A deeper look at carrier proteome effects for single-cell proteomics

Jesper Olsen (jesper.olsen@cpr.ku.dk)
University of Copenhagen https://orcid.org/0000-0002-4747-4938
Zilu Ye
University of Copenhagen https://orcid.org/0000-0001-8829-6579
Tanveer Batth
University of Copenhagen
Patrick Leopold Rüther
Novo Nordisk Foundation Center for Protein Research https://orcid.org/0000-0003-4461-9828

Brief Communication

Keywords: TMTpro, fold-change, protein copy number

Posted Date: August 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-783371/v1

Version of Record: A version of this preprint was published at Communications Biology on February 22nd, 2022. See the published version at https://doi.org/10.1038/s42003-022-03095-4.

Abstract

We probe the carrier proteome effects in single cell proteomics with mixed species TMTpro-labeled samples. We demonstrate that carrier proteomes, while increasing overall identifications, dictate which proteins are identified. We show that quantitative precision and signal intensity are limited at high carrier levels, hindering the recognition of regulated proteins. Guidelines for optimized mass spectrometry acquisition parameters and best practices for fold-change or protein copy number-based comparisons are provided.

Main

Mass spectrometry-based single cell proteomics (SCP-MS) has recently seen significant developments[1–3]. To overcome analytical barriers such as insufficient peptide ion signals for MS identification and quantification, a multiplexing strategy based on labeling tryptic peptides from single cells with isobaric tandem mass tags (TMT) alongside a labeled carrier proteome to boost MS signal has been developed[4]. Several studies have recently emphasized the importance of increasing the number of ions sampled from the single-cell channels (SCCs) with a carrier proteome channel (CPC), concluding that the depth of peptide identification needs to be balanced against accuracy of quantification[5–7]. In this study, we highlight additional crucial factors for performing SCP-MS experiments, these include: 1) proper selection of the carrier proteome; 2) unneglectable isotope impurities caused by the carrier channel; 3) balance between signal-to-noise ratio (SNR), collisional energy and resolution; 4) suitability of SNR and intensity for different data interpretation strategies.

We modeled an SCP-MS experiment using TMTpro[8] 16plex labeling reagents, where channel 126 severed as the carrier proteome channel (CPC), 127C was left empty, and the last 14 channels represented SCCs at different ratios to the CPC (Fig. 1a). To achieve this, we constructed a mixed species sample from *homo sapiens* (HeLa cells), *Saccharomyces cerevisiae* (Yeast) and *Escherichia coli* (E. coli), which were pooled at different known ratios in the SCCs, in order to elucidate the bidirectional effect on identification and quantification in SCP (Fig. 1a). We investigated the effects of CPC quantities and proteome types by designing different CPC constructs with one of three different carrier proteomes: Human only (H), *E. coli* and yeast mixed (EY), and all three species (HEY) mixed across a large range (14x to 434x) of CPC to SCC ratios (hereafter as carrier levels). Together with samples without any CPC (no carrier), these samples were analyzed by liquid chromatography tandem mass spectrometry (LC-MS) with different MS parameters (Fig. 1a, **Supplementary Note 1)**. Loading amounts (50pg to 200pg) per SCC were equivalent to single cell proteomes[5].

We first tested how different carrier proteomes affect protein identifications across the mixed species channels. We compared number of non-human and human proteins identified with EY, Y and HEY as the carrier proteome as well as without any carrier proteome (Fig. 1b). The carrier proteomes primarily dictated which proteins were identified in different SCCs. Moreover, this pronounced bias correlated directly with the carrier levels. The results suggests that carrier proteomes need to be properly weighed to

act as impartial carriers for all proteins in SCCs. We next examined total numbers of proteins identified and quantified at different carrier levels. As expected, including a CPC increased the numbers of identified proteins consistently with rising carrier levels (Fig. 1c, **Supplementary Fig. 1, Supplementary Table 1**). However, the number of human proteins with precise quantification across the 14 SCCs (CV < = 20 %) peaked at relatively lower carrier levels (42x). In the sample containing the highest carrier level (434x), the majority of identified proteins could not be reproducibly quantified (Fig. 1c).

