Evaluation of Differential Peptide Loading on Tandem Mass Tag-Based Proteomic and Phosphoproteomic Data Quality

James A Sanford

Pacific Northwest National Laboratory <https://orcid.org/0000-0001-7901-5579>

Yang Wang

Pacific Northwest National Laboratory

Joshua R Hansen

Pacific Northwest National Laboratory

Marina A Gritsenko

Pacific Northwest National Laboratory

Karl K Weitz

Pacific Northwest National Laboratory

Cristina E Tognon

Oregon Health & Science University

Vladislav A Petyuk

Pacific Northwest National Laboratory

Wei-Jun Qian

Pacific Northwest National Laboratory

Tao Liu

Pacific Northwest National Laboratory

Brian J Druker

Oregon Health & Science University

Karin D Rodland

Pacific Northwest National Laboratory

Paul D. Piehowski (✉ paul.piehowski@pnnl.gov)

Pacific Northwest National Laboratory

Research

Keywords: TMT, isobaric labeling, clinical proteomics, phosphoproteomics, Acute Myeloid Leukemia, differential loading

Posted Date: August 4th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-51123/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Global and phosphoproteome profiling has demonstrated great utility for the analysis of clinical specimens. One major barrier to the broad clinical application of proteomic profiling is the large amount of biological material required, particularly for phosphoproteomics—currently on the order of 25 mg wet tissue weight, depending on tissue type. For hematopoietic cancers such as acute myeloid leukemia (AML), the sample requirement is in excess of 10 million (1E7) peripheral blood mononuclear cells (PBMCs). Throughout the course of a prospective study, this requirement will certainly exceed what is obtainable from many of the individual patients/timepoints. For this reason, we were interested in examining the impact of differential peptide loading across multiplex channels on proteomic data quality.

Methods: To achieve this, we tested a range of channel loading amounts (20, 40, 100, 200, and 400 µg of tryptic peptides, or approximately the material obtainable from 5E5, 1E6, 2.5E6, 5E6, and 1E7 AML patient cells) to assess proteome coverage, quantification reproducibility and accuracy in experiments utilizing isobaric tandem mass tag (TMT) labeling.

Results: As expected, we found that fewer missing values are observed in TMT channels with higher peptide loading amounts compared to those with lower loading. Moreover, channels with lower loading amounts have greater quantitative variability than channels with higher loading amounts. Statistical analysis of the differences in means among the five loading groups showed that the 20 µg loading group was significantly different from the 400 µg loading group. However, no significant differences were detected among the 40, 100, 200 and 400 µg loading groups.

Conclusions: These assessment data demonstrate the practical limits of loading differential quantities of peptides across channels in TMT multiplexes, and provide a basis for designing the optimal clinical proteomics study when specimen quantities are limited.

Background

Mass spectrometry-based proteomic profiling has become a powerful tool for broad quantification of proteins and their post-translational modifications in cancer research (1–6). Due to extensive efforts in the field to benchmark and standardize complex workflows, particularly those of the Clinical Proteomics Tumor Analysis Consortium (CPTAC), proteomics is now being more broadly utilized for clinical research (7, 8). Reliable and reproducible quantification of > 10,000 proteins and > 30,000 phosphosites is now routinely attainable from mammalian tissue samples (7). Integration of deep-scale proteomic analysis of human tumors with genomic data has been shown to improve specificity for identifying pathway alterations caused by tumor associated mutations. Furthermore, phosphoproteome measurements provide information on pathway activation not discernible from genetic measurements and thus offers unique insights into potential therapeutic targets (9–14).

A substantial challenge in the field of clinical proteomics is obtaining the quantity of protein necessary for deep coverage of the proteome, as clinical specimens are often limited in size and available material.

This challenge is particularly acute in the case of phosphoproteomics, which requires 100-fold more starting material than global profiling, due to the lower number of peptides with potential phosphosites, and the low percentage of phosphorylation on those sites. The current gold standard approach for clinical proteomics employs a tandem mass tag (TMT) isobaric labeling approach to achieve deep quantitative proteomic and phosphoproteomic measurements, utilizing an enrichment step to increase phosphoproteome coverage depth. In this approach, up to 16 patient samples are tagged with unique isobaric labels, mixed in equal proportion, and fractionated prior to liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis (15–17). This creates a key challenge relating to variability in the amount of protein available per patient, due to external variables in the study. In these cases, the investigators need to decide whether to exclude sample-limited patients, reduce the protein loading per patient for the study, or to include that individual patient at reduced protein loading.

