Exploiting the multiplexing capabilities of tandem mass tags for high-throughput estimation of cellular protein abundances by mass spectrometry

Erik Ahrné[1], Amalia Martinez-Segura[2], Afzal Pasha Syed[1], Arnau Vina-Vilaseca[1], Andreas J. Gruber[1], Samuel Marguerat[2] and Alexander Schmidt[1]

1) Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland
2) Quantitative Gene Expression group, MRC Clinical Sciences Centre,, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN, UK

Corresponding authors: Alexander Schmidt (alex.schmidt@unibas.ch)

Keywords: Mass spectrometry, absolute protein quantification, tandem mass tags, Schizosaccharomyces pombe, HEK 293, iBAQ

Running title: High-throughput estimation of cellular protein levels

ABSTRACT

The generation of dynamic models of biological processes critically depends on the determination of precise cellular concentrations of biomolecules. Measurements of system-wide absolute protein levels are particularly valuable information in systems biology. Recently, mass spectrometry based proteomics approaches have been developed to estimate protein concentrations on a proteome-wide scale. However, for very complex proteomes, fractionation steps are required, increasing samples number and instrument analysis time. As a result, the number of full proteomes that can be routinely analyzed is limited. Here we combined absolute quantification strategies with the multiplexing capabilities of isobaric tandem mass tags to determine cellular protein abundances in a high throughput and proteome-wide scale even for highly complex biological systems, such as a whole human cell line. We generated two independent data sets to demonstrate the power of the approach regarding sample throughput, dynamic range, quantitative precision and accuracy as well as proteome coverage in comparison to existing mass spectrometry based strategies.

1. Introduction

To date, global and quantitative analysis of mRNA levels across conditions has been the method of choice for systematic high-throughput measurement of gene expression. This strategy has been so popular that most biological insights generated from high throughput data result from transcriptomics analysis[1-3]. Despite impressive technical advances [1,4-7] and numerous novel applications [1,4-11], transcriptomics approaches cannot inform on post-transcriptional processes, such as translational controls and regulated protein degradation [1,4-7,12,13], and therefore do not provide quantitative estimates of actual protein levels in a cell or organism. Recent large-scale studies have indeed demonstrated very different regulation at the transcript and protein levels [1,8-11,14-16], which could be assigned to variations in protein synthesis and degradation rates [17-19]. Thus, there is a strong need to supplement the transcript-centered view of biological systems with global quantitative measurements of proteins. Analyzing expression levels of proteins with similar throughput, analytical depth and quantitative accuracy as provided by transcriptomics approaches would, without doubt, provide data highly relevant for characterizing a biological system.

Recent developments in mass spectrometry (MS)-based proteomics have made possible the comprehensive and quantitative analysis of proteomes in entire organisms [20-23] including the determination of absolute protein concentrations on a system-wide level [24-29]. However, high proteome coverage in higher eukaryotes comes at a cost; it can currently only be achieved by extensive sample pre-fractionation [30-32] that multiplies instrument time and limits throughput to a few samples [21,22,33]. Recently developed chemical tags that support multiplexed analysis of up to 10 samples without increasing sample complexity, such as the tandem mass tags (TMT) [4,5,34] or isobaric tag for relative and absolute quantitation (iTRAQ) [8,30] technology, have already demonstrated to increase sample throughput several fold. In combination with state-of-the-art fast scanning high-resolution MS platforms, these techniques can increase analytical speed to a degree compatible with high-throughput proteome measurements of fractionated samples in higher eukaryotes and achieve an analytical depth and quantitative accuracy similar to transcriptomics approaches [17,19,35,36]. Notably, the quantitative information obtained by these labeling

approaches is limited to relative comparisons of protein levels between samples and does not provide actual cellular protein concentrations that are often required for modeling of biological processes.

Here, we combined the high-throughput capabilities of the TMT labeling technology with recent label-free approaches for system-wide estimation of protein abundances from MS intensities to derive comprehensive protein concentrations snapshots for large sample batches in eukaryotic organisms. We assessed our approach in two different eukaryotic model systems, the fission yeast *Schizosaccharomyces pombe*, and a human cell line. We could demonstrate an important increase in sample throughput and proteome coverage compared to existing protocols while maintaining high quantitative accuracy of protein level estimates. Our multiplexed absolute quantification strategy is therefore well suited for global high-throughput analysis of cellular protein abundances in complex eukaryotes, including human cells. Our protocol will help to increase the number of full proteome data sets and further improve, in combination with other "omics" approaches, our understanding of very complex biological systems.

2. Material and Methods

2.1. Sample preparation

In principle, most protein samples, including proteins extracted from tissues and body fluids, are amenable to the multiplexed absolute quantification strategy described here. A minimum of 25 ug of total protein amount, corresponding to around 100,000 human cells, is recommended per samples. We applied our isobaric tags based quantification method to estimate absolute global protein levels in *S. pombe* and *HEK 293* cells, and compared our data with previously published methods and data sets.

2.1.1. Cell lysis and proteolysis

Human cells, HEK 293 cell line [20,37], were exponentially grown in DMEM media with 10% FCS for 8 days to 90% confluence. Cells were collected by centrifugation at 400g at 4°C, washed twice with 2 ml ice-cold PBS buffer, harvested by centrifugation at 500g and the pellet was snap frozen in liquid nitrogen and stored at -80˚C until further processing. For S*. pombe*, cultures of wild type 972 h+ cells were grown in YE medium [24,30] at 25˚C to concentrations of $3.5x10^6$ cells/ml and around 10^8 cells of each sample were collected by centrifugation at 2,000xg, washed twice with PBS buffer and harvested by centrifugation at 2,000xg. Cells were resuspended in 100µl lysis buffer (100 mM ammoniumbicarbonate, 2% sodium deoxycholate, 5mM TCEP), disrupted by strong indirect sonication (100% amplitude, 0.5 cycle, 2 × 10 s) in a Vial Tweeter (Hielscher) and reduced for 15min at 95°C. Since the reducing reagent can interfere with the accuracy of standard protein colorimetric assays (e.g. Bradford, BCA), a novel reducing reagent tolerant BCA assay was used to determine accurate protein concentrations of the samples (Reducing Agent-Compatible BCA Assay, Thermo Fisher Scientific). Notably, the tandem mass tag labeling efficiency critically depends on accurate peptide amounts to achieve complete labeling and therefore, the BCA measurement should be done in triplicates and with various sample amounts to obtain protein concentrations of high confidence. In the meantime, samples were alkylated with 10mM iodoacetamide for 30min in the dark at 25°C. After quenching the reaction with 12mM N-acetyl-cysteine, the proteins were proteolyzed for 4h at

