UNIVERSITY OF CALIFORNIA

Los Angeles

# Novel Approaches in Bottom-Up Proteomic Sample Preparation, Acquisition, and Analysis

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biological Chemistry

by

William Dana Barshop

2018

ABSTRACT OF THE DISSERTATION

Novel Approaches in Bottom-Up Proteomic

Sample Preparation, Acquisition, and Analysis

by

William Dana Barshop

Doctor of Philosophy in Biological Chemistry

University of California, Los Angeles, 2018

Professor James Akira Wohlschlegel, Chair

The use of proteomic mass spectrometry has become a pervasive component of modern biological and biochemical research. The experimental detection and quantitation of proteins is largely accomplished via liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). In this work, we explore approaches to common problems which pervade three core facets of contemporary proteomic biochemistry by LC-MS/MS: sample preparation, data acquisition, and bioinformatic analysis workflow management.

A common step in the sample-preparatory framework is the affinity purification of protein targets, often through the use of antibodies or the protein streptavidin. Avidin proteins such as streptavidin are capable of binding the small molecule biotin with high affinity and specificity. The binding of biotin to streptavidin is oft exploited due to the extremely high affinity and near irreversibility of the interaction. Elution of biotinylated proteins remains inefficient, and many

rely on enzymatic digestion, ultimately releasing a large amount of contaminating streptavidin peptides. We explore a method of chemical derivatization which protects streptavidin from tryptic proteolysis, dramatically reducing sample contamination while retaining biotin binding. The method appears generalizable to immunoglobulins antibodies like those against the hemagglutinin epitope.

Relative quantitation of proteins and peptides is often performed by comparing the intensities of many samples in a single chromatographic run through multiplexing provided by isobaric tagging reagents. Quantitation of these isobaric tags is observed after fragmentation of a purified analyte, typically selected by Data Dependent Acquisition in a semi-stochastic manner. We explore a new method of acquisition, Sequential Windowed Acquisition of Reporter Masses (SWARM), a Data Independent Acquisition-like approach to isobaric tagged peptide quantitation. This approach biases machine acquisition toward analytes based on their quantitative trends, allowing biologists to focus instrument time on putative analytes of interest.

Data produced from the multitude of proteomic experiments must be rigorously analyzed to deconvolute the complex aggregate of mass signals before returning actionable interpretation. The expansion of computational tools the for interrogation of LC-MS/MS data has been a boon to the field, and has made many sophisticated and statistically robust analyses available. However, these tools have been left in unfortunately disjointed sets of software packages lacking convenient interoperability. To help address this problem, we created MilkyWay. MilkyWay is a label free proteomic data analysis platform for quantitative comparisons. Powered by an assemblage of utilities wrapped into the Galaxy bioinformatic workflow management system, MilkyWay contains a R/Shiny web application for the interactive definition of experimental design, file upload, and data exploration.

The dissertation of William Dana Barshop is approved.

Keriann Marie Backus

Hilary Ann Coller

Joseph Ambrose Loo

James Akira Wohlschlegel, Committee Chair

University of California, Los Angeles

2018

# TABLE OF CONTENTS

**LIST OF FIGURES**

**LIST OF TABLES**

**ACKNOWLEDGEMENTS**

I owe an immeasurable debt to my family and friends. You have all supported me throughout the duration of my education and dissertation, and have always challenged me to do my best. This is especially true of my patient and kind fiancée, Kana. None of this would have been possible without you.

When I came to UCLA, I could never have guessed the myriad resources and opportunities that would be available to me. The individuals who have passed through the Wohlschlegel lab during my time at UCLA will always have a special place in my heart. The multitude of incredible scientists I have had the pleasure of working beside truly drove me to explore further, ask harder questions, and find joy in the painstaking work of science. A special thank you to my friend and colleague, Hee Jong Kim. The experience of working with a partner in lab who is so driven, curious, and talented cannot be understated.

I would also like to acknowledge the funding sources that made my research possible: the Cellular and Molecular Biology Training Grant (Ruth L. Kirschstein National Research Service Award GM007185), the Audree Fowler Fellowship in Protein Science, and the UCLA Dissertation Year Fellowship. The funding allowed me the freedom to tinker, explore, and take risks in research. Lastly, I want to thank my thesis advisor and mentor, James Wohlschlegel. James provided me with expert advice and guidance, and always encouraged me to continue to try new things in lab. Thank you for your support.

# WILLIAM DANA BARSHOP

## Education

*AB* in Biology, Washington University in St. Louis (2008 – 2012)

## Research

University of California, Los Angeles with Dr. James Wohlschlegel (2012 – current)
Washington University in St. Louis with Dr. Sarah C.R. Elgin (2011 – 2012)

## Awards and Fellowships

**Audree Fowler Fellowship in Protein Science** (2018)
**UCLA Dissertation Year Fellowship** (2017 – 2018)
**Cellular and Molecular Biology Training Grant** (Ruth L. Kirschstein National Research Service Award GM007185) (2015 – 2017)
**National Merit Scholarship** (2008 – 2012)

## Professional Experience

**Teaching Assistant, UCLA Honors Collegium** (2014 – 2015)
- HC70A: Genetic Engineering in Medicine, Law, & Agriculture with Dr. Robert B Goldberg

**Teaching Assistant, Washington University in St. Louis** (2011 – 2012)
- Biology 4342: Research Explorations in Genomics with Dr. Sarah C.R. Elgin
- Biology 191: Phage Hunters with Dr. Kathy Hafer

## Presentations

**UCLA Molecular Biology Institute Annual Retreat** (2018)
Oral Presentation: *"Masking Streptavidin for LC-MS and Proteomic Data Analysis in MilkyWay"*

**UCLA Biological Chemistry Floor Meeting** (2015, 2018)
Oral Presentation: *"Quantitative Proteomic Approaches to Protein Complexes"*

**American Society for Mass Spectrometry Conference** (2017)
Poster: *"MilkyWay: A Galaxy platform for quantitative comparative analysis of bottom-up proteomic mass spectrometry datasets"*

# Publications

Pope WH, Jacobs-Sera D, Russell DA, Peebles CL Al-Atrache Z, …**Barshop WD**,… Zuniga MY, Hendrix RW, Hatfull GF. "Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution." PLoS One. 2011 Jan 27;6(1):e16329.

- See publication for full list of authors

Leung, W., Shaffer, C.D., Reed, L.K., Smith, S.T., **Barshop, W**., Dirkes, W., Dothager, M., Lee, P., Wong, J., Xiong, D., et al. (2015). "Drosophila muller f elements maintain a distinct set of genomic properties over 40 million years of evolution." G3 (Bethesda) 5(5): 719-740.

Shimogawa MM, Saada EA, Vashisht AA, **Barshop WD**, Wohlschlegel JA, Hill KL. (2015). "Cell Surface Proteomics Provides Insight into Stage-Specific Remodeling of the Host-Parasite Interface in Trypanosoma brucei." Mol Cell Proteomics 14(7): 1977-1988.

*Wang, Q., ***Barshop, WD**, Bian, M., Vashisht, A.A., He, R., Yu, X., Liu, B., Nguyen, P., Liu, X., Zhao, X., et al. (2015). "The blue light-dependent phosphorylation of the CCE domain determines the photosensitivity of Arabidopsis CRY2." Mol Plant 8(4): 631-643.

* denotes co-first authors

Langousis G, Shimogawa MM, Saada EA, Vashisht AA, Spreafico R, Nager AR, **Barshop WD**, Nachury MV, Wohlschlegel JA, Hill KL. (2016) "Loss of the BBSome disrupts endocytic trafficking and virulence of *Trypanosoma brucei*" Proc Natl Acad Sci U S A 113(3): 632-637.

Kim EW, Nadipuram SM, Tetlow AL, **Barshop WD**, Liu PT, Wohlschlegel JA, Bradley PJ (2016). "The Rhoptry Pseudokinase ROP54 Modulates Toxoplasma gondii Virulence and Host GBP2 Loading." mSphere 1(2).

Caslavka Zempel, K.E., Vashisht, A.A., **Barshop, W.D**., Wohlschlegel, J.A., and Clarke, S.G. (2016). "Determining the Mitochondrial Methyl Proteome in Saccharomyces cerevisiae using Heavy Methyl SILAC." Journal of proteome research 15, 4436-4451.

Nitarska, J., Smith, J.G., Sherlock, W.T., Hillege, M.M., Nott, A., **Barshop, W.D.**, Vashisht, A.A., Wohlschlegel, J.A., Mitter, R., and Riccio, A. (2016). "A Functional Switch of NuRD Chromatin Remodeling Complex Subunits Regulates Mouse Cortical Development." Cell reports 17, 1683-1698.

Leung, W., Shaffer, C.D., Chen, E.J., Quisenberry, T.J., Ko, K., Braverman, J.M., Giarla, T.C., Mortimer, N.T., Reed, L.K., Smith, S.T*., et al.* (2017). Retrotransposons Are the Major Contributors to the Expansion of the Drosophila ananassae Muller F Element. G3 *7*, 2439-2460.

- See publication for full list of authors

Liu, Q., Wang, Q., Deng, W., Wang, X., Piao, M., Cai, D., Li, Y., **Barshop, W.D.**, Yu, X., Zhou, T*., et al.* (2017). Molecular basis for blue light-dependent phosphorylation of Arabidopsis cryptochrome 2. Nature communications *8*, 15234.

Benador, I.Y., Veliova, M., Mahdaviani, K., Petcherski, A., Wikstrom, J.D., Assali, E.A., Acin-Perez, R., Shum, M., Oliveira, M.F., Cinti, S*.,* Sztalryd, C., **Barshop, W.D.**, Wohlschlegel, J.A., Corkey, B.E., Liesa, M., Shirihai, O.S. (2018). Mitochondria Bound to Lipid Droplets Have Unique Bioenergetics, Composition, and Dynamics that Support Lipid Droplet Expansion. Cell Metab *27*, 869-885 e866.

# Chapter 1: Introduction to Proteomic Mass Spectrometry

## *Liquid Chromatography and Bottom-Up Proteomic Mass Spectrometry*

In recent years, proteomic mass spectrometry has become a critical facet of cell biological research to probe the function of bioregulatory mechanisms and understand the activities of uncharacterized proteins within complex systems[1]. The analytical interrogation of such protein mixtures, though, has developed over quite some time into a complex and refined process. Modern mass spectrometers, the instruments responsible for the detection of ionized analytes by their mass-to-charge (*m/z*) ratio, have become both sensitive and have obtained high resolution and mass accuracy[2]. The ionization of protein and peptide analytes of interest is most often accomplished by means of electrospray ionization (ESI), a soft ionization method in which compounds in solution are exposed to a high voltage source and ionized in the subsequent desolvation of sprayed droplets[3]. ESI has been a successful approach due in part to both its applicability to a litany of chemically dissimilar species, but also its ability to ionize large and fragile biomolecules without fragmentation. In many highly complex mixtures, though, the number of analytes quickly becomes intractable to analyze without further biochemical separation due to the natural dynamic range of species as well as ion suppression effects[4].

The coupling of liquid chromatography to electrospray ionization and mass spectrometry has become a commonplace solution to simplify the mixture of analytes introduced into the mass analyzer at any moment in time. The chromatographic separation of intact proteins and complexes, however, has proven to be an analytical challenge and sometimes requires multiple dimensions of separation to comprehensively analyze[5]. Instead, protein mixtures are enzymatically digested to produce smaller peptide analytes which are more readily separated and identified by LC-MS/MS via easily implemented reversed phase chromatographic setups. The use of trypsin for proteolysis has dominated "bottom-up" proteomics for its high specificity, and its production of peptides with

C-terminal enrichments of the basic amino acids, lysine and arginine[6]. These tryptic peptides are more easily distinguished from chemical noise due to their tendency to ionize well at charge states at or above +2. Although the digested peptides are more amenable to chromatographic separation, the act of enzymatically cleaving a complex mixture of proteins yields a biochemical mixture with a dramatic increase in the number of chemically distinct peptide analytes. Furthermore, the gambit of digestion causes the loss of valuable proteoform level information[7]. For example, multiple distinct proteins may simultaneously give rise to the same tryptic peptide while isoform level information is made difficult or impossible to deconvolute, as individual peptides are no longer meaningfully associated to their progenitor proteins[8]. Additionally, once a protein is digested, the protein level context for post-translational modifications becomes obfuscated.

These peptide mixtures, separated chromatographically, are ionized and detected as intact charged analytes during survey (MS1) scans. When instrumentation is operated using Data Dependent Acquisition (DDA), detectable signals in the MS1 scans are ranked by intensity, filtered to remove any recently fragmented signals, and processed to only target analytes of user specified charge states. Those selected targets are sequentially purified in the gas phase, by means of quadrupole or linear ion trap isolation, and typically fragmented (MS2). Depending on the instrumentation and intent of the experimenter, various methods of fragmentation may be employed[9,10]. Most commonly, methods for rapid sequencing of peptides will utilize methods of collisional dissociation (Higher-energy Collisional Dissociation, HCD, or Collision-induced Dissociation, CID) which excite and collide ions with inert bath gasses to generate fragmentation events along the peptide backbone[11]. Scan data resulting from the fragmentation of peptide analytes may be compared against theoretical fragmentation patterns of peptides contained within *in silico* proteome tryptic digests to yield sequence identifications[12-15]. Ultimately, protein

identifications must be inferred from the corpus of peptide identifications gleaned from the peptide-spectrum-matches (PSMs)[8,16].

## *Quantitative Methods in Proteomics*

As the ability to generate confident identifications of thousands of proteins per hour has reached maturity, the initial focus on merely identifying proteins within a sample has given way to a heightened focus on quantitative measurements. The observable signal of an ionized analyte by mass spectrometry is not inherently quantitative, as every molecule will have a different ionization efficiency and detectability; researchers have placed great efforts in exploring methods of both relative and absolute quantification of peptides. The simplest methods of relative quantitation are those based on the spectral-counting (SpC) methodology, in which protein relative abundance is estimated simply by counting the number of confident PSMs which map to a protein of interest[17]. An extension of this method, the Normalized Spectral Abundance Factor (NSAF) attempts to offset the effect of protein size on spectral counts and the total number of spectra generated within the experiment[18]. Spectral counting measures of protein abundance are a crude, but widely and easily available. The use of intensity based label free quantification of peptides has become increasingly popular despite the challenges and complexities in data processing. Signal extraction of peptide analytes from DDA LC-MS/MS experiments falls into two camps: (1) those which attempt to detect peptide LC-MS elution features directly from the MS1 intensity data without seeding putative features from identifications and instead associating them after feature generation[19-21], and (2) those which use identification information to calculate expected *m/z* values and generate extracted ion chromatograms (EICs or XICs) at the retention time of the identification for several C13 isotopes[22,23].

As instrumentation has become faster and more sensitive, alternates to the popular DDA instrument acquisition cycle for label-free bottom-up proteomics have become of interest. Most notably, Data Independent Acquisition (DIA) is an approach which segments the precursor *m/z* range of interest into large, predefined windows for sequential fragmentation and detection[24,25]. The approach effectively removes the stochastic selection of analytes in favor of repeated, predictable fragmentation of all analytes in the interrogated *m/z* space. As a benefit, DIA acquisition yields MS2-level chromatographic information. These fragment level data are subject to less interference than the intact (MS1) signals used for intensity based quantitation of DDA datasets[26]. Until recently it was difficult to effectively generate peptide identifications from DIA datasets due to the wide quadrupole isolation windows for each fragmentation scan which frequently results in the co-fragmentation and detection of multiple analytes simultaneously. Advances in computational tools have enabled the deconvolution of the cofragmented peptide analytes allowing for the use of well-established database search algorithms originally developed for DDA[27,28].