In a data-dependent LC-MS/MS experiment, peptide identification is achieved by fragmenting tryptic peptides followed by database searching to obtain sequence information, that is then matched to the parent protein. To obtain confident peptide identifications, a sufficient number of peptide ions are needed to generate high quality MS/MS spectra. A notable advantage of the TMT approach is that many patient samples can be combined to increase the available peptide mass, allowing identification of low abundance species, particularly when coupled with 2-dimensional LC separation. Recently, a number of groups have demonstrated a modified TMT labeling scheme that utilizes a “boosting” or “carrier” channel that increases the sensitivity for sample limited samples (18–21). In these approaches, one or more TMT channels are used to label a larger representative sample that is then mixed with the patient samples for analysis. The high abundance of peptides in the boosting channel triggers MS/MS selection and provides the ion flux necessary for quality MS/MS spectra and confident peptide identification, while the reporter ions provide quantitative information on the samples of interest. While this approach enables profound gains in sensitivity for both global (20) and phosphoproteomics (21) workflows, the increased dynamic range of peptide concentration induced by the boosting sample(s) results in compromised quantification with issues such as reduced measurement precision and an increase in missing values (21, 22).

In this study, we set out to interrogate two common challenges of clinical proteomics experiments: the amount of protein available from patient samples of various sizes, and the impact of differential channel loading on global and phosphoproteomics results using an established clinical workflow. The overarching goal of this study is to experimentally determine boundary conditions for patient inclusion in clinical proteomics investigations. First, we determined protein yields from patient samples of varying cell count. We then used these results to guide an experimental design using 2 TMT 11-plexes to explore a range of peptide loadings that we expect to encounter in executing a clinical proteomics experiment. The results were examined for protein/phosphosite coverage and various aspects of quantitative reproducibility. Our results highlight the protein yields achievable from representative AML samples and demonstrate a thorough examination of the impacts of differential channel loading to provide researchers with a resource to make informed decisions concerning their study design.

Methods

Cell Counting

The clinical specimen used for this study was collected with informed consent from the patient according to a protocol approved by the Oregon Health & Science University institutional review board (IRB 4422; NCT01728402). Three cell pellets, each containing approximately 20 million PBMCs isolated from a single de-identified patient, were combined into a single sterile 1.5 mL microcentrifuge tube. Each pellet was transferred to the sterile tube with 250 μ L of 1x phosphate buffered saline (PBS). The original pellet tubes were rinsed with an additional 250 μ L 1x PBS that was quantitatively transferred to the pooled sample tube for a final volume of 1 mL. A 20x dilution stock was created to determine cell concentration and live cell count utilizing an Invitrogen Countess II FL cell counter and Invitrogen Countess disposable slides. The dilute cells were combined 1:1 with Trypan blue and 20 μ L was loaded on the slide and read four times; the values were then averaged. Once the total cell concentration was determined, a concentration series was created with three replicates each of 1E7, 5E6, 1E6, 5E5, 1E5, 5E4, and 1E4 cells.

Cell Lysis and Protein Extraction

Fresh lysis buffer was prepared, containing 8 M urea (Sigma-Aldrich), 50 mM Tris pH 8.0, 75 mM sodium chloride, 1 mM ethylenediamine tetra-acetic acid, 2 μ g/mL Aprotinin (Sigma-Aldrich), 10 μ g/mL Leupeptin (Roche), 1 mM PMSF in EtOH, 10 mM sodium fluoride, 100 μ L of phosphatase inhibitor cocktail 2 and 3 (Sigma-Aldrich), 20 μ M PUGNAc, and 0.01 U/ μ L Benzonase,. Lysis buffer was added to samples based on cell concentration ranging from 200 μ L for 1e7 cells to 1 μ L for 1e4 cells. Once the lysis buffer was added, the samples were vortexed for 10 seconds and then placed in thermomixer for 15 minutes at 4°C and 800 RPM. To ensure cell lysis, samples were vortexed for an additional 10 seconds and incubated again for 15 minutes utilizing the same settings. After incubation the samples were centrifuged for 10 minutes at 4°C and 18000 rcf to remove cell debris. Due to the viscous nature of the samples, additional lysis buffer to double initial volume was added to all samples. Samples were then incubated again twice for 15 minutes at 25°C and 500 RPM, vortexing samples between incubations then centrifuging to remove cell debris. A single 5x dilution BCA (ThermoFisher) was performed on the supernatant to determine the protein yields for the varying cell concentrations. After examining the results of the protein yields, all the samples were mixed into a large protein pool and concentrated with a 10 k spin filter for use in downstream sample prep and MS analysis. A final 10x dilution BCA (ThermoFisher) was performed on the supernatant to determine final protein yield for digestion.

Protein Digest

The pooled protein sample was diluted to a concentration of 8 μ g/ μ L total protein with 50 mM Tris, pH 8.0 before reducing the sample with 5 mM dithiothreitol (DTT) (Sigma-Aldrich) for 1 hour in a thermomixer set to 37°C and 800 rpm. Reduced cystine residues were alkylated with 10 mM iodoacetamide (IAA) (Sigma-Aldrich) for 45 minutes in a thermomixer set to 25°C and 800 rpm in the dark. The sample was diluted five-fold with 50 mM Tris-HCl, pH = 8.0 and then initially digested with Lys-C

(Wako) at a 1:50 enzyme:substrate ratio for 2 hours in a thermomixer set to 25°C and 800 rpm. Following the initial digest, trypsin (Promega) was added at a 1:50 enzyme:substrate ratio, followed by a 14-hour incubation in a thermomixer set to 25°C and 800 rpm. The digestion was quenched by acidifying the solution to 1% formic acid (FA), and the sample was centrifuged for 15 minutes at 1500 rcf to remove any remaining cell debris. Peptides were desalted using a C18 solid phase extraction (SPE) cartridge (Waters Sep-Pak).

