Simultaneous Label-free Quantitation of Proteoforms, Proteoform Ratios, and Total Protein up to 80kDa Split across Physiochemical Multidimensional Space

UT SOUTHWESTERN MEDICAL CENTER

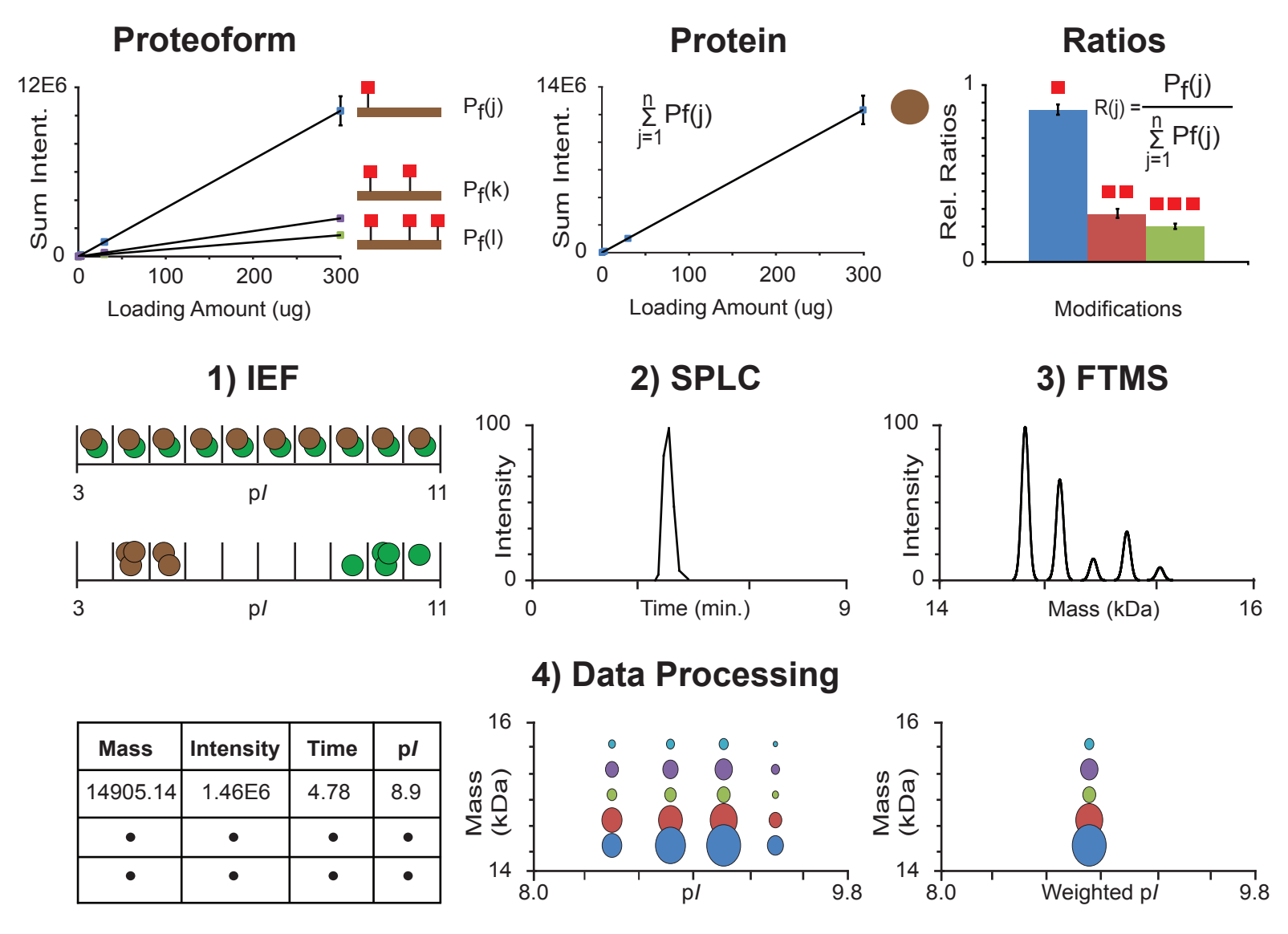
John R. Corbett, Daniel A. Plymire, Casey E. Wing, William S Phipps, and Steven M. Patrie*

