDR	1% FDR & Phospho	1% FDR & Phospho
Total	6612	4857 (73%)	4946 (75%)	5078 (77%)	5015 (76%)	5199 (79%)
Verified	5056	4486 (88%)	4551 (90%)	4623 (91%)	4605 (91%)	4689 (93%)
Uncharacterized	753	328 (44%)	353 (47%)	389 (52%)	367 (49%)	404 (54%)
Dubious	691	1 (0%)	1 (0%)	19 (3%)	1 (0%)	13 (2%)
Transposable Elements	90	33 (37%)	34 (38%)	36 (40%)	31 (34%)	85 (94%)
Pseudogenes	18	2 (11%)	2 (11%)	3 (17%)	2 (11%)	6 (33%)
Silenced	4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (50%)
Median Sequence Coverage Library		42.3%	41.7%	40.8%	41.2%	n.a.
Median Sequence Coverage		54.9%	53.9%	52.9%	55.8%	n.a.

TABLE 1. NUMBER OF PROTEIN GROUPS IDENTIFIED

Coverage of SGD database annotations and median sequence coverage for different datasets applying different FDR values.

DISCUSSION AND CONCLUDING REMARKS

The MS-based proteomic platform consists of (1) sample preparation, (2) liquid chromatography, (3) mass spectrometry, and (4) bioinformatic analysis of the measurements. This thesis focused on improvements at the level of sample preparation, liquid chromatography, and mass spectrometry. A benchmark goal was to achieve complete coverage of the *S. cerevisiae* model organism and very deep proteomic coverage of higher eukaryotic systems. Together, the two main projects aimed to improve and optimize the workflow for complete proteomics to be more robust, reproducible, and straightforward in single-shot measurements or with simple pre-fractionation techniques.

The platform developments discussed in this thesis delivered near-complete proteomic datasets of *S. cerevisiae* with 4-h single-shot measurements and essentially complete proteomic coverage with six-fraction SCX measurements. The platform further delivered datasets with accurate and deep protein copy-number estimates for the model systems *S. cerevisiae*, *S. pombe*, and the HeLa cancer cell-line model.

SAMPLE PREPARATION PIPELINES

Development of a simplified sample preparation was a major undertaking to improve the existing proteomics workflow; the aim was to streamline the overall sample handling process. The first major alteration to classical sample preparation protocols was the combination of lysis, disulfide bond reduction, and cysteine alkylation into a single step. Importantly, replacing previous multi-step protocols did not negatively affect reduction and alkylation efficiencies but accelerated the first part of sample preparation from more than one hour to 20 minutes and the number of pipetting steps from three to one. While TCEP and CAA successfully reduce and alkylate proteins during the lysis procedure the long-term stability of the chemical mixture was not determined.

Previous reports about the compatibility of TCEP and IAA were somewhat inconsistent. A study focusing on alkylation efficiencies reported that iodo-acetamide remained active in the presence of TCEP but lost activity in combination with thiol-containing DTT [129]. These results were partially contradicted by Shafer et al. reporting a severe loss of IAA after 30 minute incubation in presence of TCEP and a near-complete loss in presence of 2-mercaptoethanol under basic conditions [130]. As iodo-acetamide demonstrates higher reactivity than chloro-acetamide it can be expected to be more stable in the presence of TCEP, also reaction rates with thiol-groups are consistently reported to be higher. It can therefore be concluded that dedicated lysis buffers should not be stored at room temperature for extended time periods.

The next step of classical sample preparation workflows is lysate clarification in the presence of strong detergents and subsequent detergent removal. Most protocols perform a simple centrifugation step to pellet insoluble sample constituents; here the soluble fraction is transferred to a fresh reaction tube which necessarily leads to some sample loss since complete supernatant transfers are nearly impossible. Very recent publications reported a similar effect and argued for the removal of the sample clarification steps [71, 131].

Strong detergents need to be removed before enzymatic protein digestion and this has mostly been accomplished by protein precipitation. Classical precipitation procedures are very laborious and efficient protein precipitation typically takes many hours at 4°C. The development of the FASP protocol has made detergent removal easier, leaving only few centrifugation steps. However, while easier, FASP is sensitive to sample loss when handled incorrectly. The technology developed in this thesis avoids sample clarification and protease incompatible detergents. This simplification of the protocol has proved to be very efficient and to also be suitable for the digestion of membrane-spanning proteins.

PROTEOLYTIC DIGESTION

Chemicals used during cell lysis and therefore during the digestion are of very high importance for the overall procedure. The most prominent chaotrope used for trypsin and Lys-C digestions is urea, which is still compatible with the enzymes even at high concentrations. It is a well-known fact that extended exposure or higher temperatures cause urea to carbamylate lysine residues and protein N-termini [132]. The modified lysine residues can inhibit enzyme recognition, which can cause high missed cleavage rates and makes database search more difficult. Previous protocols therefore avoided increased temperatures, even though trypsin and Lys-C demonstrate higher digestion rates at 37°C than at room temperature. Recently guanidium-hydrochloride (GdmCl) has been shown to be a non-reactive alternative [39]. The protocol developed in this thesis successfully applied this chaotrope for lysis and protein digestion, however, we found that proteases can be sensitive to high GdmCl concentrations and that a relatively high dilution of the initial lysis buffer is necessary to achieve acceptable digestion rates.

A recent publication compared digestion efficiencies in presence of various chemicals and concluded that trypsin and Lys-C achieve maximum results in presence of sodium-deoxycholate (SDC) which can be readily removed for final peptide fractionation and clean-up [40]. Two procedures for SDC removal have been published, namely by acidic precipitation of the surfactant and by extraction using ethyl acetate [133]. Our procedure successfully applied SDC/ethyl acetate extraction and the removal of the solvent by cation-based extraction on a StageTip. This combination resulted in very high and clean peptide recoveries but further improvements of SDC removal might be necessary for sensitive samples because large quantities of ethyl acetate may elute peptides and reduce the overall sensitivity while insufficient SDC extraction can lead to carry-over of the surfactant.

Efficient and complete proteolytic digestions are a very important goal of sample preparation since reduced missed-cleavage rates translate to lower sample complexities at the peptide level. Furthermore, complete digestions are typically performed with 16h incubations, which represents the most time consuming step of the entire sample preparation. Increasing digestion rates and the stability of the protease has been an active field of exploration. The most prominent factor for more efficient digestion is increased temperature. However, one effect of higher temperature is increased auto-proteolysis and decreased life-time of the enzyme. For this reason trypsin has been chemically modified, increasing the temperature optimum [134, 135]. While these modifications greatly improve the temperature stability, simple and cost-efficient lysine modifications such as lysine methylation already reduce auto-digestion [136].