An alternative to label-free methods of comparative proteomics introduce various forms of labels which can allow for multiplexing of samples and comparison of quantities within a single LC-MS/MS run. These labels frequently rely on the introduction of isotopic labels into either the amino acids directly, or into a moiety which may be enzymatically or chemically bound to digested peptides. The former method, Stable Isotope Labeling of by Amino Acids in Cell culture (SILAC)[29], calls for the incubation of various isotopic labels of amino acids in the cell culture media. During cellular division and protein synthesis, these heavy amino acids are integrated into the proteome. Samples given different mass labels can be mixed together and jointly analyzed, with each labeled sample producing its own MS1 signals which are separated by a known mass

shift. For each peptide, these MS1 signals may be relatively quantified by generation of extracted ion chromatograms for each label/sample. A popular alternative to SILAC is chemical labeling through the use of isobaric tags, like the primary-amine reactive Tandem Mass Tags (TMT)[30]. These tags can be reacted with the samples after proteolytic digestion, with each tag having the same intact mass, such that they produce a single MS1 signal per analyte. Upon fragmentation, though, the tags are designed such that the distribution of heavy isotopes within the label generates a reporter ion signal of distinguishable mass for each sample. In this paradigm, relative quantitation is observed after the isolation and fragmentation of a putative peptide analyte, almost exclusively by DDA or through targeted methods like TOMAHAQ[31].

## *Affinity Purification-Mass Spectrometry*

Among the most popular methods in proteomics is to probe the protein composition of biologically relevant complexes by Affinity Purification-Mass Spectrometry (AP-MS). Target genes of interest are often engineered to code for a protein gene product fused with an affinity enrichment epitope tag of interest[32]. These tags have been utilized for protein biochemical experiments across many years, such as the FLAG tag, hemagglutinin (HA) tag, and the Myc tag. These tags have commercially available antibodies which may be used to purify tagged protein targets. Upon binding, the unbound fraction of proteins can be washed away, leaving the tagged protein and any proteins with which stable association is made. This approach allows for the identification and quantification of proteins in complex, facilitating studies seeking to reveal the biological function of the tagged protein of interest. Critically, proteins studied in this way must be amenable to copurification and maintain stable association with their binding partners.

To study protein complexes or biological compartments which are fragile, transient, or difficult to purify, the use of proximity biotinylation experiments have become increasingly popular. These methods necessitate the expression of proteins fused to enzymes capable of catalyzing localized, promiscuous biotinylation. Simply, proteins are expressed as fusion products with either the enzyme ascorbate peroxidase in the APEX method[33], or a mutated form of the *Escherichia coli* biotin ligase, BirA, for BioID[34]. Under ideal conditions, these fusion proteins will localize normally and can produce reactive chemical moieties containing biotin when supplied with the reactants. The biological material can be lysed, and protein content solubilized before the biotinylated proteins are purified via affinity purification with immobilized streptavidin protein. The streptavidin-biotin interaction is among the highest affinity known, and consequently one of the most often exploited in biochemical research[35]. While the high affinity of the interaction facilitates highly efficient recovery of biotinylated products, the elution of those retained analytes is a difficult task and often requires high amounts of detergent and heat[36].

# Chapter 2: Sequential Windowed Acquisition of Reporter Masses for Quantitation-First Proteomics

# *Sequential Windowed Acquisition of Reporter Masses for Quantitation-First Proteomics*

**William D. Barshop, Shima Rayatpisheh, Hee Jong Kim, James A. Wohlschlegel**

## *Abstract*

The standard approach for proteomic data acquisition of isobaric tagged samples by mass spectrometry is Data Dependent Acquisition (DDA). This semi-stochastic, identification-first paradigm generates a wealth of peptide-level data without regard to relative abundance. We introduce a data acquisition concept called Sequential Windowed Acquisition of Reporter Masses (SWARM). This approach performs quantitation-first thereby allowing subsequent acquisition decisions to be predicated on user-defined patterns of reporter ion intensities. The efficacy of this approach is validated through experiments with both synthetic mixtures of Escherichia coli ribosomes spiked into human cell lysates at known ratios, and the quantitative evaluation of the human proteome's response to the inhibition of Cullin-based protein ubiquitination via the small molecule MLN4924. We find SWARM informed PRM acquisitions display effective acquisition biasing toward analytes displaying quantitative characteristics of interest, resulting in an improvement in the detection of differentially abundant analytes. The SWARM concept provides a flexible platform for further development of new acquisition methods.

## *Introduction*

Comparative proteomic experiments have become increasingly dependent on the use of isobaric mass tags to facilitate the multiplexing of samples into single acquisitions[30]. As a bottom-up proteomic method, protein samples from multiple biological samples are digested with trypsin, and chemically tagged via reaction with either an amino or sulfhydryl reactive reagent. These tags are isobaric while intact but produce isotopically labeled reporter ions of different masses upon fragmentation. Under this regime, peptide ions are detected as intact analytes (MS1) and can be quantified after a single round of gas-phase purification through a linear ion trap or quadrupole followed by fragmentation and detection of reporter ion relative intensities (MS2). On newer instruments, multiple fragmentation products from the MS2 level may be subsequently copurified in a linear ion trap by synchronous-precursor-selection (SPS) and further fragmented to the MS3 level to produce reporter ion signals[37].

In bottom-up proteomic mass spectrometry, peptide ion targets are selected in real-time by the instrument through Data Dependent Acquisition (DDA). After intact analytes have been detected in an MS1 survey scan covering the *m/z* range of interest, putative peptide targets are ranked by signal intensity and sequentially isolated and fragmented before repeating the cycle with another MS1 survey scan. This acquisition methodology facilitates the identification of large numbers of peptides but does so in a way that is indiscriminate to quantitative trends in the isobaric tag reporter ion intensities. This results in frequent sampling of zero-fold change analytes that are likely of low biological interest. The glut of peptide and protein identifications and quantitative data can lead to thorough and deep characterization of a proteome but comes at the cost of increasing the stringency of multiple hypothesis testing correction during statistical tests for differential protein abundance[38].

To address these issues, we introduce the concept of a quantitation-first data acquisition cycle for isobaric tagged samples (**Figure 2-1**). This Data Independent Acquisition-like cycle (DIA), termed Sequential Windowed Acquisition of Reporter Masses (SWARM), leverages fragmentation of small regularly placed isolation windows tiled across the target *m/z* space. Unlike traditional DIA approaches, SWARM scans only acquire data for the reporter ion mass range, providing data about the relative abundance of isolated and fragmented analytes while lacking any identification information. These SWARM scans allow users to bias acquisition toward analytes exhibiting user-defined differences in reporter ion intensities. In this work, we utilize SWARM scanning to generate quantitative maps of isobarically tagged protein mixtures. These quantitative maps are subsequently filtered to create target lists for subsequent parallel reaction monitoring (PRM) experiments focused on the identification of the high fold change peptides.

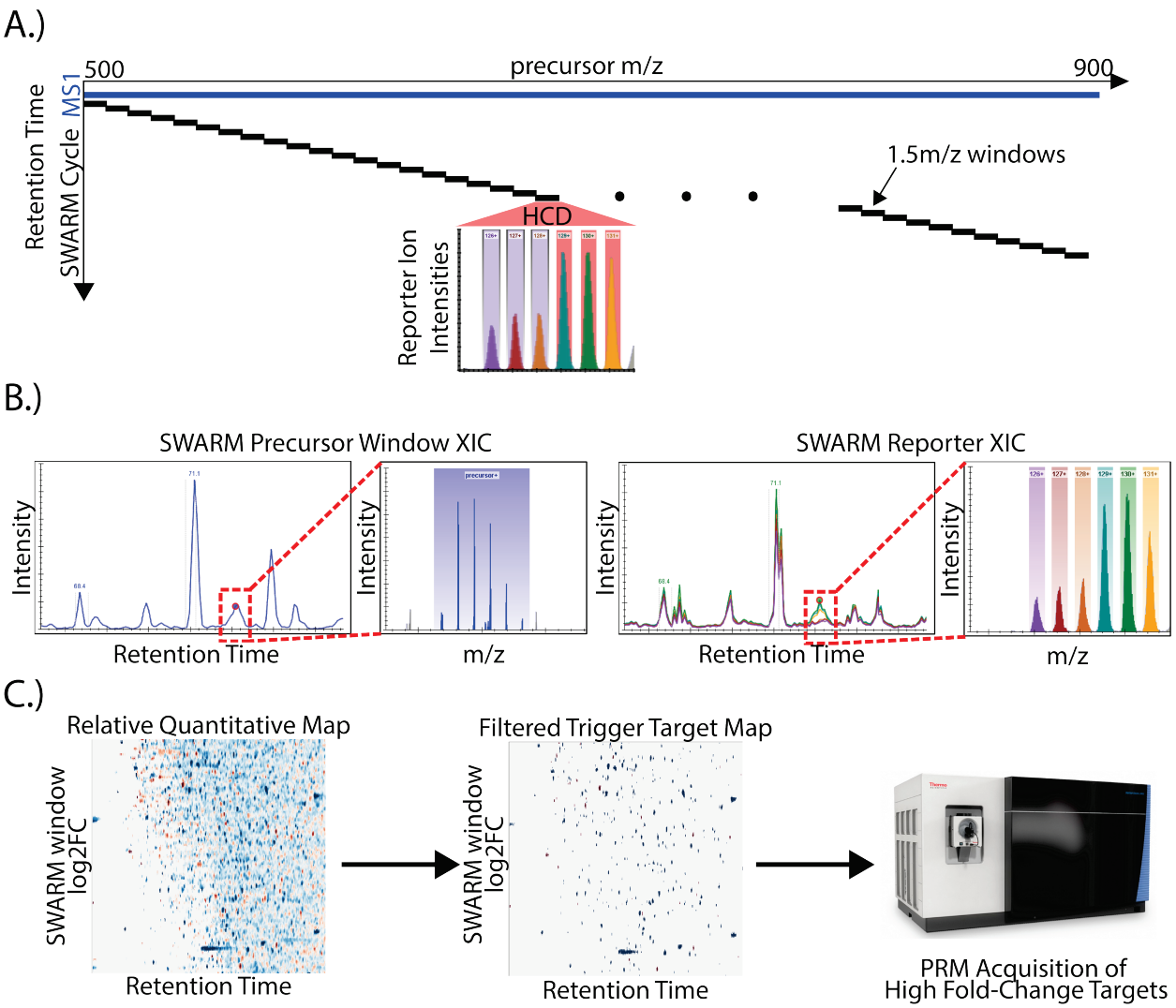


**Figure 2-1.**

A.) An overview schematic of the Sequential Windowed Acquisition of Reporter Masses (SWARM) scan cycle. This DIA-like cycle is comprised of small regularly spaced isolation windows in which HCD fragmentation is carried out and isobaric tagging reporter ions detected.

B.) An example SWARM window MS1 extracted ion chromatogram across retention time and the corresponding TMT reporter ion signals extracted from the MS2 linear ion trap scans. An example of a set of high fold change SWARM scans are highlighted.

C.) The general workflow for data acquisition under the SWARM+PRM paradigm. First, SWARM scanning is used to produce a relative fold change quantitation map of the sample. The map is filtered to retain only high fold change analytes and used to construct a set of PRM target *m/z* values and retention times. In a follow-up PRM acquisition, quantitation and identification scans are produced.

## *Experimental Procedures*

**An implementation of the SWARM scan cycle on an Orbitrap Fusion Lumos mass spectrometer:**

Intact full (MS1) scans were interlaced between SWARM (MS2) scan cycles covering from 500 *m/z* to 900 *m/z* (**Figure 2-1**). Each SWARM isolation window was purified by the quadrupole with a width of 1.5 *m/z* and placed in a DIA-like design tiled uniformly across the *m/z* space. Each cycle contained 268 windows. The ribosome experiments contained an additional SWARM scan cycle offset by a half-window width. To reduce the number of available target windows for the PRM, the data from the second SWARM windows is dropped from consideration, while the MLN4924 SWARM acquisition omits this second scan cycle entirely. SWARM scans were acquired in the linear ion trap, using a high HCD collision energy of 60%, the normal scan rate for the ribosome experiments and the rapid scan rate for the MLN4924 experiments, a 1e4 AGC target, 5ms maximum injection time, and covering only the *m/z* range 125-132. Each window was organized as a target in a tMS2 experiment, with a default *z* of 2 for collision energy calculation purposes.

All raw datasets acquired and analyzed here are available within ProteomeXchange on the MassIVE data repository, under the identifier PXD011544.

**SWARM Quantitative Map Data Processing and target selection:**

SWARM datasets were imported into a Skyline document containing small molecule targets placed at the center of each SWARM window[22]. For each of these small molecule targets representing a SWARM window, extracted ion chromatograms for each reporter mass of the Thermo TMT6Plex kit were generated with a 0.7 *m/z* extraction tolerance (±0.35 Th). Intensity

chromatograms were exported from Skyline and further processed by means of an in-house Python (v. 2.7) script. Briefly, the chromatograms were imported into a Pandas dataframe, and time indexed by the retention time of the first scan in the SWARM cycle. For each scan, SWARM reporter ion data was dropped if all reporter peak areas reported by Skyline fell below 15000. Further, individual reporter ion peak areas below 10000 were censored as NA values. The remaining reporter ion channel data was median equalized across all SWARM windows, and fold change values were calculated for every scan as the ratio of the average for the three condition replicates in each experimental condition and ultimately log2 transformed. These log2 relative fold change values were plotted across retention time as quantitative maps that enable visualization of the biological condition reporter ion fold changes throughout the sampled *m/z* space. These maps are constructed without the identities of analytes that contribute to the detected reporter ion fold changes. The maps were filtered such that only scans exhibiting fold changes that exceed a user specified threshold are retained, referred to here as the trigger condition. These high fold change SWARM scans are designated as trigger scans, as they represent *m/z*-retention time segments of the chromatographic separation from which we would desire to identify the analyte of interest. For the *E. coli* ribosome spike-in experiments, a 2-fold change minimum between the average of the triplicate conditions was designated as the trigger condition, while a 2.25-fold change threshold was used for the MLN4924 experiment. PRM target windows were generated, centered on and seeded by the triggering scan(s), and extending for 30-60 seconds depending on how many sequential scans in that window met the trigger condition.

**Testing Reporter Ion DIA/SWARM Scan Rates:**

For all scan rate testing experiments, ionization voltage was set to zero. This effectively forces scans to hit the maximum injection time limit during ion accumulation. MS2 scans were

set to only cover the region 122-135 *m/z,* and built as part of a DIA cycle with quadrupole isolation windows of 2*m/z* spanning from 500 to 900*m/z*. Under these conditions, we varied the scan resolution of the orbitrap or the scan rate of the linear ion trap. Concomitantly, we varied the maximum injection time of the scans. For each combination of scan duration/resolution and maximum injection time, three one-minute datasets were acquired. The scan rate for each run was calculated across the full run, and the triplicates averaged and plotted. A small random wiggle was added to the data to enhance visualization of overlapping datapoints.

**Nanoflow Chromatography and peptide ionization:**

Coupled with the mass spectrometric analysis, samples were separated on a 25cm C18 reversed phase column (75uM ID x 25cm), packed with 1.9uM ReproSil-Pur beads with 120A pores (Dr. Maisch GmbH). Analytical columns were connected to an Ultimate 3000 ProFlow nanoflow UHPLC (Thermo Fisher Scientific). Each acquisition was run on a 140min chromatographic gradient with water with 3% DMSO and 0.1% formic acid as buffer A, and acetonitrile with 3% DMSO and 0.1% formic acid as buffer B. Briefly, gradients began at 400nl/min flow rate, at 1% organic. Over the first 5 minutes, the flow rate drops to 200nl/min, and the percent organic increases to 5.5%B, with a Dionex curve value of 4. Until 128min, a linear gradient to 27.5%B was executed, followed by a linear increase to 35%B at 135min. Over one minute, the percent organic increases to 80%B and held for two minutes, before dropping back to 1%B over a half minute and held until the end of the chromatographic run. Ionization was carried out via electrospray ionization through a Nimbus ionization source (Phoenix S&T), with ionization voltage set to 2.2kV.

