Comparative proteome signatures of ultra-low input samples by TMT multiplexed Data-Independent Acquisition

Authors:

Claudia Ctortecka[1], Gabriela Krššáková[1,2,3], Karel Stejskal[1,2,3], Josef M. Penninger[3,4], Sasha Mendjan[2], Karl Mechtler[1,2,3], Johannes Stadlmann[2,5*].

Affiliations:

1 Research Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Campus-Vienna-Biocenter 1, 1030 Vienna, Austria.

2 Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria.

3 The Gregor Mendel Institute of Molecular Plant Biology of the Austrian Academy of Sciences (GMI), Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria.

4 Department of Medical Genetics, Life Sciences Institute, University of British Columbia, Vancouver Campus, 2350 Health Sciences Mall, Vancouver, BC Canada V6T 1Z3.

5 BOKU - University of Natural Resources and Life Sciences, Muthgasse 18, A 1190 Vienna, Austria

* Correspondence:

Johannes Stadlmann

Institute of Biochemistry

BOKU - University of Natural Resources and Life Sciences

Muthgasse 18

A-1190 Vienna

Austria

email: j.stadlmann@boku.ac.at

Abstract

Single cell proteomics strives to complement transcriptomics approaches in characterizing issues or cell populations at single cell resolution. Despite continuous technological improvements in sensitivity, mass spectrometry-based single cell proteomics ultimately faces the challenge of reproducibly comparing the protein expression profiles of thousands of individual cells. Efficient analytical strategies for the comprehensive and quantitative characterization of such extremely large sample cohorts are scarce. While conventional, data-dependent shotgun proteomics (DDA) of ultra-low input samples critically suffers from the accumulation of missing values with increasing sample-cohort size, alternative data-independent acquisition (DIA) strategies do usually not to take full advantage of isotope-encoded sample multiplexing. Here, we combine two hither opposing analytical strategies, DIA and TMT-multiplexing, to generate highly reproducible, quantitative proteome signatures from ultra-low input samples. Introducing a novel, identification-independent proteomics data analysis approach, we quantitatively compare our DIA-TMT proteome signatures across hundreds of samples, independent of their biological origin, and show that this highly scalable approach allows to reproducibly distinguish between different human cell lines and standard single protein knockouts, even from ultra-low input samples.

Main

Single-cell proteomics aims at differentiating individual cells and subpopulations based on their protein expression profiles. As the comprehensive proteomic characterization of individual mammalian cells still challenges the sensitivity of current MS/MS-based workflows, the concept of multiplexed shotgun proteomics analyses of individual cells in conjunction with a highly abundant, congruent carrier proteome has been seminal to the field.[1] Elegantly, such use of established *in vitro* stable-isotope labeling techniques (e.g. Tandem-Mass-Tag, TMT) not only increases precursor- and fragment-ion abundances for peptide identification and quantification from ultra-low input samples, but also critically increases sample throughput. Currently, such multiplex single-cell proteomics workflows allow for the quantitative analysis of up to 13 barcoded single cells in one analytical run.[2]

Nevertheless, paralleling state-of-the-art transcriptomic datasets, single-cell proteomics ultimately faces the challenge of comparatively analyzing hundreds or even thousands of ultra-low input proteomics samples.[3–5] As such sample cohort sizes vastly exceed the capacities of any currently available MS multiplexing technology[6], data aggregation from multiple individual analytical runs is inevitable. Importantly, merging an increasing number of individual quantitative shotgun proteomics (DDA) datasets into one, most often results in a considerable number of peptides which are not reliably identified in all analytical runs.[7] This method-intrinsic co-accumulation of missing values greatly limits the use of DDA strategies for the comparative analysis of protein levels in large sample cohorts which was addressed by various computational data-imputation or “match-between runs” methods.[8–10]

By contrast, data-independent acquisition (DIA) regimes, which unselectively subject all precursor ions within a defined m/z window to MS/MS analysis, have been demonstrated to allow for the robust quantification of protein expression, even across extremely large sample cohorts.[11] More recently, DIA strategies were further extended to sequentially windowed DIA-

schemes (SWATH), specifically designed to cover all theoretical mass-spectra and to provide deep proteome coverage.[12,13] In the interest of developing a massively scalable high-throughput data-acquisition strategy for comparative single-cell proteomics, we thus explored the combined use of *in vitro* multiplexing strategies for MS/MS based quantification (i.e. TMT10plex ™ Isobaric Label Reagent Set) and small window DIA data-acquisition regimes (i.e. m/z = 6 Th) for the analysis of ultra-low peptide amounts.

DIA-TMT provides highly reproducible, quantitative proteome signatures.