Besides temperature changes and chemical modifications other physical means were observed to increase protease activities. The most prominent methods are high-pressure cycling, ultrasonic treatment, and microwave assisted protein digestion [137-141]. These technologies may be especially interesting for very fast sample preparation where time is of the essence. This may be especially beneficial for certain clinically relevant tests or diagnostic assays.

SAMPLE CLEAN-UP AND PRE-FRACTIONATION

The majority of classical sample preparation protocols remove salts before LC-MS analysis. This is typically done using C_{18}-based reversed-phase SPE protocols. Even though these materials are well suited for desalting, the actual affinity and retention of peptides is lower than in the corresponding micron-sized C_{18} materials used in modern LC systems. This can be observed by a loss of small and highly hydrophilic peptides otherwise eluting in the first part of the chromatogram. Other SPE materials have been well established in the field of metabolomics because metabolites are often more difficult to retain on the reversed-phase. Results of this thesis demonstrated that SCX-based materials are somewhat better suited to bind and retain peptides. This is expected since all peptides analyzed by positive-ion mode LC-MS need to be able to carry positive charges.

We demonstrated compatibility of SCX-based StageTip materials with proteomic sample clean-up and the selected materials had the additional benefit to enable fractionation of the samples with all volatile buffer components. A protocol to clean-up and fractionate the samples into three or six fractions proved highly efficient and resulted in very good fractionation efficiencies. The volatile buffers in turn can be removed from the sample using a vacuum centrifuge without any additional desalting step decreasing loss of material during preparation.

REDUCTION OF METHIONINE SULFOXIDES

Even though methionine sulfoxides (Met(O)) are currently disregarded in nearly all sample preparation procedures, they represent important post translational *in vivo* and *in vitro* modifications [142]. Besides being described in the context of biological stress conditions, they likewise occur spontaneously during sample work up in the presence of reactive oxygen species (ROS). Extended storage and even the electrospray ionization process can therefore introduce

Met(O) modifications and these modifications increase the observed sample complexity. While these modifications are not addressed in the current protocols, they may be of very high interest for future improvements. In particular, chemical reagents capable of Met(O) reduction could be interesting in this context.

DTT has a mild reducing effect, N-methylmercaptoacetamide (MMA) and thioglycolic acid (TGA) demonstrate stronger reducing effects. Depending on pH, temperature, and incubation times all methionine sulfoxides can be reduced using stronger reagents [143]. In very recent tests, we applied MMA during proteolytic digestion and TGA for the final peptide acidification. Preliminary results have shown a decrease in Met(O)-containing peptides from typically >10% to less than 5%. While these improvements remain under development, they pose an intriguing opportunity to achieve more comprehensive proteomic coverage with identical identification rates.

HIGHER RESOLUTION LIQUID CHROMATOGRAPHY

Peptide separation by liquid chromatography is a major part of the proteomics platform and was one of the main topics for the developments described in this thesis. We set out to achieve excellent chromatographic resolution compatible with the high acquisition rates of a novel high-performance benchtop quadrupole Orbitrap instrument (Q Exactive). Employing long packed chromatography columns with sub-2-micron particle size significantly reduced the peak width and improved the overall chromatographic resolution (50 cm, 75 µm I.D., 1.8 µm C_{18} beads). Long columns with reduced bead-size resulted in higher backpressure of the column which made the use of the ultra-high pressure (UHPLC) system essential. A good compromise between electrospray stability, peak shape, and flow-rate was observed at a flow-rate of 250 nl/min (35°C, approximately 500 bar).

Higher resolution chromatography systems with high-backpressure columns often entails low utilization rates of the system – i.e. relatively long times with no data acquisition. Higher back-pressures slow down sample loading (which is typically performed at higher flow rates) and longer columns cause a delay in peptide elution. The platform employed here works with 50 cm columns with 250 nl/min flow-rate and a delay of nearly 20 minutes before first peptides elute can be observed. This in turn favors longer measurements because of the increased down-time. Shorter columns with smaller inner diameters may be able to increase the utilization rates while maintaining ESI efficiency and chromatographic resolution, but decreased column I.D.s typically lead to smaller sample capacity and even higher backpressures. Alternatively, higher flow-rates on shorter columns and higher temperatures might be beneficial for short measurements since steeper gradients cause small peak widths, but such developments should be done using even faster mass spectrometers.

HIGH PERFORMANCE MASS SPECTROMETERS

Current developments on mass spectrometry instrumentation are the most important aspect for future proteomics platforms. The Q Exactive mass spectrometer is one of the best performing instruments today and is capable of acquiring an entire Top10 cycle within 1.2s while reaching up to 60% identification rates. A combination of high-resolution chromatography with the high-performance Q Exactive was therefore a key to achieving near-complete proteomic depth in 4h measurements. Even though acquisition rates are improving, the ion-source and ion-path remain bottle-necks. Recent advances of the Q Exactive PLUS appear to have improved the ion path and these and other MS developments hold tremendous promise to achieve even better coverage in shorter time.

COMPLETE PROTEOMICS

Coupling a high-resolution nano-UHPLC system to the Q Exactive was a fortuitous choice since very high identification rates became possible. Especially single-shot measurements of the *S. cerevisiae* model systems demonstrated the deep proteomic coverage enabled by modern mass spectrometers. The simple six fraction approach discussed above resulted in the deepest proteomes measured for the model organisms *S. cerevisiae* and *S. pombe* (4,575 protein groups and 4,087 protein groups respectively) in only 24 h total gradient time. Applying the same iST fractionation approach in quadruplicated to the human HeLa cancer cell line resulted in a remarkable 9,667 protein identifications. The overall results on complete proteomes argue for the very high potential of proteomic platforms for comprehensive measurements. Future developments will likely make analysis shorter and easier.

MS-BASED CLINICAL-DIAGOSTICS

The work presented here demonstrates the capabilities of MS-based proteomics platforms. Since sample handling steps, LC-MS measurements, and bioinformatic analysis are very streamlined many clinical applications should now become feasible. The major challenges of the proteomics platforms have now been solved and only few challenges remain to broadly apply the technology. We believe that clinically relevant measurements would be possible and could potentially help to elucidate unknown disease mechanisms and diagnose known ones. With new and faster instrumentation LC-MS has the potential to become a major competitor to well established diagnostic tools.