37°C using sequencing-grade Lys-C (Wako Chemicals) at 1/200 w/w. Then, the samples were diluted with 100mM ammoniumbicarbonate buffer to a final sodium deoxycholate concentration of 1% and further digested by incubation with sequencing-grade modified trypsin (1/50, w/w; Promega, Madison, Wisconsin) over night at 37°C. The sequential double Lys-C/trypsin digestion has been recently shown to be more efficient in generating fully cleaved peptides than tryptic digest alone [30] and is therefore recommended for all protein samples analyzed by LC-MS. Sodium deoxycholate is one of a few LC-MS compatible detergents, since it precipitates after acidification [25,33,38] using 2M HCl to a final concentration of 50mM and can then be easily removed by centrifugation at 10,000g for 15min before LC-MS analysis. All peptide samples were then desalted by C18 reversed-phase spin columns according to the manufacturer's instructions (Macrospin, Harvard Apparatus), separated in aliquots of 25 ug peptides, dried under vacuum and stored at -80ºC until further use. For label-free single dimension 1D-LC/MS analysis, samples were solubilized in solvent A (98% water, 2% acetonitrile, 0.15% formic acid) at a concentration of 0.5 ug/ul and 2 ul were injected per LC-MS run.

2.1.2. TMT labeling

Sample aliquots comprising 25 ug of peptides were subsequently labeled with isobaric tandem mass tags (TMT 6-plex, Thermo Fisher Scientific) using a previously published protocol [34,38] with a few modifications. Specifically, the TMT reagents were dissolved in 21 ul of DMSO, respectively, and 5 ul of each TMT reagent was added to the individual peptide samples solubilized in 20 ul labeling buffer (2M Urea, 0.2 M HEPES, pH 8.3). It is important to note that the dissolved reagents should be stored at -20°C and used within a few weeks after solubilization to maintain good labeling efficiency. After tagging peptides for 1 hour at 25°C, the reaction was quenched by adding 1.5 ul of an aqueous 1.5M hydroxylamine solution and incubating for another 10 minutes. After pooling all labeled peptide samples, the pH of the sample pool was increased to 11.9 by adding 1M phosphate buffer (pH 12) for 20 minutes to remove TMT labels linked to peptide hydroxyl groups that also form during the labeling process as an unwanted side product. Subsequently, the peptide solution was acidified using 2M hydrochloric acid followed by sample desalting using C18 reversed-phase spin columns according to the manufacturer's instructions (Macrospin, Harvard Apparatus) and dried under vacuum.

2.1.3. Off-Gel electrophoresis

The dried TMT labeled peptides were solubilized in 1800 µl Off-Gel electrophoresis buffer according to the manufacturer`s instructions (3100 OFFGEL Fractionator, Agilent Technologies). Then, peptide mixtures were separated on a 13cm linear pH 3-10 IPG strip (GE Healthcare) using a protocol of 1h rehydration at maximum 500V, 50µA and 200mW. Peptides were separated at maximum 8000V, 100µA and 300mW until 20kVh was reached. Notably, since the TMT labeled peptides are not equally distributed across the applied pH range, fractions that contain fewer peptides (for both data sets we identified much less peptides in fractions 3 and 10, see Supplementary Figure 1) can be combined with other fractions of lower complexity to generate samples of similar complexity. Thus, the total number of fractions that have to be analyzed can be reduced while keeping fraction complexity manageable for the LC-MS platform. The peptide fractions were subsequently desalted using C18 reversed-phase columns according to the manufacturer's instructions (Microspin, Harvard

Apparatus), dried under vacuum and solubilized in 20 ul 0.1% formic acid. 2 ul of the peptide samples were subjected to LC-MS/MS analysis.

2.2. LC-MS/MS analysis

The setup of the µRPLC-MS system was as described previously [14,30]. Chromatographic separation of peptides was carried out using an EASY nano-LC 1000 system (Thermo Fisher Scientific), equipped with a heated RP-HPLC column (75 µm x 50 cm) packed in-house with 1.9 µm C18 resin (Reprosil-AQ Pur, Dr. Maisch). Aliquots of 1 µg total peptides were analyzed per LC-MS/MS run using a linear gradient ranging from 95% solvent A (0.15% formic acid, 2% acetonitrile) and 5% solvent B (98% acetonitrile, 2% water, 0.15% formic acid) to 30% solvent B over 180 minutes at a flow rate of 200 nl/min. Mass spectrometry analysis was performed on a dual pressure LTQ-Elite (for all human samples) or LTQ-Velos (for all *S. pombe* samples) Orbitrap mass spectrometer equipped with a nanoelectrospray ion source (both Thermo Fisher Scientific) and a custom made column heater set to 60°C. For TMT samples, each MS1 scan (acquired in the Orbitrap) was followed by high-collision-dissociation (HCD, acquired in the Orbitrap) of the 10 most abundant precursor ions with dynamic exclusion for 60 seconds. Total cycle time was approximately 2s. For MS1, 10E6 ions were accumulated in the Orbitrap cell over a maximum time of 300ms and scanned at a resolution of 120,000 (LTQ-Velos 30,000) FWHM (at 400 m/z). MS2 scans were acquired at a target setting of 50,000 ions, accumulation time of 100ms and a resolution of 15,000 (LTQ-Velos 7,500) FWHM (at 400 m/z). To minimize ratio distortion for reporter ion quantification [35,36,39], the mass selection window was reduced to 1 Da. For LFQ samples, each MS1 scan (acquired in the Orbitrap) was followed by collision-induced-dissociation (CID, acquired in the linear ion trap) of the 20 most abundant precursor ions with dynamic exclusion for 60 seconds. Total cycle time was approximately 2s. For MS1, 10E6 ions were accumulated in the Orbitrap cell over a maximum time of 300ms and scanned at a resolution of 240,000 (LTQ-Velos 60,000) FWHM (at 400 m/z). MS2 scans were acquired at a target setting of 10,000 ions, accumulation time of 25ms and normal scan rate. The preview mode was activated and the mass selection window was set to 2 Da. For all LC-MS measurements, singly charged ions and ions with unassigned charge state were excluded from triggering MS2 events. Besides, the normalized collision energy was set to 35% and one microscan was acquired for each spectrum.