TMT Labeling

Pooled peptides were aliquoted in discrete amounts, with four replicates each of 20 µg, 40 µg, 100 µg, 200 µg, and six replicates of 400 µg for a total of 3.84 mg for TMT labeling (Thermo Fisher). The samples were then divided evenly into two TMT-11 plex sets. After drying the peptides down in a speed-vac, each sample was reconstituted with 50 mM HEPES, pH 8.5 to a concentration of 5 µg/µL. Each isobaric tag aliquot was dissolved in 40 µL anhydrous acetonitrile to a final concentration of 20 µg/µL. The tag was added to the sample at a 1:1 µg/µg peptide:label ratio (23) and incubated in a thermomixer for 1 hour at 25°C and 400 rpm, then diluted to 2.5 µg/µL with 50 mM HEPES pH 8.5, 20% acetonitrile (ACN). Finally, the reaction was quenched with 5% hydroxylamine and incubated on the thermomixer for 15 minutes at 25°C and 400 rpm. The samples for each multiplex set were then combined and concentrated in a speed-vac before a final C18 SPE cleanup. Each 11-plex experiment was fractionated into 96 fractions by high pH reversed phase separation, followed by concatenation into 12 fractions (24).

Phosphopeptide enrichment using IMAC

A small aliquot (5% volume) of each of the 12 fractions from both multiplex sets was removed and vialled for MS analysis of global protein abundance. For phosphopeptide enrichment, the remaining 95% of the 12 fractions were further combined to create six fractions per plex and dried by speed-vac. Fe³⁺-NTA-agarose beads were freshly prepared for phosphopeptide enrichment using Ni-NTA-agarose beads (Qiagen). Sample peptides were reconstituted to a 0.5 µg/µL concentration with 80% ACN, 0.1% TFA and incubated with 40 µL of the bead suspension for 30 minutes at RT in a thermomixer set at 800 rpm. After incubation the beads were washed with 100 µL 80% ACN, 0.1% TFA and 50 µL 1% FA to remove any non-specific binding. Phosphopeptides were eluted off beads with 210 µL 500 mM K₂HPO₄, pH 7.0 directly onto C18 stage tips and eluted from C18 material with 60 µL 50% ACN, 0.1% FA. Samples were dried in speed-vac concentrator and reconstituted with 12 µL 3% ACN, 0.1% FA prior to analysis.

LC-MS/MS analysis

The pooled proteomics fractions were separated using a Waters nano-Aquity UPLC system (Waters) equipped with a homemade 75 µm I.D. × 50 cm length C18 PicoFrit (New Objective) column packed with ReproSil-Pur 120 Å, C18-AQ, 1.9 µm. A 110-min gradient of 100% mobile phase A (0.1% (v/v) formic acid in water) to 60% (v/v) mobile phase B (0.1% (v/v) FA in acetonitrile) was applied to each fraction. The column was equipped with a 20 cm Nanospray column heater (Phoenix S & T). The separation was coupled to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher) for MS/MS analysis. MS Spectra were collected from 350 to 1800 m/z at an MS1 resolution setting of 60,000. The top 20 most

intense ions were selected with an isolation width of 0.7 m/z for higher energy collision dissociation (HCD); + 1 charged species were excluded, and the dynamic exclusion window was set at 45 s. MS2 spectra were acquired at a mass resolution of 15,000. Phosphoproteomics fractions were separated as described above, with the LC gradient length extended to 200 min for each fraction. The separation was coupled to the same Lumos mass spectrometer with the following adjustments to the acquisition method: the mass resolution of the MS2 scan was increased to 50,000, maximum ion time for MS2 was increased to 105 s, the top 20 was replaced with a 2 s cycle time, and the dynamic exclusion window was decreased to 20 s.

Proteomics data processing

The obtained MS/MS spectra were searched by the MS-GF + tool (25), against the UniProt human database (downloaded in October 2018) for peptide sequence identification. Carbamidomethylation on cysteine residues, and TMT-11 modifications on lysine residues and the N termini were set as fixed modifications, with oxidation on methionine residues as a dynamic modification. For phosphoproteomics, phosphorylation on serine, threonine, and tyrosine residues were set as dynamic modifications. A target-decoy approach was used to control false discovery. Criteria for filtering peptide spectrum match (PSM) include precursor mass tolerance as 10 ppm and PepQvalue < 0.01. In both datasets, this resulted in a less than 1% FDR at the unique peptide level. For identified peptides, the TMT reporter ion intensities were extracted by MASIC (26). Data from all fractions of a multiplex (12 fractions for global abundance and 6 fractions for phosphopeptide abundance per plex) were aggregated based on common peptides or proteins (global proteomics) or phosphopeptides (phosphoproteomics). After summing reporter ion intensities to the peptide, protein, or phosphopeptide level, data was log2-transformed.

