Published in final edited form as:

Proteomics. 2017 March ; 17(6): . doi:10.1002/pmic.201600300.

Relative Protein Quantification and Accessible Biology in Lung Tumor Proteomes from Four LC-MS/MS Discovery Platforms

Paul A. Stewart¹, Bin Fang², Robbert J.C. Slebos¹, Guolin Zhang¹, Adam L. Borne¹, Katherine Fellows¹, Jamie K. Teer³, Y. Ann Chen³, Eric Welsh³, Steven A. Eschrich³, Eric B. Haura¹, and John M. Koomen^{4,*}

¹Thoracic Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612

²Proteomics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612

³Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612

⁴Molecular Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612

Abstract

Discovery proteomics experiments include many options for sample preparation and mass spectrometry data acquisition, which are capable of creating datasets for quantifying thousands of proteins. To define a strategy that would produce a dataset with sufficient content while optimizing required resources, we compared 1) single sample liquid chromatography-tandem mass spectrometry (LC-MS/MS) with data dependent acquisition (DDA) to single sample LC-MS/MS with data independent acquisition (DIA) and 2) peptide fractionation with label-free quantification (LF) to isobaric labeling followed by peptide fractionation (TMT 6-plex). These strategies were applied to the same set of 4 frozen lung squamous cell carcinomas and 4 adjacent tissues, and the overall outcomes of each experiment were assessed. We identified 6,656 unique protein groups with LF, 5,535 using TMT, 3,409 proteins from single sample analysis with DIA, and 2,219 proteins from single sample analysis with DDA. Pathway analysis indicated the number of proteins per pathway was proportional to the total protein identifications from each method, suggesting limited biological bias between experiments. The results suggest the use of single sample experiments as a rapid tissue assessment tool and digestion quality control or as a technique to maximize output from limited samples and use of TMT or LF quantification as methods for larger amounts of tumor tissue with the selection being driven mainly by instrument time limitations. Data are available via ProteomeXchange with identifiers PXD004682, PXD004683, PXD004684, and PXD005733.

Keywords

data independent acquisition; discovery proteomics; label free quantification; lung squamous cell carcinoma; TMT

*Corresponding author: John Koomen, PhD, Moffitt Cancer Center, SRB3, 12902 Magnolia Drive, Tampa, FL 33612, (813) 745 8524, (813) 745 3829 FAX, John.Koomen@moffitt.org.

1 Introduction

Proteomics has the potential to further our understanding of tumor biology which can provide actionable therapeutic targets and new biomarkers.[1] Proteomic characterization of non-small cell lung cancer tissues has been performed on different sample sets with a variety of experimental workflows. These studies have shown the utility of using data dependent acquisition (DDA), including label-free quantification (LF),[2–8] relative quantification using chemical labeling (TMT/iTRAQ),[9–15] and SILAC.[16, 17] Data independent acquisition (DIA) proteomic strategies have also been implemented for molecular phenotyping of lung cancers. [18] To date, there have been few head-to-head comparisons between proteomics methods run on the same human tumor samples,[19, 20] so this direct comparison of DDA and DIA using lung cancer specimens can provide useful information for planning large scale proteomics experiments using archived human tumors.

In data dependent acquisition (*e.g.* LF, TMT), the highest intensity peptide signals are selected for fragmentation.[21] This approach produces tandem mass spectra that are compared to a database for peptide and protein identification. DDA is known for undersampling; some observed peptides will not be selected for tandem mass spectrometry. Despite this issue, DDA has been the gold standard for expression proteomics.[22] In contrast, DIA samples all detectable peptides across mass-to-charge (m/z) windows (or SWATHs) [23] and subjects them to fragmentation,[24] creating chimeric tandem mass spectra representing sequence information from multiple peptides. This process should maximize the ability to identify peptide sequences, because MS/MS data could theoretically be produced for each available precursor ion. Several implementations of DIA, ordered from widest mass selection window to narrowest, include all ion fragmentation (AIF),[25] MS^E, [26] SWATH-MS,[23] P-SMART,[27] DIA,[24] and PACIFIC.[28] Recent studies have applied DIA to clinically oriented studies;[29] furthermore, improved methodologies [30] and software tools [30] are increasing the data content of DIA experiments.[31] While the DIA approach is still being evaluated for different instruments (*e.g.* hybrid quadrupole-orbital ion traps like Thermo's Q Exactive) and samples, existing data in the literature are promising.[32]

Here, we compare 1) single sample analysis of unfractionated tryptic digests using DDA and DIA for protein identification and relative quantification using the peak areas of peptide ion signals and 2) LF quantification to a multiplexed chemical labeling approach (TMT 6-plex) with twelve concatenated bRPLC fractions using the same set of four lung squamous cell carcinoma (SCC) and four adjacent lung tissue samples (workflow shown in Figure 1A). Given the technological advances of proteomics in recent years, we first wanted to understand the coverage provided by single experiment without any peptide fractionation. Results from several groups have shown that significant biological content can be obtained without extensive fractionation. [33–36] Compared to a LF strategy with peptide fractionation, single sample analyses require of a minimal amount of total protein (< 1 μ g) and the least amount of instrument time. Our motivation was to define the proteome that would be detectable during the examination of larger cohorts (containing 100 or more patient samples). We then wanted to assess the differences in proteome coverage provided by LF and multiplexed chemical labeling approaches. Although TMT uses similar sample

amounts and the same peptide fractionation as LF, it requires less instrument time due to sample multiplexing (e.g. sixplex TMT with four samples and two pooled samples for batch-to-batch comparison requires ¼ of the instrument time). Thus, these alternate methods may be more appropriate for large-scale studies where instrument availability and overall time for data acquisition are important issues. From previous experience, we hypothesized that LF would outperform TMT for the number of peptide and protein identifications, but here we explore the biological content of each dataset (e.g. pathway representation and differentially expressed proteins) and compare the outcomes for each method.