**PRM Follow-Up Data Acquisition:**

The PRM acquisitions were collected using orbitrap MS2 scans at 15,000 resolution at 400 *m/*z after CID fragmentation at 35% collision energy, with a 10ms activation time and an AGC target of 5e4 and a maximum injection time of 150ms. The data from the MS2 scans were used to target 10 fragments per scan for SPS-MS3 (Synchronous Precursor Selection) occurring from 400-1500*m/z*, and omitting targets occurring in a window from -18 *m/z* to +5 *m/z* from the PRM target/SWARM window center. MS3 scans were generated using HCD fragmentation at 60% collision energy and detecting the reporter ion intensities in the linear ion trap with a rapid scan from 125 to 132 *m/z*, an AGC target of 1e4 and 150ms maximum injection time. Quadrupole isolation window was set to 1.5 *m/z* to match the placement of the SWARM window responsible for triggering the PRM. MS1 precursor scans covering 495-905 *m/z* were acquired between each PRM cycle at a resolution of 120,000 with a maximum injection time of 50ms and an AGC target of 4e5.

**Paired DDA Data Acquisition:**

DDA acquisitions were carried out utilizing the manufacturer's default SPS-MS3 acquisition method but matched to the same precursor range (500-900 *m/z*) sampled in the SWARM acquisitions. Briefly, Orbitrap MS1 scans were acquired at 120,000 resolution in a 3 second cycle time, followed by selection of precursors in the 500-900 *m/*z space, with monoisotopic precursor selection set to peptide, an intensity threshold of 5e3, charge states from +2 through +7, and a 30 second dynamic exclusion. These scans were carried out with an AGC target of 4e5 and a maximum injection time of 50ms. Selected targets were subjected to CID fragmentation at 35% collision energy, with a 10ms activation time. The AGC target was set to 1e4, with a maximum injection time of 35ms and scanned in the linear ion trap at the turbo scan rate. Ten fragments were selected for SPS-MS3 after filtering for those from 400-1500 *m/z*, and

not occurring in a window from -18 *m/z* to +5 *m/z* from the precursor and excluding isobaric tag losses from TMT fragmentation. The quantitation scans were obtained in the orbitrap, after HCD fragmentation at 65% collision energy, at a resolution of 50,000 with a maximum injection time of 105ms and an AGC target of 1e5.

***E. coli* ribosome sample preparation:**

A single 15cm plate of HEK293 cells was harvested by scraping in cold PBS and pelleted at 800g. The cell pellet was lysed in 8M Urea with 100mM Tris-HCl at pH 8.0, with 1mM of AEBSF, DTT, leupeptin and pepstatin A. Pierce Universal Nuclease for Cell Lysis was added to the lysate at 1:10000 by volume, followed by centrifugation at 20,000g to clarify lysate and recovery of the soluble fraction. The soluble lysate was precipitated in TCA, washed in ice-cold acetone, and the precipitate resuspended in 1M Urea, 50mM HEPES pH 8.5. Protein concentration was measured by BCA, and 300ug taken for digestion and TMT labeling, described below.

The protein content was split equally between all six channel labels in the 6-plex kit, yielding a constant protein background. Separately, 120ug of purified *E. coli* ribosome (New England Biosciences, cat #P0763S) was suspended in 1M Urea, 50mM HEPES and digested and TMT labeled in the same manner described. The final mixing of each TMT labeled sample allowed for defined ratios to be controlled for the *E. coli* ribosomes. Each TMT channel contained 8.3ug HEK293 lysate, while the *E. coli* ribosomes were added in at either a 1:1:1:5:5:5 ratio (.2ug or 1ug per channel) or 1:1:1:10:10:10 ratio (.1ug or 1ug per channel). After the final desalting, the sample was resuspended in 50uL, and 1uL used for each injection.

**MLN4924 sample preparation:**

Six 10cm plates of HeLa cells were cultured, with three of them treated for 24 hours before harvest with 1uM final concentration of MLN4924 (Chemietek, cat #CT-M4924), and the remaining three treated for an equivalent time with a DMSO vehicle control. Cells were harvested by scraping in cold PBS and pelleted at 800g. Each cell pellet was lysed in 8M Urea, 100mM Tris-HCl at pH 8.0, with 1mM of AEBSF, DTT, leupeptin and pepstatin A and 1:10000 Universal Nuclease (Pierce). After 30 minutes of rotation, samples were centrifuged at max speed to remove the cellular debris. Proteins were precipitated in 20% TCA followed by 3 cold acetone washes. Samples were resuspended in 1M urea, 50mM HEPES and protein content quantified and normalized by BCA, with 50ug from each sample taken for digestion and TMT labeling, as described below.

**Proteomic Sample Digestion and TMT labeling:**

For each sample, protein was reduced via addition of TCEP to a final concentration of 5mM and incubated for 20 minutes at room temperature. Iodoacetamide, to alkylate cysteine residues, was added to 10mM, followed by an additional 20-minute incubation in the dark. Endopeptidase Lys-C was added to each sample at a 1:100 ratio of enzyme to substrate, followed by a 4-hour incubation at 37°C in the dark. After the incubation, a 100mM stock of $CaCl_2$ was used to bring the solution's working concentration of $CaCl_2$ to 1mM. Trypsin protease was added to each sample at a 1:20 enzyme to substrate ratio, and the samples incubated overnight at 37°C in the dark. Finally, digestion was quenched by the addition of formic acid to a final concentration of 5%, and each sample centrifuged at maximum speed for 5 minutes to pellet any insoluble components, and the soluble fraction moved into new tubes. Samples were desalted by binding to C18 tips, washed twice in 200uL of 5% formic acid, and eluted in 60% ACN with 5% formic acid.

Desalted, digested samples were dried by SpeedVac, before being resuspended in TMT labeling buffer of 200mM HEPES pH 8.5 and 30% ACN. Samples were briefly sonicated, while isobaric labels from Thermo's TMTsixplex™ kit (Cat. #90066) were brought to room temperature and resuspended in 41uL of anhydrous ACN per 800ug of TMT label. 5uL of the appropriate TMT label was added to each sample, mixed thoroughly, and incubated at room temperature for 1 hour. The labeling reaction was quenched with 6uL of 5% hydroxylamine, 200mM HEPES per 50ug of protein in each sample. After 15 minutes at room temperature, formic acid was added to a final concentration of 5%. Each TMT labeled sample was kept separately prior to appropriate mixing for its experiment, and a final round of C18 tip desalting, as described above. The final labeled, mixed, and desalted samples were resuspended in 5% formic acid prior to chromatographic and mass spectrometric analysis.

**Proteomic Database Search:**

All DDA and PRM data were searched via the Andromeda search engine provided within the MaxQuant platform (v 1.6.2.10). For each analysis, database searching was performed against the EMBL-EBI Human Reference Proteome, with the *E. coli* reference proteome additionally appended when appropriate. For data acquired using linear ion trap MS2 scans, fragment mass tolerance was set to 0.5Da, while Orbitrap MS2 scans were searched with 20ppm fragment mass tolerances. DDA acquisitions were searched with a first search tolerance of 20ppm and a main search tolerance of 4.5ppm. The PRM searches were allowed a 6Da window for both searches, as their scans are targeted to the center of the triggering SWARM window and therefore do not provide accurate precursor *m/z*. Carbamidomethylation was specified as a fixed modification, and Trypsin/P digestion of the database specified allowing for a maximum of 2 missed cleavage sites

and using the "specific" digestion mode. Search results were filtered by MaxQuant's provided reversed-sequence target-decoy FDR calculation at both a PSM and Protein level FDR of 1%[21,39].

**Reporter Ion Quantification:**

For each experiment, quantitative scans were filtered to require precursor isolation specificities of at least 70%. TMT reporter ion signals measured in the linear ion trap were extracted with a ±0.35*m/z* tolerance, or ±.0035*m/z* when measured in the Orbitrap. Reporter ion intensities for the DDA experiments, which utilize the orbitrap for MS3 quantitative scans, were generated by MaxQuant and read from the evidence table output of each runs' respective analysis. For the PRM experiments, we extracted reporter ion intensities directly from the vendor '.raw' files for every MS3 scan, and associated those values with the identifications made available in the MaxQuant evidence tables for each identified MS2 scan. Direct scan data and metadata access from the vendor data was provided by the RawDiag package (v 0.0.10) in R (v 3.5.0)[40]. Reporter ion intensities were calculated as the sum of the signal detectable within the extracted tolerance window relative to the position of the monoisotopic mass of each reporter ion.

When reporter ion intensities were extracted through RawDiag, we also calculated the precursor isolation purity for the associated MS2 scan from the appropriate MS1 master scan. To calculate precursor isolation purity, high resolution MS1 scans were converted from vendor format into mzML via msconvert (v. 3.0.11252) with vendor peak picking turned on[41]. The resulting mzML files were read into R via the MSnbase package (v. 2.6.2)[42]. For each PSM reported by MaxQuant, signal contained within the relevant MS2 scan's precursor isolation window was extracted with a 15ppm tolerance around the monoisotopic mass of the identified peptide and from any C13 isotope that would fall within the isolation window at the given charge state of the analyte.

This extracted signal was divided by the sum of all signal within the extraction window to represent the precursor isolation purity. The R script responsible for the TMT signal extraction and integration with MaxQuant results for the SWARM+PRM samples has been included as part of the supplementary materials.

**Statistical Analysis of TMT Intensities:**

Reporter ion intensity outputs were used to provide relative quantitative values to the R package MSstatsTMT (v. 0.4.3), after normalizing each run's TMT channels by equalization of median intensities calculated from the subset of all human derived PSMs. Protein intensities for each experiment and channel were calculated via log transformation of the summation of TMT reporter ion intensities across the protein's filtered PSMs. The requirement of a minimum of two uniquely mapping quantitative features for a protein to be carried into the comparative analysis was imposed through MSstatsTMT's "removeProtein_with1Feature" and "useUniquePeptide" options. Fold changes were calculated between biological conditions, and p-values for differential abundance testing generated by means of the *t*-test implementation provided. The Benjamini-Hochberg method was employed to adjust p-values for multiple hypothesis testing. For the MLN4924 experiments, Cullin-associated genes were annotated by a list of 549 Cullin-Associated genes aggregated by a prior study, after mapping to 551 Uniprot Swissprot-reviewed protein accessions[43,44].

## *Results and Discussion*

For each sample, SWARM scans were acquired from 500-900 m/z in 1.5 m/z windows (**Figure 2-1A**). These scans were collected in the linear ion trap, where we can leverage the benefit of a small ion capacity of the trap for rapid fill times and simultaneously exploit the linear ion

trap's scan duration proportionality to the width of the *m/z* range of the scan. Under these conditions, the linear ion trap is capable of scanning in excess of 65Hz on a DIA cycle when scans are limited to the TMT reporter m/z range while the Orbitrap scans at a rate of approximately 37.5Hz (**Figure 2-6 and 2-7**). The SWARM data was imported into Skyline, and TMT reporter extracted ion chromatograms generated for each precursor window (**Figure 2-1B**)[45]. Reporter ion channels were filtered to remove low-intensity regions of the chromatography, and log transformed fold change ratios were calculated for the two conditions. The reporter ion relative intensities and sample condition fold change values that are calculated for each SWARM precursor window and resampled during each SWARM cycle represent a relative quantitative map of the sample. After applying fold change filters between the sample condition triplicate sets, we are able to generate a map of only the high fold change scans (Trigger Map), which seed target *m/z* and retention times for follow-up PRM acquisition (**Figure 2-1C**). Each sample was additionally acquired by DDA, utilizing the vendor default SPS-MS3 quantitative method for TMT samples (Experimental Procedures).

We first tested the SWARM methodology using a model protein mixture in which equal amounts of HEK293 whole cell lysate were labeled across all six channels using 6-plex TMT isobaric tags. Purified *E. coli* ribosomes were labeled and spiked into each channel at ratios of either 1:1:1:5:5:5 or 1:1:1:10:10:10. These samples were acquired by SWARM and a follow-up PRM (SWARM+PRM) targeting windows with at least two-fold change with SPS-MS3 for quantitation (**Figure 2-2**, **Figure 2-8**). The datasets were searched using the Andromeda search algorithm, as part of MaxQuant (v. 1.6.2.10) against the EMBL Human and *Escherichia coli* reference proteome databases concatenated together[21,39]. Precursor mass tolerance for DDA searches was set to 20 ppm, while a 6 Da window was allowed for the PRM scans which contain

no assigned target precursor to provide an accurate *m/z*. This necessity stems from the targeting of the PRM isolation windows at the same *m/z* of the SWARM window which produced the triggering scan(s) instead of targeting the accurate mass of a single precursor as in DDA.

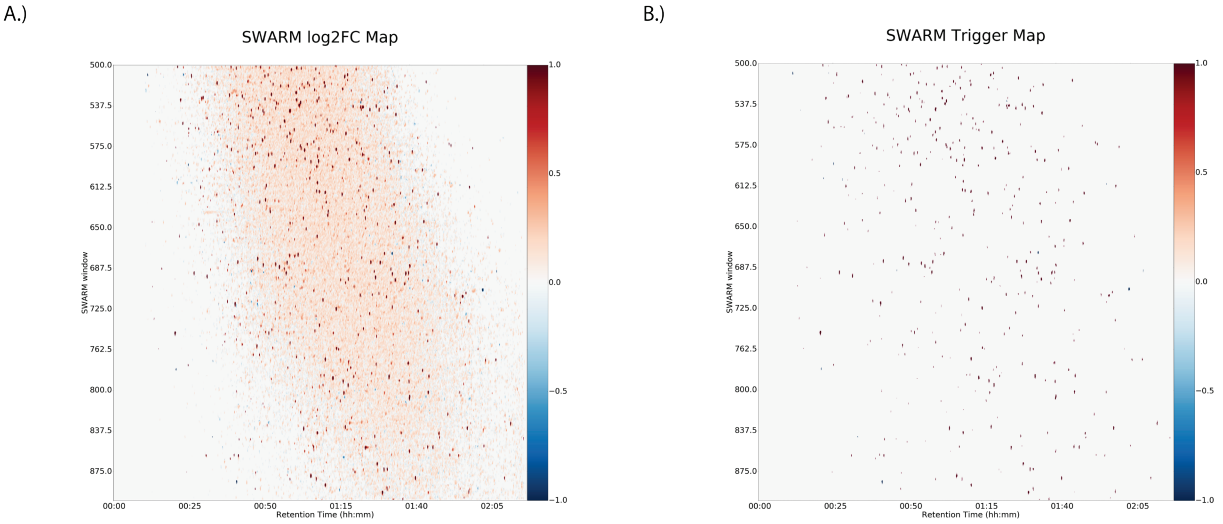


**Figure 2-2.**

A.) The SWARM quantitative map for the *E. coli* ribosome 5:1 sample shows the log2(FC) between the two conditions across chromatographic time for each SWARM window's *m/z*.

B.) The quantitative map was filtered for a minimum 2-fold change to produce the trigger map. From these scan events, PRMs were scheduled for targeted reacquisition.

As expected, the PRM runs identify many fewer confident peptide identifications when compared to the DDA acquisitions (**Figure 2-3A, Supporting Information Table 2-1**). However, we observed more similar counts of confident *E. coli* peptide identifications between the PRM and DDA runs, with slightly higher counts in the DDA runs (**Figure 2-3B).** Concomitantly, we observe a dramatic increase in the number of *E. coli* peptide-spectrum matches (PSMs) in the PRM datasets corresponding to ribosome-derived peptides (**Figure 2-3C**).

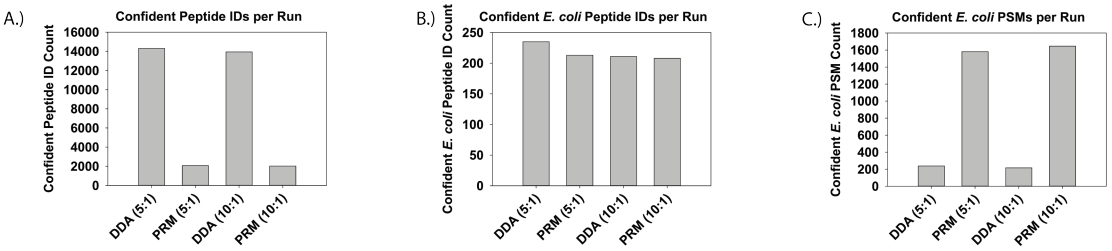


**Figure 2-3.**

A.) Peptide counts passing the 1% FDR filter for the *E. coli* ribosome spike-in experiments.
B.) Confident *E. coli* peptide counts passing the FDR filter.
C.) Confident *E. coli* PSM counts passing the FDR filter.