For this, we acquired DDA and DIA MS/MS-data sets in technical triplicates of TMT10plex-labelled tryptic digests derived from two human cell lines, serially diluted to total peptide amounts similar to those expected for single mammalian cells (i.e. 0.3 ng and lower). Two commercially available tryptic digests of whole human cell lysates (i.e. HeLa and K562) were labeled with all available TMT10plex-tags, respectively. The labeled standards were then combined into three different two-proteome mixes equipped with different sets of isobaric tags (detailed in the methods section). Each of the three mixes was then analyzed at four different input levels (i.e. 10, 5, 1 and 0.5 ng total peptide) by nano-HPLC-ESI-MS/MS, using an Orbitrap Exploris™ 480 Mass Spectrometer, fitted with FAIMS Pro (detailed acquisition parameters in the methods section).

Manual inspection of the raw DIA MS/MS-data confirmed a robust and input independent number of MS/MS scans. In DDA, however, the number of MS/MS scans markedly decreased with lower amounts of peptide subjected to analysis (Fig. 1b). Nevertheless, subsequent peptide identification by dedicated MS/MS search engines (i.e. Spectronaut™ for DIA and SpectroMine™ for DDA data) yielded twice as many PSMs and unique peptide identifications in the lower concentrations from DDA data, when compared to corresponding DIA data. At 0.5 ng total peptide input we identified 1.302 and 3.843 PSMs or 1.202 and 3.348 unique peptides in DIA and DDA acquisition mode, respectively. The 1 ng total peptide input samples yielded

1.982 and 4.856 PSMs or 1.817 and 4.296 unique peptides from DIA and DDA strategies, respectively (Fig. 1b-c). These findings demonstrate that, despite steadily generating an enormous amount of MS/MS data, especially in DIA, the number of identified peptides markedly decreases with lower sample input.

To evaluate scalability of DIA strategies for peptide identification in larger sample cohorts by data-aggregation, we merged quantitation data from multiple analytical runs, based on peptide identity. For this, we only considered peptides which were consistently identified in all three technical replicates at a given total peptide input level. As expected, at 10 ng total peptide input, we observed that DIA indeed provided more consistent peptide identifications across multiple analytical runs (i.e. > 90 % central overlap), when compared to DDA (i.e. < 50 % central overlap) (Fig. 1d-e). Importantly, however, this key benefit of DIA strategies was gradually lost with decreasing sample input (Fig. 1d), presumably because of decreasing total ion current. This suggests that, like DDA, also DIA fails to provide consistent peptide identifications across multiple ultra-low input samples, and gives rise to a similarly detrimental accumulation of missing values without data imputation.

Next, we assessed the quantitative information generated by the two MS/MS-data acquisition strategies. For this, we first employed a conventional, identification-based TMT-data analysis workflow, and analyzed unique peptide-derived TMT reporter-ion intensities by principal component analysis (PCA). Although based on very different numbers of data points (i.e. DDA: 1471, DIA: 503) used for PCA, both unique peptide sets allowed a distinction between the two human cell lines via the first two principal components, even at 0.5 ng total peptide level. The first principal component (PC) of the DDA data, which is displayed on the x-axis separates the cell lines with 63% explained variance, while the second PC representing only 25% of the variance discriminates between the analytical runs (y-axis) (Fig. 1f). In DIA, however, the main difference shown in PC1 is attributed to the analytical runs and only PC2 distinguishes the two

sample sets with 9% explained variance (Fig. 1g). This indicates that through identification-dependent aggregation of DDA data the biological variance is the main characteristic, while in DIA it is the technical variability (Fig. 1f-g). Nevertheless, we investigated the clustering dimensions by calculating the Davies-Bouldin (DB) index, that considers both intra and inter-cluster distances (detailed in the methods section – in brief, smaller values indicate superior clustering).[14] Identification-dependent PCA of unique peptides from three analytical runs show comparable cluster dimensions with DB index of 0.2 and 0.14 for DIA and DDA, respectively. Even though identification-dependent aggregation of DDA data allows for better distinction of the two cell types, comparable DB indices suggest an additional driver of variance for both datasets.

Identification-independent analysis of DIA-TMT data precludes missing data

This observation prompted us to further investigate the vast amount of additional quantitative information generated by DIA, independently of peptide identifications. We hypothesized that, the generation of comprehensive, quantitative proteome signatures rather than peptide based profiles would capitalize subtle expression changes in trace samples. For this, at first, all datasets were retention-time (RT) aligned, based on the elution time-points of eight doubly charged peptide precursor-ions, evenly distributed across the entire analytical gradient and consistently detected in all samples (detailed in the methods section). The RT-aligned DIA data were then aggregated using a dual indexing approach, based on the central m/z of the respective isolation window and the assigned duty cycle number as indices. For direct comparison of identification-independent clustering DDA MS/MS data were binned according to the central m/z of the isolation window (bin-size = 1 Th.) and RT (bin-size = 0.2 min: i.e. approx. two times the average chromatographic peak-width), and were matched between files (with a tolerance of +/- 1 bin, respectively). The extracted raw TMT-reporter-ion intensities from all aggregated MS/MS spectra, irrespective of peptide identification (using the in-house developed PD-Node IMP-Hyperplex) were then analyzed by PCA. Intriguingly, this identification-

independent data-analysis approach consistently yielded more than 36.000 data-points (Fig. 2b) from each DIA run and resulted in PCA clustering of the expected cell populations, for all samples, even at 0.5 ng total peptide input (Fig. 2d). Of note, albeit based on substantially less data-points, PCA of corresponding DDA datasets provided similar clustering of HeLa and K562 cells (Fig. 2c) (i.e. DB index DDA: 1.05; DIA: 0.99).