2.3. Data analysis

2.3.1. Database searching and relative protein quantification (TMT dataset)

The acquired raw-files were converted to the mascot generic file (mgf) format using the msconvert tool (part of ProteoWizard, version 3.0.4624 (2013-6-3)). Using the MASCOT algorithm (Matrix Science, Version 2.4.0), the mgf files were searched against a decoy database containing normal and reverse sequences of the predicted entries of Homo sapiens (SwissProt, www.uniprot.org, release date 30/10/2014) or *Schizosaccharomyces pombe* (ftp://ftp.sanger.ac.uk/, release date 10/08/2011) including commonly observed contaminants (in total 41,250 sequences for Homo sapiens and 10,584 for *S. pombe*) generated using the SequenceReverser tool from the MaxQuant software (Version 1.0.13.13). The precursor ion tolerance was set to 10 ppm and fragment ion tolerance was set to 0.02 Da. The search criteria were set as follows: full tryptic specificity was required (cleavage

after lysine or arginine residues unless followed by proline), 3 missed cleavages were allowed, carbamidomethylation (C), TMT6plex (K and peptide n-terminus) were set as fixed modification and oxidation (M) as a variable modification. Next, the database search results were imported to the Scaffold Q+ software (version 4.3.2, Proteome Software Inc., Portland, OR) and the protein false identification rate was set to 1% based on the number of decoy hits. Specifically, peptide identifications were accepted if they could achieve an FDR less than 1.0% by the scaffold local FDR algorithm. Protein identifications were accepted if they could achieve an FDR less than 1.0% and contained at least 1 identified peptide. Protein probabilities were assigned by the Protein Prophet program [14,37]. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters. Acquired reporter ion intensities in the experiments were employed for automated quantification and statically analysis using a modified version of our in-house developed SafeQuant R script [30,40]. This software including a detailed description is publicly available via GitHub (https://github.com/eahrne/SafeQuant/). The single steps include correction of reporter ion intensities for isotopic impurities according to the manufacturer's instructions followed by summing the reporter ion intensities for each peptide and protein identification, global normalization across all acquisition runs and ratio calculation and statistical analysis.

2.3.2. Database searching and relative protein quantification (LFQ dataset)

For label-free quantification (LFQ), the generated raw files were imported into the Progenesis LC-MS software (Nonlinear Dynamics, Version 4.0) and analyzed using the default parameter settings. MS/MS-data were exported directly from Progenesis in mgf format and searched against the same decoy databases as above using MASCOT. The search criteria were set as follows: full tryptic specificity was required (cleavage after lysine or arginine residues); 3 missed cleavages were allowed; carbamidomethylation (C) was set as fixed modification; oxidation (M) as variable modification. The mass tolerance was set to 10 ppm for precursor ions and 0.6 Da for fragment ions. Results from the database search were imported into Progenesis and the peptide false discovery rate (FDR) was set to 1% using the number of reverse hits in the dataset. In particular, the mass error, peptide length and mascot score were used to reduce FDR rates. Reducing the number of false hits on the peptide level is crucial to maximize the number of quantified proteins. The final protein lists containing the summed peak areas of all identified peptides for each protein, respectively, were exported from Progenesis LC-MS and further statically analyzed using an in-house developed R script (SafeQuant) [30,35,36].

2.3.3. Estimation of cellular protein abundances (TMT dataset)

The generated raw files obtained from TMT labeled and fractionated peptide samples were subjected to LFQ as described above using the Progenesis LC-MS software tool. Each raw file was separately imported and analyzed applying the appropriate database search parameters for TMT labeled peptides as explained above. The result files of all 10 fractions were combined and a single protein list comprising the summed precursor ion abundances of all quantified peptide ions for each protein was generated. Next, these protein abundances representing the combined MS1 intensities of all 6 TMT labeled samples were distributed across the individual 6 samples according to the relative reporter ion intensities (Figure 1). Specifically, the reporter ion intensity of each sample was

divided by the sum of all reporter ion intensities and the calculated ratio was multiplied with the corresponding MS1 protein intensity. In this way, sample specific MS1 intensities and iBAQ values could be calculated and absolute protein concentrations estimated in each of the 6 samples as recently specified [25,38,41]. In detail, the MS1 intensities were divided by the number of possible tryptic peptides amenable for LC-MS analysis of each protein to control protein size and sequence differences and log transformed. Importantly, the generated iBAQ values have to be correlated with accurate levels of at least 10 proteins ideally spanning the whole concentration range of interest to generate an unbiased abundance estimation model [25,38]. To do so, we deployed protein concentration as copies per cell from two recently published studies comprising 34 *S. pombe* [14,42] and 25 human proteins [14,28,39]. The accuracy of the approach was evaluated by comparing the results to those obtained from a standard label-free quantification analysis of the same un-fractionated samples. Specifically, we compared fold errors derived by leave one out cross validation (LOOCV) as well as proteome coverage using recently published protein concentrations for *S. pombe* [14] and human cells [40]. All quantified proteins employed in this manuscript including calculation of iBAQ values are displayed in Supplementary Table S1 (*S. pombe* data set) and S2 (HEK 293 cell data set).

2.4. Accession code

All raw mass spectrometry data files and the resulting Mascot output files can be downloaded from http://www.proteomexchange.org/ using the following accession key: (data submitted, px-submission #39730, final key will be provided when ready).

3. Results

To demonstrate the power of the approach, including its speed, analytical depth, accuracy and throughput, we applied it to a simple eukaryote, *S. pombe*, and to a complex model system, human embryonic kidney cells.