REFERENCES

1. GOFFEAU, A., ET AL., *LIFE WITH 6000 GENES.* SCIENCE, 1996. **274**(5287): P. 546, 563-7.

2. LANDER, E.S., ET AL., *INITIAL SEQUENCING AND ANALYSIS OF THE HUMAN GENOME.* NATURE, 2001. **409**(6822): P. 860-921.

3. VENTER, J.C., ET AL., *THE SEQUENCE OF THE HUMAN GENOME.* SCIENCE, 2001. **291**(5507): P. 1304-51.

4. CLAMP, M., ET AL., *DISTINGUISHING PROTEIN-CODING AND NONCODING GENES IN THE HUMAN GENOME.* PROC NATL ACAD SCI U S A, 2007. **104**(49): P. 19428-33.

5. GRADA, A. AND K. WEINBRECHT, *NEXT-GENERATION SEQUENCING: METHODOLOGY AND APPLICATION.* J INVEST DERMATOL, 2013. **133**(8): P. E11.

6. MORTAZAVI, A., ET AL., *MAPPING AND QUANTIFYING MAMMALIAN TRANSCRIPTOMES BY RNA-SEQ.* NAT METHODS, 2008. **5**(7): P. 621-8.

7. VELCULESCU, V.E., ET AL., *CHARACTERIZATION OF THE YEAST TRANSCRIPTOME.* CELL, 1997. **88**(2): P. 243-51.

8. LAURENT, J.M., ET AL., *PROTEIN ABUNDANCES ARE MORE CONSERVED THAN MRNA ABUNDANCES ACROSS DIVERSE TAXA.* PROTEOMICS, 2010. **10**(23): P. 4209-12.

9. VOGEL, C. AND E.M. MARCOTTE, *INSIGHTS INTO THE REGULATION OF PROTEIN ABUNDANCE FROM PROTEOMIC AND TRANSCRIPTOMIC ANALYSES.* NAT REV GENET, 2012. **13**(4): P. 227-32.

10. COX, J. AND M. MANN, *IS PROTEOMICS THE NEW GENOMICS?* CELL, 2007. **130**(3): P. 395-8.

11. ALTELAAR, A.F., J. MUNOZ, AND A.J. HECK, *NEXT-GENERATION PROTEOMICS: TOWARDS AN INTEGRATIVE VIEW OF PROTEOME DYNAMICS.* NAT REV GENET, 2013. **14**(1): P. 35-48.

12. PANDEY, A. AND M. MANN, *PROTEOMICS TO STUDY GENES AND GENOMES.* NATURE, 2000. **405**(6788): P. 837-46.

13. GHAEMMAGHAMI, S., ET AL., *GLOBAL ANALYSIS OF PROTEIN EXPRESSION IN YEAST.* NATURE, 2003. **425**(6959): P. 737-41.

14. HUH, W.K., ET AL., *GLOBAL ANALYSIS OF PROTEIN LOCALIZATION IN BUDDING YEAST.* NATURE, 2003. **425**(6959): P. 686-91.

15. NEWMAN, J.R., ET AL., *SINGLE-CELL PROTEOMIC ANALYSIS OF S. CEREVISIAE REVEALS THE ARCHITECTURE OF BIOLOGICAL NOISE.* NATURE, 2006. **441**(7095): P. 840-6.

16. AEBERSOLD, R. AND M. MANN, *MASS SPECTROMETRY-BASED PROTEOMICS.* NATURE, 2003. **422**(6928): P. 198-207.

17. MILO, R., ET AL., *BIONUMBERS--THE DATABASE OF KEY NUMBERS IN MOLECULAR AND CELL BIOLOGY.* NUCLEIC ACIDS RES, 2010. **38**(DATABASE ISSUE): P. D750-3.

18. BENESCH, J.L., ET AL., *PROTEIN COMPLEXES IN THE GAS PHASE: TECHNOLOGY FOR STRUCTURAL GENOMICS AND PROTEOMICS.* CHEM REV, 2007. **107**(8): P. 3544-67.

19. HECK, A.J., *NATIVE MASS SPECTROMETRY: A BRIDGE BETWEEN INTERACTOMICS AND STRUCTURAL BIOLOGY.* NAT METHODS, 2008. **5**(11): P. 927-33.

20. COMPTON, P.D. AND N.L. KELLEHER, *SPINNING UP MASS SPECTROMETRY FOR WHOLE PROTEIN COMPLEXES.* NAT METHODS, 2012. **9**(11): P. 1065-6.

21. YATES, J.R., C.I. RUSE, AND A. NAKORCHEVSKY, *PROTEOMICS BY MASS SPECTROMETRY: APPROACHES, ADVANCES, AND APPLICATIONS.* ANNU REV BIOMED ENG, 2009. **11**: P. 49-79.

22. HARRISON, S.T., *BACTERIAL CELL DISRUPTION: A KEY UNIT OPERATION IN THE RECOVERY OF INTRACELLULAR PRODUCTS.* BIOTECHNOL ADV, 1991. **9**(2): P. 217-40.

23. GOLDBERG, S., *MECHANICAL/PHYSICAL METHODS OF CELL DISRUPTION AND TISSUE HOMOGENIZATION.* METHODS MOL BIOL, 2008. **424**: P. 3-22.

24. CLARKE, P.R. AND C.R. HILL, *PHYSICAL AND CHEMICAL ASPECTS OF ULTRASONIC DISRUPTION OF CELLS.* J ACOUST SOC AM, 1970. **47**(2): P. 649-53.

25. HOPKINS, T.R., *PHYSICAL AND CHEMICAL CELL DISRUPTION FOR THE RECOVERY OF INTRACELLULAR PROTEINS.* BIOPROCESS TECHNOL, 1991. **12**: P. 57-83.

26. VON HAGEN, J., *PROTEOMICS SAMPLE PREPARATION*. 2011: WILEY.

27. LUNDBLAD, R.L., *BIOCHEMISTRY AND MOLECULAR BIOLOGY COMPENDIUM*. 2010: TAYLOR & FRANCIS.

28. IVANOV, A.R. AND A.V. LAZAREV, *SAMPLE PREPARATION IN BIOLOGICAL MASS SPECTROMETRY*. 2011: SPRINGER.

29. DE GODOY, L.M., ET AL., *COMPREHENSIVE MASS-SPECTROMETRY-BASED PROTEOME QUANTIFICATION OF HAPLOID VERSUS DIPLOID YEAST.* NATURE, 2008. **455**(7217): P. 1251-4.

30. WISNIEWSKI, J.R., ET AL., *UNIVERSAL SAMPLE PREPARATION METHOD FOR PROTEOME ANALYSIS.* NAT METHODS, 2009. **6**(5): P. 359-62.