2 Materials and Methods

Clinical Samples

Samples were collected under the Moffitt's Total Cancer Care protocol (Liberty IRB # 12.11.0023) and Moffitt's General Banking protocol (IRB #: USF 101642), and these experiments were performed under protocol MCC #50083. De-identified clinical attributes including gender, race, ethnicity, tumor cellularity, tumor necrosis, smoking status, and TNM staging can be found in Table S1. Six of the eight clinical samples are paired sets of tumor and adjacent lung tissues from the same patients, and the unpaired samples are designated as patients 4 and 5 or "P4/5."

Protein Extraction and Digestion

Tissue samples were homogenized with a BioPulverizer (Biospec) and resuspended in lysis buffer containing 20 mM HEPES, pH 8.0, 9 M urea, 1 mM sodium orthovanadate, 2.5 mM sodium pyrophosphate, and 1 mM β -glycerophosphate. After brief sonication, the lysate was cleared by centrifugation at 10,000 $\times$ g at 15 °C for 30 minutes. The proteins were reduced with 4.5 mM DTT at 60 °C for 20 minutes followed by alkylation with 11 mM iodoacetamide at room temperature for 15 minutes in the dark. The sample was then diluted 4-fold to a final concentration of 2 M urea, 20 mM HEPES, pH 8.0, and trypsin digestion was carried out overnight at 37 °C with an enzyme/substrate ratio of 1/50.

Sep-Pak C18 Peptide Cleanup

The digested peptide solution was acidified with 20% TFA to a final TFA concentration of 1%. After incubation at room temperature for 10 minutes, the solution was cleared by centrifugation at 10,000 $\times$ g at 15 °C for 15 minutes. Sep-Pak cartridges were washed with 5 ml acetonitrile followed by 3 ml and 4 ml washes with Sep-Pak solvent A (aqueous 0.1% TFA). After acidified peptides were loaded, the cartridge was washed with 1 ml, 5 ml and 6 ml of Sep-Pak solvent A. Elution was carried out 3 times using 2 ml of Sep-Pak solvent B (aqueous 40% acetonitrile with 0.1% TFA). After freezing, the peptides were lyophilized to dryness over 2 days.

For LF, an aliquot (600 μ g) of total protein digest was fractionated using basic pH reversed phase HPLC, adapted from the method published by Mertins *et al.*[37] For TMT, an aliquot (100 μ g) of total protein digest was labeled for each reporter ion channel, combined and fractionated using basic pH reversed phase HPLC, *ibid.* For DIA, aliquots (500 ng) of total protein digest were injected for each sample.

bRPLC Separation

Dried peptides were reconstituted in 20 mM ammonium formate, pH 10. bRPLC separation was carried out on Dionex Ultimate3000 UPLC using a 2.1 mm ID×50 mm column (XBridge, Waters) packed with C18, 3.5 µm particle size at a 0.2 ml/min flow rate. The gradient setting was: 100% bRPLC A (aqueous 2% acetonitrile with 5 mM ammonium formate, pH 10) for 9 minutes, then the concentration of bRPLC solvent B (aqueous 90% acetonitrile with 5 mM ammonium formate, pH 10) was increased to 6% over 4 minutes, to 28.5% over 50 minutes, to 34% over 5.5 minutes, and to 60% over 13 min, and then kept constant for 8.5 minutes prior to re-equilibration at the original conditions. Twelve concatenated fractions were analyzed with LC-MS/MS for the LF and TMT experiments.

DDA LC-MS/MS Analysis of Single Samples and Fractionated Peptides for the Label-Free Approach

A nanoflow ultra high performance liquid chromatograph (RSLC, Dionex, Sunnyvale, CA) interfaced with an electrospray quadrupole-orbital ion trap mass spectrometer (QExactive Plus, Thermo, San Jose, CA) was used for tandem mass spectrometry peptide sequencing experiments. The sample was first loaded onto a pre-column (2 cm×100 µm ID packed with C18 reversed-phase resin, 5µm particle size, 100Å pore size) and washed for 8 minutes with aqueous 2% acetonitrile and 0.04% TFA. The trapped peptides were eluted onto the analytical column, (C18, 75 µm ID×50 cm, 2 µm, 100Å, Dionex, Sunnyvale, CA). The 90-minute gradient was programmed as: 95% solvent A (aqueous 2% acetonitrile + 0.1% formic acid) for 8 minutes, solvent B (aqueous 90% acetonitrile + 0.1% formic acid) from 5% to 38.5% in 60 minutes, then from 38.5% to 90% in 7 minutes, and held at 90% for 5 minutes, followed by solvent B from 90% to 5% in 1 minute and re-equilibration for 10 minutes. The flow rate on analytical column was 300 nl/min. Sixteen tandem mass spectra were collected in a data-dependent manner following each survey scan using 60 second exclusion for previously sampled peptide peaks and normalized collision energy set to 28. A standard peptide retention time calibrator mixture (PRTC, Pierce #88320) was spiked in each LC-MS/MS analysis at the level of 10 fmol of each peptide per injection to monitor instrument performance (Figure S1). Total LC-MS/MS time for the single sample DDA and DIA experiments was 2 days each. Total LC-MS/MS time for the LF experiment was 24 days. The acquisition sequences can be found in Supplementary Methods.