This increase in the number of PSMs comes with the benefit of increasing the quantitative sampling of these *E. coli* peptides. At the protein level, the wealth of data produced by unbiased data dependent sampling can come at the cost of harsh correction when adjusting for multiple hypothesis testing. Proteins with an adjusted p-value less than or equal to 0.05 with a minimum 2-fold change were considered significantly differentially abundant. The focused acquisition of the SWARM+PRM approach yields more sensitive comparisons, producing a list of 34 significantly enriched *E. coli* proteins in the 5:1 ribosome sample (**Figure 2-4A, Supporting Information Table 2-2**) while no *E. coli* proteins were deemed significantly enriched in the DDA acquisition under these conditions (**Figure 2-4B, Supporting Information Table 2-3**). At more extreme fold changes, the advantage of SWARM+PRM is less pronounced although still clearly evident. In the 10:1 ratio ribosome sample, SWARM+PRM is able to detect 32 *E. coli* proteins as differentially enriched while DDA offers a close 27 significant proteins (**Figure 2-4C-D, Supporting Information Tables 2-4 and 2-5)**. In these synthetic samples, SWARM+PRM provides effective biasing of instrument acquisition focus toward *E. coli* ribosomes, substantially increasing *E. coli* PSMs while maintaining controlled false positive rates during differential abundance testing.

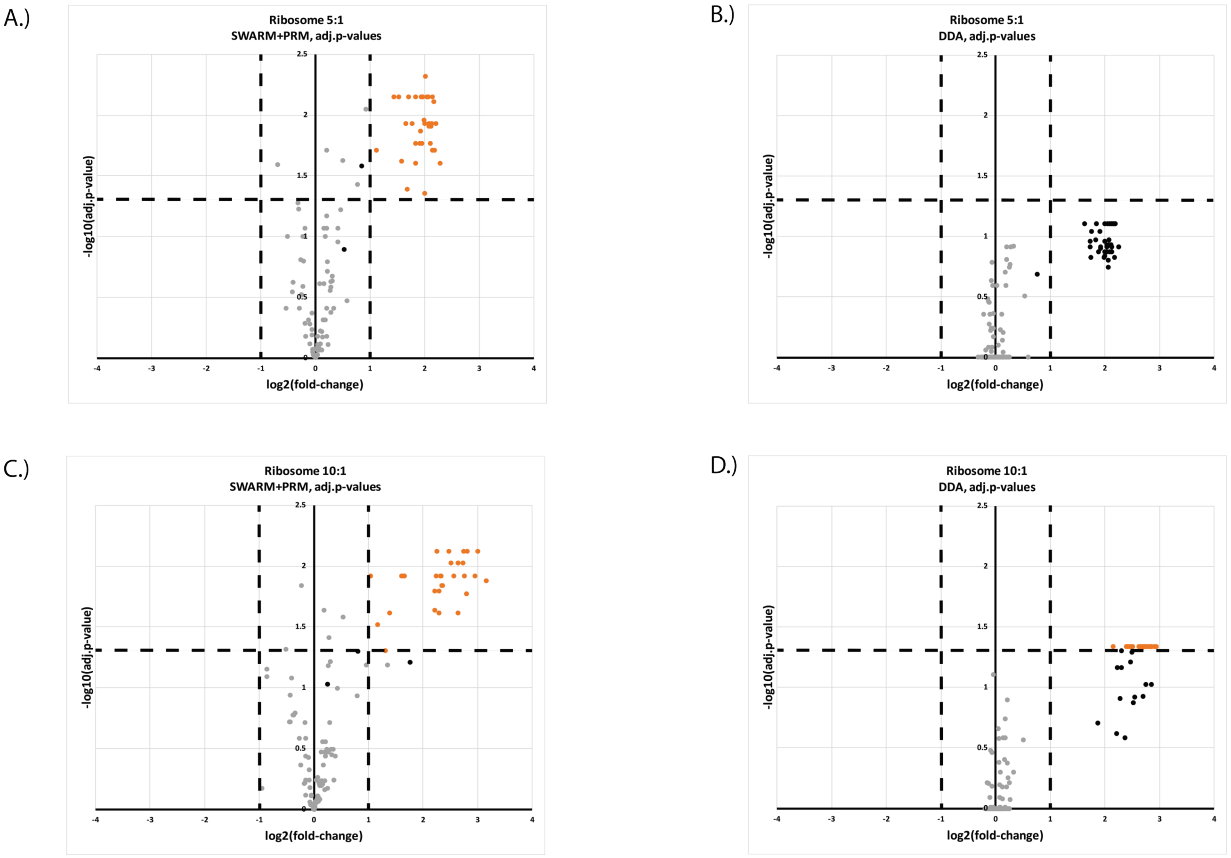


**Figure 2-4.**

A.) Volcano plot for the SWARM+PRM acquisition of the *E. coli* 5:1 spike-in sample. A total of 33 proteins were significantly differentially enriched (adj. p-value <=0.05) above a 2-fold change threshold in either direction. True positive, high fold change, significant *E. coli* proteins are shown in orange. False negative *E. coli* proteins are denoted in black, while true negative human proteins are gray.

B.) Volcano plot for the DDA acquisition of the *E. coli* 5:1 sample. No proteins were deemed significantly differentially under our criteria.

C.) Volcano plot for the SWARM+PRM acquisition of the *E. coli* 10:1 sample. 32 *E. coli* proteins were deemed to be significantly enriched in this analysis.

D.) Volcano plot for the DDA acquisition of the *E. coli* 10:1 sample. 27 *E. coli* proteins were deemed to be significantly enriched.

We then applied SWARM+PRM toward surveying proteins whose abundance is altered upon the reduction of Cullin RING Ligase (CRL) activity via inhibition of the Nedd8-activating enzyme (NAE) via the small molecule MLN4924[46]. Proteins dysregulated upon CRL inhibition would be candidates for association with CRL E3 ubiquitin ligase complexes. Triplicate cultures

of HeLa cells were treated with either 1uM MLN4924 or DMSO for 24 hours. The six cultures were lysed, and the soluble protein content of the whole cell lysate digested and labeled with TMT 6-plex isobaric tags. These samples were analyzed using SWARM+PRM design similar to the ribosome samples, but targeting SWARM windows at retention times exhibiting at least 2.25-fold change in either direction between the DMSO and MLN4924 treated triplicate channels (**Figure 2-9A-B**). A corresponding DDA acquisition was also acquired under the same parameters as in the ribosome experiments. Quantified proteins were considered significantly differentially abundant at a maximum adjusted p-value of 0.05, and a minimum fold change of 1.75 in either direction. We additionally compared the identified proteins to a list of previously annotated Cullin-associated genes. The SWARM+PRM results showed an improvement in the detection of differentially abundant proteins after MLN4924 treatment compared to the DDA acquisition (**Figure 2-5A-B**)[43]. We identified a population of 71 proteins meeting these thresholds from the SWARM+PRM experiment, where no proteins were deemed significant from the analysis of the DDA data (**Supporting Information Tables 2-6 and 2-7**). Of the proteins classified as significantly differentially abundant, 3 were previously annotated as being Cullin-associated, including the well characterized Cullin target β-catenin[47].

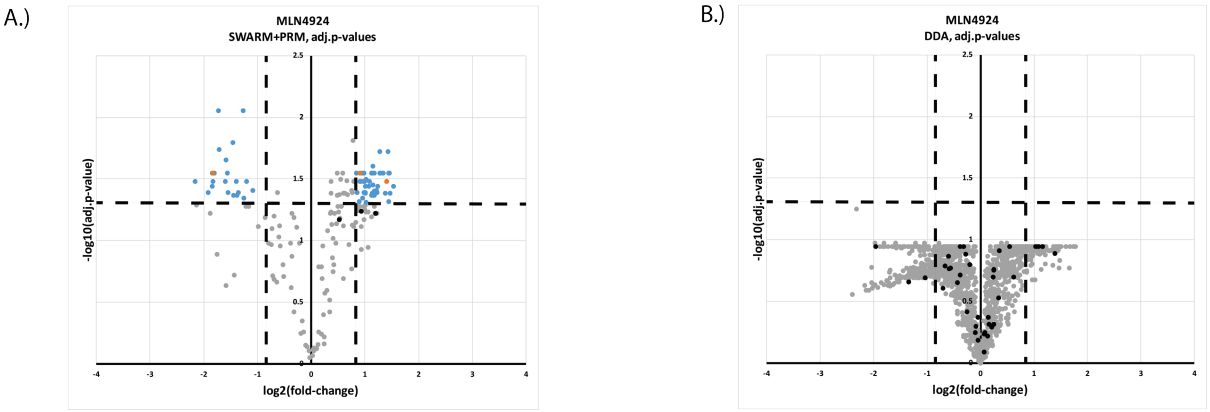


**Figure 2-5.**

A.) Volcano plot for the SWARM+PRM acquisition of the MLN4924 sample. 71 Proteins remained significant below an adjusted p-value of 0.05 and above 1.75-fold change. Three Cullin associated proteins which are enriched upon MLN4924 treatment are included in this population. High fold change, significant Cullin-associated proteins are shown in orange and significant non-associated proteins in blue. Non-significant Cullin-associated proteins are denoted in black, while non-significant non- associated proteins are gray.

B.) Volcano plot for the DDA acquisition of the MLN4924 sample. No proteins were deemed significantly differentially regulated under our criteria.

## *Conclusions*

We have explored the feasibility and utility of the inchoate implementation of the SWARM cycle, a novel approach to data acquisition for samples tagged by isobaric mass tags. This quantitation-first approach allows for acquisition behaviors to be predicated on the quantitative results of the SWARM scans. These experiments demonstrate the utility of the SWARM scan cycle as an effective way to bias acquisition toward higher fold change targets in subsequent PRM runs. By focusing acquisition, the SWARM methodology can impactfully reduce the burden of multiple hypothesis testing correction as we demonstrated using complex mixtures of ribosomes spiked into whole cell lysates (**Figure 2-4**) and in cultured cells treated with the small molecule MLN4924 (**Figure 2-5**). This increased capacity to detect statistically significant differences between the composition of protein samples is critical for maximizing the biological meaningful conclusions that can be gleaned from these datasets. Further, we expect that this benefit to multiple

hypothesis testing correction will be particularly relevant to peptide-level quantitation where the number of hypotheses tested scales with the number of identified peptides

The initial implementation of SWARM necessitates use of the lower resolution linear ion trap for reporter ion scanning in order to facilitate sufficiently small and numerous quadrupole isolation windows at a duty cycle compatible with modern nanoflow chromatographic separations of chemically complex mixtures. This mass analyzer choice effectively constrains the approach to isobaric tags with at least unit mass separations. We hope that newer generations of high-resolution mass analyzers will enable speeds and sensitivities capable of SWARM scanning higher multiplex isobaric tag kits with sub-integer mass separations.

While the current incarnation of our approach necessitates a second acquisition, and therefore highly reproducible chromatography, we envision that a major advantage of future implementations of SWARM will be the ability to incorporate real-time decision making into trigger scan events in order to enable a wide range of flexible acquisition strategies. Initially, this would likely include the capacity to trigger a data-dependent MS2 scan on any SWARM reporter scan with a fold-change exceeding a user-defined threshold which would remove the necessity to perform follow-up PRM acquisitions as is done in the current workflow. Longer term implementations could include real-time decisions based on thermal stability performed in the context of thermal proteome profiling experiment or changes in signaling kinetics measured in phosphopeptide profiling analyses. We expect the adaptability of SWARM to different proteomic methodologies will be a key asset in its future development.

**Supplementary Figures and Tables**

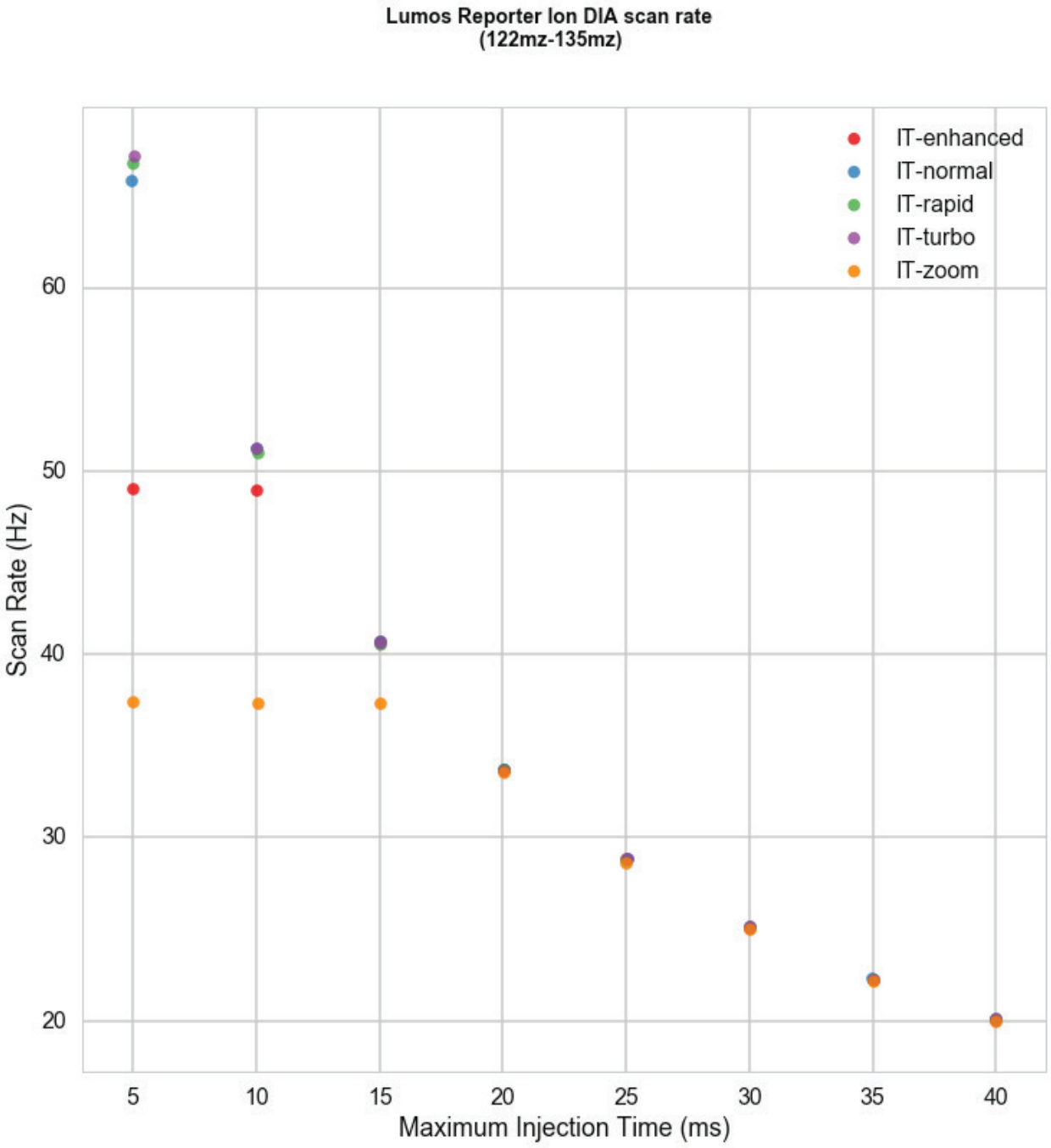


**Figure 2-6.** DIA scan rate measurements for various linear ion trap scan modes and maximum injection times for scans. Injection times ranged from 5ms to 40ms in 5ms increments for zoom, enhanced, normal, rapid, and turbo scan modes. A small amount of wiggle was added to the plot data to facilitate visualization of overlapping datapoints.