Furthermore, we tested for scalability of our identification-independent single-cell proteomics approach. To evaluate weather PCA clustering was independent of sample amount, we merged all twelve DIA and DDA datasets, respectively. Intriguingly, for both datasets, we observed mix-independent cell type clustering, with 52% explained variance in PC1 (Fig. 2e-f). Detailed investigation of the cluster dimensions highlighted the reduced intra-cluster dimensions of DDA experiments with a DB index of 0.54 compared to 0.94 for DIA data. While precursor-ion sampling stochasticity in DDA reduced the number of MS/MS scans which were consistently measured and matched across 120 samples to merely 226 scans (Fig. 2e), DIA afforded the consistent accumulation of 36.367 data-points across all samples measured (Fig. 2f). This highlights that, despite a similar accumulation of missing values in conventional peptide identification workflows, DIA does indeed generate more robust and highly congruent proteome signatures from larger cohorts, also when analyzing ultra-low input samples. Most importantly, in contrast to identification dependent data aggregation (Fig. 1g), identification-independent clustering of DIA datasets were found to be largely independent of sample batches (Fig. 2f). This corroborated the applicability of identification-independent, quantitative proteome signatures for cell-type differentiation by PCA.

We pursued investigating the two cell-type populations, clearly separated by PC1 in PCA of the identification-independently aggregated DIA and DDA datasets (Fig. 2e-f). For this, we merged the respective raw MS/MS datasets, re-analyzed them using appropriate MS/MS search engines, projected the resulting peptide identifications onto the respective previously indexed

data-tables and calculated fold changes in protein abundance between cluster A and B. This data-analysis allowed us to compare the protein expression levels of 380 and 1.741 proteins by DDA and DIA, respectively. To compare differential protein expression to transcriptome data, we then calculated fold changes of fragments per kilobase of transcript per million mapped reads (FPKM) values of HeLa and K562 cells, publicly available via the ENCODE project (GEO accession: GSE33480).[15] Using the Uniprot database, Ensembl GeneIDs from the transcript data and Protein Accession numbers from SpectroMine™ or Spectronaut™ data, were converted to gene names and joined for subsequent analysis (DDA: 163 proteins/transcripts; DIA: 1.707 proteins/transcripts). Transcript fold changes of HeLa and K562 were plotted against protein fold changes and the top 300 (DIA) or top 50 (DDA) protein fold changes of cluster A and B are indicated by the respective color (Fig. 2g-h). We observed ratio compression in both DDA and DIA protein fold changes compared to transcript data (DDA: protein: -3.4 to 3.9, transcript: -27.3 to 24.6; DIA: protein -2.5 to 4.1, transcript: -27.3 to 27.4 in $\log_2$ space), as expected.[16] The fold changes between the two sample sets (transcriptomics or proteomics) are in strong agreement, confirming the cell type specific clusters shown in Fig. 2e-f.

DIA-TMT discriminates samples independent of origin or prevalence.

The file-specific batch effects of DDA (Fig. 2e), presumably are introduced by the stochastic precursor sampling and amplified with our identification-independent aggregation method. Numerous approaches aim at correcting such batch effects pre- or post-acquisition.[7,17,18] However, correction methods are devoted to specific types of missingness and rely on selecting the most appropriate for each experiment.[19] In identification-based aggregation approaches this is essential for proper analysis of underrepresented samples in a multi-batch TMT experiment.[20,21] Peptide precursors only identified in a subset of all analytical runs are prone to over-normalization or exclusion and are easily attributed as missing data. To artificially introduce such an underrepresented sample we independently generated data sets using the

yeast TKO standard at four input levels (i.e. 10, 5, 1 and 0.5 ng total peptide) and merged them post-acquisitively using both the identification-dependent and identification-independent strategies. The commercial TMT11-plex labeled tryptic TKO standard, comprised of three different single knockout (met6, his4, ura2) and wild-type yeast strains, is regularly used for TMT benchmarking experiments.[21] The TKO standard in conjunction with our human cell line mixes comprises a dual species analysis aiming at simulating e.g. xenograft, host-pathogen experiments, or mimicking infiltrated tumor samples with largely underrepresented or unexpected cell types, inherent to single cell analysis.