TMT Quality Control

After chemical labeling, LC-MS/MS analysis was performed on each TMT channel followed by Mascot/Sequest searches and the results were summarized in Scaffold. Spectral counting was used to calculate percentage of peptides labeled with TMT reagent. In addition, MaxQuant analysis was also performed on each TMT channel to verify correct TMT channel was labeled for each sample and to examine the crosstalk between channels. [38] The TMT experimental design can be found in Supplementary Methods.

LC-MS/MS Analysis of TMT Samples

The same UPLC and mass spectrometry instrument parameters were used as described above except in MS² where the first mass was set to m/z 100 to include the reporter ions and

the normalized collision energy value was set to 30. Twenty-five femtomoles of PRTC was spiked in each LC-MS/MS analysis to monitor instrument performance (Figure S1). Total LC-MS/MS time was 6 days. The acquisition sequence can be found in Supplementary Methods.

DIA Analysis of Label-Free Unfractionated Samples

The same UPLC conditions were used as described above for the DDA experiments. The mass spectrometry method utilized MS¹ scans alternated with looped 18 narrow window DIA scans covering m/z range from 450 to 1,080. The MS/MS isolation windows from 450 to 900 were set at 5 Th and the isolation windows from 900 to 1,080 were set at 8 Th. Resolution was set at 70,000 for MS¹ and 17,500 for MS/MS. Twenty-five femtomoles of PRTC peptide mix was spiked in each LC-MS/MS analysis to monitor instrument performance (Figure S1). The acquisition sequence can be found in Supplementary Methods.

MaxQuant Data Analysis

MaxQuant (version 1.5.2.8) was used for the peptide and protein identification from LF and TMT results.[38] Database searches were conducted against human entries in UniProt (downloaded 05/25/2015). The “align across runs” option was selected with a mass error of 10 ppm and a retention time window of 4 minutes. We set less stringent preliminary false discovery rate (FDR) thresholds in order to assess content that would otherwise be missed. For LF, the false discovery rate (FDR) thresholds for peptides and proteins were set to 0.05% and 0.2%, respectively. For TMT, the FDR threshold for peptides was 0.05% and 0.1% for proteins. Concordance of LF and single sample DDA technical replicates were measured using Pearson’s correlation of log₂-transformed intensities (Figures S2 and S3, respectively). For TMT, we found that MaxQuant automatically combined the technical replicates, so we examined the correlation of the MS data between experimental replicates (Figure S4). Peptide and protein group intensities were extracted from MaxQuant output and input into R/RStudio[39, 40] with corresponding metadata. We filtered the data using a posterior error probability (PEP) cutoff of 0.05. PEP can be used as a data quality filter similar to FDR [41]. Reverse-only sequences, non-human peptides, and peptides that were identified but had no corresponding intensities were also filtered out. Intensity values for peptides and protein groups that passed filtering were then log₂-transformed for subsequent analyses. Missing values were replaced with *NA* so that they were excluded from plots and calculations in the R environment. The differential expression threshold was defined as 1.5 times the standard deviation (SD) of the average log₂ ratio of tumor and adjacent protein group intensities. Summarized results can be found in Table S2.

IDPicker Data Analysis

MSConvert was first used to convert Thermo .RAW data files into mzML and mzXML formats [42]. DIA-Umpire (<http://www.nesvilab.org/software.html>) was used to create 3 pseudo-tandem MS spectra from each mzXML data file. DIA pseudo-spectra and DDA spectra were then searched with MS-GF+ against the UniProt database (Sep 2015 release) [43]. Search parameters included a 15 ppm mass tolerance; carbamidomethylation (cysteine), oxidation (methionine), and pyroglutamic acid (glutamic acid) as allowable

modifications; and no limitation on missed cleavages. IDPicker version 3 was used to create a final protein assembly from the DDA and DIA experiments [44]. A maximum of 1% FDR for peptide spectrum matches were allowed with a minimum of one spectrum per peptide and one peptide per protein. The final assembly had a spectrum FDR of 0.27%, a peptide FDR of 1.27%, and protein FDR 6.84%. Summarized results can be found in Table S3.

Pinnacle Data Analysis

Label free files were searched against human entries in UniProt (as above) using both Sequest and Mascot in Protein Discoverer (v. 1.4 with Percolator). Two missed tryptic cleavages were allowed, precursor mass tolerance was set to 10 ppm, MS/MS mass tolerance was 0.05 Da, and dynamic modifications included cysteine carbamidomethylation and methionine oxidation. The search result files were loaded into Pinnacle version 1.0.26.0 (OPTYS Tech Corporation), and the spectral library threshold settings included a Percolator q-value less than 0.05, Sequest XCorr scores above 1.5/2.0/2.2 for singly/doubly/triply protonated peptides, or a Mascot Ion score > 29. Pinnacle was then used for the peptide and protein identification from the DIA results with cutoff values for CV within the group (50%) and missing values (< 20% across the dataset).[27] Concordance of technical replicates was measured using Pearson's correlation of log₂-transformed peptide peak areas (Figure S5). Pinnacle-assigned peptides from the spectral library that were not observed with sufficient confidence in the LF data (the source of the spectral library) were removed as along with their corresponding proteins. Pinnacle proteins with less than a quality score of 3 were also removed (full Pinnacle scoring can be found in Table S2). Peak areas were then log₂-transformed for subsequent analyses. Missing values were replaced with NA, so that they were excluded from plots and calculations in the R environment. The differential expression threshold was defined as 1.5 times the SD of the average log₂ ratio of tumor and adjacent lung tissue protein peak areas.