3.1. System-wide estimation of protein abundances in S. pombe

For the first data set, we prepared six peptide samples from proliferating *S. pombe* cells, labeled them with 6-plex TMT reagents, respectively, and fractionated the labeled peptide pool using Off-gel electrophoresis (for details see Figure 1). Each fraction and an aliquot of the unlabeled samples were analyzed by LC-MS/MS followed by label-free quantification (LFQ) to determine precursor ion intensities (MS1 quantification). Since the MS intensities derived from isobaric labeled TMT peptides comprise the sum of all six samples included in the experiment, an additional, sample specific MS2 quantification of the reporter ions present in the tandem mass spectra was performed to split the precursor intensities according to the levels present in the original samples. To test, if these calculated MS intensities (here referred to as TMT-2D data) actually represent those in the single samples, we first correlated them with those directly determined by LFQ from unlabeled samples (here referred to as LFQ-1D data). Protein MS intensities obtained by the two different strategies were very similar as indicated by a high squared Pearson correlation coefficient (R^2=0.814, Figure 2A). It is important to note that TMT quantification was not corrected for any ratio interference that has been shown to suppress ratio values [28,29,35,36]. However, since this only affects a small portion of the quantified proteins, in fact those that are regulated, its global impact on the quantification accuracy is rather small. Nonetheless, recently developed multi-notch approaches

[38,41], which are fully compatible with our workflow, but require special MS instrumentation that was not available for this study, have shown to reduce ratio distortion and should further increase quantification accuracy of our approach. We next tested if these calculated MS intensities can also be employed for estimating absolute protein concentration within the single samples using the recently introduced iBAQ approach [14,25,38]. Therefore, we determined iBAQ values for each protein and samples and correlated them with published protein concentrations (in copies/cell) accurately determined by stable isotope dilution [42]. We observed a good correlation of predicted and actual protein concentrations (Figure 2B) and low median fold errors. The quantification error was determined by 'leave one out' cross validation (LOOCV) and visibly both TMT-2D and LFQ-1D methods have similar quantification accuracies (Figure 2C). It is important to note that these quantities are measured values with errors and therefore the calculated accuracies are slight underestimations. Spiking in more accurately quantified reference proteins, like the commercially available universal protein standard (UPS, Sigma-Aldrich) consisting of 48 absolutely quantified human proteins, will provide more accuracy assessment. However, since these proteins need to be quantified within the same analytical background to be accurate[38], its application is limited to non-human and mouse samples to avoid interferences with peptides derived from endogenous proteins. Having shown that the TMT-2D approach is capable of estimating absolute protein levels within complex samples with a similar accuracy as conventional methods, we next had a closer look at the proteome coverage and penetration achieved by the two methods. As shown in Figure 2D, considerably more proteins (77.5%) were quantified by our TMT-2D method (in total 2739 proteins) compared to the standard LFQ-1D workflow (in total 1543 proteins). Additionally, the concentration range of the quantified proteins covered was much wider for the TMT-2D dataset spanning 5 orders of magnitude compared to around 4 for the LFQ-1D approach. Notably, the linear dynamic range of the TMT quantification is limited [43] and consequently quantities determined for highly variable proteins outside this range should be carefully checked. However, since only a tiny fraction of the quantified proteins are usually outside this range (in this data set only 3 of the 3000 quantified *S. pombe* proteins), their quantities can be manually investigated and verified with relatively little efforts. For samples with strong protein concentration differences, MS methods that increase reporter ion intensities and therefore the overall linear dynamic quantification range should be considered[44,45].

To conclude, the TMT approach presented here was capable of system-wide estimation of protein abundances at accuracies similar to the commonly used LFQ only approaches, even when the sample is extensively fractionated. Furthermore, the proteome coverage and dynamic concentrations range achieved rivals those of unlabeled samples [14,28,29] while increasing sample throughout 6-fold.

3.2. Extensive estimation of protein abundances in a human cell line

The comprehensive quantitative analysis of highly complex proteomes, coming from human cell lines for instance, is a very challenging task even when using state-of-the-art mass spectrometry approaches. Yet, human cells are a very relevant model system for biological studies and effective proteomics strategies that allow high throughput, comprehensive, and quantitative analysis of cellular protein inventories in complex samples, such as the method described here, are required to level the scope of proteomics with that of transcriptomics.

To test, if our strategy is capable of this challenging task, we applied it to estimate absolute protein abundances in a commonly studied human cell line (HEK 293 cells). Since we did include biological

replicate samples in this data set, we employed them to evaluate the precision of the determined protein levels. As shown in Supplementary Figure 2, we observed very high Pearson correlation coefficients (R^2) of more than 0.99 (A-C) and very low coefficients of variances (median around 5%, D) for the triplicate measurements performed, indicating that the overall precision of our method is very high. In line with the *S. pombe* data set, we evaluated the quantitative accuracy and proteome coverage of the quantitative data set by comparing it to an additional data set generated form the same samples using a conventional LFQ-1D workflow. In agreement with the previous results, we observed a very good correlation of TMT-2D with LFQ-1D MS1 protein intensities (Figure 3A) and with published cellular concentrations of human proteins [39] (Figure 3B), even when analyzing this very complex protein mixture. Furthermore, the fold error distributions calculated by LOOCV were in a similar range for both datasets (Figure 3C), indicating that our workflow is also applicable to highly complex protein samples such as a whole human cell lysate. Strikingly, we could quantify 6818 proteins spanning a concentration range of 7 orders of magnitude. Compared to the LFQ-1D data set that included 1741 proteins, we could increase proteome coverage by around 3-fold and analytical depth by 3 orders of magnitude with only a slight increase in measurement time. It is important to note that the use of the novel TMT 10-plex reagents [19] will bring MS measurement time in the same range as LFQ-1D while maintaining a similar extended proteome coverage and dynamic range. Moreover, the proteome depth achieved was in the same range as existing large-scale proteomics studies of human cell lines analyzing fractionated samples [1,9-11,40], however, the multiplexing abilities of the TMT tag considerably increased throughput by 6-fold.

Overall, the conclusions drawn from the human cell line study match those of the *S. pombe* experiments. Once again, when compared to a previously developed LFQ approach, the TMT-2D method allowed for multiplexed system-wide estimation of protein abundances at a similar accuracy and depth, but with much higher throughput than reported previously for highly complex samples [28,29].