Using identification dependent data analysis strategies, only up to 17 peptides in both DDA and DIA, were consistently identified in all analytical runs combined. Of note, the amino-acid sequences of these few peptides are conserved between human and yeast. Furthermore, this small group of peptides did not exhibit sufficiently large differences in abundance between the cell types, and thus did not allow us to observe distinct clustering (Supplementary Fig 1a-b).

Next, we performed standard missing value imputation based on commonly used Perseus parameters.[22] Missing values in both, DIA and DDA data, were replaced with random numbers from a normal distribution shifted into the noise. Based on the largely computationally generated quantitative data the two species were separated via the first PC (Fig. 3a; DDA with 93% and Fig. 3b; DIA with 84%explained variance). Importantly, by contrast, the two human cell lines did not separate within the first two components, yet a trend towards input dependent separation was observed.

To preserve actual quantitative differences between the samples, we thus evaluated our identification-independent data aggregation method on the 16 dual-species analytical runs. Strikingly, this DIA data analysis retained critical cell type specific characteristics based on all 36.367 quantitative MS/MS scans across 164 samples without the need for any imputation whatsoever (Fig. 3d). We conclude that PC1 with 73% explained variance separates the main

differences between the datasets, the two species and PC2 with 13% explained variance differentiates the two human cell lines. For direct comparison we performed PCA analysis of the remaining 19 quantitative data points after identification-independent data aggregation of DDA data and observed no separation, as expected (Fig. 3c).

Interestingly, DIA analysis allows to observe clustering of the individual strains even down to 1 ng peptide input from one analytical run (Fig. 3.f), as opposed to DDA analysis (Fig. 3.e). Clustering evaluations demonstrated the superiority of the DIA datasets with a DB index of 0.34, compared to DDA with 2.07. Additionally, despite the large variance between two species or two human cell lines, in Fig. 3.d a zoom in the TKO cluster displays trends of separating the single knockouts (Fig. 3.g). We speculate that the DIA acquisition scheme results in a homogenous dataset highlighting small differences within samples (i.e. single knockout, representative spectra in Supplementary Fig. 2a-b). By contrast, the stochastic precursor picking and the consequential need for missing value imputation in DDA dominate sample variances.

Taken together, we here demonstrate that DIA identification-independently is scalable and yields meaningful clusters of both, closely related cell types (Hela vs K562) and completely different composite of distinct cell types (human versus yeast). Aggregation of several hundred ultra-low input samples without the accumulation of missing data is a major advantage over standard DDA identification-dependent approaches. Furthermore, unexpected and underrepresented cell populations can be identified post-acquisition using our method, which is increasingly important when analyzing limited sample material other than homogenous cell lines. The prospect of species-independent analysis would allow for studying diverse samples with no *a priori* knowledge about the specimen. We do acknowledge, however, that solely cell type clustering without the identification of the respective drivers challenges biological interpretation. Still, we anticipate future developments in single-cell sample processing and data

analysis in combination with DIA identification-independent analysis, which will lead to a universally applicable workflow for the study of protein expression at single cell resolution.

Methods

Sample preparation

Tryptic digests were obtained from Promega (K562, V6951) and Thermo Fisher (HeLa, 88328) were labeled according to manufacturer's instructions. Briefly, samples were labeled in 100 mM TEAB and 10% ACN for 1 hr at room temperature. Unreacted TMT reagent was quenched with 5% hydroxylamine/HCl for 20 minutes at RT and subsequently mixed corresponding to each TMT10 plex. Mixes were compiled as follows (Mix 1: 126, 127N, 127C, 128N, 128C - K562; 129N, 129C, 130N, 130C, 131 – Hela; Mix 2: inverted Mix 1; Mix 3: 126, 127C, 128C, 129C, 130C – K562, 127N, 128N, 129N, 130N, 131 - Hela to exclude any label specific effects.

LC MS/MS analysis

Samples were measured on an Orbitrap Exploris 480 Mass Spectrometer (Thermo Fisher Scientific) with a reversed phase Dionex UltiMate 3000 high-performance liquid chromatography (HPLC) RSLCnano System coupled via a Nanospray Flex ion source equipped with FAIMS Pro. Chromatographic separation was performed on nanoEase M/Z Peptide BEH C18 Column (130Å, 1.7 µm, 75 µm X 150 mm, Waters, Germany) developing a two-step solvent gradient ranging from 2 to 20% over 47 min and 20 to 32% ACN in 0.08% formic acid within 15 min, at a flow rate of 250 nL/min.