Statistical analysis

Quantification reproducibility was assessed by calculating the percent coefficient of variation (%CV) among replicates of each loading group using the raw TMT reporter ion intensities aggregated to the peptide or protein level. Following reporter ion intensity aggregation and log2-transformation, data for each sample within a TMT set was divided by the universal reference channel of that plex (designated as channel 131C). Within each sample, the central tendency method was used for data normalization for the global data based on medians (27). For phosphoproteomics, each TMT channel was normalized based on the coefficients derived from global data (28). Principal component analysis (PCA) was applied to demonstrate the clustering of the different loading groups. Analysis of Variance (ANOVA) testing was employed to determine the statistical significance between the five loading groups and Tukey's HSD (honestly significant difference) test was used for post-hoc analysis and comparison of individual groups. Statistical results were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure, and adjusted p-value < 0.05 was considered as the statistical significance cutoff. Data analysis was implemented in R language for statistical computing.

Results

Experimental design, protein yield, and peptide identifications

In clinical proteomic studies, biospecimen availability can be a constraining factor depending on the tissue of interest being analyzed. In this experiment, we used peripheral blood mononuclear cells (PBMCs) isolated from AML patients as a model and followed a modified version of the CPTAC TMT-based clinical proteomics pipeline (7). An illustration of the general workflow used to process the samples, from cell lysis through LC-MS/MS analysis, is presented in Fig. 1. To determine the amount of protein obtainable from patient samples of varying sizes and volumes, we set up three replicates each of AML patient cells representing 1E7, 5E6, 1E6, 5E5, 1E5, 5E4, and 1E4 total cell counts. Samples were lysed in proportional amounts of lysis buffer; protein concentration was determined by BCA assay and total protein yield was calculated. As shown in Fig. 2, achieving the current gold standard of 400 µg total protein requires greater than 10 million cells (Fig. 2A). Lower numbers of cells resulted in a linear decrease in protein yields, with less than 1 µg of total protein extracted from smaller amounts of cells (5E4 and 1E4 cells). While these lower protein yields are substantially less than what is preferred for multiplexed proteomic studies, the sample sizes represent total amounts of biomaterial isolated in real-world clinical settings where available material is constrained. In these cases, researchers may face difficult decisions about inclusion of samples in experimental designs; thus, we set out to evaluate the impact of lower and varying protein amounts on LC-MS/MS peptide and protein identification.

As technical variability in proteomics workflows (i.e. protein extraction and digestion) is unavoidable—and likely amplified when sample size, protein amount and buffer volumes differ—we opted to pool all protein extracted from the different cell pellets and carry out digestion in a single reaction (Fig. 1). From this homogenous pool of starting material, we generated aliquots containing either 400 µg (the conventional amount loaded into each channel of clinical TMT proteomics experiments), 200 µg, 100 µg, 40 µg, or 20 µg of total peptides to assess the effect of variable peptide amount. We then prepared 2 sets of samples for labeling with TMT 11-plex reagents, with each set containing 2 replicates of each peptide aliquot distributed randomly through the first 10 TMT channels (Fig. 2B). An additional aliquot of 400 µg of peptides was included in each plex for use as a universal reference and was assigned to the 131C channel. This experimental design—standardizing the source of peptides in each channel from a large pool homogenized after digest and clean up—ensures that differences detected in downstream data processing and analysis are based on peptide loading and not sample processing. Following TMT labelling, the samples were mixed into their respective multiplexes, desalted by C18 SPE, fractionated by high pH reverse-phase HPLC, and concatenated into 12 total fractions. An aliquot of each fraction was removed for global proteomics analysis, and the remaining material was further concatenated into 6 fractions and underwent IMAC phosphopeptide enrichment. Global and phosphopeptide-enriched samples were analyzed by LC-MS/MS, and data was used to evaluate reporter ion intensities for each TMT channel as well as calculate total numbers of peptide and protein identifications.

Figure 2C displays the total reporter ion intensities calculated for each channel across both TMT11 multiplexes. In general, channels with equivalent peptide loadings showed consistent reporter ion intensities within and across the 2 plexes (Fig. 2C). Furthermore, the media intensities for different channels increase linearly with the amount of peptide loaded per channel, demonstrating the general quantitative information achievable through utilization of TMT methodology (Fig. 2D). Data from global and phospho-enriched fractions were used to determine the number of unique peptides and proteins identified from these samples (Table 1). The numbers of peptides and proteins identified from both datasets are consistent with previous studies of similar samples, despite the variations in channel loading and lower quantity of overall peptides in each multiplex (unpublished data). Additionally, Table 1 displays the number of peptides/proteins from global and phosphoproteomic datasets that are quantified in 25%, 33%, 50%, and 100% of samples—cut-offs commonly applied to proteomics datasets prior to statistical analysis.

Table 1
Peptide and protein identifications.