SCC Marker Identification

In a previous publication, we analyzed two publically available SCC datasets and compared the findings with our own experiment.[9] We took the list of SCC markers from these three datasets and examined their expression levels in these LF, TMT, and DIA datasets. We defined a marker as having at least 1.5 times higher expression in SCC compared to lung adenocarcinoma and a p-value of less than 0.05 from a Welch two-sample t-test of the relative protein expression levels (Table S4).

Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifiers PXD004682, PXD004683, PXD004684, and PXD005733.[45]

3 Results

Peptide, Protein, and SCC Marker Identifications

We analyzed the single sample DDA, single sample DIA, LF, and TMT 6-plex data according to the workflow in Figure 1A with the resulting identifications found in Table 1

and Figure 1B (summarized data in Tables S2 and S3). We found good agreement between the two single sample experiments using IDPicker: DDA was able to identify 104,117 spectra, 13,764 peptides, and 2,219 protein groups, while DIA was able to identify 70,094 spectra, 12,976 peptides, and 2,155 protein groups (Tables S3 and S5; Figure S6). Noticeably, DDA identified about 1.5 times the amount of spectra compared to DIA even though the total numbers of protein groups were nearly identical. We next wanted to understand if more identifications could be found using an alternative computational approach. Using Pinnacle (OPTYS Tech Corporation), we were able to identify 1,587 proteins not found in the IDPicker results (Figure S6). This result is likely due to the fact that Pinnacle in this version identifies all proteins separately (no parsimony), and it also has the ability to match lower quality spectra and chimeric spectra to the MS/MS data in the spectral libraries more effectively than traditional database searching would match these experimental data to theoretical tandem mass spectra. In addition, Pinnacle also makes use of the accurate mass measurement of the intact m/z values and reproducibility of the peptide retention time, which are not included in other searches. Therefore, we used the Pinnacle DIA results for all subsequent DIA analysis and comparisons.

We next compared a LF approach with prefractionation to a comparable 6-plex TMT approach. Not surprisingly, LF had the most protein identifications (6,656 protein groups or 6,964 proteins) compared to TMT (5,535 TMT protein groups or 5,830 proteins). We then used a list of 1,316 lung SCC candidate biomarkers from a previous work (Table S4) to assess each method's ability to recapitulate the underlying biological differences between tumors and adjacent tissues,[9] and these identifications were proportional to the total number of identifications with LF identifying the most ($n = 1,095$) followed by TMT ($n = 1,015$), DIA ($n = 681$), and DDA ($n = 562$).

We expected the LF experiment to have the most content, but we speculate the higher number of TMT peptide identifications may have been due to the hydrophobicity of the isobaric tags causing extremely hydrophilic peptides to be recovered more reliably than in LF; chemical labeling may also create more balance in peptide basicity and modulate ion suppression effects leading to more peptide identifications. Variations in instrument settings and performance between the two datasets may have also contributed. We removed all DIA peptides not found in the filtered spectral library generated by Sequest/Mascot (Figure S7), but there were still 1,178 peptides unique to the Pinnacle DIA dataset that Andromeda/MaxQuant did not identify in LF or TMT. This difference carried over into the identification of proteins and accounted for 295 (68%) of the 434 proteins unique to the DIA dataset. Differences between Sequest/Mascot and Andromeda/MaxQuant identifications, differences in filtering of processed data, and FDR all contributed to the presence of unique peptides and proteins in the Pinnacle DIA dataset.

Pathway Analysis Recapitulates SCC Biology Regardless of Experiment Type

GO-Slim functional classifications of the identified proteins using PANTHER [46] were proportional to the total number of identifications from each method (Figure S8). We had similar findings with the protein classifications of unique proteins from each method (Figure S9), and these results show that there is not an inherent bias towards a particular protein type

in a given experiment. Pathway enrichment was next performed separately on proteins identified from each experiment using GeneGO MetaCore (Thomson Reuters, New York, NY; Table S6).[47] Four of the top 10 pathways ranked by FDR from each experiment were shared: *Cytoskeleton remodeling_Keratin filaments*; *Cytoskeleton remodeling_Cytoskeleton remodeling*; *Cell adhesion_Chemokines and adhesion*; and *Immune response_Classical complement pathway*. We then compared the pathway content from LF against the DIA and DDA experiments and against the TMT experiment. As expected, LF had the most content per pathway in a majority of cases followed by TMT, DIA, and DDA (Figure 1C, Figure 1D). All four experiments were able to identify proteins in selected lung cancer-related pathways (Table 2). Again as expected, the LF experiment performed the best in identifications within these lung cancer pathways and pathway FDR, followed by TMT, then DIA, and DDA. Notably, these pathway results were proportional to the number of identifications from each method, and this outcome seems to suggest that there is not an inherent bias towards certain biology or pathways in a given experiment.