For both, DIA and DDA experiments FAIMS was operated at -50 CV. In DDA LC-MS/MS experiments, full MS data were acquired in the range of 370-1200 m/z at 120.000 resolution. The maximum automatic gain control (AGC) and injection time was set to 3e6 and automatic maximum injection time. Multiply charged precursor ions (2-5) were isolated for higher-energy collisional dissociation (HCD) MS/MS using a 2 Th wide isolation window and were accumulated until they either reached an AGC target value of 2e5 or a maximum injection time of 118 ms. MS/MS data were generated with a normalized collision energy (NCE) of 34, at a resolution of 60.000, with the first mass fixed to 100 m/z. Upon fragmentation precursor ions were dynamically excluded for 120 s after the first fragmentation event.

DIA experiments were acquired in the range of 400–800 m/z at a resolution of 45,000. The AGC was set to 2e5 and the maximum injection time was automatically determined for each scan. DIA windows were acquired with a 5 Th isolation windows (5 Th windows, 1 Th overlap) with stepped NCE 35, 37.5 and 45 at compensation voltages of -50V.

Data analysis

Reporter ion quantification was performed within the Proteome Discoverer environment (version 2.3.0.484) using the in-house developed, freely available PD node "IMP-Hyperplex" (pd-nodes.org) with a reporter mass tolerance of 10 ppm. The software extracts raw reporter ion intensities from respective spectra for quantification.

Peptide identification was performed using the standard parameters in SpectroMine™ 2.0 against the human reference proteome sequence database (UniProt; version:2018-11-26 accessed April 2019) and yeast reference proteome sequence database (Uniprot; version: 2019-07-25; accessed November 2019). TMT spectral libraries were generated from the DDA files with the highest input and adapted a customized script, kindly provided by Oliver Bernhard from Biognosys. Identification-dependent data aggregation was performed using standard parameters via Spectronaut™ or SpectroMine™ for DIA or DDA, respectively. By default, global median normalization is performed across all experiments. Reporter ion intensities were directly imported into R for further processing. PSMs were filtered to unique peptides using the best scoring (Q-value) PSM for subsequent analysis. Venn Diagrams are based on unique peptide sequences and are calculated using BioVenn.[23]

Identification-independent retention time alignment was based on eight doubly charged precursors (466.7743, 660.3674, 464.779, 437.2732, 587.3539, 587.864, 706.4723) observed in all runs prior to identification-independent data aggregation using the dual indexing approach described above. In detail, in DDA and DIA individual MS/MS scans were indexed according to isolation window and RT. This dual indexing approach for DIA results in a complete grid

with no missing data, for DDA we binned precursors within the isolation window (1 Th) and 2 times the FWHM (14 sec) and matched them with +/- 1 bin tolerance, to establish a similar grid. Multi-batch TMT identification-independent datasets were normalized by scaling to equal signals per channel. If indicated missing value imputation was performed based on random numbers from normal distribution shifted into the noise by 1.8 in $\log_{10}$ space (Lazar et al., 2016). Principal component analysis was internally evaluated considering intra- and inter cluster dimensions using the Davies-Bouldin (DB) index $DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$. Where n is the number of clusters, c_i is the centroid of cluster i, o_i is the average Euclidean distance of all elements in cluster i to c_i. $d(c_i, c_j)$ is the distance between centroids of the clusters c_i and c_j.[14]

FPKM values were extracted directly from the publicly available ENCODE project (GEO accession: GSE33480) from Hela and K562 datasets, from which fold changes were calculated. For both DIA and DDA data, loadings of the displayed principal components in Fig 2B were extracted, respective MS/MS scans were intersected with the SpectroMine™/Spectronaut™ derived identifications. PSMs were grouped according to their Master Protein Accession (reporter ion intensity mean) and fold changes between the proteins were calculated. Using the Uniprot database, protein accessions were converted to gene names, which were then intersected with the transcriptome derived FPKM fold changes. Remaining proteins/transcripts were plotted according to their transcript fold change and top 50 (DDA) or top 300 (DIA) protein fold changes are indicated with the respective color. Gene names are indicated for the top 20 (DDA) or top 30 (DIA) transcripts.

Data Availability

All mass spectrometry-based proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier [PXD023574].

All R-scripts are deposited via GitHub [DIA-TMT].

References

1. 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**, (2018).

2. Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* **22**, 50 (2021).

3. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* **26**, 1271–1279 (2020).

4. Miller, A. J. *et al.* In Vitro and In Vivo Development of the Human Airway at Single-Cell Resolution. *Dev. Cell* **53**, 117-128.e6 (2020).

5. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).

6. Thompson, A. *et al.* TMTpro: Design, Synthesis, and Initial Evaluation of a Proline-Based Isobaric 16-Plex Tandem Mass Tag Reagent Set. *Anal. Chem.* **91**, 15941–15950 (2019).

7. Brenes, A., Hukelmann, J., Bensaddek, D. & Lamond, A. I. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol. Cell. Proteomics* **18**, 1967–1980 (2019).

8. Karpievitch, Y. V., Dabney, A. R. & Smith, R. D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **13**, S5 (2012).

9. Liu, H., Sadygov, R. G. & Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).

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

11. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).

12. Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (2015).

13. 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* **11**, (2012).

14. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).

15. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

16. Savitski, M. M. *et al.* Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *J. Proteome Res.* **12**, 3586–3598 (2013).

17. Karp, N. A. *et al.* Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell. Proteomics MCP* **9**, 1885–1897 (2010).

18. Schwacke, J. H., Hill, E. G., Krug, E. L., Comte-Walters, S. & Schey, K. L. iQuantitator: A tool for protein expression inference using iTRAQ. *BMC Bioinformatics* **10**, 342 (2009).

19. Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **15**, 1116–1125 (2016).

20. Paulo, J. A., Navarrete-Perea, J., Guha Thakurta, S. & Gygi, S. P. TKO6: A Peptide Standard to Assess Interference for Unit-Resolved Isobaric Labeling Platforms. *J. Proteome Res.* (2018) doi:10.1021/acs.jproteome.8b00902.

21. Paulo, J. A., O'Connell, J. D. & Gygi, S. P. A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J. Am. Soc. Mass Spectrom.* **27**, 1620–1625 (2016).

22. Tyanova, S. & Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. in *Cancer Systems Biology* 133–148 (Humana Press, New York, NY, 2018). doi:10.1007/978-1-4939-7493-1_7.

23. Hulsen, T., de Vlieg, J. & Alkema, W. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488 (2008).

Acknowledgements

We thank all members of our laboratories for helpful discussions. We specifically thank Florian Stanek, Gerhard Dürnberger and Andé Lüttig for bioinformatics support. We thank Oliver Bernhardt and Lynn Verbeke from Biognosys for software support. We want to especially thank Rebecca Beveridge, Johannes Griss, Markus Hartl and Elisabeth Roitinger for critical input on the manuscript. This work has been supported by EPIC-XS, project number 823839, funded by the Horizon 2020 program of the European Union and the Austrian Science Fund by ERA-CAPS I 3686-B25-MEIOREC international project.

Author contributions

J.S. and C.C. conceptualized the project, interpreted data and wrote the manuscript. G.K., K.S. and C.C. optimized equipment and acquired data. C.C. performed sample preparation and data analysis. K.M., J.P., S.M. and J.S. supervised experiments. All authors read and approved the manuscript.

Competing interest

All authors declare no conflict of interest.