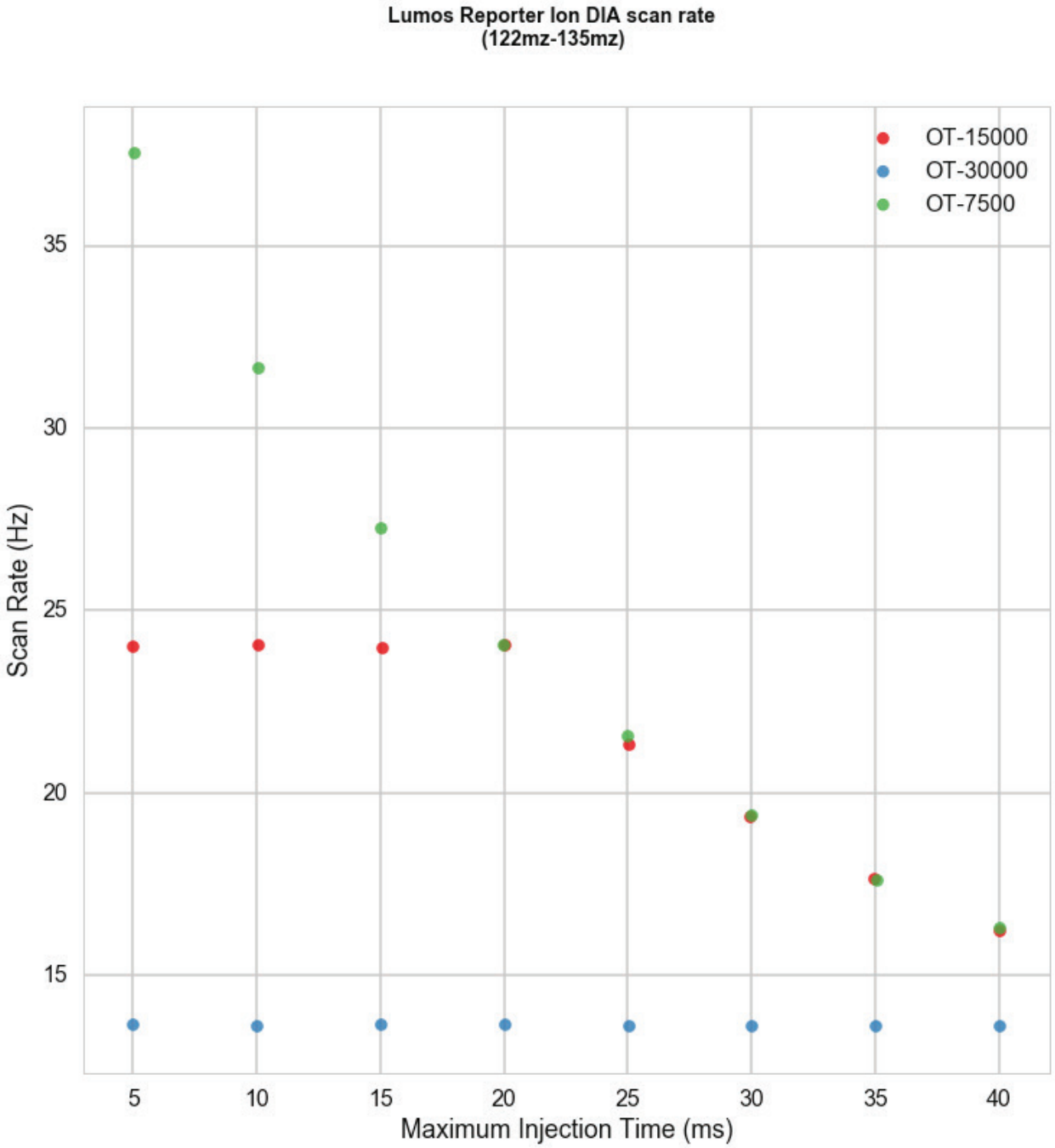


**Figure 2-7.** DIA scan rate measurements for various linear Orbitrap scan modes and maximum injection times. Injection times ranged from 5ms to 40ms in 5ms increments for the 7500, 15000, and 30000 resolution Orbitrap scans. A small amount of wiggle was added to the plot data to facilitate visualization of overlapping datapoints.

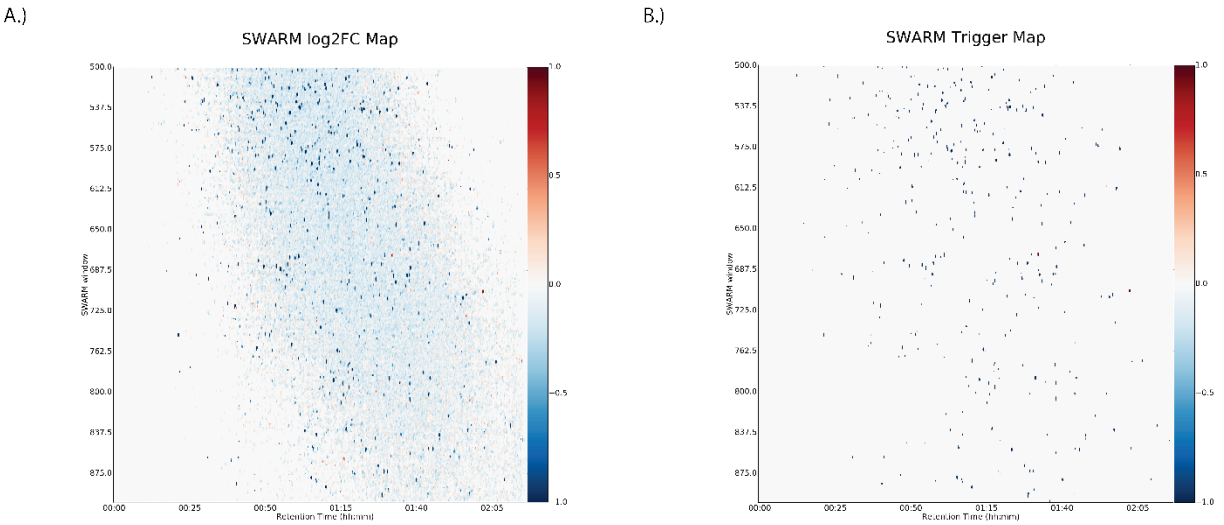


**Figure 2-8.**

A.) The SWARM quantitative map for the *E. coli* Ribosome 10:1 sample shows the log2(FC) between the two labeled conditions. The color-direction of the fold change for this plot was inverted when compared to the 5:1 trigger map, but does not represent a change in the label positions.

B.) The quantitative map was filtered for a minimum fold change of 2 to produce the trigger map. From these scan events, PRMs were scheduled for targeted reacquisition.

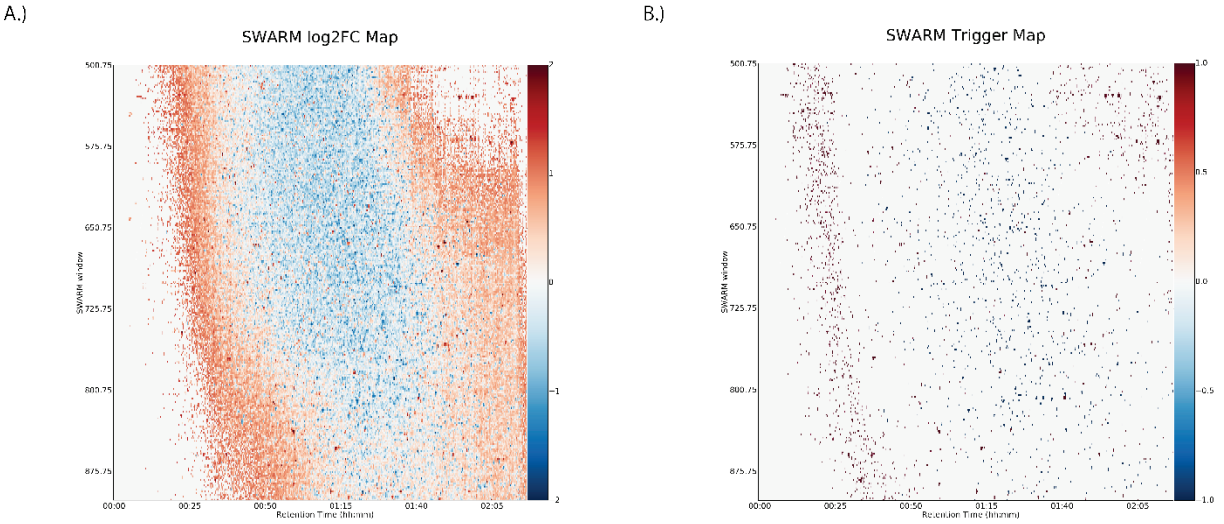


**Figure 2-9.**

A.) The SWARM quantitative map for the MLN4924 experiment displays the log2(FC) difference between the treated and untreated cells. Red indicates a relative increase in abundance, while blue indicates a decrease in abundance upon exposure to MLN4924.

B.) The quantitative map was filtered for a minimum fold change of 2.25 to produce the trigger map. From these scan events, PRMs were scheduled for targeted reacquisition.

# Chapter 3: Chemical Derivatization of Streptavidin Provides Protection from Tryptic Proteolysis

## *Chemical Derivatization of Streptavidin Provides Protection from Tryptic Proteolysis*

**William D. Barshop, Hee Jong Kim, Xiaorui Fan, Shima Rayatpisheh, James A. Wohlschlegel**

## *Abstract*

The enrichment of biotinylated proteins using immobilized streptavidin has become a staple methodology for affinity purification based proteomics. Many of these workflows rely upon tryptic digestion to elute streptavidin-captured moieties from the beads. The concurrent release of high amounts of streptavidin-derived peptides into the digested sample, however, can significantly hamper the effectiveness of downstream proteomic analyses by increasing the complexity and dynamic range of the mixture. Here, we describe a strategy for the chemical derivatization of streptavidin that renders it largely resistant to proteolysis by trypsin and thereby dramatically reduces the amount of streptavidin contamination in the sample. This rapid and robust approach improves the effectiveness of mass spectrometry-based characterization of streptavidin-purified samples making it broadly useful for a wide variety of applications. In addition, we show that this chemical protection strategy can also be applied to other affinity matrices including immobilized antibodies against HA epitope tags.

## *Introduction*

The ability to enrich a specific protein or class of peptides or proteins using affinity-based purification techniques is the foundation for of a wide range of biochemical methods. The subsequent characterization of these affinity-purified mixtures is often done using proteomic mass spectrometry which has the capacity to elucidate the composition, abundance, and post-translational modification state of the sample in a largely unbiased manner. Although these workflows are well-established in the field, these enrichment methods still face technical challenges that can limit their overall effectiveness. For example, the salt content, surfactant, or solvent composition required for elution from specific affinity matrices may be incompatible with mass spectrometry necessitating further clean up of the sample[36]. Similarly, the elution from certain affinity supports may be inefficient or compromised by the co-elution of contaminants that interfere with the analysis. This is commonly the case when biotinylated proteins are isolated from biological mixtures using immobilized streptavidin[48]. The extremely high affinity of the biotin-streptavidin interaction prevents facile elution of the proteins of interest and requires either extremely harsh chemical conditions or more commonly the use of trypsin to digest the proteins directly from the beads[49]. Although effective for elution, this second option releases high amounts of streptavidin-derived peptides into the sample upon tryptic digestion which can compromise the overall effectiveness of the downstream analysis.

To document these technical limitations, we demonstrate that the high levels of streptavidin-derived peptides present in a typical on-bead digestion of streptavidin-bound samples reduces overall peptide identification rates in the region of the chromatography corresponding to their elution. In addition, the elution of these abundant streptavidin-derived peptides leads to local chromatographic disturbances that results in both ion suppression and retention time shifts for co-

eluting peptides of interest. To overcome these challenges, we have developed a novel strategy for the chemical derivatization of streptavidin which renders it largely resistant to trypsinization without affecting its biotin binding character. We show that the use of these derivatized streptavidin beads in standard proteomics workflows prevents the reduction in peptide identification rates and chromatographic shifts observed in purifications using underivatized streptavidin beads. In addition, we show that this chemical derivatization strategy can also limit digestion of antibody-based supports without interfering with target binding using immobilized α-HA antibody as an example. Together, these data suggest that this strategy is robust, generalizable, and has the capacity to improve the effectiveness of a wide range of proteomic workflows.

## *Materials and Methods*

### Plasmids and Cell Culture:

Gene sequences of interest were amplified from purchased plasmids with primers integrating AttB1/2 sites using the Phusion TaqDNA polymerase (New England Biolabs) as described previously[50]. Amplified gene sequences were cloned into the pDONR221 vector via the Gateway cloning system (Invitrogen). Genes were recombined into destination vectors encoding the sequences for affinity purification tags containing either 3×HA-3×FLAG or BioID-FLAG sequences. Plasmids for MMS19 was acquired previously, while plasmids for PCNA, CIAPIN1, and BOLA2 were purchased from Dharmacon[50]. MMS19 and PCNA were used to generate BioID and BioID2 fusion products, respectively. CIAPIN1 and BOLA2 were tagged with the 3×HA-3×FLAG tag sequence. HEK293 Flp-In T-Rex cells with stable, doxycycline inducible integrands of the various gene fusions mentioned above were cultured in a mixture of Dulbecco's Modified Eagle's Medium (DMEM) with 10% Fetal Bovine Serum (FBS), and 2mM glutamine, into which

antibiotic-antimycotic (Gibco™ 15240062) was added. Cells were cultured at 37°C in 5% $CO_2$. Induction for expression was carried out by the addition of 500ng/mL of doxycycline into the cell culture media for 24 hours prior to induction. For BioID experiments, the cells were additionally cultured for the duration of induction in the presence of a final concentration of 50µM biotin. Cells were harvested by scraping, and the pellets washed 3 times in 50mL PBS with spins at 800g to pellet in between washes. The cell pellets were snap frozen and stored at -80°C until further use.

**Reductive Methylation of Affinity Purification Matrices:**

Pierce™ High capacity streptavidin agarose (20359) or Pierce™ α-hemagglutinnin agarose (α-HA, 26181) was reductively methylated using the Hampton Research Reductive Alkylation Kit (HR2-434). Briefly, 1mL of bead slurry was washed and equilibrated 5 times with 1mL of Phosphate Buffered Saline (PBS, Gibco™ 10010023) on ice. After the final wash, the beads were resuspended in 1mL cold PBS and 20µL of 1M dimethylamine borane complex and 40µL of 1M formaldehyde were added. The beads were placed on a laboratory rotator for 2 hours at 4°C. The addition of dimethylamine borane complex and formaldehyde was repeated, and the beads left for an additional 2-hour incubation on rotation at 4°C. A final addition of 10µL of 1M dimethylamine borane complex was carried out, and the beads left to rotate overnight. Finally, the reaction was quenched with the addition of 125µL of 1M glycine (pH 8.6) and 125µL of 50mM dithiothreitol along with a final 2-hour incubation and rotation. The derivatized beads were washed 10 times with 1mL of PBS and finally resuspended in PBS to a final combined slurry volume of 1mL and stored at 4°C.

**Methylglyoxal Derivatization of Affinity Purification Matrices:**

Affinity purification bead slurries were resuspended, and 1mL taken, and washed 5 times with 1mL PBS on ice. The beads were exchanged into 1mL of 100mM methylglyoxal (Sigma Aldrich, M0252) in PBS and placed on rotation at 37°C. After 24 hours, the derivatized beads were washed 10x in ice cold PBS and stored at 4°C. For beads modified by both reductive methylation and methylglyoxal derivatization, the reductive methylation was invariably performed first.

**Streptavidin-Biotin Binding Colorimetric Assay:**

Derivatized streptavidins were interrogated to determine their biotin binding capabilities by colorimetric assay to determine biotinylated-HRP retention on the beads. For each of the four relevant bead types, 800µL of bead slurry was washed with 1mL of PBS with the beads placed on a laboratory rotator for 3 minutes between each wash. 50µL bead slurry aliquots were moved into separate Eppendorf tubes for each bead type, in duplicates, for each of 5 steps of a 10-fold dilution series. Biotinylated HRP (Pierce™ 29139) was introduced to each aliquot of beads at 1ng, 10ng, 100ng, 1ug, or 10ug at a fixed volume of 200µL and placed on rotation at room temperature for 30 minutes. The beads were washed 5 times with 1 mL PBS, allowing for 5 minutes on rotation between washes. Peroxidase activity was measured using the colorimetric Slow TMB ELISA (Thermo Scientific, 34024) substrate solution at 450nm according to the manufacturer's directions on a Thermo Scientific NanoDrop 2000 spectrophotometer.