Comparison of Differentially Expressed Proteins between Methods

We then compared differentially expressed proteins between tumor and adjacent tissue from each MS experiment. The differential expression threshold was defined as 1.5 times the standard deviation (SD) of the average $\log_2$ ratio of tumor and adjacent relative expression. There were 773 differentially expressed proteins in the LF data, 562 in the TMT data, 385 in the DIA data, and 137 in the DDA data (Table 1, Table S2). We then looked at a subset of SCC candidate biomarkers in common with LF and TMT or LF and DIA that were differentially expressed in at least one sample pair to look at the similarities and differences between MS experiments (Figure 2). It is important to note that these proteins may not be differentially expressed in the other pairs, which may contribute in part to differences between the LC-MS/MS measurement approaches.

We noticed several proteins in the LF and DIA heat map (Figure 2B) did not agree in the direction of change between adjacent and tumor. We investigated further and found that DIA did not agree with LF in 233 out of 540 comparisons (135 proteins times four samples) or 43% of the time, while TMT did not agree with LF 394 out of 1,816 comparisons (454 proteins times four samples) or 22% of the time. If we instead consider differentially expressed proteins instead of the proteins from the heat map, DIA still disagrees with LF more (53 out of 144 times or 37%) when compared to TMT (119 out of 496 or 24%). Since a protein with a single peptide identification is more likely to be error prone than a protein with multiple peptides, we checked the number of single peptide identifications from the differentially expressed proteins, but this factor did not contribute to the differences between these datasets. Since the discordance here was measured by the sign of the ratio, differences could occur in $\log_2$ ratios that are near zero. In addition, DIA may be challenged in a single sample to measure low abundance analytes that have stronger ion signals with peptide fractionation and higher levels of peptide loading for LC-MS/MS in the LF experiments (12-fold enrichment); in single sample LC-MS/MS, these ion signals would also be more prone to interference using MS¹ precursor ion signal quantification in this DIA method and parallel reaction monitoring is likely to be a better approach to quantification of low abundance analytes with DIA. Finally, the observed disagreement of DIA and LF may be an

artifact of working in ratios since there is good correlation between LF and DIA measurements in tumor-adjacent tissues ($R = 0.55$) and in the tumors ($R = 0.66$; Figure S10).

Because the DIA results were expected to be similar to the LF results as both use peptide precursor measurements for quantification, we plotted the peptide mean intensities or peak areas against the peptide SD of the intensities or peak areas for both the LF and DIA experiments (Figures 3A, 3B). In both datasets, the variance decreases with increasing ion signal intensity, but the distributions are very similar. We then repeated this for the LF and DIA proteins (Figures 3C, 3D) and found some differences in the LF intensity distribution compared to the DIA peak area distribution. This variation is related to the difference between quantification using peak apex (MaxQuant) and total peak area (Pinnacle), but could also arise from the way MaxQuant and Pinnacle handle value assignment of protein groups since the distribution of the peptides is fairly similar. We then looked at the number of single peptide identifications from both experiments, but it did not appear that these significantly contributed to the variability of DIA since there were similar single peptide identifications in LF (578 proteins or 9% of 6,656 protein groups) compared to DIA (260 proteins or 8% of 3,409 proteins).

Using a quartile system, the ion signals for DIA and LF were compared globally (Figures 1E, 1F, 1G). As expected, peptide fractionation and higher loading in the LF dataset could be observed to bring peptides from lower quartiles in DIA to higher quartiles in LF. In addition, the relative standard deviation values of peptide measurements were then examined by quartile of ion signal for DIA and LF datasets. Consistent with expectation, the RSD values decreased with increasing peptide ion signal. RSD values in the DIA experiments were generally lower, which is expected because these peptides are the most abundant and less variability is predicted based on the use of peak area as compared with the peak apex in LF.

4 Discussion

Selection of LC-MS/MS strategy depends on a number of factors including experimental design, sample type and amount, sample availability, instrument time, instrument availability, and budget. We found the protein functional classifications were proportional to the number of total protein identifications from each MS experiment (Figure S8, Figure S9). This result, along with the finding that average pathway representation of each MS experiment was proportional to the number of total protein identifications from each experiment (Figures 1C, 1D), suggesting that these methods are not biased towards or against a particular biology. LF is clearly the best in terms of protein identifications (Table 1), but its biggest drawback is the amount of required instrument time, which ultimately results in added cost. Multiplexing with chemical labeling strategies can save time, reduce cost, and capture nearly the same amount of identifications of LF, but multiplexing methods in our hands have had biases towards proteins that have a common background level (*e.g.* most, if not all, peptides are detected from all 6 channels in TMT), which may not be optimal for detection of features in each individual tumor. The major drawback shared by both LF and TMT are that these methods require fractionation prior to LC-MS/MS. The single LC-MS/MS experiments in DIA mode obtained 50–60% of the proteome coverage as

24 LC-MS/MS runs from a fractionated sample in DDA mode. DIA needed fewer peptide assignments per protein, and it was the most efficient as measured by proteins identified per day (Table 1). Regardless of whether a single sample dataset is sufficient to address a biological or clinical question, a single DDA or DIA LC-MS/MS experiment does provide benefit as a rapid quality control check for sample digestion and tissue histological content. Additionally, the small quantities of material needed for the single sample experiments (here we used 1,000 fold less material: 500 ng of proteolytic peptides for DIA and DDA per LC-MS/MS experiment compared to 600 µg for LF and TMT due to the offline analytical scale fractionation, though minimum amounts needed for these experiments could be ~25 µg) can provide a means for analyzing limited clinical samples such as laser capture microdissected samples, tissue microarrays, core biopsies, and fine needle aspirates.