	Global Proteomics	Phosphoproteomics
Unique peptides (total)	138373	27351
Unique peptides (> 25% present)	137302 (99.2%)	26753 (97.8%)
Unique peptides (> 33% present)	134399 (97.1%)	24230 (88.6%)
Unique peptides (> 50% present)	118418 (85.6%)	17409 (63.7%)
Unique peptides (no missing values)	45925 (33.2%)	3366 (12.3%)
Unique proteins (total)	8926	4861
Unique proteins (no missing values)	7641 (85.6%)	1538 (31.6%)

Table 1. Number of unique peptides and proteins identified from global proteomics datasets and phosphoproteomics datasets across the two experimental TMT11 multiplexes. Unique peptide counts were filtered based on presence in > 25%, 33%, 50%, and 100% of sample channels, and percentages displayed represent the fraction of overall unique peptide identifications that pass the filtering criteria.

Effects of differential peptide loading on missing data

While our results demonstrate that reporter ion intensities correlate strongly with peptide loadings and multiplex experiments with differentially loaded channels yield good proteome coverage, a larger question remains regarding the impact of differential loading on quantitative data reproducibility and reliability. In multiplex proteomic experiments, missing data—that is, identified spectra where reporter ion intensities are not detected for one or more TMT channels—pose a significant challenge, especially when comparing across multiple TMT plexes (22). Indeed, when evaluating these datasets, an obvious trend of increasing missingness was evident in the data as peptide loading amounts decreased (Fig. 3A). In general, this issue was more pronounced in the phosphoproteomic datasets, as the differences in samples was likely exacerbated by the phosphopeptide enrichment protocol. The effects of missing data in global proteomics datasets can be largely mitigated by rolling peptide identifications/ quantifications up to the protein level (Fig. 3A); however, in cases where comparisons are to be made between individual peptide intensities (i.e. phosphoproteomic datasets), missing data can have tremendous implications on downstream statistical analysis. Consistent with the increased levels of missing data as peptide loadings decrease, comparison of the peptides identified in all replicates of each loading group illustrates that as channel loading decreases to 40 μ g or 20 μ g we begin to see increasing numbers of peptides that are not quantified in these samples (Fig. 3B-C).

Statistical differences induced by differential channel loading

Before making any comparisons across differential loading groups, data for each sample were normalized by the central tendency method based on median values (27, 28), a standard approach in proteomics data analysis that accounts for technical variations between samples (Supplemental Fig. 1). Following median normalization, principal component analysis (PCA) of both global and phosphoproteomic datasets indicate that peptide loading influences data quantification at a certain threshold: while 400 μ g, 200 μ g, and 100 μ g samples all group reasonably close to one another post-normalization, samples with 40 μ g or 20 μ g of peptides drift away from the other samples and show more variation within the replicates (Fig. 4A-B). Additionally, the reproducibility within each loading group decreases as a function of peptide quantity, demonstrated by plotting the percent coefficient of variation calculated from raw TMT reporter ion intensities among replicates in both global and phosphoproteomic datasets (Fig. 4A-B). These data indicate that as the amount of peptide loaded per channel decreases, precision of reporter ion intensity measurement decreases. In settings where comparisons are to be made between channels (i.e. when comparing patient samples), large variations in amount of material loaded will likely impair the ability to discern statistically relevant biological differences.

While preliminary visualization by principal component analysis (PCA) plots suggest that samples with lower peptide loading cluster less tightly than samples with higher loadings, we sought to gain a better understanding of the effects of differential channel loading on data analysis. To this end, we used ANOVA testing to compare the five different sample loading groups, treating the four samples in every group as biological replicates. Tukey's HSD post-hoc test was used to individually compare each sample

loading group to the 400- μ g sample group, and p-values were corrected using the Benjamini-Hochberg procedure. As samples were all derived from a common pool of peptide digest, we would not expect any differences between the samples. In all cases—global proteomics data at the peptide level or protein level, and phosphoproteomics data at the peptide level—no features were statistically different after p-value correction when comparing the 200 μ g, 100 μ g, or 40 μ g samples with the 400 μ g samples (Fig. 4C). However, the 20 μ g samples, representing a 20-fold difference in peptide loading, began to show significant differences from the 400 μ g samples (Fig. 4C). The rate of false positives was consistent between the global and phosphoproteomics data, and while relatively low, these data demonstrate that at this scale of differential loading even constitutively identical samples begin to appear different. Combined, these data demonstrate that using standard data normalization methods, up to 10-fold differences in channel loading can be effectively corrected and not have significant impacts on quantification accuracy, while more drastic differences in channel loading (i.e. 20-fold) may cause difficulties when trying to detect differences between patient samples.

Discussion

In this study, we explored the practical limits for loading differential amounts of peptides across channels in a TMT11 multiplex experiment. Biologically identical material derived from a common peptide pool generated from the white blood cells of a single patient was analyzed in aliquots ranging from the “gold standard” for clinical proteomics experiments—400 μ g of peptide per TMT channel—down to 20 μ g per channel (representing 20-fold less peptide). Samples were randomly divided across 2 TMT11 plexes and processed for LC-MS/MS-based global proteomics and phosphoproteomics measurements using a standard clinical proteomics protocol developed under the CPTAC. Analysis of the data generated from this work aimed to address two aspects of concern for quantitative proteomics experiments: missing data and data reproducibility.