UT DALLAS

Overview

In contrast to conventional bottom-up methods for protein quantitation, top-down proteomics enables simultaneous label-free quantitation (LFQ) of proteoforms, proteoform ratios, and total protein. To date, integrated access to these levels in a single experiment has been hindered by limited reliability of multidimensional chromatography, mismatches in procedures that maximize peak capacity for protein or proteoform detection across a proteome, and limitation in data processing tools. Utilizing optimized chromatography and a fully automated data processing pipeline, we demonstrate that IEF-SPLC-FTMS can reliably perform simultaneous LFQ on the three quantitative levels across a chromatographic multidimensional space for masses up to 80kDa.

Methods



Mass	Intensity	Time	pI
14905.14	1.46E6	4.78	8.9
•	•	•	•
•	•	•	•

Standard proteins ranging in size and physiochemical properties and complex mixtures were isoelectric focused (IEF) on an Offgel 3100 (Agilent) at different loading amounts. Focused fractions were subsequently DTT incubated and separated with RPLC Poroshell (Agilent). The LC eluate was split using a TriVersa NanoMate (Advion) with ~0.5 µL/min directed to a LTQ-Orbitrap-XL/ETD (Thermo) at resolving powers of either 60,000 or 15,000 depending on protein size. Collected files were processed using modified THRASH and MassDecon algorithms facilitating parallel processing and sliding-window functionality. Extracted data files were introduced to an automated pipeline for data binning using nearest neighbor queries based upon empirically determined tolerances for mass and weighted hydrophobicity/p*I*. LFQ is completed using the summed intensity values within each respective bin (1,2).