31. ETHIER, M., ET AL., *THE PROTEOMIC REACTOR: A MICROFLUIDIC DEVICE FOR PROCESSING MINUTE AMOUNTS OF PROTEIN PRIOR TO MASS SPECTROMETRY ANALYSIS.* J PROTEOME RES, 2006. **5**(10): P. 2754-9.

32. CLELAND, W.W., *DITHIOTHREITOL, A NEW PROTECTIVE REAGENT FOR SH GROUPS.* BIOCHEMISTRY, 1964. **3**: P. 480-2.

33. RUEGG, U.T. AND J. RUDINGER, *REDUCTIVE CLEAVAGE OF CYSTINE DISULFIDES WITH TRIBUTYLPHOSPHINE.* METHODS ENZYMOL, 1977. **47**: P. 111-6.

34. HAN, J.C. AND G.Y. HAN, *A PROCEDURE FOR QUANTITATIVE DETERMINATION OF TRIS(2-CARBOXYETHYL)PHOSPHINE, AN ODORLESS REDUCING AGENT MORE STABLE AND EFFECTIVE THAN DITHIOTHREITOL.* ANAL BIOCHEM, 1994. **220**(1): P. 5-10.

35. CRANKSHAW, M.W. AND G.A. GRANT, *MODIFICATION OF CYSTEINE.* CURRENT PROTOCOLS IN PROTEIN SCIENCE, 2001: P. 15.1. 1-15.1. 18.

36. SMEJKAL, G.B., ET AL., *SIMULTANEOUS REDUCTION AND ALKYLATION OF PROTEIN DISULFIDES IN A CENTRIFUGAL ULTRAFILTRATION DEVICE PRIOR TO TWO-DIMENSIONAL GEL ELECTROPHORESIS.* J PROTEOME RES, 2006. **5**(4): P. 983-7.

37. ROMBOUTS, I., ET AL., *IMPROVED IDENTIFICATION OF WHEAT GLUTEN PROTEINS THROUGH ALKYLATION OF CYSTEINE RESIDUES AND PEPTIDE-BASED MASS SPECTROMETRY.* SCI REP, 2013. **3**: P. 2279.

38. NIELSEN, M.L., ET AL., *IODOACETAMIDE-INDUCED ARTIFACT MIMICS UBIQUITINATION IN MASS SPECTROMETRY.* NAT METHODS, 2008. **5**(6): P. 459-60.

39. POULSEN, J.W., ET AL., *USING GUANIDINE-HYDROCHLORIDE FOR FAST AND EFFICIENT PROTEIN DIGESTION AND SINGLE-STEP AFFINITY-PURIFICATION MASS SPECTROMETRY.* J PROTEOME RES, 2013. **12**(2): P. 1020-30.

40. LEON, I.R., ET AL., *QUANTITATIVE ASSESSMENT OF IN-SOLUTION DIGESTION EFFICIENCY IDENTIFIES OPTIMAL PROTOCOLS FOR UNBIASED PROTEIN ANALYSIS.* MOL CELL PROTEOMICS, 2013. **12**(10): P. 2992-3005.

41. GLATTER, T., ET AL., *LARGE-SCALE QUANTITATIVE ASSESSMENT OF DIFFERENT IN-SOLUTION PROTEIN DIGESTION PROTOCOLS REVEALS SUPERIOR CLEAVAGE EFFICIENCY OF TANDEM LYS-C/TRYPSIN PROTEOLYSIS OVER TRYPSIN DIGESTION.* J PROTEOME RES, 2012. **11**(11): P. 5145-56.

42. 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.

43. ISHIHAMA, Y., J. RAPPSILBER, AND M. MANN, *MODULAR STOP AND GO EXTRACTION TIPS WITH STACKED DISKS FOR PARALLEL AND MULTIDIMENSIONAL PEPTIDE FRACTIONATION IN PROTEOMICS.* J PROTEOME RES, 2006. **5**(4): P. 988-94.

44. RAPPSILBER, J., M. MANN, AND Y. ISHIHAMA, *PROTOCOL FOR MICRO-PURIFICATION, ENRICHMENT, PRE-FRACTIONATION AND STORAGE OF PEPTIDES FOR PROTEOMICS USING STAGETIPS.* NAT PROTOC, 2007. **2**(8): P. 1896-906.

45. KOCHER, T., R. SWART, AND K. MECHTLER, *ULTRA-HIGH-PRESSURE RPLC HYPHENATED TO AN LTQ-ORBITRAP VELOS REVEALS A LINEAR RELATION BETWEEN PEAK CAPACITY AND NUMBER OF IDENTIFIED PEPTIDES.* ANAL CHEM, 2011. **83**(7): P. 2699-704.

46. THAKUR, S.S., ET AL., *DEEP AND HIGHLY SENSITIVE PROTEOME COVERAGE BY LC-MS/MS WITHOUT PREFRACTIONATION.* MOL CELL PROTEOMICS, 2011. **10**(8): P. M110 003699.

47. VANDEEMTER, J.J., F.J. ZUIDERWEG, AND A. KLINKENBERG, *LONGITUDINAL DIFFUSION AND RESISTANCE TO MASS TRANSFER AS CAUSES OF NONIDEALITY IN CHROMATOGRAPHY (REPRINTED FROM CHEM ENGNG SCI, VOL 5, PG 271-289, 1956).* CHEMICAL ENGINEERING SCIENCE, 1995. **50**(24): P. 3869-3882.

48. MCNAUGHT, A.D. AND A. WILKINSON, *I{UPAC}. COMPENDIUM OF CHEMICAL TERMINOLOGY, 2ND ED. (THE "GOLD BOOK")*. WILEYBLACKWELL; 2ND REVISED EDITION EDITION.

49. MEYER, V.R., *PRACTICAL HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY*. 2013: WILEY.

50. DOLE, M., L.L. MACK, AND R.L. HINES, *MOLECULAR BEAMS OF MACROIONS.* JOURNAL OF CHEMICAL PHYSICS, 1968. **49**(5): P. 2240-&.

51. TAYLOR, G., *DISINTEGRATION OF WATER DROPS IN ELECTRIC FIELD.* PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON SERIES A-MATHEMATICAL AND PHYSICAL SCIENCES, 1964. **280**(1380): P. 383-+.

52. FENN, J.B., ET AL., *ELECTROSPRAY IONIZATION FOR MASS-SPECTROMETRY OF LARGE BIOMOLECULES.* SCIENCE, 1989. **246**(4926): P. 64-71.

53. GRIMM, R.L. AND J.L. BEAUCHAMP, *EVAPORATION AND DISCHARGE DYNAMICS OF HIGHLY CHARGED MULTICOMPONENT DROPLETS GENERATED BY ELECTROSPRAY IONIZATION.* JOURNAL OF PHYSICAL CHEMISTRY A, 2010. **114**(3): P. 1411-1419.