It is readily clear from our results that peptide quantity and missing data are inversely correlated at the channel level—that is, as lower amounts of peptide are labeled in a channel, the number of features for which no quantitative information is extracted increases. While it is well-established that missing data is a major challenge in proteomics analyses (29, 30), this typically only arises when comparing data acquired across multiple TMT batches, and there are generally very low rates of missingness within a single TMT plex when all channels contain equivalent amounts of labeled peptide (31). Our data illustrates an exacerbation of the cross-plex missing data problem, with more than 50% of phosphopeptide identifications failing to be quantified in channels loaded with only 20 μ g. Furthermore, within either of the two multiplexes, only 27% and 45% of phosphopeptides were quantified in all 10 channels, and only 79% and 87% of phosphopeptides were quantified in more than 6 channels (Supplemental Table 1). These rates of missing data across channels within a single plex are significantly higher than in experiments where all channels are loaded equally—on average, greater than 95% of phosphopeptides are observed in all channels, and over 99% are observed in more than 6 channels (Supplemental Table 1). While the issue of missing data is more apparent in phosphoproteomics measurements likely due to the low abundance of enriched phosphopeptides, the

problem still exists in global proteomics. At the peptide level, only 76% of observations were quantified in all channels of either multiplex, while 96% and 97% of observations were quantified in more than 6 channels (Supplemental Table 2). Global proteomics measurements benefit from the aggregation of data to the protein level; when evaluating quantification at the protein level, 95% and 96% of observations have values in all channels of either plex (Supplemental Table 3). Again, these values are higher than standard, equally loaded TMT experiments, where 99% of peptides and proteins are typically observed in all channels (Supplemental Tables 2 and 3). In all cases, the higher levels of missing data occur in channels with lower peptide loading, which we attribute to the reduced signal-to-noise ratio for these channels.

In addition to higher levels of missing data, TMT channels loaded with less peptide also displayed increased variation among the replicates. Coefficient of variation values steadily increased as peptide loadings were reduced, with noticeable differences present in the 40 μ g and 20 μ g samples, particularly in phosphoproteomics data. Additionally, PCA analysis illustrated that samples in the 40 μ g and 20 μ g groups began to separate away from the 400 μ g, 200 μ g, and 100 μ g samples, which all clustered tightly together. We performed ANOVA statistical tests to compare each different loading group to the 400- μ g group and observed statistical differences in the 20- μ g group that persisted even after multiple hypothesis testing correction. Together, this data demonstrates that while 10-fold sample loading differences lead to increased variation and separation from other groups, the effects are negligible from a statistical standpoint; however, by 20-fold sample loading differences the impact on data quantitation can no longer be corrected.

The impacts of differential peptide loading on missing data and quantitative reproducibility have a compounding effect when one is interested in comparing samples. First, the amount of missing data present in samples with lower loadings reduces the number of observations for which statistical analysis can be performed: in this case, ANOVA testing required data to be present in all samples, which drastically decreased the number of phosphopeptides observed (27,351) to those which could be analyzed statistically (3,366). Relaxing the criteria for missing data- i.e. only requiring 2 or 3 observations of the 4 samples in a group—will allow for more features to be compared, but a decreased n will lower the statistical power. Second, among the features that have enough observations for statistical testing, quantitation is negatively impacted in samples with lower peptide loadings—likely due to signals closer to the noise level of the mass spectrometer, leading to increased variation when compared with higher loading groups. Importantly, the samples in this experiment were all derived from a common biological source; in true clinical studies, negative impacts on reproducibility will increase quantitative variability and reduce the statistical power, hindering the ability to confidently detect differences between patients, tissue types, or normal vs. diseased samples (28, 32).

To illustrate this point, we calculated and plotted the standard deviations for each phosphopeptide when using only the samples loaded with 400 μ g of peptide, or when combining any of the other loading groups with the 400 μ g samples. As lower loading groups are combined with the 400 μ g samples, standard deviation measurements increase (Supplemental Fig. 2A). Using the mean standard deviation for each sample set, we estimated the sample size per group necessary to measure fold changes ranging

from 1.2 to 2-fold with statistical significance (Supplemental Fig. 2B). While larger fold changes are detectable in any sample set with a reasonable number of patients per group (< 2 patients per group for equal 400 µg loadings; 7.8 patients per group with 400 µg and 20 µg samples), smaller fold changes (1.2 and 1.4 fold) require a dramatic increase in the number of patients per group to detect statistical significance when combining the more variable 20 µg loading group with the 400 µg loading group. Compared to equal 400 µg loading, where only 4.8 or 2.4 patients per group are required to detect fold changes of 1.2 and 1.4, respectively, combining 400 µg and 20 µg loadings increase the necessary sample sizes to 100 and 29 patients per group. From a clinical standpoint, obtaining and analyzing samples from this number of subjects can be a major challenge.

Reproducible and in-depth proteomic analysis of samples smaller than 500,000 cells requires significant improvements in sensitivity over standard approaches. The necessary technological improvements are actively being pursued in our lab (21, 33–36) and others (20, 37–39) with great success in recent years. However, much more work is needed to make these technologies available to less specialized laboratories.