LF identified more proteins than any of the other experiments (Table 1). However, Figure 1B indicates that all experiments identified unique proteins. Given the numbers of unique proteins from TMT and DIA, this finding suggests that the combination of TMT and DIA data may provide a reasonable alternative to LF for protein identification if instrument run-time is limited. Further work could address the question of maximizing proteome coverage through combining identification and/or quantification between the two technologies as well as assessing peptide fractionation prior to DIA LC-MS/MS to maximize the ability to detect and quantify proteins using a minimum of instrument time. Additionally, repeated measures between the two technologies (DIA and TMT) offer the possibility of imputation of missing values between platforms or more robust estimation of protein abundance. Additional studies of tumors collected from larger cohorts of lung cancer patients are ongoing. Based on these findings, we are moving forward with a proteogenomic characterization of 116 lung SCC patient samples using a single sample DIA approach for quality control after tryptic digestion and chemical labeling with sixplex TMT as a time and cost-saving alternative to LF analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project is possible due to the support of the Moffitt Lung Cancer Center of Excellence and the Moffitt Foundation. This work was supported by Proteomics, Cancer Informatics, and the Tissue Core at the Moffitt Cancer Center, which are funded in part by a Cancer Center Support Grant (P30 CA076292) from the National Cancer Institute. This project has been funded by the American Lung Association through a Lung Cancer Discovery Award (LCD-257857-N to JK) and the NCI (R21-CA169980 to JK). We would like to thank Amol Prakash (OPTYS Tech) and Scott Peterman (Thermo) for thoughtful discussion of DIA and assistance with Pinnacle; Michelle Fournier, Marek Wloch, and Katherine Fellows provided assistance with obtaining clinical samples and data.

Conflict of Interest Statement

The Koomen laboratory is funded in part by a sponsored research agreement with Proteome Sciences (Electrophoretics), which produces TMT reagents; those funds were not used for this project.

Abbreviations

DDA data dependent acquisition

DIA	data independent acquisition
FDR	false discovery rate
GO	gene ontology
LF	label-free
PRTC	peptide retention time calibrator standards
SCC	lung squamous cell carcinoma
SD	standard deviation
TMT	tandem mass tags

References

1. Koomen JM, Haura EB, Bepler G, Sutphen R, et al. Proteomic contributions to personalized cancer care. *Mol Cell Proteomics*. 2008; 7:1780–1794. [PubMed: 18664563]
2. Kikuchi T, Hassanein M, Amann JM, Liu Q, et al. In-depth proteomic analysis of nonsmall cell lung cancer to discover molecular targets and candidate biomarkers. *Mol Cell Proteomics*. 2012; 11:916–932. [PubMed: 22761400]
3. Li L, Wei Y, To C, Zhu CQ, et al. Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. *Nat Commun*. 2014; 5:5469. [PubMed: 25429762]
4. Wei Y, Tong J, Taylor P, Strumpf D, et al. Primary tumor xenografts of human lung adeno and squamous cell carcinoma express distinct proteomic signatures. *J Proteome Res*. 2011; 10:161–174. [PubMed: 20815376]
5. Birse CE, Lagier RJ, FitzHugh W, Pass HI, et al. Blood-based lung cancer biomarkers identified through proteomic discovery in cancer tissues, cell lines and conditioned medium. *Clin Proteomics*. 2015; 12:18. [PubMed: 26279647]
6. Okayama A, Miyagi Y, Oshita F, Nishi M, et al. Proteomic analysis of proteins related to prognosis of lung adenocarcinoma. *J Proteome Res*. 2014; 13:4686–4694. [PubMed: 25004236]
7. Tan F, Jiang Y, Sun N, Chen Z, et al. Identification of isocitrate dehydrogenase 1 as a potential diagnostic and prognostic biomarker for non-small cell lung cancer by proteomic analysis. *Mol Cell Proteomics*. 2012; 11 M111 008821.
8. Kawamura T, Nomura M, Tojo H, Fujii K, et al. Proteomic analysis of laser-microdissected paraffin-embedded tissues: (1) Stage-related protein candidates upon non-metastatic lung adenocarcinoma. *J Proteomics*. 2010; 73:1089–1099. [PubMed: 19948256]
9. Stewart PA, Parapatics K, Welsh EA, Müller AC, et al. A Pilot Proteogenomic Study With Data Integration Identifies MCT1 and GLUT1 as Prognostic Markers in Lung Adenocarcinoma. *PLoS One*. 2015; 10:e0142162. [PubMed: 26539827]
10. Pernemalm M, De Petris L, Branca RM, Forshed J, et al. Quantitative proteomics profiling of primary lung adenocarcinoma tumors reveals functional perturbations in tumor metabolism. *J Proteome Res*. 2013; 12:3934–3943. [PubMed: 23902561]
11. Li Y, Wang X, Ao M, Gabrielson E, et al. Aberrant Mucin5B expression in lung adenocarcinomas detected by iTRAQ labeling quantitative proteomics and immunohistochemistry. *Clin Proteomics*. 2013; 10:15. [PubMed: 24176033]
12. Zhang PF, Zeng GQ, Yi LZ, Liu JP, et al. Identification of integrin beta1 as a prognostic biomarker for human lung adenocarcinoma using 2D-LC-MS/MS combined with iTRAQ technology. *Oncol Rep*. 2013; 30:341–349. [PubMed: 23677397]
13. Zhang PF, Zeng GQ, Hu R, Li C, et al. Identification of flotillin-1 as a novel biomarker for lymph node metastasis and prognosis of lung adenocarcinoma by quantitative plasma membrane proteome analysis. *J Proteomics*. 2012; 77:202–214. [PubMed: 22982323]