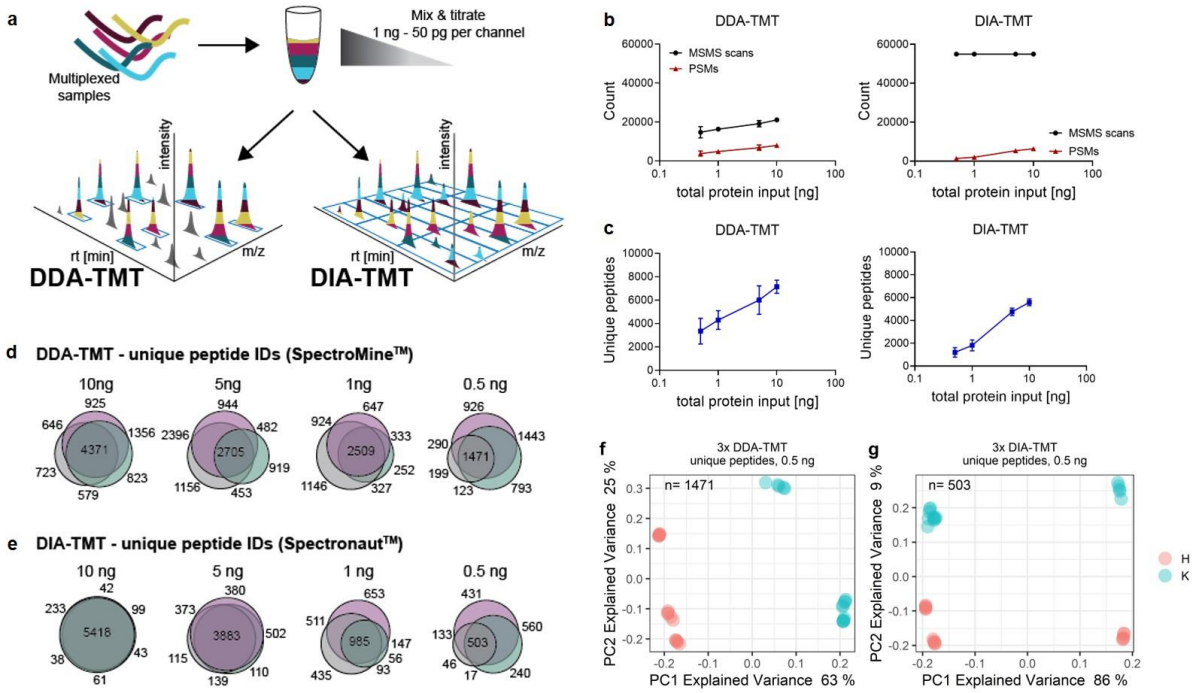

Figure 1: Evaluation of DIA and DDA schemes. **(a)** Graphical abstract of the workflow, **(b)** comparison of DDA and DIA experiments, showing the numbers of MS/MS scans (black), PSMs (brown) and **(c)** unique peptides (blue) using Spectromine^{TM} database search (DDA) or Spectronaut^{TM} library search (DIA) at various peptide input levels (i.e. 0.5 ng, 1 ng, 5 ng, 10 ng). Mean and standard deviations are shown. Venn Diagrams of unique peptide sequences identified across three technical replicates at indicated peptide input of **(d)** DDA and **(e)** DIA data (mix 1 = purple, mix 2 = grey, mix 3 = blue). PCA clustering of three analytical runs at 0.5 ng total peptide input acquired with **(f)** DDA and **(g)** DIA acquisition schemes, aggregated via the standard identification-dependent method. The two different colors represent channel loadings, n= number of unique peptides included in PCA. H = Hela, K = K562.