Reproducible Chromatography: Figures of Merit

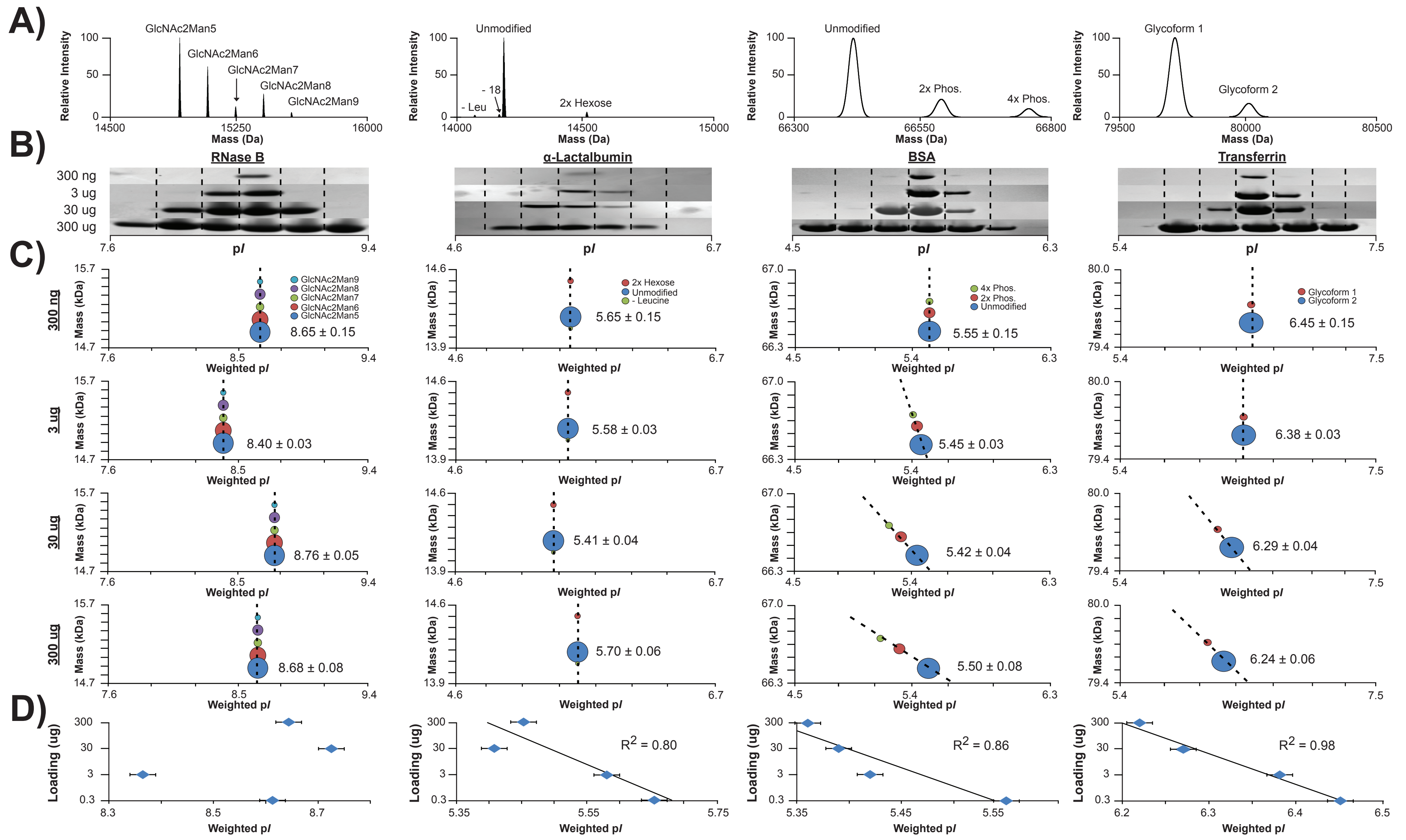


Figure 1: A) Standard proteins covering a wide physiochemical property range were analyzed via LC-MS in order to assign proteoforms that will be used for figures of merit generation. B) To assess the IEF-SPLC-FTMS platform, standards were subjected to IEF (n=3) at 300 ng, 3 ug, 30 ug, and 300 ug sample loadings and visualized with silver stain. Representative gels highlight that with increased loading, chromatographic peak broadening occurs requiring binning of mass spectra across the p*I* domain. C) Average weighted p*I* bubble plots for the standard proteins at each sample loading with the average weighted p*I* value shown for the base (most abundant) proteoform. Trends in p*I* changes among related proteoforms is shown (dashed lines) with proteoforms containing charge changing PTMs (i.e. phosphorylation, sialylation) having different weighted p*I* values compared to the base form. D) Assessment of load vs. total protein weighted p*I* values for each respective standard protein (n=3) across the four different loading amounts suggests a leftward (p*I* domain) trend between weighted p*I* and loading.