54. MANN, M. AND M. WILM, *ELECTROSPRAY MASS SPECTROMETRY FOR PROTEIN CHARACTERIZATION.* TRENDS BIOCHEM SCI, 1995. **20**(6): P. 219-24.

55. HAHNE, H., ET AL., *DMSO ENHANCES ELECTROSPRAY RESPONSE, BOOSTING SENSITIVITY OF PROTEOMIC EXPERIMENTS.* NAT METHODS, 2013. **10**(10): P. 989-91.

56. GROSS, J.H. AND P. ROEPSTORFF, *MASS SPECTROMETRY: A TEXTBOOK*. 2011: SPRINGER.

57. 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.

58. BARNER-KOWOLLIK, C., ET AL., *MASS SPECTROMETRY IN POLYMER CHEMISTRY*. 2012: WILEY.

59. MARSHALL, A.G. AND C.L. HENDRICKSON, *HIGH-RESOLUTION MASS SPECTROMETERS.* ANNU REV ANAL CHEM (PALO ALTO CALIF), 2008. **1**: P. 579-99.

60. WELLS, J.M. AND S.A. MCLUCKEY, *COLLISION-INDUCED DISSOCIATION (CID) OF PEPTIDES AND PROTEINS.* METHODS ENZYMOL, 2005. **402**: P. 148-85.

61. 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.

62. OLSEN, J.V., ET AL., *HIGHER-ENERGY C-TRAP DISSOCIATION FOR PEPTIDE MODIFICATION ANALYSIS.* NAT METHODS, 2007. **4**(9): P. 709-12.

63. MICHALSKI, A., ET AL., *A SYSTEMATIC INVESTIGATION INTO THE NATURE OF TRYPTIC HCD SPECTRA.* J PROTEOME RES, 2012. **11**(11): P. 5479-91.

64. MIKESH, L.M., ET AL., *THE UTILITY OF ETD MASS SPECTROMETRY IN PROTEOMIC ANALYSIS.* BIOCHIM BIOPHYS ACTA, 2006. **1764**(12): P. 1811-22.

65. SCHWARTZ, J.C., M.W. SENKO, AND J.E. SYKA, *A TWO-DIMENSIONAL QUADRUPOLE ION TRAP MASS SPECTROMETER.* J AM SOC MASS SPECTROM, 2002. **13**(6): P. 659-69.

66. MAKAROV, A., ET AL., *DYNAMIC RANGE OF MASS ACCURACY IN LTQ ORBITRAP HYBRID MASS SPECTROMETER.* J AM SOC MASS SPECTROM, 2006. **17**(7): P. 977-82.

67. ZUBAREV, R.A. AND A. MAKAROV, *ORBITRAP MASS SPECTROMETRY.* ANAL CHEM, 2013. **85**(11): P. 5288-96.

68. MAKAROV, A., ET AL., *PERFORMANCE EVALUATION OF A HYBRID LINEAR ION TRAP/ORBITRAP MASS SPECTROMETER.* ANAL CHEM, 2006. **78**(7): P. 2113-20.

69. 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.* MOL CELL PROTEOMICS, 2012. **11**(3): P. O111 013698.

70. DOMON, B. AND R. AEBERSOLD, *OPTIONS AND CONSIDERATIONS WHEN SELECTING A QUANTITATIVE PROTEOMICS STRATEGY.* NAT BIOTECHNOL, 2010. **28**(7): P. 710-21.

71. HEBERT, A.S., ET AL., *THE ONE HOUR YEAST PROTEOME.* MOL CELL PROTEOMICS, 2014. **13**(1): P. 339-47.

72. MICHALSKI, A., ET AL., *MASS SPECTROMETRY-BASED PROTEOMICS USING Q EXACTIVE, A HIGH-PERFORMANCE BENCHTOP QUADRUPOLE ORBITRAP MASS SPECTROMETER.* MOL CELL PROTEOMICS, 2011. **10**(9): P. M111 011015.

73. OLSEN, J.V. AND M. MANN, *IMPROVED PEPTIDE IDENTIFICATION IN PROTEOMICS BY TWO CONSECUTIVE STAGES OF MASS SPECTROMETRIC FRAGMENTATION.* PROC NATL ACAD SCI U S A, 2004. **101**(37): P. 13417-22.

74. GEIGER, T., J. COX, AND M. MANN, *PROTEOMICS ON AN ORBITRAP BENCHTOP MASS SPECTROMETER USING ALL-ION FRAGMENTATION.* MOL CELL PROTEOMICS, 2010. **9**(10): P. 2252-61.

75. DISTLER, U., ET AL., *DRIFT TIME-SPECIFIC COLLISION ENERGIES ENABLE DEEP-COVERAGE DATA-INDEPENDENT ACQUISITION PROTEOMICS.* NAT METHODS, 2014. **11**(2): P. 167-70.

76. GILLET, L.C., ET AL., *TARGETED DATA EXTRACTION OF THE MS/MS SPECTRA GENERATED BY DATA-INDEPENDENT ACQUISITION: A NEW CONCEPT FOR CONSISTENT AND ACCURATE PROTEOME ANALYSIS.* MOL CELL PROTEOMICS, 2012. **11**(6): P. O111 016717.

77. KAO, A., ET AL., *DEVELOPMENT OF A NOVEL CROSS-LINKING STRATEGY FOR FAST AND ACCURATE IDENTIFICATION OF CROSS-LINKED PEPTIDES OF PROTEIN COMPLEXES.* MOL CELL PROTEOMICS, 2011. **10**(1): P. M110 002212.

78. DOERR, A., *MASS SPECTROMETRY-BASED TARGETED PROTEOMICS.* NAT METHODS, 2013. **10**(1): P. 23.

79. MALLICK, P., ET AL., *COMPUTATIONAL PREDICTION OF PROTEOTYPIC PEPTIDES FOR QUANTITATIVE PROTEOMICS.* NAT BIOTECHNOL, 2007. **25**(1): P. 125-31.

80. LANGE, V., ET AL., *SELECTED REACTION MONITORING FOR QUANTITATIVE PROTEOMICS: A TUTORIAL.* MOL SYST BIOL, 2008. **4**: P. 222.

81. COX, J. AND M. MANN, *QUANTITATIVE, HIGH-RESOLUTION PROTEOMICS FOR DATA-DRIVEN SYSTEMS BIOLOGY.* ANNU REV BIOCHEM, 2011. **80**: P. 273-99.