Conclusions

From this work, we conclude that when designing TMT multiplex experiments, researchers should aim to keep peptide loadings within 10-fold differences across all channels. At this level of variation, it appears possible to minimize the effects on downstream statistical analysis with common data normalization techniques such as global median-centering. While we recommend from our analysis to maintain loading differences of 10-fold or less, researchers can use this data to independently assess the tradeoffs associated with including sample-limited patients in their study design. For example, in cases where only a small number of patients have limited material and n is large, it may be most prudent to leave out these patients to preserve the highest quality data. Conversely, the use of sample-limited patients may be required to create a properly balanced study design and the researcher needs to account for loss of statistical power resulting from increased variance and data missingness.

Abbreviations

AML: acute myeloid leukemia

PBMCs: Peripheral blood mononuclear cells

TMT: tandem mass tag

LC-MS/MS: liquid chromatography tandem mass spectrometry

PBS: phosphate buffered saline

DTT: dithiothreitol

IAA: iodacetamide

SPE: solid phase extraction

IMAC: immobilized metal affinity chromatography

CPTAC: Clinical Proteomics Tumor Analysis Consortium

PCA: principal component analysis

Declarations

Ethics approval and consent to participate:

The clinical specimen used for this study was collected with informed consent from the patient according to a protocol approved by the Oregon Health & Science University institutional review board (IRB 4422; NCT01728402).

Consent for publication:

Not applicable.

Availability of data and material:

Mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (40) via the PRIDE partner repository (41) with the dataset identifier PXD019820 and 10.6019/PXD019820.

Competing interests:

The authors have no conflicts of interest to declare

Funding:

This work was supported by The Proteogenomic Translational Research Center for Clinical Proteomics Tumor Analysis Consortium, U01 CA214116, from NIH NCI to BJD and KDR and U24CA210955 from NIH NCI to TL, and R01 DK122160 from NIH NIDDK to WQ.

Authors' contributions:

The study design was conceived by PDP, WQ, TL, BJD and KDR. Patient samples were provided by BJD. Patient samples were prepared by CET, JRH, and MAG. Mass spectrometry analysis was carried out by KKW. Data analysis was done by YW, JAS, VAP and PDP. The manuscript was written by JAS and PDP, with critical review from KDR and CET.

Acknowledgements:

Parts of this research were performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at PNNL.

References

1. Doll S, Gnad F, & Mann M (2019) The Case for Proteomics and Phospho-Proteomics in Personalized Cancer Medicine. *PROTEOMICS – Clinical Applications* 13(2):1800113.
2. Piehowski PD, *et al.* (2018) Residual tissue repositories as a resource for population-based cancer proteomic studies. *Clinical Proteomics* 15(1):26.
3. Bradshaw RA, Hondermarck H, & Rodriguez H (Cancer Proteomics and the Elusive Diagnostic Biomarkers. *PROTEOMICS* 0(ja):1800445.
4. Jimenez CR, Zhang H, Kinsinger CR, & Nice EC (2018) The cancer proteomic landscape and the HUPO Cancer Proteome Project. *Clinical Proteomics* 15(1):4.
5. Shah P, *et al.* (2015) Integrated Proteomic and Glycoproteomic Analyses of Prostate Cancer Cells Reveal Glycoprotein Alteration in Protein Abundance and Glycosylation. *Mol Cell Proteomics* 14(10):2753-2763.
6. Zhou Y, *et al.* (2020) An Integrated Workflow for Global, Glyco-, and Phospho-proteomic Analysis of Tumor Tissues. *Anal Chem* 92(2):1842-1849.
7. Mertins P, *et al.* (2018) Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography–mass spectrometry. *Nature Protocols* 13(7):1632-1661.
8. Mertins P, *et al.* (2013) Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature Methods* 10:634.
9. Mertins P, *et al.* (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534:55.
10. Zhang B, *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* 513:382.
11. Zhang H, *et al.* (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* 166(3):755-765.
12. Vasaikar S, *et al.* (2019) Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* 177(4):1035-1049 e1019.
13. Clark DJ, *et al.* (2020) Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* 180(1):207.
14. Dou Y, *et al.* (2020) Proteogenomic Characterization of Endometrial Carcinoma. *Cell* 180(4):729-748 e726.
15. McAlister GC, *et al.* (2012) Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem* 84(17):7469-7478.