2D Label-free Top-down Quantitation

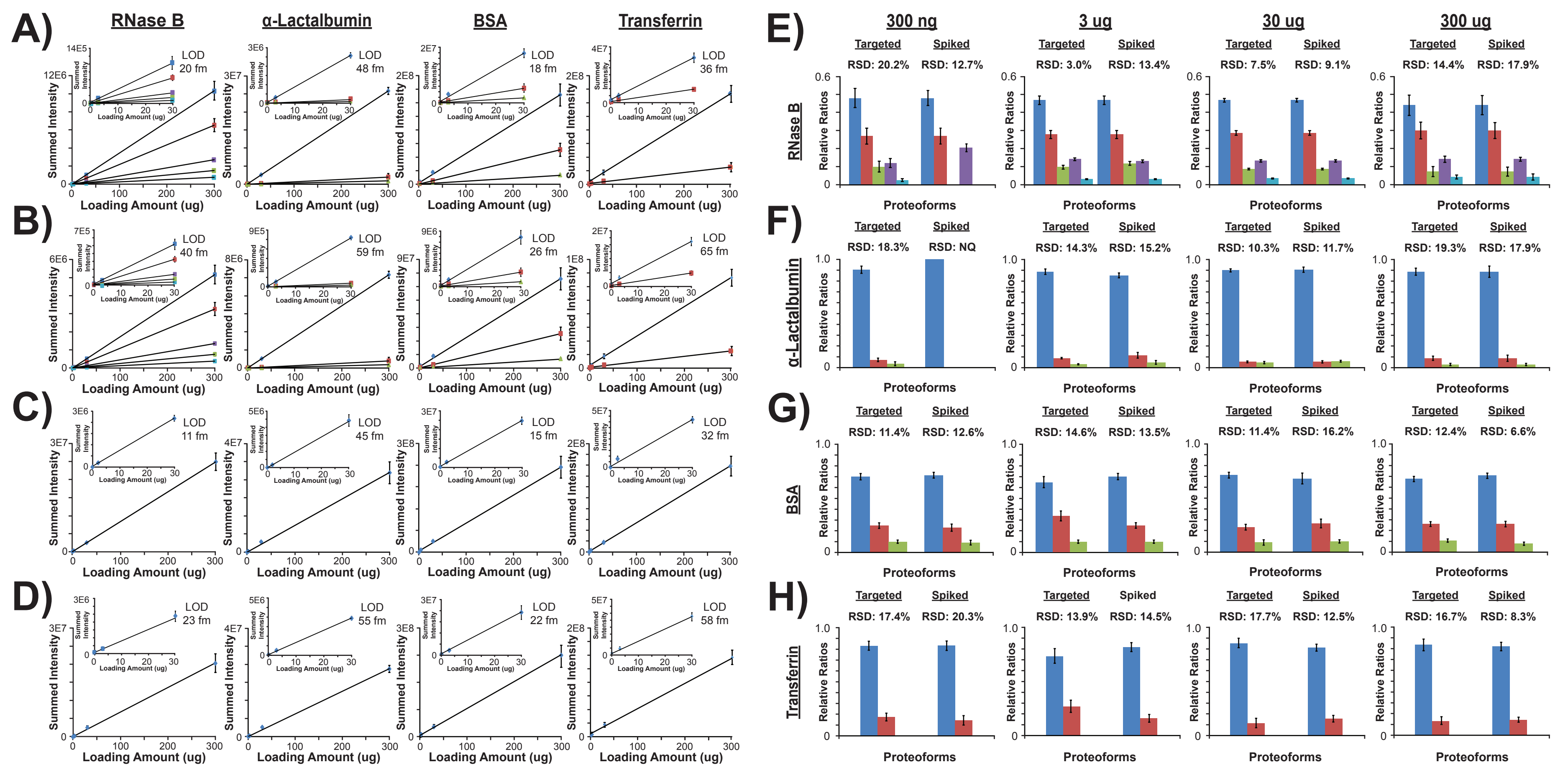


Figure 2: A-B) Analysis of standards at the four sample loadings was completed under targeted and spiked conditions with the latter occurring in the presence of background 1 mg *E. coli* lysate (analytical matrix effect). Proteoform calibration curves based on binned summed intensity values for each respective proteoform at both targeted (A) and spiked (B) experiments illustrate a linear dynamic range beyond three orders of magnitude. C-D) Total protein calibration curves at both targeted (C) and spiked (D) experiments mirror the proteoform three orders of magnitude linear dynamic range. E-H) Analysis of the standard proteins proteoform ratios at both targeted and spiked conditions across the four loading amounts highlight that ratios are constant across sample loadings for both targeted and spiked conditions with minor matrix effects observed at 300 ng.