82. MOORE, R.E., M.K. YOUNG, AND T.D. LEE, *PROTEIN IDENTIFICATION USING A QUADRUPOLE ION TRAP MASS SPECTROMETER AND SEQUEST DATABASE MATCHING.* CURR PROTOC PROTEIN SCI, 2001. **CHAPTER 16**: P. UNIT 16 10.

83. LISTGARTEN, J. AND A. EMILI, *STATISTICAL AND COMPUTATIONAL METHODS FOR COMPARATIVE PROTEOMIC PROFILING USING LIQUID CHROMATOGRAPHY-TANDEM MASS SPECTROMETRY.* MOL CELL PROTEOMICS, 2005. **4**(4): P. 419-34.

84. NESVIZHSKII, A.I., O. VITEK, AND R. AEBERSOLD, *ANALYSIS AND VALIDATION OF PROTEOMIC DATA GENERATED BY TANDEM MASS SPECTROMETRY.* NAT METHODS, 2007. **4**(10): P. 787-97.

85. COX, J. AND M. MANN, *MAXQUANT ENABLES HIGH PEPTIDE IDENTIFICATION RATES, INDIVIDUALIZED P.P.B.-RANGE MASS ACCURACIES AND PROTEOME-WIDE PROTEIN QUANTIFICATION.* NAT BIOTECHNOL, 2008. **26**(12): P. 1367-72.

86. COX, J., ET AL., *ANDROMEDA: A PEPTIDE SEARCH ENGINE INTEGRATED INTO THE MAXQUANT ENVIRONMENT.* J PROTEOME RES, 2011. **10**(4): P. 1794-805.

87. MOORE, R.E., M.K. YOUNG, AND T.D. LEE, *QSCORE: AN ALGORITHM FOR EVALUATING SEQUEST DATABASE SEARCH RESULTS.* J AM SOC MASS SPECTROM, 2002. **13**(4): P. 378-86.

88. ELIAS, J.E. AND S.P. GYGI, *TARGET-DECOY SEARCH STRATEGY FOR INCREASED CONFIDENCE IN LARGE-SCALE PROTEIN IDENTIFICATIONS BY MASS SPECTROMETRY.* NAT METHODS, 2007. **4**(3): P. 207-14.

89. KALL, L., ET AL., *ASSIGNING SIGNIFICANCE TO PEPTIDES IDENTIFIED BY TANDEM MASS SPECTROMETRY USING DECOY DATABASES.* J PROTEOME RES, 2008. **7**(1): P. 29-34.

90. ONG, S.E. AND M. MANN, *MASS SPECTROMETRY-BASED PROTEOMICS TURNS QUANTITATIVE.* NAT CHEM BIOL, 2005. **1**(5): P. 252-62.

91. BANTSCHEFF, M., ET AL., *QUANTITATIVE MASS SPECTROMETRY IN PROTEOMICS: A CRITICAL REVIEW.* ANAL BIOANAL CHEM, 2007. **389**(4): P. 1017-31.

92. ONG, S.E., ET AL., *STABLE ISOTOPE LABELING BY AMINO ACIDS IN CELL CULTURE, SILAC, AS A SIMPLE AND ACCURATE APPROACH TO EXPRESSION PROTEOMICS.* MOL CELL PROTEOMICS, 2002. **1**(5): P. 376-86.

93. ONG, S.E., *UNBIASED IDENTIFICATION OF PROTEIN-BAIT INTERACTIONS USING BIOCHEMICAL ENRICHMENT AND QUANTITATIVE PROTEOMICS.* COLD SPRING HARB PROTOC, 2010. **2010**(3): P. PDB PROT5400.

94. GEIGER, T., ET AL., *SUPER-SILAC MIX FOR QUANTITATIVE PROTEOMICS OF HUMAN TUMOR TISSUE.* NAT METHODS, 2010. **7**(5): P. 383-5.

95. SCHWANHAUSSER, B., ET AL., *GLOBAL ANALYSIS OF CELLULAR PROTEIN TRANSLATION BY PULSED SILAC.* PROTEOMICS, 2009. **9**(1): P. 205-9.

96. GYGI, S.P., ET AL., *QUANTITATIVE ANALYSIS OF COMPLEX PROTEIN MIXTURES USING ISOTOPE-CODED AFFINITY TAGS.* NAT BIOTECHNOL, 1999. **17**(10): P. 994-9.

97. BOERSEMA, P.J., ET AL., *MULTIPLEX PEPTIDE STABLE ISOTOPE DIMETHYL LABELING FOR QUANTITATIVE PROTEOMICS.* NAT PROTOC, 2009. **4**(4): P. 484-94.

98. HSU, J.L., ET AL., *STABLE-ISOTOPE DIMETHYL LABELING FOR QUANTITATIVE PROTEOMICS.* ANAL CHEM, 2003. **75**(24): P. 6843-52.

99. THOMPSON, A., ET AL., *TANDEM MASS TAGS: A NOVEL QUANTIFICATION STRATEGY FOR COMPARATIVE ANALYSIS OF COMPLEX PROTEIN MIXTURES BY MS/MS.* ANAL CHEM, 2003. **75**(8): P. 1895-904.

100. ROSS, P.L., ET AL., *MULTIPLEXED PROTEIN QUANTITATION IN SACCHAROMYCES CEREVISIAE USING AMINE-REACTIVE ISOBARIC TAGGING REAGENTS.* MOL CELL PROTEOMICS, 2004. **3**(12): P. 1154-69.

101. DAYON, L., ET AL., *RELATIVE QUANTIFICATION OF PROTEINS IN HUMAN CEREBROSPINAL FLUIDS BY MS/MS USING 6-PLEX ISOBARIC TAGS.* ANAL CHEM, 2008. **80**(8): P. 2921-31.

102. ZIESKE, L.R., *A PERSPECTIVE ON THE USE OF ITRAQ REAGENT TECHNOLOGY FOR PROTEIN COMPLEX AND PROFILING STUDIES.* J EXP BOT, 2006. **57**(7): P. 1501-8.

103. BONDARENKO, P.V., D. CHELIUS, AND T.A. SHALER, *IDENTIFICATION AND RELATIVE QUANTITATION OF PROTEIN MIXTURES BY ENZYMATIC DIGESTION FOLLOWED BY CAPILLARY REVERSED-PHASE LIQUID CHROMATOGRAPHY-TANDEM MASS SPECTROMETRY.* ANAL CHEM, 2002. **74**(18): P. 4741-9.