14. Zeng GQ, Zhang PF, Deng X, Yu FL, et al. Identification of candidate biomarkers for early detection of human lung squamous cell cancer by quantitative proteomics. *Mol Cell Proteomics*. 2012; 11 M111 013946.
15. Hsu CH, Hsu CW, Hsueh C, Wang CL, et al. Identification and characterization of potential biomarkers by quantitative tissue proteomics of primary lung adenocarcinoma. *Mol Cell Proteomics*. 2016
16. Clark DJ, Fondrie WE, Yang A, Mao L. Triple SILAC quantitative proteomic analysis reveals differential abundance of cell signaling proteins between normal and lung cancer-derived exosomes. *J Proteomics*. 2016; 133:161–169. [PubMed: 26739763]
17. Zhang W, Wei Y, Ignatchenko V, Li L, et al. Proteomic profiles of human lung adeno and squamous cell carcinoma using super-SILAC and label-free quantification approaches. *Proteomics*. 2014; 14:795–803. [PubMed: 24453208]
18. Tenzer S, Leidinger P, Backes C, Huwer H, et al. Integrated quantitative proteomic and transcriptomic analysis of lung tumor and control tissue: a lung cancer showcase. *Oncotarget*. 2016
19. Tabb DL, Wang X, Carr SA, Clauser KR, et al. Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *J Proteome Res*. 2016; 15:691–706. [PubMed: 26653538]
20. Mertins P, Udeshi ND, Clauser KR, Mani DR, et al. iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Mol Cell Proteomics*. 2012; 11 M111 014423.
21. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science*. 2006; 312:212–217. [PubMed: 16614208]
22. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003; 422:198–207. [PubMed: 12634793]
23. Gillet LC, Navarro P, Tate S, Rost H, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics*. 2012; 11 O111 016717.
24. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods*. 2004; 1:39–45. [PubMed: 15782151]
25. Geiger T, Cox J, Mann M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics*. 2010; 9:2252–2261. [PubMed: 20610777]
26. Silva JC, Denny R, Dorschel CA, Gorenstein M, et al. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem*. 2005; 77:2187–2200. [PubMed: 15801753]
27. Prakash A, Peterman S, Ahmad S, Sarracino D, et al. Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis. *J Proteome Res*. 2014; 13:5415–5430. [PubMed: 25244318]
28. Panchaud A, Scherl A, Shaffer SA, von Haller PD, et al. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem*. 2009; 81:6481–6488. [PubMed: 19572557]
29. Sajic T, Liu Y, Aebersold R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics Clin Appl*. 2015; 9:307–321. [PubMed: 25504613]
30. Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, et al. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics*. 2015; 15:964–980. [PubMed: 25430050]
31. Ting YS, Egertson JD, Payne SH, Kim S, et al. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics*. 2015; 14:2301–2307. [PubMed: 26217018]
32. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, et al. Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods*. 2013; 10:744–746. [PubMed: 23793237]
33. Megger DA, Padden J, Rosowski K, Uszkoreit J, et al. One Sample, One Shot - Evaluation of sample preparation protocols for the mass spectrometric proteome analysis of human bile fluid without extensive fractionation. *J Proteomics*. 2016; 154:13–21. [PubMed: 27932303]

34. Nagaraj N, Kulak NA, Cox J, Neuhauser N, et al. System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol Cell Proteomics*. 2012; 11 M111 013722.
35. Thakur SS, Geiger T, Chatterjee B, Bandilla P, et al. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol Cell Proteomics*. 2011; 10 M110 003699.
36. Zhu G, Sun L, Yan X, Dovichi NJ. Single-shot proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with production of more than 1250 *Escherichia coli* peptide identifications in a 50 min separation. *Anal Chem*. 2013; 85:2569–2573. [PubMed: 23394296]
37. Yang F, Shen Y, Camp DG 2nd, Smith RD. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev Proteomics*. 2012; 9:129–134. [PubMed: 22462785]
38. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008; 26:1367–1372. [PubMed: 19029910]
39. RStudio Team. RStudio, Inc.; Boston, MA: 2015.
40. R Core Team. R Foundation for Statistical Computing; Vienna, Austria: 2015.
41. Kall L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res*. 2008; 7:40–44. [PubMed: 18052118]
42. Chambers MC, Maclean B, Burke R, Amodei D, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 2012; 30:918–920. [PubMed: 23051804]
43. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*. 2014; 5:5277. [PubMed: 25358478]
44. Ma ZQ, Dasari S, Chambers MC, Litton MD, et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res*. 2009; 8:3872–3881. [PubMed: 19522537]
45. Vizcaino JA, Csordas A, Del-Toro N, Dienes JA, et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res*. 2016; 44:11033. [PubMed: 27683222]
46. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*. 2016; 44:D336–342. [PubMed: 26578592]
47. Thomson Reuters. 2016.