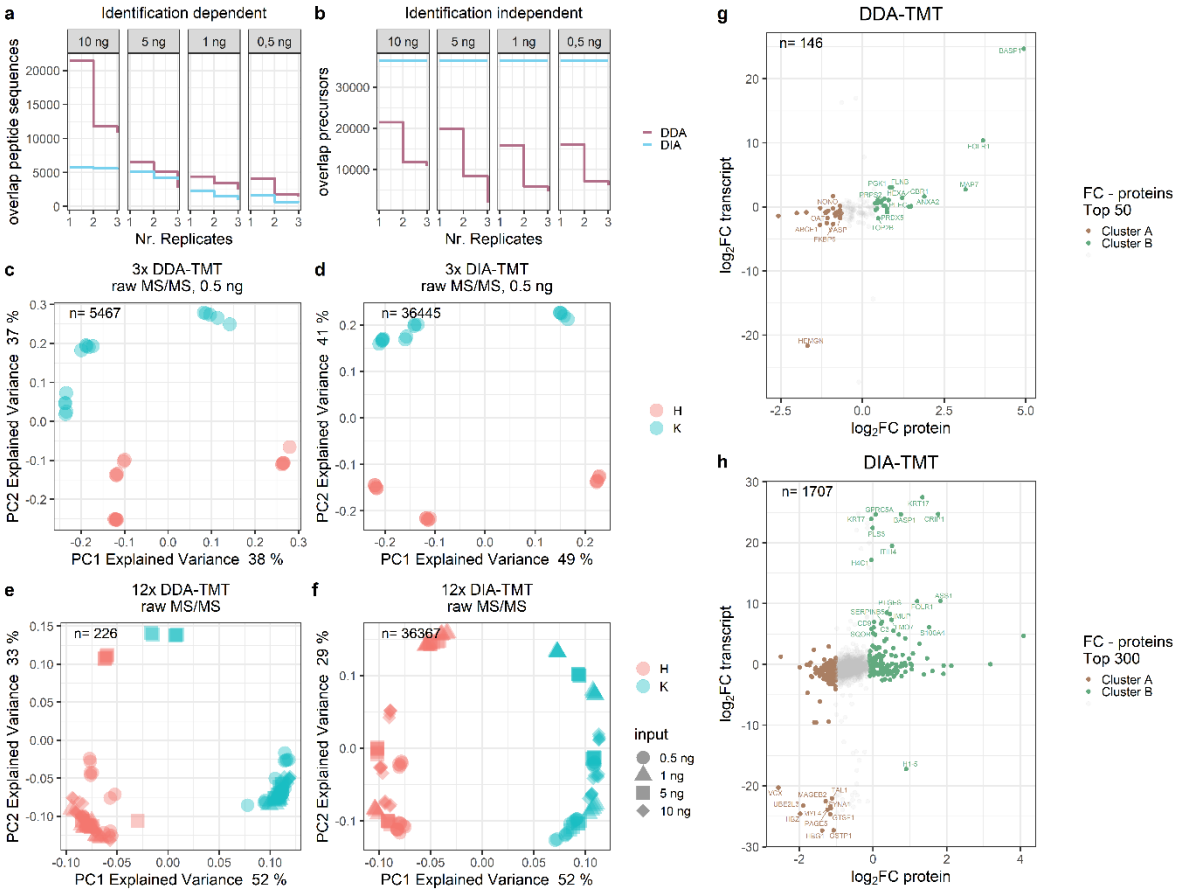

Figure 2: Data aggregation in DDA and DIA acquisition methods. **(a)** Identification-dependent accumulation of non-overlapping peptides and **(b)** identification-independent aggregation showing quantitative data points across 12 analytical runs at decreasing peptide input (i.e. 0.5, 1, 5, 10 ng) using the DDA (purple) and DIA (blue) acquisition schemes. PCA of three analytical runs (30 multiplexed samples) at 0.5 ng total peptide input were acquired with **(c)** DDA or **(d)** DIA identification-independent data aggregation. PCA of twelve analytical runs (120 multiplexed samples) at four peptide inputs (0.5, 1, 5, 10 ng) using **(e)** DDA or **(f)** DIA acquisition schemes, identification-independent aggregated. Samples are colored according to channel loadings and the respective peptide inputs are indicated with different symbols. H = Hela cells, K = K562 cells. n= number of MSMS scans included in PCA. Intersection of transcriptome (FPKM) with proteome data from both **(g)** DDA and **(h)** DIA. Top 300 or 50 proteins are colored according to cluster contributions. Top 60 or 40 proteins are labeled for DIA and DDA, respectively.

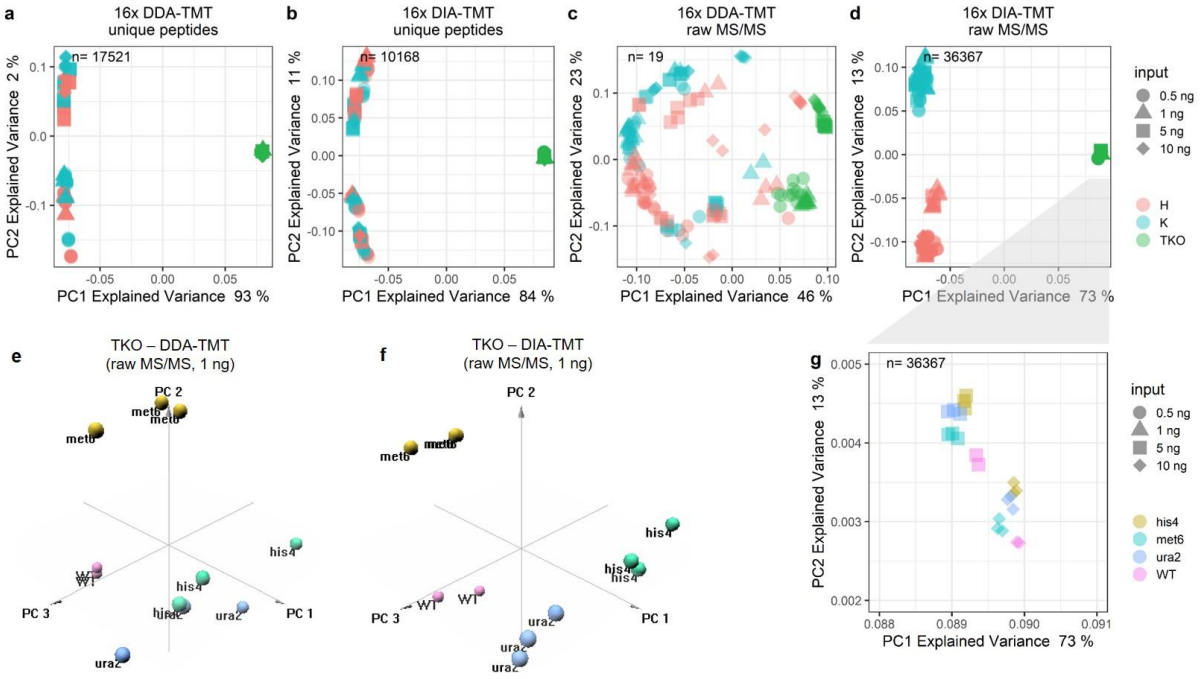

Figure 3: Identification-independent data aggregation allows for the analysis of underrepresented and unrelated cell types. PCA of 16 analytical runs (164 multiplexed samples) with 0.5, 1, 5, or 10 ng peptide inputs. Standard identification dependent data aggregation with missing value imputation of **(a)** DDA or **(b)** DIA acquisition schemes. n= number of unique peptides included in PCA. PCA of identification-independent aggregation based on 16 analytical runs (164 multiplexed samples) without imputation of **(c)** DDA or **(d)** DIA data. Samples are colored according to channel loadings and the respective peptide input is indicated with different symbols. H = Hela cells, K = K562 cells, TKO = TKO11 yeast. n= number of MSMS scans included in PCA. PCA of one representative analytical run of yeast TKO11 at 1 ng peptide input. Colors indicate the respective strains (i.e. WT, knockouts: *met6*, *his4*, *ura2*,). **(e)** DDA-TMT explained variance PC1: 64%, PC2: 30% and PC3: 3%. **(f)** DIA-TMT explained variance PC1: 79%, PC2: 17% and PC3: 2%. **(g)** Yeast cluster of the aggregated 16 analytical runs shown in panel (d). Samples are colored according to channel loadings and the respective peptide inputs are indicated with different symbols. n= number of MSMS scans included in PCA.