104. WIENER, M.C., ET AL., *DIFFERENTIAL MASS SPECTROMETRY: A LABEL-FREE LC-MS METHOD FOR FINDING SIGNIFICANT DIFFERENCES IN COMPLEX PEPTIDE AND PROTEIN MIXTURES.* ANAL CHEM, 2004. **76**(20): P. 6085-96.

105. STRITTMATTER, E.F., ET AL., *PROTEOME ANALYSES USING ACCURATE MASS AND ELUTION TIME PEPTIDE TAGS WITH CAPILLARY LC TIME-OF-FLIGHT MASS SPECTROMETRY.* J AM SOC MASS SPECTROM, 2003. **14**(9): P. 980-91.

106. SILVA, J.C., ET AL., *QUANTITATIVE PROTEOMIC ANALYSIS BY ACCURATE MASS RETENTION TIME PAIRS.* ANAL CHEM, 2005. **77**(7): P. 2187-200.

107. 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.* MOL CELL PROTEOMICS, 2012. **11**(3): P. O111 009613.

108. BEYNON, R.J., ET AL., *MULTIPLEXED ABSOLUTE QUANTIFICATION IN PROTEOMICS USING ARTIFICIAL QCAT PROTEINS OF CONCATENATED SIGNATURE PEPTIDES.* NAT METHODS, 2005. **2**(8): P. 587-9.

109. BROWNRIDGE, P., ET AL., *GLOBAL ABSOLUTE QUANTIFICATION OF A PROTEOME: CHALLENGES IN THE DEPLOYMENT OF A QCONCAT STRATEGY.* PROTEOMICS, 2011. **11**(15): P. 2957-70.

110. 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. 265-73.

111. RAPPSILBER, J., ET AL., *LARGE-SCALE PROTEOMIC ANALYSIS OF THE HUMAN SPLICEOSOME.* GENOME RES, 2002. **12**(8): P. 1231-45.

112. 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.* MOL CELL PROTEOMICS, 2005. **4**(9): P. 1265-72.

113. SILVA, J.C., ET AL., *ABSOLUTE QUANTIFICATION OF PROTEINS BY LCMSE: A VIRTUE OF PARALLEL MS ACQUISITION.* MOL CELL PROTEOMICS, 2006. **5**(1): P. 144-56.

114. SCHWANHAUSSER, B., ET AL., *GLOBAL QUANTIFICATION OF MAMMALIAN GENE EXPRESSION CONTROL.* NATURE, 2011. **473**(7347): P. 337-42.

115. WISNIEWSKI, J.R., ET AL., *EXTENSIVE QUANTITATIVE REMODELING OF THE PROTEOME BETWEEN NORMAL COLON TISSUE AND ADENOCARCINOMA.* MOL SYST BIOL, 2012. **8**: P. 611.

116. MARCO Y. HEIN, K.S., JÜRGEN COX, MATTHIAS MANN, *PROTEOMIC ANALYSIS OF CELLULAR SYSTEMS*, IN *HANDBOOK OF SYSTEMS BIOLOGY*. 2013. P. 3-25.

117. ANDERSON, N.L. AND N.G. ANDERSON, *THE HUMAN PLASMA PROTEOME: HISTORY, CHARACTER, AND DIAGNOSTIC PROSPECTS.* MOL CELL PROTEOMICS, 2002. **1**(11): P. 845-67.

118. NAGARAJ, N., ET AL., *DEEP PROTEOME AND TRANSCRIPTOME MAPPING OF A HUMAN CANCER CELL LINE.* MOL SYST BIOL, 2011. **7**: P. 548.

119. LUQUE-GARCIA, J.L. AND T.A. NEUBERT, *ON-MEMBRANE TRYPTIC DIGESTION OF PROTEINS FOR MASS SPECTROMETRY ANALYSIS.* METHODS MOL BIOL, 2009. **536**: P. 331-41.

120. PICOTTI, P., ET AL., *FULL DYNAMIC RANGE PROTEOME ANALYSIS OF S. CEREVISIAE BY TARGETED PROTEOMICS.* CELL, 2009. **138**(4): P. 795-806.

121. WISNIEWSKI, J.R., A. ZOUGMAN, AND M. MANN, *COMBINATION OF FASP AND STAGETIP-BASED FRACTIONATION ALLOWS IN-DEPTH ANALYSIS OF THE HIPPOCAMPAL MEMBRANE PROTEOME.* J PROTEOME RES, 2009. **8**(12): P. 5674-8.

122. WANG, Y., ET AL., *REVERSED-PHASE CHROMATOGRAPHY WITH MULTIPLE FRACTION CONCATENATION STRATEGY FOR PROTEOME PROFILING OF HUMAN MCF10A CELLS.* PROTEOMICS, 2011. **11**(10): P. 2019-26.

123. PENG, M., ET AL., *PROTEASE BIAS IN ABSOLUTE PROTEIN QUANTITATION.* NAT METHODS, 2012. **9**(6): P. 524-5.

124. DELAHUNTY, C.M. AND J.R. YATES, 3RD, *MUDPIT: MULTIDIMENSIONAL PROTEIN IDENTIFICATION TECHNOLOGY.* BIOTECHNIQUES, 2007. **43**(5): P. 563, 565, 567 PASSIM.

125. CRISTOBAL, A., ET AL., *IN-HOUSE CONSTRUCTION OF A UHPLC SYSTEM ENABLING THE IDENTIFICATION OF OVER 4000 PROTEIN GROUPS IN A SINGLE ANALYSIS.* ANALYST, 2012. **137**(15): P. 3541-8.

126. LIEBERMANN, R.A., *EVALUATION OF MICRON SIZED SILICA BASED PACKING MATERIAL FOR ULTRA HIGH PRESSURE CAPILLARY LIQUID CHROMATOGRAPHY*, IN *DEPARTMENT OF CHEMISTRY, UNIVERSITY OF NORTH CAROLINA*2009.

127. KELSTRUP, C.D., ET AL., *OPTIMIZED FAST AND SENSITIVE ACQUISITION METHODS FOR SHOTGUN PROTEOMICS ON A QUADRUPOLE ORBITRAP MASS SPECTROMETER.* J PROTEOME RES, 2012. **11**(6): P. 3487-97.

128. PICOTTI, P., ET AL., *A COMPLETE MASS-SPECTROMETRIC MAP OF THE YEAST PROTEOME APPLIED TO QUANTITATIVE TRAIT ANALYSIS.* NATURE, 2013. **494**(7436): P. 266-70.

129. GETZ, E.B., ET AL., *A COMPARISON BETWEEN THE SULFHYDRYL REDUCTANTS TRIS(2-CARBOXYETHYL)PHOSPHINE AND DITHIOTHREITOL FOR USE IN PROTEIN BIOCHEMISTRY.* ANAL BIOCHEM, 1999. **273**(1): P. 73-80.