We explored the relationship between quantitative precision and number of fragment ions, and found very high carrier levels led to worse correlations and inferior quantification performance despite higher number of total ions accumulated for MS/MS scans (**Supplementary Fig. 2)**. We compared averaged SNR of the 14 single cell reporters (Av14) or a subset of 12 reporters with the least isotope impurities (Av12) with the respective CV values in the 14 or 12 SCCs for human peptide spectral matches (PSMs) and proteins (Fig. 1d, **Supplementary Fig. 2, Supplementary Fig. 3**). The CV values displayed negative correlations with the average SNR at both PSM and protein levels in agreement with the findings in other studies[5]. Higher carrier ratios limited the maximum single cell SNR, as the signal of both reporter ions and peptide fragment ions primarily derived from the CPC (Fig. 1d). It should be noted that the reporter ion intensities, which are normalized by injection time, correlated worse with the CV values than SNR (**Supplementary Fig. 2, Supplementary Fig. 3**).

Furthermore, we observed significant contribution of isotopic impurities, particularly from carrier channel 126. We calculated protein CV's in all 14 SCCs with either raw or impurity corrected SNR of reporter ions and found impurity correction substantially increased the number of proteins with CV < = 20% in samples with carrier levels higher than 98x (Fig. 1e). At increased ratios, we observed that channel 126 produced noteworthy isotopic impurities in addition to those in the empty 127C channel (**Supplementary Note 2**), which affect channel 128C (126 + 2x13C), and importantly, 127N (126 + 15N). Of note, Cheung *et al.* also noticed this negative impact of channel 127N but ascribed it to ion coalescence[5]. The impurities explain worse quantification at high booster ratios if ignored. In fact, TMT can accurately quantify ratios higher than 400 even for channel 127N after impurity correction (Fig. 1f). Unfortunately, impurity correction also led to higher variations of quantified ratios and the correction for 15N is not available in most data processing tools (**Supplementary Note 3**). Due to the negative impact by 127N and 128C, we calculated the CV of all PSMs and proteins without these two channels, resulting in much more accurate and reproducible quantifications on the 12 unaffected channels (Fig. 1d, 1e, **Supplementary Fig. 3**). Next, we aimed to evaluate the overall quantitative accuracy across all 14 SCCs in yeast peptides by comparing their relative intensities against the expected values (**Supplementary Fig. 4**). Similar to quantitative precision, accuracy was highly dependent on SNR. Distributions of relative intensities in SCCs were highly dispersed at low SNR and they converged to expected values with increased SNR. As the Av14 values in samples with high carrier levels was limited, the abundance ratios were distributed almost randomly and led to the poor quantification accuracy.

To assess the impact of the CPC for detecting significantly regulated proteins, we took advantage of the known protein ratios in our mixed species samples and examined the sensitivity and specificity of the

CPC approach. We utilized the predefined relative species abundances between channels of a ratio of 2 for yeast peptides, 0.5 for *E. coli* peptides, and 1 for human peptides (**Supplementary Note 4**). We used the four ratio estimates to perform t-test (represented by a volcano plot) analysis to identify significantly regulated PSMs with log2-fold change higher than 0.5 at $p < 0.05$. In all cases, less than one percent of human PSMs were wrongly assigned as significantly regulated, suggesting a high specificity (Fig. 1g, 1h). Despite lower number of PSMs identified, samples without any carrier were most likely to assign highest percentage of identified yeast and E. coli peptides as correctly regulated, however this sensitivity decreased as carrier levels increased. Ultimately, samples with 98x carrier were detected the highest number of regulated peptides. Conversely, only a small percentage of yeast and E. coli peptides were accurately quantified in samples with very high carrier levels (210x and 424x) despite the highest numbers of identified peptides.

We tested the most direct MS parameters, the normalized collisional energy (NCE) and MS/MS resolution to enhance SNR for quantification accuracy (**Supplementary Note 1**). We found elevated NCE levels (35%-38%) at lower MS/MS resolution is the best compromise between quantification accuracy and identification. In accordance with a previous study[8], NCE levels between 32% and 35% gave most PSMs (Fig. 1i). However, numbers of PSMs with CV < = 20% generally increased as NCE was correspondingly increased, particularly at high carrier levels. This was due to a consistent increases of reporter SNR with higher NCE (Fig. 1j), despite the Sequest HT score function XCorr and MaxQuant Andromeda[9] scores peaking at lower NCE (**Supplementary Fig. 5**). We simultaneously observed a steady decrease of single cell SNR values as carrier levels increased (Fig. 1j) indicating that even NCE level at 38% could not overcome the limits caused by the CPC. Higher MS/MS resolution resulted in higher fraction of identifications with CV < = 20% however this came at the cost of significantly reduced number of PSMs (Fig. 1k, **Supplementary Fig. 6**) due to the slower scan speed.