**BioID Sample Lysis and Streptavidin Affinity Purification:**

BioID-fusion protein expressing cell pellets were lysed in the pellet's volume equivalent of 8M Urea, 100mM Tris pH 8.0 and thoroughly mixed at room temperature. After complete resuspension, 1µL of Benzonase nuclease was added to reduce sample viscosity via degradation of nucleic acids. Samples were placed on rotation for 30 minutes at room temperature, and spun at

20,000rcf for 15 minutes to pellet any insoluble debris. The soluble fraction for each sample was taken and normalized for protein quantity by a BCA assay.

After normalization, 125µL of each relevant streptavidin bead slurry was equilibrated in the urea lysis buffer via 3x 1mL washes. For each wash, the previous buffer was removed and replaced, and the beads placed on rotation at room temperature for 5 minutes before the beads were pelleted by a slow centrifugation at 31rcf. Normalized lysates were split equally between each of the modified bead types and left on rotation for 30 minutes at room temperature. Samples were centrifuged and washed 5x with 1mL urea lysis buffer, in a similar method to the slurry equilibration. Finally, all liquid was removed from the beads using narrow bore gel-loading tips (Eppendorf, 022351656) and replaced with 50µL of the urea lysis buffer for digestion.

**HA Tagged Sample Lysis and Immunoprecipitation:**

HA-tagged fusion protein expressing cell pellets were lysed in native lysis buffer containing 2mL 100mM Tris-HCl pH 8.0, 150 mM NaCl, 5% Glycerol, 0.1% NP-40, 1µM leupeptin, 1µM pepstatin, and 1µM AEBSF. Into each sample, 1µL of Benzonase nuclease was added and each sample placed on rotation for 30 minutes at 4°C, and clarified by a 15-minute spin at 15,000rcf with retention of only the soluble supernatant. Sample content was normalized via measurement of absorption at 280nm on a Thermo Scientific NanoDrop 2000.

For each sample, 100µL of bead slurry was equilibrated with three buffer exchanges of native lysis buffer, in the same manner as the streptavidin affinity purification beads. Normalized protein extracts were split between each derivatized bead type and bound during a 2-hour rotation at 4°C. Beads were washed thrice with buffer exchanges of 1mL native lysis buffer, and a final wash of native lysis buffer lacking protease inhibitors. After the final wash, all liquid was removed

from the beads with Eppendorf gel-loading tips, and the beads resuspended in 50µL of 8M Urea, 100mM Tris pH 8.0.

**Sample Digestion and Desalting:**

Each sample was reduced and cysteines alkylated via addition of 1.25µL of 200mM TCEP and 1.2µL of 500mM iodoacetamide prior to a 20-minute dark incubation while shaking at 1300rpm at room temperature. 2.5µL of 0.1µg/µL endoproteinase Lys-C (Wako Chemicals, 125-05061) was added to each sample and allowed to continue to shake in the dark for 4 hours at 37°C. Urea content of each sample was reduced from 8M to 2M via the addition of 150µL of 100mM Tris-HCl pH 8.5, and the addition of 2µL of 100mM $CaCl_2$. Trypsinization was carried out with the addition of 4µL of 0.4µg/µL Trypsin per immunoprecipitation, and incubated, shaking, in the dark, at 37°C overnight. Digestion was quenched via the addition of formic acid to bring the final concentration to 5% by volume. Each digestion was desalted via binding to C18 desalting tips, washing twice with 200uL of 5% formic acid, and elution in 50µL of 60% acetonitrile with 5% formic acid. Eluates were dried via SpeedVac and resuspended in 15µL of 5% formic acid prior to chromatographic separation and mass spectrometric acquisition.

**Streptavidin LC-MS Acquisition:**

Samples generated for the streptavidin affinity purification bead comparison were interrogated by Data Dependent Acquisition (DDA) on a Thermo Q-Exactive classic instrument. Mass spectrometric acquisition was coupled to a nanoflow liquid chromatographic separation delivered by a Thermo easy nLC-1000 over a 30-minute gradient on a 100uM ID, 12cm column home-packed with 1.9µM C18 particles (Dr. Maisch GmbH). For buffer A, water with 0.1% formic acid while buffer B contained acetonitrile with 0.1% formic acid. To both buffer A and B,

3% DMSO was added. Gradient delivery started at a flow rate of 450nl/min and 3% B. Over the first 2 minutes, gradient flow rate was reduced to 300nl/min while the gradient organic content increased to 9% B. Over the 23 subsequent minutes, the gradient increased linearly to 38% B, at which point the gradient rapidly increased to 80% B over 2 minutes time. The column was held at 80% B for the remaining 3 minutes of the gradient delivery, completing in 30 minutes.

During this gradient delivery, peptides were ionized by an electrospray ionization voltage of 2.2kV in the positive mode. The data dependent acquisition included MS1 scans of 70,000 resolution and MS2 scans of 17,500 resolution. Maximum injection time for MS1 and MS2 scans was set to 120ms, with an MS1 and MS2 AGC target of 1e6 and 5e4, respectively. MS1 scan range was set from 400 to 1800 *m/z* and data acquired in profile mode, while the MS2 scan range set from 200 to 2000 *m/z*. Precursors were selected for fragmentation provided that they were charge +2 to +6, allowing for fragmentation of multiple charge states, but excluding isotopes. Selected precursors were fragmented in a top-12 cycle. Dynamic exclusion for the shorter 30-minute gradients was set at 12 seconds, with a minimum AGC target of 5e2. The quadrupole isolation width was set to 2.1 *m/z* and HCD fragmentation collision energy set to 25NCE. Samples for the 293 control acquisition using RMMG beads, 293 control using WT beads, PCNA-BioID2 using WT beads and MMS19-BioID using WT beads were all acquired with two technical replicate acquisitions. The MMS19-BioID using RMMG beads and PCNA-BioID2 using RMMG beads were only acquired in a single technical replicate acquisition.

**α-Hemagglutinin LC-MS Acquisition:**

Data acquisition for the α-HA bead comparisons was performed on a Thermo Orbitrap Fusion Lumos mass spectrometer through DDA. Chromatographic gradient delivery was

performed by a Thermo Dionex Ultimate 3000 nanoLC ProFlow pump system. Peptides were separated on a 70-minute gradient through a 75uM ID, 18cm C18 column packed with 1.9µM C18 particles (Dr. Maisch GmbH). Buffer compositions matched the chromatographic apparatus utilized for the acquisition of the streptavidin samples. Gradient delivery began at a flow rate of 400nl/min and 1%B. In the first quarter minute, organic content increased to 4%B and 8.2%B at 4 minutes when the gradient flow rate was lowered to 200nl/min. Organic buffer composition increased linearly to 29%B at 65 minutes, and 80%B at 67 minutes. At 68 minutes, the organic buffer composition was dropped to 1%B and held there until the end of the 70-minute chromatographic separation. Before sample loading, columns were washed by introduction of 6µL 60% acetonitrile, 20% 2-propanol, 20% $H_2O$ and equilibration to aqueous condition.

Peptides were ionized by the application of 2.0kV ionization voltage, in the positive mode. DDA contained MS1 scans generated in the Orbitrap at 500,000 resolution in profile mode, and MS2 scans acquired in the linear ion trap in the rapid scan mode. Maximum injection time for MS1 scans was set to 100ms and to 35ms for the linear ion trap MS2 scans. The MS1 Orbitrap AGC target was set to 2e5, and the MS2 scans with an AGC target of 2e3. MS1 scan range was set to 400-1600 *m/z*, using quadrupole isolation, and with the easy-IC internal calibrant turned on. Peptide precursors were selected from charges +2 to +6, with an intensity threshold of 4e3 with monoisotopic precursor selection turned on in a 3 second cycle time between MS1 scans. Quadrupole isolation for MS2 scans was set to a width of 1.6 *m/z*, with HCD fragmentation utilized with 35% collision energy. Dynamic exclusion was set to 25 seconds, with ±10 ppm tolerances and isotope exclusion turned on. All conditions were acquired in technical replicates on two separate chromatographic columns.

**Bioinformatic Analysis:**

Each experiments' LC-MS raw data was converted to mzML format by ProteoWizard's msconvert (v. 3.0.11348) with vendor peak picking enabled[41,51]. Each run was searched against the EMBL human reference proteome, appended with both mouse IgG heavy chain sequence (P01868) and the Streptavidin sequence (P22629). Database searching was carried out by MSGF+ (v. 2016.06.29) considering peptides with a precursor mass tolerance of 15ppm and an allowable isotope error in the range [-1,2], requiring candidate peptides to be within 6-40 amino acids in length and obeying tryptic enzymatic digestion rules at both termini[13-15]. The high resolution MS2 scans from the Streptavidin (QE) dataset was searched with the "Q-Exactive" instrument ID, while the low resolution MS2 scans from the linear ion trap of the Lumos was searched using the High-res LTQ instrument ID in MSGF+ with carbamidomethylation added as a fixed modification on cysteine residues for both experiments. Target/decoy searching was carried out by means of database protein sequence reversal, and separate target/decoy searches[52]. For each of the two experimental sets, the target and decoy searches for the corresponding runs were combined and fed to the crux (v. 3.1) wrapper of percolator (v. 3.01.nightly-18-1e0fbeb)[53,54]. The resulting PSMs were fed into the standalone version of FIDO (v. 1.0) to produce protein level probabilities, which were subsequently converted to q-values[16]. Identifications were filtered at both PSM and protein level q-value thresholds of 0.01. Spectral counts were calculated by the crux spectral-counts function.

For label free, intensity-based comparisons, confident identifications were converted into spectral libraries and MS1 extracted ion chromatograms generated by Skyline (v. 4.1.0.18169)[22]. Skyline's peptide database background was set to a digestion setting of "Trypsin/P", allowing for no missed cleavage sites, and disallowing ragged-ended peptides. Extracted ion chromatogram windows were generated with a 2-minute retention time tolerance for the Lumos runs containing

a longer gradient and 1-minute for the shorter QE datasets. For both experiments, an 8ppm mass tolerance window around three isotopic peaks per analyte was extracted. For each of the experiments Skyline analyses, an mProphet peak picking model was trained on all available scores and used to assign confidence to the integrated peaks[55]. Peptide intensity values exported by Skyline were filtered by mProphet q-value at a threshold of 0.01, and protein intensities modeled and compared by the MSstats package (v. 3.9.2) after filtering to require that all peptides used for quantitation mapped uniquely within the background proteome and requiring proteins to have two quantifiable peptides[56]. Protein intensities were summarized by means of the Tukey Median Polish implementation within MSstats, with model-based imputation turned on and the "maxQuantileforCensored" set to NULL. Normalization for the α-HA experiment was set to include all peptides belonging to 5 human proteins selected for their universal identification amongst every acquisition in the dataset: IRS4, THRAP3, GTF2I, BCLAF1, and LSM14A. For the streptavidin comparisons, intensities were normalized by means of median equalization. Statistical differential protein abundance testing was provided by means of the linear mixed model implementation within the MSstats package, and p-values adjusted for multiple hypothesis testing by the Benjamini-Hochberg correction[57].

To determine the signal intensity impact of streptavidin-peptide derived chromatographic shifting, mProphet filtered peptide intensity values were grouped by those without chromatographic shifts and those peptides for which the median Skyline determined retention time of the peptide differed between RMMG and WT runs by at least 30 seconds. Only peptides which were confidently detected in both conditions were included in this analysis. The median normalized intensity values were made into log2 transformed ratios comparing the WT and RMMG peptide intensities. The two populations were compared by the Mann-Whitney U test

implementation within Python's (v2.7) scipy stats module[58]. Histograms were rendered within R (v.3.5.0) using the ggplot2 package, and basepeak chromatograms through the MSnbase interface to mzML files[42]. All raw data acquired and analyzed here are available through the MassIVE repository via the ProteomeXchange identifier "PXD011858"[59,60].

## *Results and Discussion*

The tryptic digestion of protein samples bound to immobilized streptavidin beads is routinely performed during the course of a wide range of proteomic experiments. Despite the fact that the compact structure of streptavidin makes it naturally resistant to trypsin-mediated proteolysis, we have frequently observed high amounts of streptavidin-derived peptides in these samples. A representative basepeak chromatogram from the LC-MS/MS run of one such sample prepared by digesting streptavidin beads with trypsin is shown in **Figure 3-1A**. Multiple high intensity peaks are observed in the chromatogram that we hypothesized were generated by the digestion of streptavidin by trypsin. We confirmed this by LC-MS/MS analysis which identified a large number of peptides mapping to the streptavidin sequence and the subsequent plotting of extracted ion chromatograms (XICs) of those peptides which clearly demonstrated that the dominant peaks observed in the basepeak chromatogram traces correspond to streptavidin-derived peptides (**Figure 3-1B**). Given the prevalence of these peptides in the sample, it seemed likely that they were lowering the quality of these LC-MS/MS datasets by increasing the overall complexity and dynamic range of the sample.

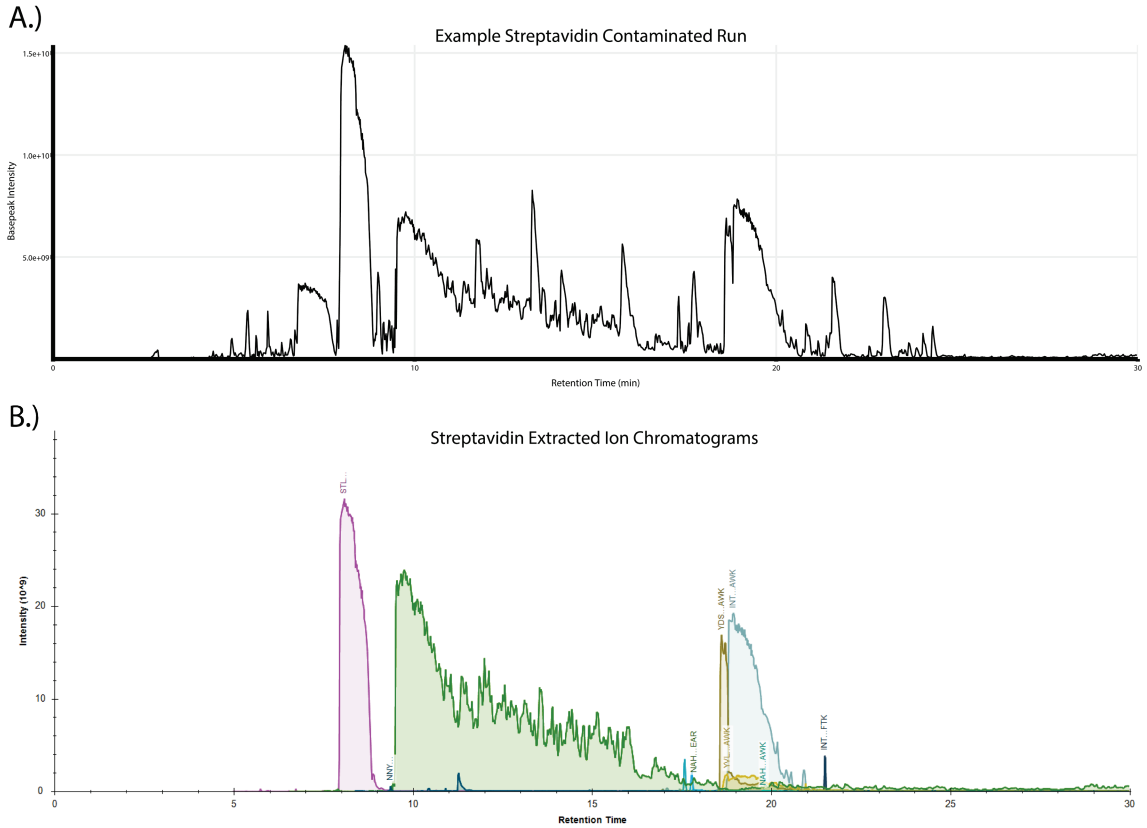


**Figure 3-1.**

On-bead tryptic digestion of streptavidin affinity purification samples leads to substantial streptavidin peptide contamination of samples.