4. Conclusion

Our multiplexed protein abundance estimation approach demonstrated high proteome coverage, sample throughput (6-fold higher than existing methods with the potential to further increase using novel 10-plex TMT tags) and quantitative precision and an accuracy that is similar to established label-free approaches. Compared to existing label-free approaches, the TMT protocol requires only a few additional steps during sample preparation and the extra costs for reagents are largely compensated by savings in instrument time. This is the first quantitative MS-based approach that is capable to determine cellular protein concentrations for higher eukaryotes in a high throughput and system-wide manner, and will be an important instrument in the “omics” toolbox. Latest fast scanning mass spectrometers, like quadrupole Orbitrap or time-of-flight instruments, are ideally suited for analyzing isobaric labeled peptides and will further boost proteome coverage and quantification accuracy and dynamic range of the TMT approach described here. Defining the actual concentration of proteins within cells and across a high number of samples is key for the generation of quantitative models in system biology. The ability to derive these numbers from higher eukaryotes in a high-throughput manner will enable, in combination with other quantitative data,

the development of novel algorithms that model biological processes in humans during health and disease.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council, UK (AMS, SM). APS, AVV and AJG are supported in part by SystemsX StoNets RTD and in part by ERC grant (Project #310510 WHYMIR). We also thank Mihaela Zavolan for helping to design and plan the HEK 293 experiment.

Figure legends

Figure 1:
Overview of the multiplexed system-wide estimation of cellular protein levels strategy using tandem mass tags (TMT). The cells of up to 10 different samples are lysed, proteins extracted and proteolyzed using trypsin. After desalting, the generated peptides are labeled with different tandem mass tag versions, respectively, pooled, fractionated using OFFGEL electrophoresis (OGE) and analyzed by shotgun LC-MS/MS. The acquired MS data are subjected to two independent data analysis workflows focusing on the quantification of the sample specific reporter ions present in the MS2 spectra to derive relative quantitative differences between samples (MS2 quantification). Additionally, a precursor ion centered label-free quantification (MS1 quantification) is performed to determine precursor ion intensities of all quantified peptide ions. Since all TMT labeled peptides have the exact same precursor ion masses, the MS1 quantification determines the summed peptide ion intensities of all samples included in the experiment. In a next step, the ratio of the sample specific reporter ion intensities (MS2 quantification) can be employed to distribute pooled precursor ion abundances (MS1 quantification) across the individual samples and generate sample specific precursor ion intensities. These can then be applied for label-free estimation of absolute protein concentrations of each sample, respectively, using well-established protocols [38].

Figure 2:
Evaluation of the multiplexed protein abundance estimation workflow using *S. pombe* samples. (A) Correlation of the MS1 precursor ion based protein abundances for the *S. pombe* sample using the new multiplexed TMT approach including peptide fractionation (TMT-2D) and a standard label-free quantification workflow of the unfractionated sample (LFQ-1D). The linear regression line (black dashed line) together with the corresponding equation and Pearson correlation coefficient (R^2) are indicated. (B) The correlation of the actual cellular abundances of 34 selected proteins (in copies/cell) recently determined by stable isotope dilution [14] and the estimated concentrations (also in copies/cell) using our TMT-2D workflow is shown including the corresponding Pearson correlation coefficient (R^2) and median fold error determined by leave one out cross validation (LOOCV). (C) Box plots illustrating the fold error determined by LOOCV for the protein concentrations estimated by LFQ-1D (red) and our multiplexed TMT-2D workflow (blue). The black bar indicates the median fold error. (D) Line chart indicating the abundance distribution (in copies/cell) [12-14] of the proteins quantified by TMT-2D (blue line) and LFQ-1D (red line).

Figure 3:
Evaluation of the multiplexed TMT protein abundance estimation workflow using *a human* cell line. (A) Correlation of the MS1 precursor ion based protein abundances for the same *HEK 293* sample analyzed in triplicates using our multiplexed TMT approach including peptide fractionation (TMT-2D) and a standard label-free quantification workflow of the unfractionated sample (LFQ-1D). The linear regression line (black dashed line) together with the corresponding equation and Pearson correlation coefficient (R^2) are indicated. (B) The correlation of the actual cellular abundances of 23 selected proteins (in copies/cell) recently determined by heavy labeled reference polypeptides [14-16,39] and the estimated concentrations (also in copies/cell) using our TMT-2D workflow is shown including the corresponding Pearson correlation coefficient (R^2) and median fold error determined by leave one out cross validation (LOOCV). The standard deviations observed between biological

triplicates are also indicated as error bars for each protein, respectively. (C) Box plots illustrating the fold error determined by LOOCV for the protein concentrations estimated by LFQ-1D (red) and our multiplexed TMT-2D workflow (blue). The black bar indicates the median fold error. (D) Line chart indicating the concentration distribution (in copies/cell) [18,40] of the proteins quantified by TMT-2D (blue line) and LFQ-1D (red line).

References

Figure 1:

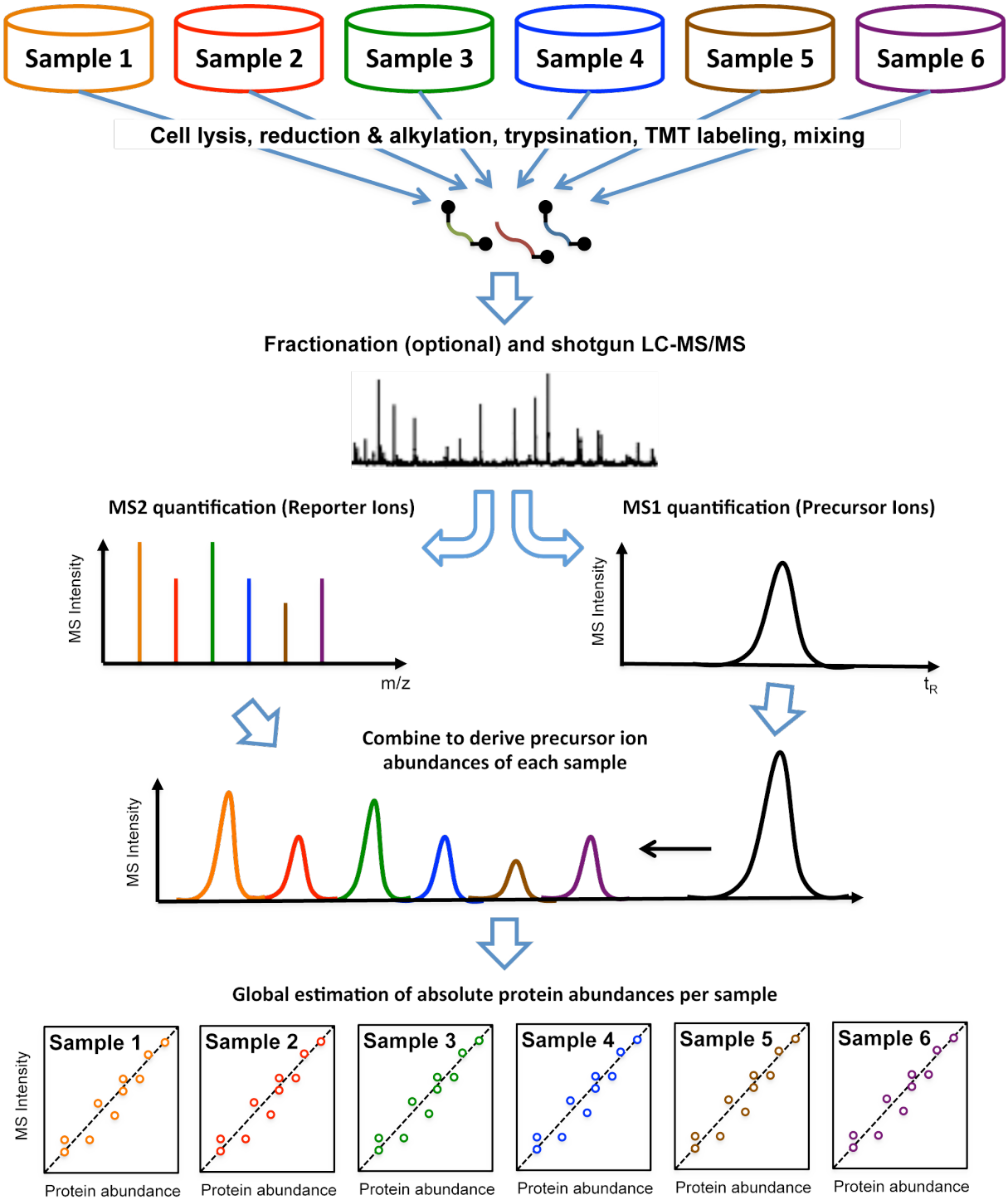


Figure 2:

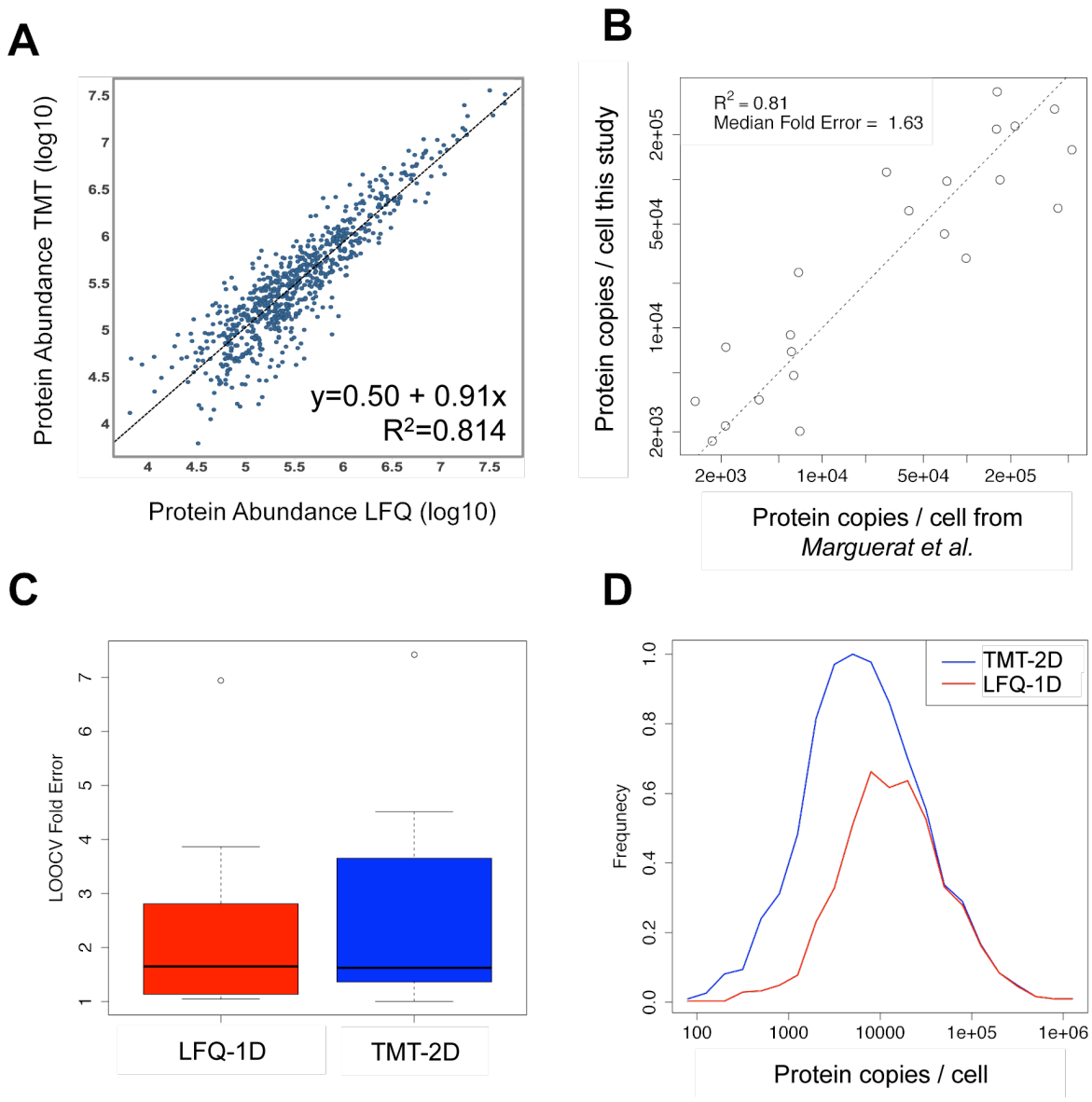
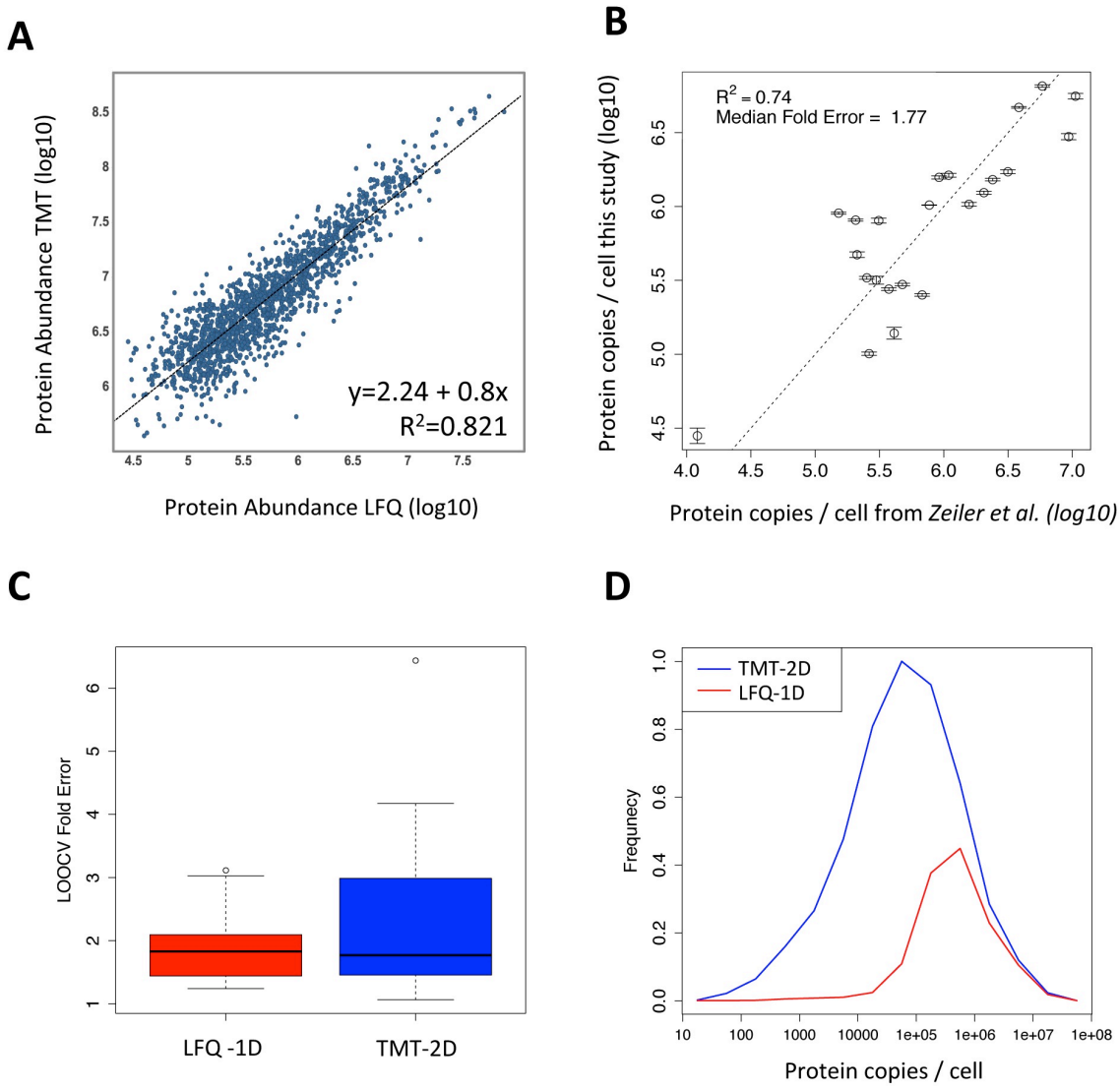