Reproducible Chromatography: Proteome Scale

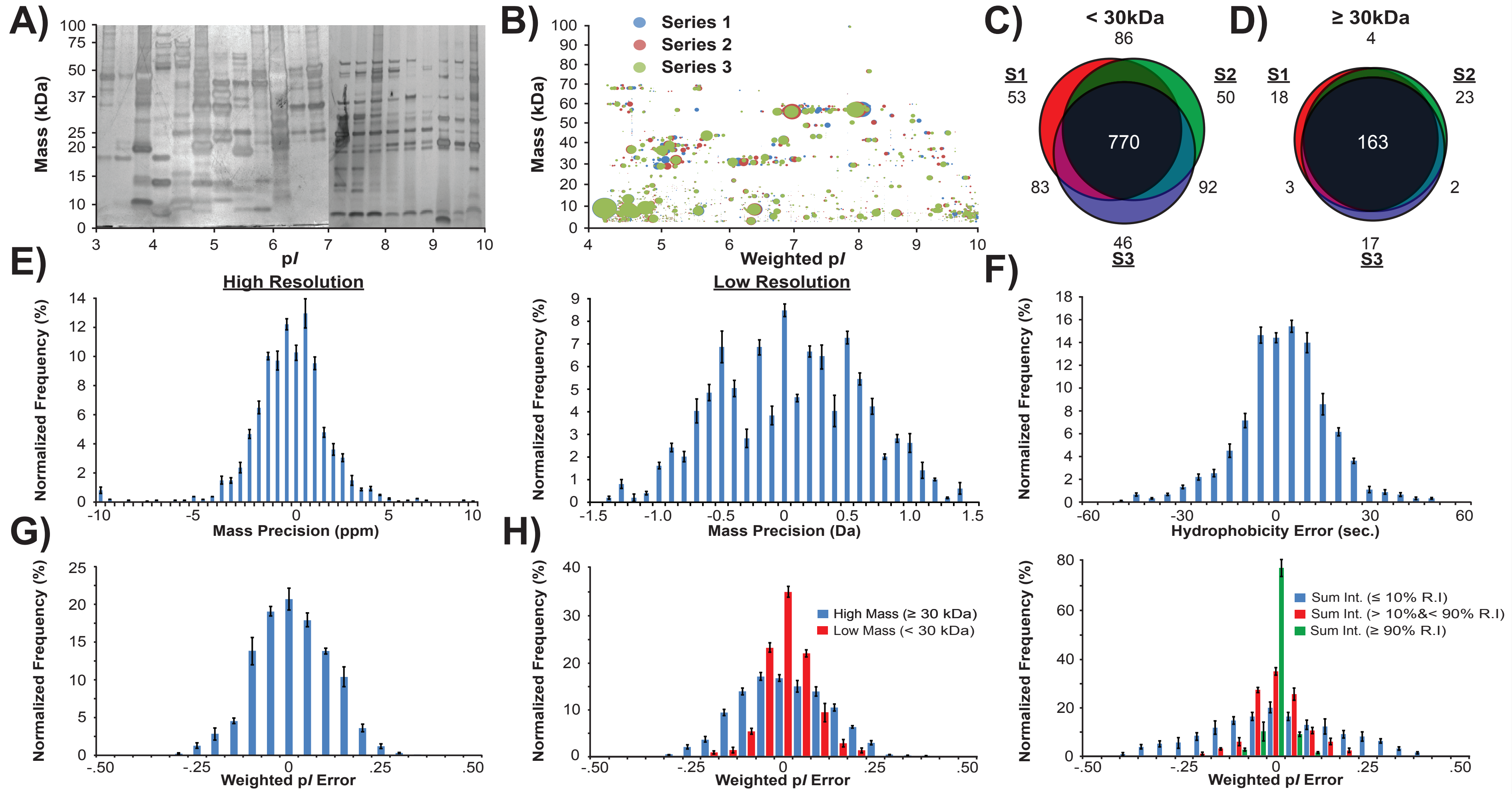


Figure 3: A-B) To assess the IEF-SPLC-FTMS platforms chromatographic reproducibility across an entire proteome run, *E. coli* lysate was analyzed (n=3) under both high (60k) and low (15k) FTMS resolution conditions in order to observe proteoforms < 30 kDa and ≥ 30 kDa respectively with bubble plot results showing good correlation to silver stained gels. C-D) Venn diagram analysis of proteoforms observed indicate ~77% (770) and ~86% (163) of species are observed in all replicates with the species that are unique to one or two datasets < 0.5% relative summed intensity. E) Analysis for the 770 and 163 species observed in the high and low FTMS resolution runs reveals that ~95% of species had a mass precision within 5 ppm and 1 Da respectively. F-G) A similar based analysis completed for hydrophobicity and weighted p*I* error show that ~95% of species fall within a 30 second and 0.25 p*I* unit window, respectively. H) Further investigation of weighted p*I* error shows that different error trends were observed depending on proteoform mass and summed intensity value. (R.I.-relative intensity).