For profiling cellular heterogeneity based on global protein expressions[10] with the isobaric carrier approach, protein abundances from reporter ions are first extracted and then subjected to dimensionality reduction methods, such as principal component analysis (PCA). To estimate relative protein copy numbers in proteomes, the intensity-based absolute quantification (iBAQ)[11] is the method of choice. Therefore, it is essential to evaluate accuracy of protein abundances derived from reporter ions. We compared 4 different reporter ion abundance values (SNR and intensities both as raw and impurity-corrected, **Supplementary Note 5**) and demonstrated that they resulted in different protein abundance estimates especially in samples with low AGC target (**Supplementary Fig. 7**). Since the protein copy number estimates in our SCP model should match between the CPC and SCCs for each species, we tested the correlations between iBAQ values computed from full scans (MS1) with the CPC and SCCs (Fig. 1l). Protein abundances at MS1 were calculated as summed abundances of identified peptides, where the Minora algorithm in Proteome Discoverer was used to perform untargeted feature detection for the peptides. Protein abundances on reporter ions were calculated as summed quantities of identified peptides from reporter ion abundances. Unlike the quantification of a single protein across TMT channels, the intensity values correlated better with MS1 abundances than SNR values, especially with

low AGC settings (**Supplementary Note 6**). This is likely due to the fact that both reporter ion intensity values and MS1 abundances are scaled based on injection times. Furthermore, carrier levels of 42x and 98x showed best correlations in almost all settings.

In conclusion, our study systematically explored the effects of the isobaric carrier approach using a defined mixed species model and provides a guideline for future SCP experiments (Table 1). Our finding that the carrier proteome specifically boosts the identification of the proteins contained within it opens the door for a variety of “targeted” SCP experiments. As we studied the tradeoff between identifications and quantitation at large carrier levels (> 100x), we observed that the underlying reasons were a compression of the dynamic range of single cell SNR at high carrier levels, and for specific channels the impurities from the carrier channel. Therefore, we suggest excluding channels 127N and 128C for SCP experiments with extreme carrier levels. We tested the sensitivity and specificity of identifying significantly regulated proteins and our model suggests an optimal carrier level of ~ 100x when analyzing 14 SCCs. We recommend using reporter ion SNR for fold-change-based quantifications across channels and reporter intensities for protein copy number estimation within each channel. A higher NCE of up to 35% achieves better quantification performance by enhancing reporter abundances while maintaining peptide identifications. In the future, the performance of SCP will benefit from the development of isobaric tags with higher multiplexing capacity (18-plex [12]), more sensitive instrumentation, and higher dynamic range of mass analyzers. This study provides a roadmap to benchmarking such new developments.

Table 1
Recommended settings in Orbitrap instruments for isobaric labeling based single-cell proteomics

Parameter	Recommended setting	Rationale
AGC	>= AGC300%	Higher AGC target allows more ions in MS2 scans
NCE	35%	Slightly higher NCE leads to higher reporter ion SNR without reducing in peptide fragment quality
Resolution in MS2	>= 60K	● To resolve isobaric reporter ions ● To make use of long fill time needed to reach the high AGC target
Carrier levels	127N and 128C included: <= 100x	Impurities from TMTpro126 are substantial with very high carrier levels
	127N and 128C excluded: > 100x	

References

1. Kelly, R.T. Single-Cell Proteomics: Progress and Prospects. *Mol. Cell. Proteomics* **19**, 1739-1748 (2020).

2. Cong, Y. et al. Ultrasensitive single-cell proteomics workflow identifies> 1000 protein groups per mammalian cell. *Chemical Science* **12**, 1001-1006 (2021).

3. Hartlmayr, D. et al. An automated workflow for label-free and multiplexed single cell proteomics sample preparation at unprecedented sensitivity. *bioRxiv* (2021).

4. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 1-12 (2018).

5. Cheung, T.K. et al. Defining the carrier proteome limit for single-cell proteomics. *Nat. Methods* **18**, 76-83 (2021).

6. Specht, H. & Slavov, N. Optimizing accuracy and depth of protein quantification in experiments using isobaric carriers. *J. Proteome Res.* **20**, 880-887 (2020).

7. Tsai, C.-F. et al. An improved Boosting to Amplify Signal with Isobaric Labeling (iBASIL) strategy for precise quantitative single-cell proteomics. *Mol. Cell. Proteomics* **19**, 828-838 (2020).

8. Li, J. et al. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399-404 (2020).

9. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301-2319 (2016).

10. Yang, L., George, J. & Wang, J. Deep Profiling of Cellular Heterogeneity by Emerging Single-Cell Proteomic Technologies. *Proteomics* **20**, 1900226 (2020).

11. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).

12. Li, J. et al. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.* **20**, 2964-2972 (2021).

13. Bekker-Jensen, D.B. et al. A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Mol. Cell. Proteomics* **19**, 716-729 (2020).

14. Batth, T.S. et al. Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation. *Mol. Cell. Proteomics* **18**, 1027-1035 (2019).

15. Wickham, H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* **3**, 180-185 (2011).

16. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442-D450 (2019).

Methods

Sample preparation. Human epithelial cervix carcinoma HeLa cells were cultured in DMEM (Gibco, Invitrogen) as previously described[13]. Cells were harvested at ~80% confluence by washing twice with PBS (Gibco, Life technologies). *E.coli* were grown on LB medium plates and colonies were scrapped manually and transfered to 1.5ml tubes. E. coli were resuspended in PBS buffer and washed 3 times followed by the centrifugation to pellet the cells and discard the supernatant. For HeLa and E.coli cells, boiling 4% SDS in 50mM Tris pH 8.5 was added added to the cells. The tube was heated for 10 minutes at 95 degrees, and DNA/RNA were sheared by sonication with a tip. Tryptophan assay was utilized to determine protein concentration followed by reduction and alkylation with TCEP and CAA. Sample prep was performed using protein aggregation capture [14] during which proteins were aggregated onto magnetic beads and digested overnight sequentially with Lys-C (1:200 protease to protein ratio) for 2 hours at 37C and Trypsin (1:50) overnight. Mass spec-compatible yeast intact (undigested) extracts were brought from Promega (Catalog number: V7341) and processed according to the technical manual. All the digest supernatant was cleaned using C18 solid phase extraction and the peptide concentration was determined using nano-drop. Digested peptides were labeled with TMTpro following manufacturer's protocol. TMTpro-labeled peptides from different species were pooled with different ratios as described in **Supplementary Table 1**.

LC-MS/MS. All samples were analyzed on an Orbitrap Exploris 480 mass spectrometer coupled with the Evosep One system using an in-house packed capillary column with the pre-programmed 30 samples-per-day gradient in data dependent acquisition mode. The column temperature was maintained at 60 °C using an integrated column oven (PRSO-V1, Sonation, Biberach, Germany). Spray voltage were set to 2 kV, funnel RF level at 40, and heated capillary temperature at 275 °C. Full MS resolutions were set to 120,000 at m/z 200 and full MS AGC target was 300% with an IT of 25 ms. Mass range was set to 350–1400. Intensity threshold was kept at 1E5. Isolation width was set at 0.8 m/z. All data were acquired in profile mode using positive polarity and peptide match was set to off, and isotope exclusion was on. AGC target value, resolution and normalized collision energy (NCE) were set differently for individual samples. A full description of the parameters for each sample was listed in **Supplementary Table 1**.

Data processing and analysis. All raw files were processed in Proteome Discoverer 2.4 (Thermo Fisher Scientific) and MaxQuant with the human, yeast and E.coli Uniprot Reference Proteome database without isoforms (January 2019 release). Trypsin was set as the digest enzyme and up to one missed cleavages was allowed. TMTpro was specified as a fixed modification on lysine and peptide N-terminus, carbamidomethylation of cysteine was specified as fixed modification and methionine oxidation was specified as a variable modification. Precursor and fragment mass tolerances were set to 10ppm and 0.02Da in Sequest HT, respectively. Specifically, reporter abundance was based on either SNR or intensity,

both with raw and impurity-corrected values. No normalization or scaling was applied. All the files were processed in batch mode to get result files individually. A modified modification.xml file was used in MaxQuant to enable TMTpro based database search. All the statistical analysis was conducted with in-house written R-scripts. All the boxplot elements were defined according to default parameters in ggplot2[15].