A
Protein Abundance TMT (log10)
Protein Abundance LFQ (log10)
y=0.50 + 0.91x
R²=0.814
B
Protein copies / cell this study
R² = 0.81
Median Fold Error = 1.63
Protein copies / cell from
Marguerat et al.
C
LOOCV Fold Error
LFQ-1D
TMT-2D
D
Frequnecy
TMT-2D
LFQ-1D
Protein copies / cell

Figure 3:


A
Protein Abundance TMT (log10)
Protein Abundance LFQ (log10)
y=2.24 + 0.8x
R2=0.821
B
R2 = 0.74
Median Fold Error = 1.77
Protein copies / cell this study (log10)
Protein copies / cell from Zeiler et al. (log10)
C
LOOCV Fold Error
LFQ -1D
TMT-2D
D
TMT-2D
LFQ-1D
Frequnecy
Protein copies / cell

References:

[1] E. Arner, C.O. Daub, K. Vitting-Seerup, R. Andersson, B. Lilje, F. Drabløs, et al., Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells, Science. 347 (2015) 1010–1014.

[2] A. Fort, K. Hashimoto, D. Yamada, M. Salimullah, C.A. Keya, A. Saxena, et al., Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance, Nat Genet. 46 (2014) 558–566.

[3] S.A. Morris, P. Cahan, H. Li, A.M. Zhao, A.K. San Roman, R.A. Shivdasani, et al., Dissecting engineered cell types and enhancing cell fate conversion via CellNet, Cell. 158 (2014) 889–902.

[4] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, et al., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS, Anal Chem. 75 (2003) 1895–1904.

[5] T. Werner, I. Becher, G. Sweetman, C. Doce, M.M. Savitski, M. Bantscheff, High-resolution enabled TMT 8-plexing, Anal Chem. 84 (2012) 7188–7194.

[6] S. Marguerat, B.T. Wilhelm, J. Bähler, Next-generation sequencing: applications beyond genomes, Biochem Soc Trans. 36 (2008) 1091–1096.

[7] B.T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, et al., Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, Nature. 453 (2008) 1239–1243.

[8] R.D. Unwin, A. Pierce, R.B. Watson, D.W. Sternberg, A.D. Whetton, Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells, Mol Cell Proteomics. 4 (2005) 924–935.

[9] J. Hausser, M. Zavolan, Identification and consequences of miRNA-target interactions - beyond repression of gene expression, Nat Rev Genet. 15 (2014) 599–612.

[10] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, et al., A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, Science. 321 (2008) 956–960.

[11] G.-W. Li, D. Burkhardt, C. Gross, J.S. Weissman, Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources, Cell. 157 (2014) 624–635.

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

[13] P. Lu, C. Vogel, R. Wang, X. Yao, E.M. Marcotte, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation, Nat Biotechnol. 25 (2006) 117–124.

[14] S. Marguerat, A. Schmidt, S. Codlin, W. Chen, R. Aebersold, J. Bähler, Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells, Cell. 151 (2012) 671–683.