A.) A basepeak chromatogram from a typical streptavidin affinity purification on-bead digestion demonstrates the high dynamic range in peptide intensities, with a small number of peaks dominating the chromatographic separation.

B.) Skyline generated extracted ion chromatograms (XICs) of confidently identified streptavidin derived peptides were manually integrated and plotted against retention time showing the level of high-intensity contamination which originates from the matrix.

To address this issue of streptavidin peptide contamination, we hypothesized that generating chemically derivatized streptavidin that was resistant to proteolysis by trypsin would improve the coverage of proteomic analyses. To test this idea, we used two chemical derivatization strategies. First, we methylated the lysine residues in streptavidin in using standard reductive methylation strategies that utilize dimethylamine borane and formaldehyde (**Figure 3-2A**). We then further modified the reductively methylated streptavidin by treatment with methylglyoxal

(MGO) to form dihydroxyimidazolidine or hydroimidazolone adducts on arginine residues (**Figure 2A**)[61]. We next tested whether covalent modification of the lysine and arginine residues in streptavidin (1) impairs its binding to biotin and (2) renders it more resistant to digestion by trypsin. To test whether biotin binding by the doubly derivatized streptavidin (RMMG) was impaired relative to wildtype, we used biotin-horseradish peroxidase (HRP) to assay its binding activity. Wildtype or modified streptavidin was incubated was biotin-HRP and washed before measuring the amount of biotin-HRP retained on the beads using the Pierce 1-Step Slow TMB ELISA colorimetric assay. **Figure 3-2B** shows that biotin binding was unaffected by modification of the lysines and arginines in streptavidin across a large range of biotin-HRP concentrations. We further validated this observation by performing a streptavidin pulldown using wildtype or modified beads from cell extracts prepared from stable cell lines expressing BioID-fused to either MMS19 or PCNA. The BioID-MMS19 and BioID2-PCNA fusion proteins non-specifically biotinylate proteins within their vicinity which can then be affinity purified using streptavidin beads, digested with trypsin, and the analyzed by LC-MS/MS. As shown in **figure 3-2C**, label-free quantitation of the amount of the BioID-MMS19 or BioID2-PCNA fusion protein was unaffected by modification of the streptavidin. In order to examine whether the derivatization of the lysines and arginines in streptavidin made it more resistant to trypsin digestion, we examined the wildtype and modified beads by LC-MS/MS after proteolytic digestion using our standard trypsin-based workflows. Extracted ion chromatograms of streptavidin-derived peptides generated from wildtype or modified streptavidin samples after tryptic digestion are shown in **figure 3-2D**. These chromatograms clearly demonstrate a major reduction in streptavidin-derived peptides upon chemical derivatization and support the hypothesis that these modifications impair tryptic digestion. Additionally, label-free quantitation of streptavidin abundance shows a substantial

decrease in streptavidin abundance for the modified streptavidin (RMMG) relative to the underivatized (WT) beads (**Figure 3-2E**). Together these data argue that the derivatized streptavidin beads are not measurably impaired in their ability to bind biotinylated proteins but are highly resistant to proteolysis by trypsin.

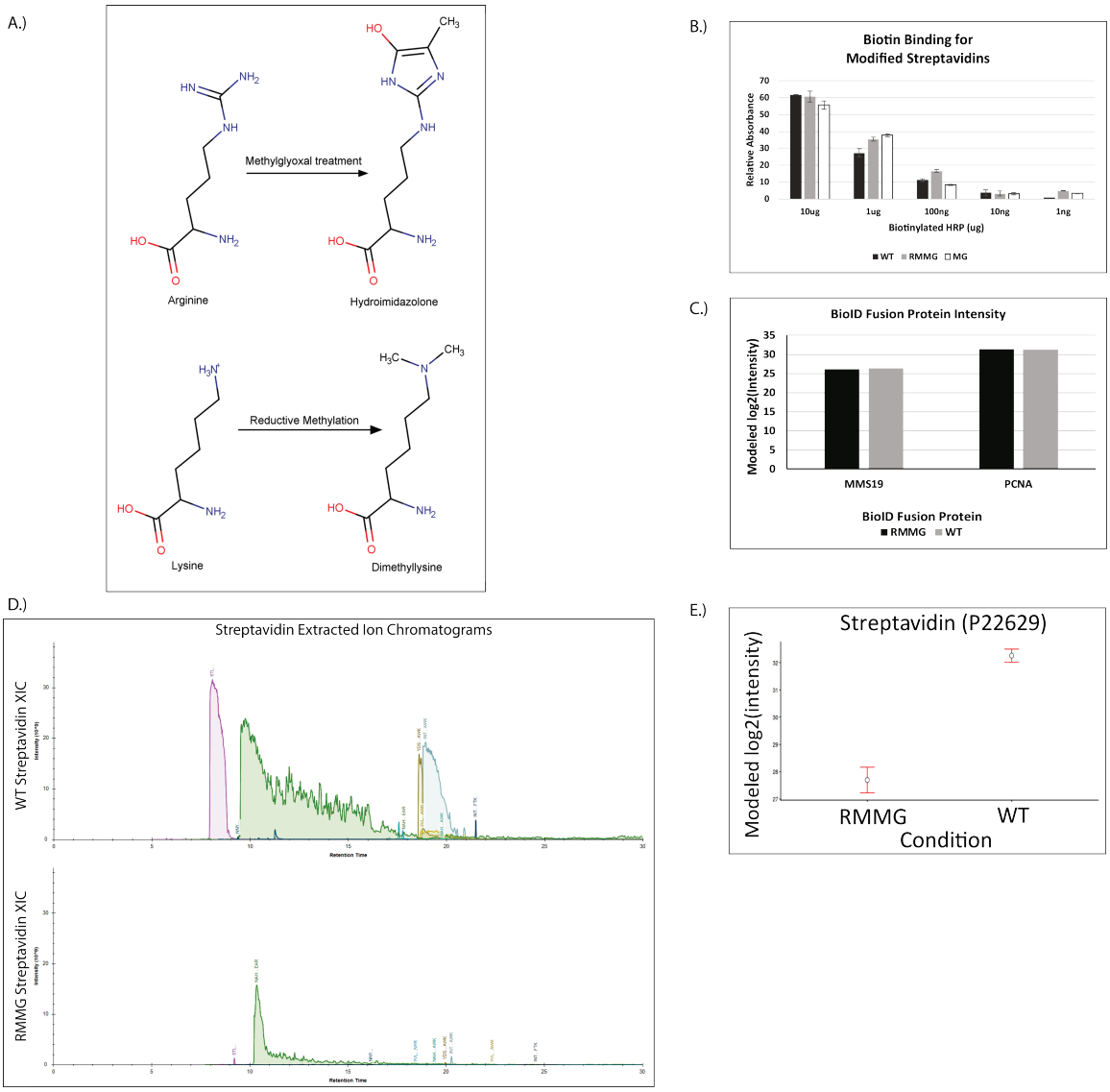


**Figure 3-2.**

Streptavidin derivatization provides protection from enzymatic digestion without disrupting biotin binding.

A.) Schematic showing the derivatization of arginine and lysine residues of polypeptides to protect against enzymatic digestion.

B.) Biotinylated HRP activity was measured after binding to various forms of derivatized streptavidin. HRP enzymatic activity was not impacted by the lysine and arginine derivatization, suggesting the binding of biotinylated proteins would not be disrupted. Values are given in arbitrary units of absorbance, with error bars showing the standard deviation of replicates for each bead type and HRP mass.
C.) Auto-biotinylated BioID fusion proteins are detected with similar label-free intensities when purified by either derivatized (RMMG) or underivatized (WT) streptavidin beads.
D.) Manually integrated extracted ion chromatograms were generated for uniquely mapping tryptic peptides derived from Streptavidin within the Skyline software. A representative pair of WT/RMMG pulldowns display the impact of derivatization on the detectable streptavidin peptide signals, dramatically reducing the overall streptavidin intensity.
E.) Label free protein intensities comparing the overall streptavidin intensity of the aggregate of all BioID data acquired here with WT beads, and those runs prepared with RMMG beads. Values displayed are modeled protein intensities, with error bars representing the reported 95% confidence intervals.

Having established the suitability of the modified streptavidin beads for affinity purification experiments, we next compared proteomic analyses of streptavidin pulldowns performed using wildtype and derivatized beads in order to assess their relative contribution to overall proteomic data quality. First, we plotted the normalized PSM identification rate across chromatographic runs for LC-MS/MS analyses of streptavidin pulldowns done using wildtype (WT) or derivatized (RMMG) beads (**Figure 3-3A**, top). Wildtype streptavidin purifications displayed regions of the chromatography in which PSM identification were reduced. These regions correspond to the elution of major streptavidin-derived tryptic peptides (**Figure 3-3B**, bottom) suggesting that the increased dynamic range generated by the elution of these high abundance contaminating peptides hampers peptide identification in those stretches. Strikingly, this reduction in peptide identifications was restored in purifications done using protected streptavidin beads highlighting the benefit of these derivatized streptavidin beads in limiting the elution of streptavidin-derived peptide contaminants and preventing the masking of signals belonging to non-streptavidin peptide analytes of interest.

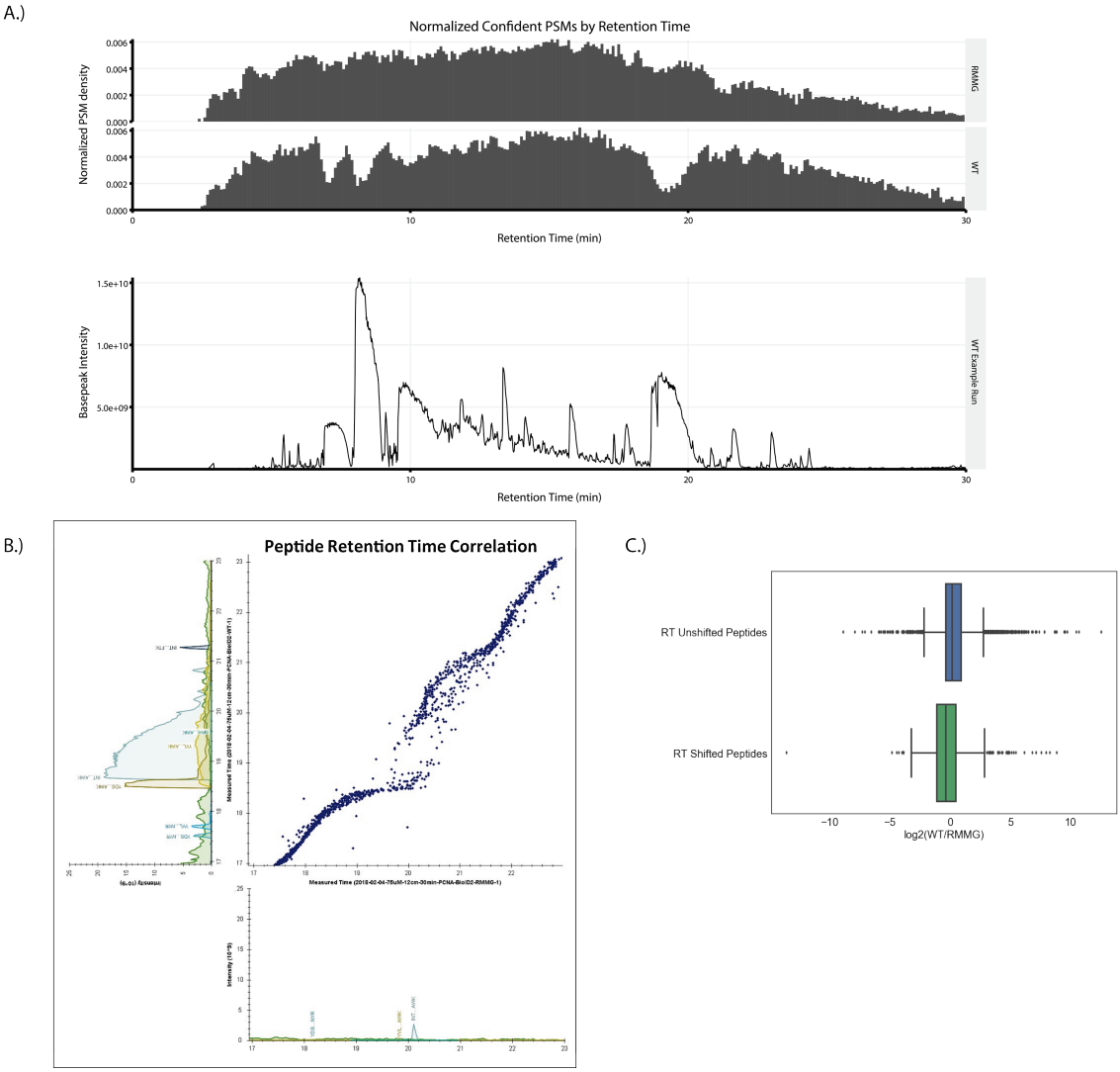


**Figure 3-3.**

Streptavidin derivatization ameliorates the local chromatographic perturbation due to streptavidin peptide elution.

A.) Top: Peptide-Spectrum Matches (PSMs) for the aggregate of all samples prepared using RMMG streptavidin beads and the aggregate of all samples prepared using WT streptavidin beads were plotted as normalized PSM density in histograms with a bin width of 0.1 min. Bottom: An example affinity purification performed using on-bead digestion of WT streptavidin beads shows high intensity streptavidin contamination which correlates in time with the drops in PSM rates for data acquired using the WT streptavidin beads.

B.) A Skyline generated plot of a paired set of peptide retention times from affinity purifications performed using WT (y-axis) and RMMG (x-axis) streptavidin beads from 17 to 23 minutes of retention time. The WT streptavidin runs exhibit local perturbation of the chromatography during elution of the high intensity streptavidin peptides.

C.) Comparison of the log2 fold-change of peptides compared between the WT and RMMG conditions. Peptides which did not exhibit a chromatographic shift between the two bead

types retained an approximately zero fold-change (n=5,872), while the peptides with retention time shifts were suppressed in intensity in the WT runs compared to the RMMG runs (n=1,059). Differences in log transformed fold-change ratios distributions between the two groups were assessed by the Mann-Whitney U test, showing a significant difference ($p<1E-54$).

We also examined the streptavidin-contaminated samples to determine whether the streptavidin-derived peptides might negatively impact the chromatography. **Figure 3-3B** shows the peptide retention time correlation between analyses of pulldowns using wildtype or protected streptavidin. Strikingly, the elution of major streptavidin species leads to a marked disruption in peptide elution times with streptavidin peptides effectively pushing other peptides out of their typical elution window. Importantly, the population of peptides displaying shifted retention times also display reduced intensity relative to the unshifted peptides suggesting that the streptavidin-derived peptides suppress the ionization of these retention time shifted peptides. Based on these data, we conclude that streptavidin-derived peptides contribute to reduced peptide identification rates, shifted retention times, and ion suppression of co-eluting peptides and that these negative effects are alleviated when digestion-resistant streptavidin beads are used for the affinity purification.