Declarations

Acknowledgements

This work was supported by Novo Nordisk Foundation (NNF14CC0001 and NNF17SA0027704), and the program of excellence from the University of Copenhagen (CDO2016).

Author contributions

Z.Y., T.B. and J.V.O. conceived and designed the study; Z.Y. and T.B. contributed with experimental data; Z.Y., T.B., P.R. and J.V.O. contributed with data interpretations; Z.Y. wrote the manuscript; and all authors edited and approved the final version.

Competing interests

All authors declare no conflicts of interest.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[16] partner repository with the dataset identifier PXD027742.

Figures

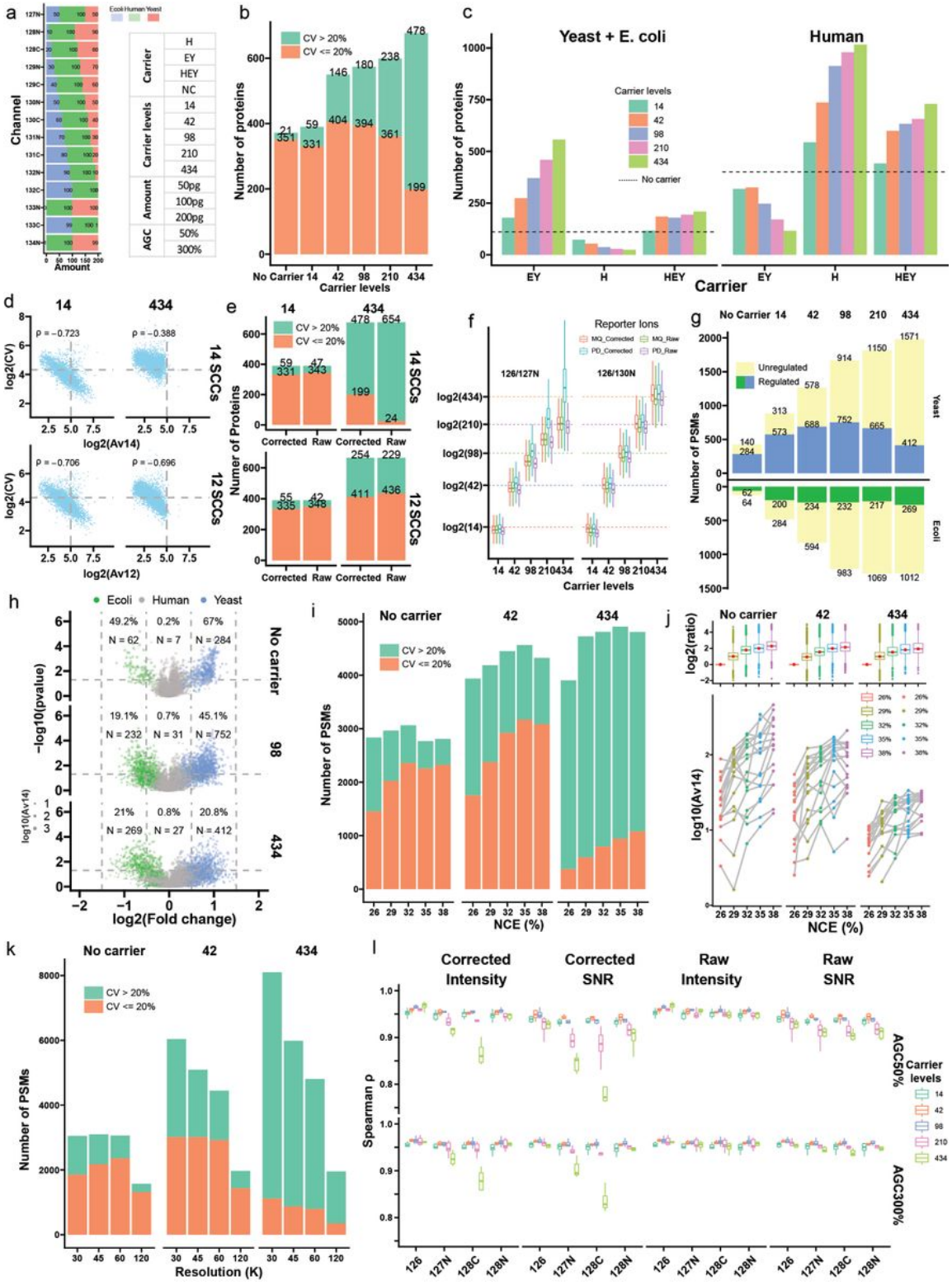

Figure 1

Explore the carrier effect with TMTpro in mixed species samples. a) Depiction of TMTpro labeled mixed species samples and mass spectrometry acquisition settings. Digested peptide mixtures from yeast, E.coli and human HeLa cells were labeled with TMTpro in 14 SCCs and mixed in different proportions. b) Number of identified proteins with different carrier channel ratios. Samples with the following settings were selected: no carrier and carrier proteome; AGC300%; 50pg per SCC; replicate 2. c) Number of

identified proteins from different organisms. Samples with the following settings were selected: no carrier and carrier proteome as HEY, H and EY; AGC300%; 50pg per SCC; replicate 2. d) Relationship between average reporter ion SNR of 14 SCCs (upper panel) or 12 SCCs (lower panel) and CV in identified PSMs from human proteins. Samples with the following settings were selected: carrier proteome as HEY; carrier levels at 14 and 434; AGC300%; 50pg per single cell channel. PSMs with Av14 >= 1 were used. Spearman's rank correlation coefficient was used as a measure of rank correlation. e) Number of identified proteins with CV > 20% and CV <= 20% using either raw or corrected SNR values. CV values were calculated with either 14 SCCs (upper panel) or 12 SCCs (lower panel). Samples with the following settings were selected: carrier proteome as HEY; carrier levels at 14 and 434; AGC300%; 50pg per single cell channel; replicate 2. f) Distribution of log2(126/130N) and log2(126/127N) in all human PSMs using reporter ion values from different