16. Kreuzer J, Edwards A, & Haas W (2019) Multiplexed quantitative phosphoproteomics of cell line and tissue samples. *Methods Enzymol* 626:41-65.
17. Paulo JA, Jedrychowski MP, Chouchani ET, Kazak L, & Gygi SP (2018) Multiplexed Isobaric Tag-Based Profiling of Seven Murine Tissues Following In Vivo Nicotine Treatment Using a Minimalistic Proteomics Strategy. *Proteomics* 18(10):e1700326.
18. Leoni E, *et al.* (2019) Combined Tissue-Fluid Proteomics to Unravel Phenotypic Variability in Amyotrophic Lateral Sclerosis. *Scientific Reports* 9(1):4478.
19. Russell CL, *et al.* (2017) Combined tissue and fluid proteomics with Tandem Mass Tags to identify low-abundance protein biomarkers of disease in peripheral body fluid: An Alzheimer's Disease case study. *Rapid Communications in Mass Spectrometry* 31(2):153-159.
20. Budnik B, Levy E, Harmange G, & Slavov N (2018) SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology* 19(1):161.
21. Yi L, *et al.* (2019) Boosting to Amplify Signal with Isobaric Labeling (BASIL) Strategy for Comprehensive Quantitative Phosphoproteomic Characterization of Small Populations of Cells. *Analytical Chemistry* 91(9):5794-5801.
22. Brenes A, Hukelmann J, Bensaddek D, & Lamond AI (2019) Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Molecular & cellular proteomics : MCP* 18(10):1967-1980.
23. Zecha J, *et al.* (2019) TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Molecular & Cellular Proteomics* 18(7):1468-1478.
24. Wang Y, *et al.* (2011) Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* 11(10):2019-2026.
25. Kim S & Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 5:5277.
26. Monroe ME, Shaw JL, Daly DS, Adkins JN, & Smith RD (2008) MASIC: A software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Computational Biology and Chemistry* 32(3):215-217.
27. Callister SJ, *et al.* (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5(2):277-286.
28. Piehowski PD, *et al.* (2013) Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. *Journal of proteome research* 12(5):2128-2137.
29. Webb-Robertson B-JM, *et al.* (2015) Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *Journal of Proteome Research* 14(5):1993-2001.
30. Lazar C, Gatto L, Ferro M, Bruley C, & Burger T (2016) Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of Proteome Research* 15(4):1116-1125.

31. Chen LS, Wang J, Wang X, & Wang P (2017) A MIXED-EFFECTS MODEL FOR INCOMPLETE DATA FROM LABELING-BASED QUANTITATIVE PROTEOMICS EXPERIMENTS. *Ann Appl Stat* 11(1):114-138.
32. Levin Y (2011) The role of statistical power analysis in quantitative proteomics. *PROTEOMICS* 11(12):2565-2567.
33. Zhu Y, *et al.* (2018) Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nature Communications* 9(1):882.
34. Piehowski PD, *et al.* (2020) Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100- μ m spatial resolution. *Nature Communications* 11(1):8.
35. Dou M, *et al.* (2019) High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform. *Analytical Chemistry* 91(20):13119-13127.
36. Dou M, *et al.* (2019) Automated Nanoflow Two-Dimensional Reversed-Phase Liquid Chromatography System Enables In-Depth Proteome and Phosphoproteome Profiling of Nanoscale Samples. *Analytical Chemistry* 91(15):9707-9715.
37. Lombard-Banek C, Moody SA, & Nemes P (2016) Single-Cell Mass Spectrometry for Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog (*Xenopus*) Embryo. *Angewandte Chemie International Edition* 55(7):2454-2458.
38. Myers SA, *et al.* (2019) Streamlined Protocol for Deep Proteomic Profiling of FAC-sorted Cells and Its Application to Freshly Isolated Murine Immune Cells. *Molecular & Cellular Proteomics* 18(5):995-1009.
39. Wu R, Pai A, Liu L, Xing S, & Lu Y (2020) NanoTPOT: Enhanced Sample Preparation for Quantitative Nanoproteomic Analysis. *Analytical Chemistry* 92(9):6235-6240.
40. Deutsch EW, *et al.* (2020) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res* 48(D1):D1145-D1152.
41. Perez-Riverol Y, *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47(D1):D442-D450.

Figures

Figure 1

Clinical proteomics workflow. Diagram illustrating the steps involved in sample processing and data acquisition for our clinical proteomics workflow.

Figure 2

Design of TMT multiplexes with differential peptide loading. A) Protein extraction yields obtained from cell pellets of decreasing cell counts. B) Design of the two TMT11 multiplexes with differential amounts of peptide loaded per channel. C) Total TMT reporter ion intensity obtained per channel from global proteomics datasets. D) Relationship between the amount of peptide loaded per channel and the median TMT reporter ion intensities acquired.

Figure 3

Missing data increases in channels with lower peptide loadings. A) Number of features (peptides or protein in global proteomics, phosphopeptides in phosphoproteomics) that are detected (present) versus undetected (missing) across all TMT channels. B-C) Venn diagrams demonstrating the overlap of peptides that were quantified (in all four replicates) of each differential peptide loading group.

Figure 4

Impact of loading quantity on data reproducibility and statistics. A-B) Visualization of intra-group data reproducibility through principal component analysis (PCA) and coefficients of variation (%CV) plots calculated across the four replicates of each loading group. C) Results of statistical analysis (ANOVA followed by Tukey's HSD) comparing individual loading groups with the gold standard loading of 400 µg

per channel. Table indicates number of features (peptides or proteins) from global proteomics of phosphoproteomics that demonstrate statistical significance (adjusted p-value <0.05).

Supplementary Files

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

- [SuppFigure1.pdf](#)
- [SuppFigure2.pdf](#)
- [SupplementalTables.docx](#)