Towards Automation

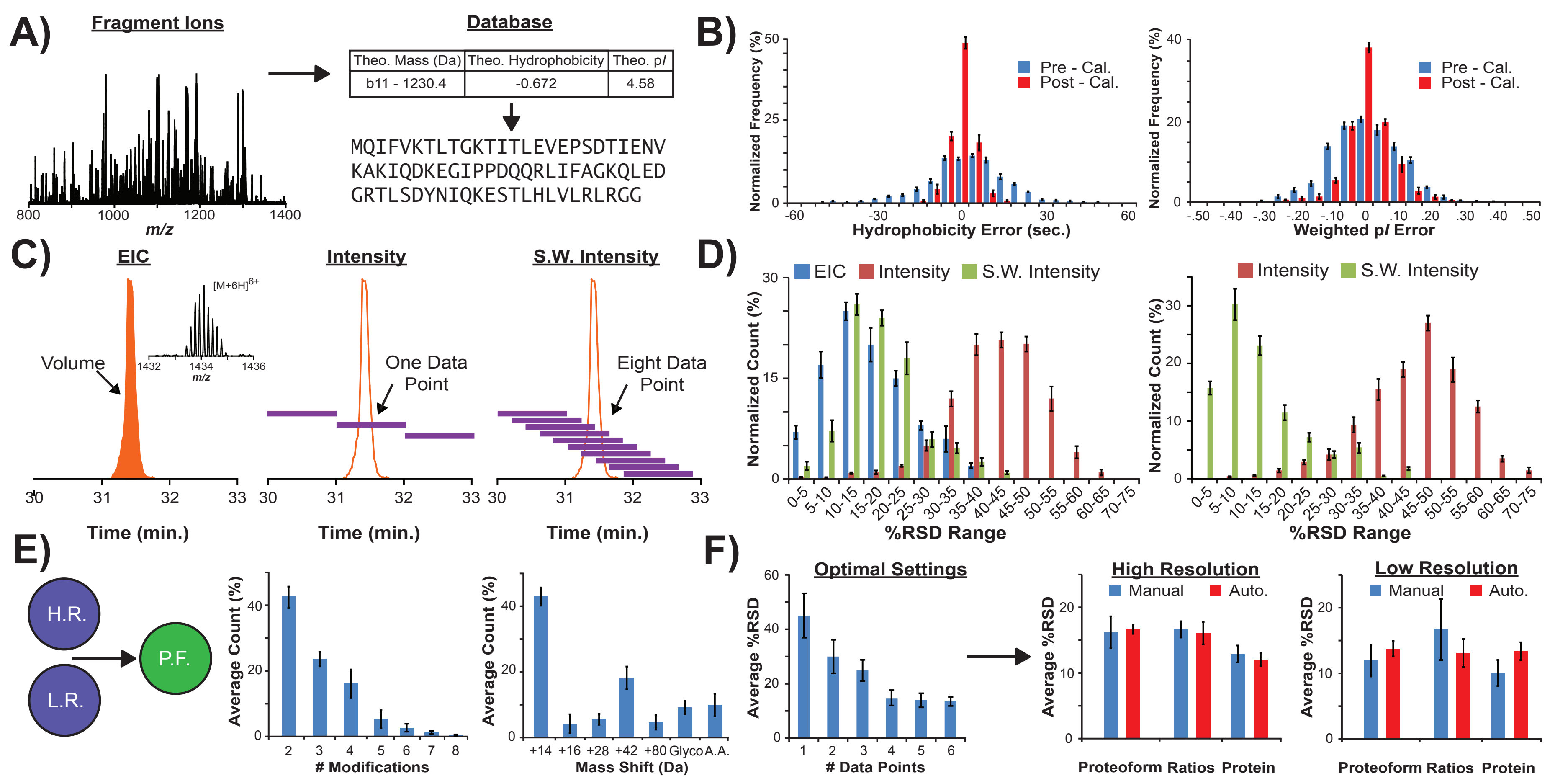


Figure 4: A) MS/MS experiments on *E. coli* lysate were completed with data searched using in-house software (Proteoformer) that includes additional theoretical physiochemical properties (p*I* and hydrophobicity) alongside fragment ions. B) Calibration curves comparing theoretical and observed physiochemical properties have been created (*E. coli* datasets) with calibrated data showing an improvement in hydrophobicity and weighted p*I* error with values of 12 seconds and 0.2 p*I* units respectively. C-D) Extracted ion current (EIC) volumes, incremental time marches, and sliding window (S.W.) approaches were tested to determine the optimal approach for quantitative studies across a proteome for both high and low FTMS resolutions. Comparison of the summed intensities for the observed *E. coli* proteoforms highlight that a S.W. approach provides good quantitative reproducibility (~17% and ~13%) at both high and low FTMS resolutions. E) Proteoform families (P.F.) were created for the triplicate *E. coli* datasets with most families having only two related proteoforms with the most commonly observed modification being methylation (A.A. - amino acids). F) Procedures have been automated and includes an aid for the users that defines optimal data extraction procedures for quantitative studies based on the number of data points across a peak and average %RSD observed. Using these automated procedures on the *E. coli* datasets reveal comparable quantitative reproducibility results for the three quantitative metrics of top-down, but in a much higher throughput (~1 hour).

Conclusions