methods. Methods for reporter ion values included raw and impurity corrected values from MaxQuant and Proteome Discoverer. Samples with the following settings were selected: carrier proteome as HEY; AGC300%; 200pg per single cell channel. g) Numbers of PSMs accurately quantified with selected TMT channels. From the 14 SCCs, we calculated log2 ratios including, 131C/132N, 131C/132N, 131C/132N and 131C/132N for yeast and human PSMs, 128N/128C, 129N/130C, 129C/131C and 130N/132C for E.coli PSMs. PSMs with Parent intensity fraction >= 0.98 and Av14 >= 1 were kept. Samples with the following settings were selected: carrier proteome not as H; AGC300%. h) Volcano plots of number of PSMs accurately quantified with selected TMT channels. Calculations of numbers were the same as g). PSMs with log2-fold change higher than 0.5 at $p<0.05$ were designated as regulated. N: number of PSMs. i) Number of identified PSMs with CV > 20% and CV <= 20% with different normalized collisional energies (NCE). j) Distribution of Av14 in scans with different NCEs from randomly selected identical precursors (lower panel); boxplot of the log2 ratios of Av14 at different NCEs divided by Av14 at NCE 26%. In both i) and j), samples with the following settings were selected: no carrier and carrier proteome as HEY; no carrier and carrier levels at 14 and 434; NCE at 26%, 29%, 32%, 35% and 38%. Resolution was set to 60K in all samples. k) Number of identified PSMs with CV > 20% and CV <= 20% with different resolutions. Samples with the following settings were selected: no carrier and carrier proteome as HEY; no carrier and carrier levels at 14 and 434; Resolution at 30K, 45K, 60K and 120K. NCE was set to 32% in all samples. l) Correlation between MS1 abundances and abundances from selected reporter ion channels. MS1 abundances were calculated from Minora node in Proteome Discoverer. 4 different kinds of reporter ion abundances were calculated including raw and impurity corrected intensities, raw and impurity corrected SNR. The Spearman’s rank correlation coefficient was used as a measure of rank correlation. A full list of sample descriptions, identified proteins and PSMs can be found in Supplementary Table 1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement.docx