Statement of Significance

Expression proteomics methods are compared using 8 lung tissue samples (4 squamous cell carcinoma and 4 tumor-adjacent tissues) to examine the utility of each method and the biology that can be sampled based on the depth of proteome information in each analysis. To date, few head-to-head comparisons between proteomics methods have been made using the same human tissue samples. This work shows the benefit of using TMT as a timesaving alternative to LF with similar proteome coverage and suggests single sample DDA and DIA LC-MS/MS methods are appropriate for rapid quality control or in instances with extremely limited sample availability (*e.g.* biopsies or tissue microarrays).

Figure 1. Comparison of Discovery Proteomics Experiments

A) The data processing and analysis workflows for the four experiments. *IDPicker and MS-GF+ were used for pseudo-DDA analysis of the DIA dataset through a traditional database searching approach. [†]Required instrument time is an additional 15 hours when LC-MS/MS is used to check labeling efficiency. **B)** Number of protein identifications from the DDA, DIA, LF, and TMT experiments. **C)** Pathway content of the LF experiment compared to the DDA and DIA experiments. **D)** Pathway content of the LF experiment compared to the TMT experiment.

Figure 2. Comparison of Expression Ratios for SCC Tumor Markers

SCC marker protein ratios altered in at least one patient (> 1.5 SD of change compared to the average log_2 ratio). **A)** Comparison of common LF and TMT proteins, which show similar trends despite dampened protein expression ratios in the TMT experiment. **B)** Comparison of LF and DIA proteins, which includes unexpected disagreement between the LF and DIA datasets. P1, P2, and P3 are patients 1–3 while P4/5 is an unmatched adjacent lung tissue-tumor pair. Missing values are colored as dark grey.

A) LF peptide SD vs. mean $\log_2$ peptide intensity with best fit line. **B)** DIA peptide SD vs. mean $\log_2$ peptide peak area with best fit line. **C)** LF protein SD vs. mean $\log_2$ protein intensity with best fit line. **D)** DIA protein SD vs. mean $\log_2$ protein peak area with best fit line. **E)** LF peptide RSD organized by mean peptide quartile. Ranges: 0%: 16.35, 25%: 20.19, 50%: 21.27, 75%: 22.66, 100%: 34.37. **F)** DIA peptide RSD organized by mean peptide quartile. Ranges: 0%: 18.27, 25%: 24.11, 50%: 25.41, 75%: 26.78, 100%: 37.62. **G)** Number of peptides contained within each quartile range.

Table 1

Peptide and Protein Identifications for each Expression Proteomics Experiment.

	DDA	DIA (IDPicker)	DIA (Pinnacle)	LF	TMT
Peptides	13,764	12,976	11,783	42,594	50,754
Proteins					
	2,219	2,155	3,409	6,964	5,830
Total LC-MS/MS Time	2 days	2 days	2 days	24 days	6 days
Proteins Per Day	1,110	1,075	1,705	290	972
Single Peptide Protein Identifications	571	530	260	578	705
Peptides per Protein	6.23	6.04	3.46	6.12	8.71
Squamous Protein Markers	562	560	681	1,095	1,015
Differentially Expressed Proteins (Tumor vs. Adjacent Lung Tissue)	137	147	385	773	562

Table 2

Selected Pathways Assembled from Observed Lung Cancer-Related Proteins in the Discovery Datasets.

Pathway	LF Proteins	LF FDR	TMT Proteins	TMT FDR	DIA Proteins	DIA FDR	DDA Proteins	DDA FDR
Apoptosis and survival_Apoptotic TNF-family pathways	18	4.43E-02	16	6.02E-02	7	6.18E-01	1	1.00E+00
Apoptosis and survival_p53-dependent apoptosis	14	2.76E-02	11	1.19E-01	8	1.33E-01	5	4.39E-01
Development_Dopamine D2 receptor transactivation of EGFR	16	3.07E-04	14	1.40E-03	8	5.92E-02	8	2.65E-02
Development_EGFR signaling via PIP3	13	7.98E-03	10	6.28E-02	6	2.26E-01	3	7.31E-01
Development_EGFR signaling via small GTPases	24	1.00E-06	20	6.90E-05	10	5.92E-02	7	2.17E-01
Development_ERBB-family signaling	19	9.56E-03	17	1.45E-02	7	5.32E-01	7	3.39E-01
G-protein signaling_K-RAS regulation pathway	14	6.49E-03	10	1.04E-01	4	6.92E-01	2	9.31E-01
Immune response_NEAT in immune response	26	1.15E-03	16	2.34E-01	12	1.58E-01	6	7.77E-01
Stimulation of TGF-beta signaling in lung cancer	24	2.44E-03	21	6.10E-03	11	1.96E-01	11	8.47E-02
Testing CDK2, ERK1/2, Rb protein, STAT1, ERK2 (MAPK1)	26	8.00E-04	24	8.66E-04	13	8.10E-02	9	2.87E-01