1. IEF separation of standards across a wide loading amount shows chromatographic peak broadening; however, weighted p*I* values are reproducibly observed at each loading amount due to the data binning procedures. Trends in weighted p*I* values among related proteoforms due to sample loading and charge changing modifications is observed.
2. Calibration curves under targeted and spiked conditions suggest a linear dynamic range beyond three orders of magnitude for both proteoform and total protein. Proteoform ratios were consistent across loading amounts and spiked conditions with minor matrix effects observed.
3. Triplicate IEF-SPLC-FTMS analysis on *E. coli* lysate was completed under both high and low FTMS conditions with proteoforms beyond 80 kDa observed. ~77% and 86% of the observed high and low FTMS datasets were reproducibly observed in all runs.
4. Precision values for the different physiochemical properties were determined with weighted p*I* error dependent on proteoform mass and intensity.
5. Proteoformer was implemented on collected MS/MS data with calibration curves based on theoretical and observed physiochemical properties created to improve hydrophobicity and weighted p*I* error.
6. A sliding window approach provides comparable summed intensity reproducibility as an EIC approach at both high and low FTMS resolution. Proteoform families have been created to determine proteoform ratios and total protein with automated procedures providing quantitative results comparable to a manual approach.

References

1. **Corbett, J.R**.; Zhang, J.; Plymire, D.A; *et. al. Proteomics*, **2014**, 14, 1223-1231.
2. Zhang, J.; Roth, M.J.; **Corbett, J.R.**; *et. al. Anal Chem*, **2013**, 85, 10377-10384.