130. SHAFER, D.E., J.K. INMAN, AND A. LEES, *REACTION OF TRIS(2-CARBOXYETHYL)PHOSPHINE (TCEP) WITH MALEIMIDE AND ALPHA-HALOACYL GROUPS: ANOMALOUS ELUTION OF TCEP BY GEL FILTRATION.* ANAL BIOCHEM, 2000. **282**(1): P. 161-4.

131. PIRMORADIAN, M., ET AL., *RAPID AND DEEP HUMAN PROTEOME ANALYSIS BY SINGLE-DIMENSION SHOTGUN PROTEOMICS.* MOL CELL PROTEOMICS, 2013. **12**(11): P. 3330-8.

132. CEJKA, J., Z. VODRAZKA, AND J. SALAK, *CARBAMYLATION OF GLOBIN IN ELECTROPHORESIS AND CHROMATOGRAPHY IN THE PRESENCE OF UREA.* BIOCHIM BIOPHYS ACTA, 1968. **154**(3): P. 589-91.

133. MASUDA, T., M. TOMITA, AND Y. ISHIHAMA, *PHASE TRANSFER SURFACTANT-AIDED TRYPSIN DIGESTION FOR MEMBRANE PROTEOME ANALYSIS.* J PROTEOME RES, 2008. **7**(2): P. 731-40.

134. VENKATESH, R. AND P.V. SUNDARAM, *MODULATION OF STABILITY PROPERTIES OF BOVINE TRYPSIN AFTER IN VITRO STRUCTURAL CHANGES WITH A VARIETY OF CHEMICAL MODIFIERS.* PROTEIN ENG, 1998. **11**(8): P. 691-8.

135. FINEHOUT, E.J., J.R. CANTOR, AND K.H. LEE, *KINETIC CHARACTERIZATION OF SEQUENCING GRADE MODIFIED TRYPSIN.* PROTEOMICS, 2005. **5**(9): P. 2319-21.

136. RICE, R.H., G.E. MEANS, AND W.D. BROWN, *STABILIZATION OF BOVINE TRYPSIN BY REDUCTIVE METHYLATION.* BIOCHIM BIOPHYS ACTA, 1977. **492**(2): P. 316-21.

137. LOPEZ-FERRER, D., J.L. CAPELO, AND J. VAZQUEZ, *ULTRA FAST TRYPSIN DIGESTION OF PROTEINS BY HIGH INTENSITY FOCUSED ULTRASOUND.* J PROTEOME RES, 2005. **4**(5): P. 1569-74.

138. CHEN, W.Y. AND Y.C. CHEN, *ACCELERATION OF MICROWAVE-ASSISTED ENZYMATIC DIGESTION REACTIONS BY MAGNETITE BEADS.* ANAL CHEM, 2007. **79**(6): P. 2394-401.

139. RIAL-OTERO, R., ET AL., *ULTRASONIC ASSISTED PROTEIN ENZYMATIC DIGESTION FOR FAST PROTEIN IDENTIFICATION BY MATRIX-ASSISTED LASER DESORPTION/IONIZATION TIME-OF-FLIGHT MASS SPECTROMETRY. SONOREACTOR VERSUS ULTRASONIC PROBE.* J CHROMATOGR A, 2007. **1166**(1-2): P. 101-7.

140. LEE, B., ET AL., *RAPID AND EFFICIENT PROTEIN DIGESTION USING TRYPSIN-COATED MAGNETIC NANOPARTICLES UNDER PRESSURE CYCLES.* PROTEOMICS, 2011. **11**(2): P. 309-18.

141. DYCKA, F., ET AL., *RAPID AND EFFICIENT PROTEIN ENZYMATIC DIGESTION: AN EXPERIMENTAL COMPARISON.* ELECTROPHORESIS, 2012. **33**(2): P. 288-95.

142. LEE, B.C. AND V.N. GLADYSHEV, *THE BIOLOGICAL SIGNIFICANCE OF METHIONINE SULFOXIDE STEREOCHEMISTRY.* FREE RADIC BIOL MED, 2011. **50**(2): P. 221-7.

143. HOUGHTEN, R.A. AND C.H. LI, *REDUCTION OF SULFOXIDES IN PEPTIDES AND PROTEINS.* ANAL BIOCHEM, 1979. **98**(1): P. 36-46.

ACKNOWLEDGEMENTS

My very special gratitude goes to Matthias Mann, who has been much more than a mentor and advisor, but also a friend. His amazing support and his believe in my abilities made this work true joy despite the hard labor. I hope we'll have many more years of exciting research and lots of reasons to celebrate together.

The most important supporter (besides Matthias) in our lab was and to some extend is Nagarjuna Nagaraj. Naga taught me how to run a measurement, how to write a paper without going crazy and how much fun lab work can be. I will always be thankful and in a huge debt to him.

Many thanks go to Richie (Richard Scheltema) who's taught me a lot about proper quality controls and how to celebrate these good measurements. He's been a tremendous help and true friend since I started my Ph.D. Along these lines, I would also like to thank Herbert Schiller for being an inspiring researcher and a good friend who somehow never gets tired and keeps on rocking.

I would also like to thank my very good friends and colleagues Garwin Picher, Sean Humphrey, Mario Oroshi, and Igor Paron for making me laugh and distractions when I needed them most. I don't know what I would have done without you guys.

One major backbone of my work and someone I could always count on is Korbi Mayr. Besides being a genius with the mass spectrometers, he is always there to fulfil my crazy ideas.

Jürgen Cox deserves a very special acknowledgement. Not only for providing me and others with the amazing MaxQuant software but for his excellent taste for food. The many lunches at Autobahn Thai were a true joy.

The best discussions about new technology and when I'm stuck are the ones I have with Marco Hein. He inspired me to keep on thinking; I hope I'll never have to miss his great input and smart humor.

I would like to also acknowledge all the rest of the department for the good times and all the support.

My very special thanks go to Felix Spira for his never ending friendship, for always being there for me in tough situations, and the hilarious times we always have together. With him everything seems a bit better and I hope it will always stay this way.

I would also like to thank Charlotte Spira for reminding me that there still is a world without work. She’s always been there for me and I hope she always will. She is family and friend and her support means the world to me.

Last but definitely not least I would like to thank my family. My brother Jan Kulak has always been a major influence on my life and without him, I would never have achieved to get so far. His family, Anita, Momo, and Jona, who I love dearly and who make my life so much better. My parents Petra and Franz Kulak were always there for me and supported all my decisions. They have been a great inspiration and a source of energy and happiness.