[15] T. Maier, A. Schmidt, M. Güell, S. Kühner, A.-C. Gavin, R. Aebersold, et al., Quantification of mRNA and protein and integration with protein turnover in a bacterium, Mol Syst Biol. 7 (2011) 511.

[16] M.V. Lee, S.E. Topper, S.L. Hubler, J. Hose, C.D. Wenger, J.J. Coon, et al., A dynamic model of proteome changes reveals new roles for transcript alteration in yeast, Mol Syst Biol. 7 (2011) 514.

[17] A.R. Gruber, G. Martin, P. Müller, A. Schmidt, A.J. Gruber, R. Gumienny, et al., Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells, Nat Commun. 5 (2014) 5465.

[18] A.R. Kristensen, J. Gsponer, L.J. Foster, Protein synthesis rate is the predominant regulator of protein expression during differentiation, Mol Syst Biol. 9 (2013) 689.

[19] J.A. Paulo, F.E. McAllister, R.A. Everley, S.A. Beausoleil, A.S. Banks, S.P. Gygi, Effects of MEK inhibitors GSK1120212 and PD0325901 in vivo using 10-plex quantitative proteomics and

phosphoproteomics, Proteomics. (2014).
[20] J. Hausser, A.P. Syed, N. Selevsek, E. van Nimwegen, L. Jaskiewicz, R. Aebersold, et al., Timescales and bottlenecks in miRNA-dependent gene regulation, Mol Syst Biol. 9 (2013) 711.
[21] K. Baerenfaller, J. Grossmann, M.A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, et al., Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics, Science. 320 (2008) 938–941.
[22] L.M.F. de Godoy, J.V. Olsen, J. Cox, M.L. Nielsen, N.C. Hubner, F. Fröhlich, et al., Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast, Nature. 455 (2008) 1251–1254..
[23] E.L. Huttlin, M.P. Jedrychowski, J.E. Elias, T. Goswami, R. Rad, S.A. Beausoleil, et al., A tissue-specific atlas of mouse protein phosphorylation and expression, Cell. 143 (2010) 1174–1189.
[24] S. Moreno, A. Klar, P. Nurse, Molecular genetic analysis of fission yeast Schizosaccharomyces pombe, Meth. Enzymol. 194 (1991) 795–823.
[25] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, et al., Global quantification of mammalian gene expression control, Nature. 473 (2011) 337–342.
[26] J. Malmstrom, M. Beck, A. Schmidt, V. Lange, E.W. Deutsch, R. Aebersold, Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans, Nature. 460 (2009) 762–765.
[27] Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F.U. Hartl, M.J. Kerner, et al., Protein abundance profiling of the Escherichia coli cytosol, BMC Genomics. 9 (2008) 102.
[28] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, et al., The quantitative proteome of a human cell line, Mol Syst Biol. 7 (2011) 549.
[29] N. Nagaraj, J.R. Wiśniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, et al., Deep proteome and transcriptome mapping of a human cancer cell line, Mol Syst Biol. 7 (2011) 548.
[30] T. Glatter, C. Ludwig, E. Ahrné, R. Aebersold, A.J.R. Heck, A. Schmidt, 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. 11 (2012) 5145–5156.
[31] S. Elschenbroich, V. Ignatchenko, P. Sharma, G. Schmitt-Ulms, A.O. Gramolini, T. Kislinger, Peptide separations by on-line MudPIT compared to isoelectric focusing in an off-gel format: application to a membrane-enriched fraction from C2C12 mouse skeletal muscle cells, J. Proteome Res. 8 (2009) 4860–4869.
[32] Y. Wang, F. Yang, M.A. Gritsenko, Y. Wang, T. Clauss, T. Liu, et al., Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells, Proteomics. 11 (2011) 2019–2026.
[33] J. Zhou, T. Zhou, R. Cao, Z. Liu, J. Shen, P. Chen, et al., Evaluation of the application of sodium deoxycholate to proteomic analysis of rat hippocampal plasma membrane, J. Proteome Res. 5 (2006) 2547–2553.
[34] A. Schmidt, J. Kellermann, F. Lottspeich, A novel strategy for quantitative proteomics using isotope-coded protein labels, Proteomics. 5 (2005) 4–15.
[35] L. Ting, R. Rad, S.P. Gygi, W. Haas, MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics, Nat Meth. 8 (2011) 937–940.
[36] M.M. Savitski, T. Mathieson, N. Zinn, G. Sweetman, C. Doce, I. Becher, et al., Measuring and managing ratio compression for accurate iTRAQ/TMT quantification, J. Proteome Res. (2013).
[37] A.I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, A statistical model for identifying proteins by tandem mass spectrometry, Anal Chem. 75 (2003) 4646–4658.
[38] E. Ahrné, L. Molzahn, T. Glatter, A. Schmidt, Critical assessment of proteome-wide label-free absolute abundance estimation strategies, Proteomics. 13 (2013) 2567–2578.

[39] M. Zeiler, W.L. Straube, E. Lundberg, M. Uhlén, M. Mann, 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 (2012) O111.009613.
[40] N.A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells, Nat Meth. (2014).
[41] G.C. McAlister, D.P. Nusinow, M.P. Jedrychowski, M. Wühr, E.L. Huttlin, B.K. Erickson, et al., MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes, Anal Chem. (2014).
[42] S.A. Gerber, J. Rush, O. Stemman, M.W. Kirschner, S.P. Gygi, Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS, Proc Natl Acad Sci USA. 100 (2003) 6940–6945.
[43] A.F.M. Altelaar, C.K. Frese, C. Preisinger, M.L. Hennrich, A.W. Schram, H.T.M. Timmers, et al., Benchmarking stable isotope labeling based quantitative proteomics, J Proteomics. 88 (2013) 14–26.
[44] J.K. Diedrich, A.F.M. Pinto, J.R. Yates, Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot, J Am Soc Mass Spectrom. 24 (2013) 1690–1699.
[45] L. Dayon, C. Pasquarello, C. Hoogland, J.-C. Sanchez, A. Scherl, Combining low- and high-energy tandem mass spectra for optimized peptide quantification with isobaric tags, J Proteomics. 73 (2010) 769–777.

Supplementary Figures

Click here to download Supplementary Material: Supplemental_Figures.pdf

Supplementary Tables

Click here to download Supplementary Material: Supplemental_Tables.xlsx