Given the effectiveness of this protection strategy in improving the analysis of on-bead digested streptavidin pulldowns, we explored the possibility that this approach could be generalizable and potentially extended to other affinity purification matrices. We first tested this using anti-HA antibodies coupled to an agarose support. The anti-HA resin was chemically derivatized using either lysine reductive methylation (RM), methylglyoxal modification of arginines (MG), or both (RMMG). Digestion of these chemically protected beads with trypsin dramatically reduced the amount of IgG-derived peptides in the digest. This is evident in **Figure 3-4A** which shows basepeak chromatograms from WT α-HA and RMMG α-HA digested samples

and the prominent loss of high abundance IgG-derived peptides specifically in the derivatized sample. These chromatograms are consistent with the LFQ analysis of these samples shown in **Figure 3-4B** in which both the RM and RMMG protected beads show dramatically reduced Igg abundance in the sample after on-bead tryptic digestion. Interestingly, reductive methylation of lysines appears to be sufficient for this effect with little to no contribution from methylgloxal treatment (MG) being observed. Importantly, we also confirmed that the chemical protection of anti-HA did not significantly impair its ability to immunoprecipitate HA-tagged proteins. Control, RM, MG, or RMMG treated anti-HA beads were used to immunoprecipitate either 3HA-3FLAG-tagged CIAPIN1 or 3HA-3FLAG-tagged BOLA2 from protein lysates generated from HEK293 cell lines stably expressing those fusion proteins. The spectral counts obtained for each bait after LC-MS/MS analysis of the immunoprecipitated sample was used to assess the effectiveness of the immunoprecipitation. **Figure 3-4C** clearly shows that HA-tagged CIAPIN1 and BOLA2 were both similarly enriched in these samples irrespective of whether untreated or protected beads were used.

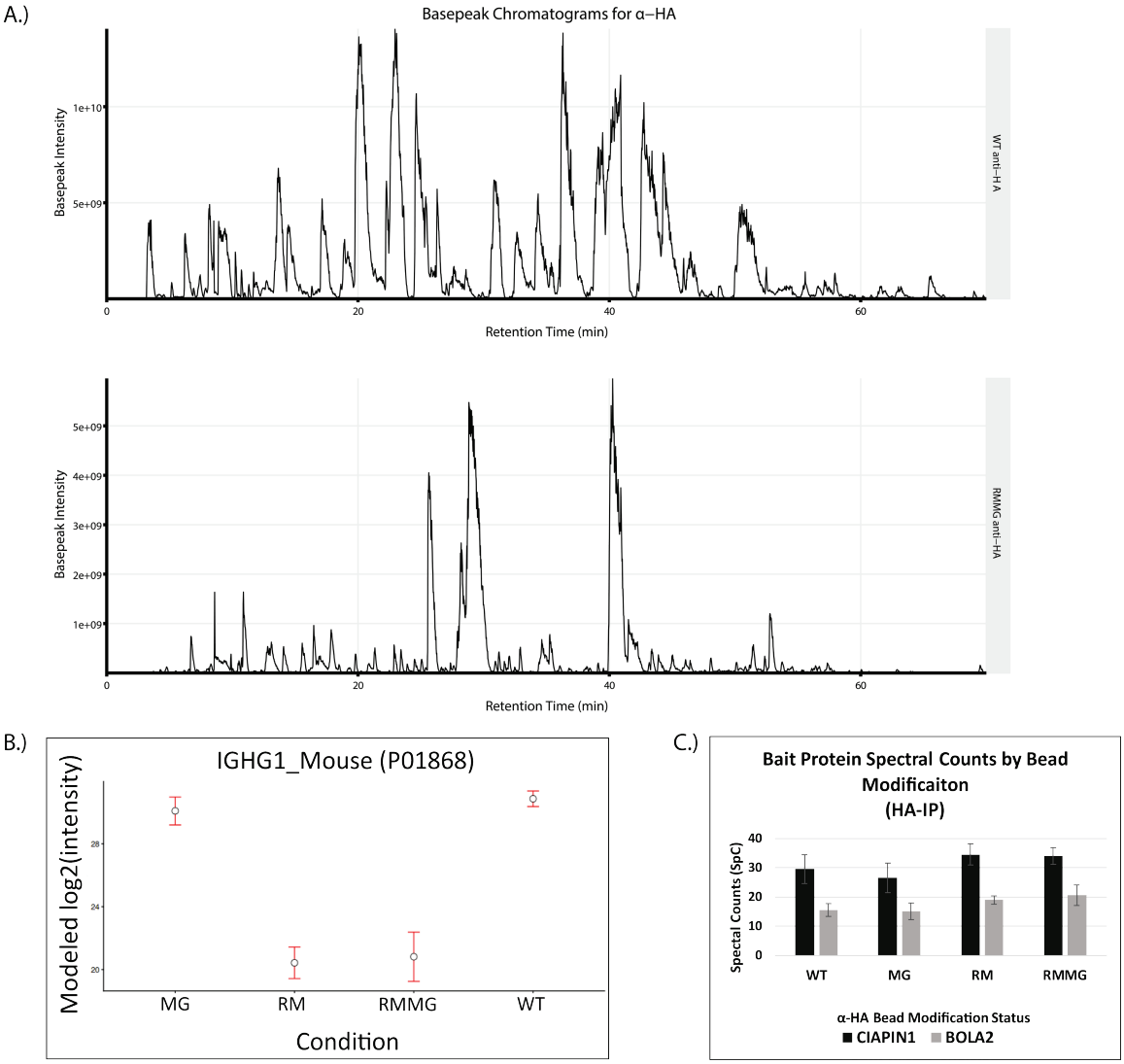


**Figure 3-4.**

The chemical derivatization method is extensible to antibody based affinity purification matrices.

A.) Example basepeak chromatograms of a pair of pulldowns generated on WT (top) and derivatized (RMMG, bottom) α-HA beads.

B.) IgG heavy chain protein intensity is dramatically reduced upon chemical derivatization of the beads prior to enzymatic digestion. The majority of signal reduction is granted by the reductive methylation and not the methylglyoxal treatment. Values are MSstats modeled protein intensities with error bars representing the 95% confidence interval.

C.) α-HA beads retain binding of epitope tagged protein targets despite modification status. Confident spectral counts (SpC) shown for both of the bait proteins from their respective pulldowns. Each pulldown was acquired with a two technical replicate on separate chromatographic columns. Values shown as the mean with error bars displaying the standard deviation.

## *Conclusions*

The on-bead digestion of streptavidin to elute proteins during affinity purification workflows results in the release of streptavidin-derived peptides into the sample. The high abundance of these peptides limits the subsequent mass spectrometric analysis by suppressing identification of co-eluting peptides and reducing peptide identification rates during their elution. We report a novel strategy for reducing the production of these peptides by chemically derivatizing lysine and arginine residues in streptavidin to render it largely resistant to trypsinization. Proteomic analysis of affinity purifications performed using these modified beads restores the loss of peptide IDs and aberrant chromatography observed in purifications done using wildtype streptavidin beads. Finally, we also demonstrate that this approach for generating digestion resistant beads is potentially generalizable using the immunoprecipitation of HA-tagged proteins with immobilized anti-HA antibodies as an example.

A wide range of elution strategies are currently employed to facilitate streptavidin-based affinity purifications. These range from engineered streptavidin resins in which their reduced biotin affinity enables elution by excess free biotin, to techniques that separate the biotinylated peptides of interest away from contaminating peptides[62,63]. Based on the effectiveness and facile implementation of our approach, we anticipate that it will become a robust alternative to these options that that can be incorporated into different workflows as needed based on their analytical requirements. The availability of a suite of sample preparation approaches will offer flexibility and adaptability as new applications are developed.

A key advantage to the presented method is its potential to be generalized to other affinity purification matrices. The ability to elute proteins or other biological analytes directly from affinity

supports using trypsin simplifies the sample preparation workflow and minimizes the opportunity for sample loss. Like streptavidin, however, these strategies are difficult for antibody-based affinity resins which will release IgG-derived peptides into the sample after proteolysis. Our results indicate that this chemical derivatization strategy can be adapted for α-HA resin opens up new sample preparation options for immunoaffinity chromatography and highlights the broad utility of this method. Although we have focused on the utility of these beads in the context of bottom-up proteomics, we anticipate potential uses for the derivatized affinity beads in other experimental workflows. For example, immobilized streptavidin is often used to purify biotinylated nucleic acids from biological mixtures. Using derivatized streptavidin in these workflows would provide an option for deproteinizing these samples without eluting the nucleic acid from the bead. Similarly, methods exist for the purification of specific cell types from mixtures using biotinylated antibodies. Protection of the streptavidin beads in these experiments would enable elution of specific cells without the concurrent proteolysis and release of substantial streptavidin contaminants.

# Chapter 4: MilkyWay, a Galaxy Proteomics Platform for Label Free Quantitative Comparisons

## *MilkyWay, a Galaxy Proteomics Platform for Label Free Quantitative Comparisons*

### *Abstract*

The interrogation of data produced by modern mass spectrometric data acquisitions has continued to grow in complexity and sophistication. With every new analytical tool, new methods of pulling critical biological inference from biological experimentation has become available. As the glut of these computational resources become available, they unfortunately exist as disjoint and standalone tools among the proteome informatics landscape. To support the production of reproducible and scientifically meaningful data interpretation in the 'omcs age, workflow management platforms have become a popular mechanism to package powerful assemblies of independent tools for distribution and wider use. Here, we present MilkyWay, a proteomic analysis platform built atop the Galaxy workflow platform which integrates several cutting-edge utilities for the analysis of label-free DIA and DDA datasets. MilkyWay provides an R/Shiny web-app interface facilitating the upload of required files, definition of experimental topology, and an interactive exploratory analysis tool for the resultant data.

### *Introduction*

For several years, modern proteomic mass spectrometric instrumentation has become capable of producing identifications for more than 4,000 proteins per hour[64]. Accordingly, the volume of data which a single laboratory can produce has dramatically increased. The bioinformatic processing and ultimate interpretation of these data has been a longstanding point of difficulty and bottleneck for facilities and groups providing proteomics services[65,66]. The analysis

of proteomic mass spectrometry datasets is a multi-step endeavor, in which discrete portions of the informatic analysis is often provided by disjoint utilities. Standardization of open file formats has accelerated the interchange and exchange of datasets[67,68], but they have yet to reach full adoption. The creation of proteome informatic workflows which provide coherent toolkits have become very popular, but each occupies its own niche offering a different portion of the available algorithmic approaches[19,21,53,69-72].

We present here a platform built atop the Galaxy bioinformatics workflow management system[73]. MilkyWay provides a toolkit capable of generating comparisons of label-free proteomic datasets, either using intensity-based measures or spectral counts/normalized spectral abundance factors[18,22]. MilkyWay takes input raw data from bottom-up liquid chromatography mass spectrometry (LC-MS) experiments in the open mzML format, or the Thermo proprietary raw file format. We provide support for comparative analysis of data dependent acquisition (DDA) datasets, data independent acquisition (DIA) datasets, or datasets comprised of both.

Raw data is converted into the mzML open format through ProteoWizard's msconvert. After file conversion, identifications can be generated directly from a search of DDA data, or a search against pseudospectra generated from DIA data through the DIAUmpire algorithm[27]. In the situation where a mixed dataset is provided for analysis, identifications are derived from the DDA acquisitions to construct a spectral library which is subsequently used to directly search the DIA data through MSPLIT-DIA and ultimately extract fragment level intensity data from the DIA acquisitions[74,75]. Database searching is provided by the MSGF+ search algorithm, with peptide-spectrum-match (PSM) confidences rescored by the percolator implementation within the crux toolkit[15,41,53,54,76]. Protein inference is handled by the FIDO standalone implementation, with posterior error probabilities converted to protein level q-values[16]. Intensity based signal extraction

is performed by the Skyline software package, and chromatographic peak confidence estimated by the mProphet model implementation therein[22,55]. Affinity purification enrichment probabilities can be calculated by spectral counts through SAINTexpress, or with intensity metrics from DIA or DDA within SAINTq[77,78]. For other comparative intensity-based experiments, relative quantitation and statistical testing is provided by the MSstats package[56]. In situations where PTM localization confidence estimation is desired, we provide support for the LuciPHOr2 model, as well as a wrapper for phosphoRS to score phosphosite localization confidence[79,80].

| Tool | Wrapper Origin | Tool Origin | Purpose |
|---|---|---|---|
| MSconvert[41] | Galaxy-P | Proteowizard | File format conversion |
| Decoy Database Generation[81] | Galaxy-P | OpenMS | Reverse/Shuffle Decoy FASTA generation |
| DIAUmpire[27] | MilkyWay | Nesvizhskii Lab | DIA pseudo-MS2 generation |
| MSGF+[15] | ProtK | Pevzner Lab | Database Search |
| msgf2pin converter[54] | MilkyWay | Percolator (Käll lab) | File format conversion |
| crux percolator[53] | MilkyWay | crux (Noble lab) | ID statistical validation |
| LuciPHOr2[79] | MilkyWay | Nesvizhskii Lab | PTM localization |
| phosphoRS[80] | MilkyWay | Mechtler Lab | PTM localization |
| FIDO[16] | MilkyWay | FIDO (Noble lab) | Protein Inference |
| crux spectral-counts[53] | MilkyWay | crux (Noble lab) | Spectral Counting |
| SAINTexpress[77] | MilkyWay | Nesvizhskii Lab | AP-MS PSM enrichment |

| | | | |
|---|---|---|---|
| SAINTq[78] | MilkyWay | Choi Lab | AP-MS intensity enrichment |
| BiblioSpec[75] | MilkyWay | MacCoss Lab | Spectral Library Building |
| MSPLIT-DIA[74] | MilkyWay | Bandeira Lab | DIA Spectral Library Searching |
| Skyline[22] | MilkyWay | MacCoss Lab | Intensity Based Data Interrogation |
| MSstats[56] | MilkyWay | Vitek Lab | Differential Expression Statistical Testing |

**Table 4-1.**

The tools integrated into the MilkyWay differential proteome analysis environment and the uses. The wrapper origin column specifies whether a tool or workflow had previously provided a Galaxy tool definition file rendering it capable to be run from Galaxy. The tool origin column contains information about the workflow from which the utility was derived, or the lab if the tool is a standalone utility.

The tools included within MilkyWay are summarized within **Table 4-1**. Scripts handling the staging, input file processing, tool execution, and output file cleanup were wrapped into Galaxy tools and structured into pre-packaged workflows (**Table 4-2**) inside the distributed MilkyWay Galaxy flavor. These workflows cover a set of basic proteomic computational analyses which serve as immediately utilizable entry points to the relatively complex network of tools implemented. Finally, each workflow ends in a data packaging step which scrapes the relevant output data, Galaxy tool execution parameters, and stores them into an ".Rdata" environment to be interrogated within the MilkyWay R/Shiny web-app, or manually in any R instance.

| Workflow | DIA-Umpire | MSGF+ | crux percolator | FIDO | PTM localization | SAINT | Skyline | MSstats |
|---|---|---|---|---|---|---|---|---|
| LFQ DDA | ☒ | ✓ | ✓ | ✓ | ☒ | ✓ | ✓ | ✓ |
| LFQ PTM DDA | ☒ | ✓ | ✓ | ✓ | ✓/☒ | ✓ | ✓ | ✓ |
| Qualitative DDA | ☒ | ✓ | ✓ | ✓ | ✓/☒ | ✓/☒ | ✓/☒ | ☒ |
| LFQ DIA | ✓ | ✓ | ✓ | ✓ | ☒ | ✓ | ✓ | ✓ |
| ID DDA+LFQ DIA | ☒ | ✓ | ✓ | ✓ | ✓/☒ | ✓ | ✓ | ✓ |

**Table 4-2.**
An overview of the basic workflows included in MilkyWay, covering several uses cases in label free bottom-up proteomic informatics. Tools specified with a check and an X are optional and can be included in the workflow execution if desired by the user.

## *Platform Architecture and Distribution*

MilkyWay is distributed as a set of docker containers available on the docker hub, with code and deployment instructions and requirements provided on GitHub (http://github.com/wohllab/milkyway_compose) along with a simplified user guide (https://milkyway-compose.readthedocs.io/en/latest/). Docker containers are a useful mechanism for the packaging and distribution of runtime environments and code complete with all necessary dependencies for operation[82]. Additionally, containers have become a central component of high performance computing and cloud environments, though they can be run on consumer grade hardware. The default deployment configuration runs atop the Docker Swarm container orchestration utility, illustrated in **Figure 4-1**. For full functionality, MilkyWay requires nodes running both Windows and Linux as workers within the swarm to provide environments for execution of all tools.

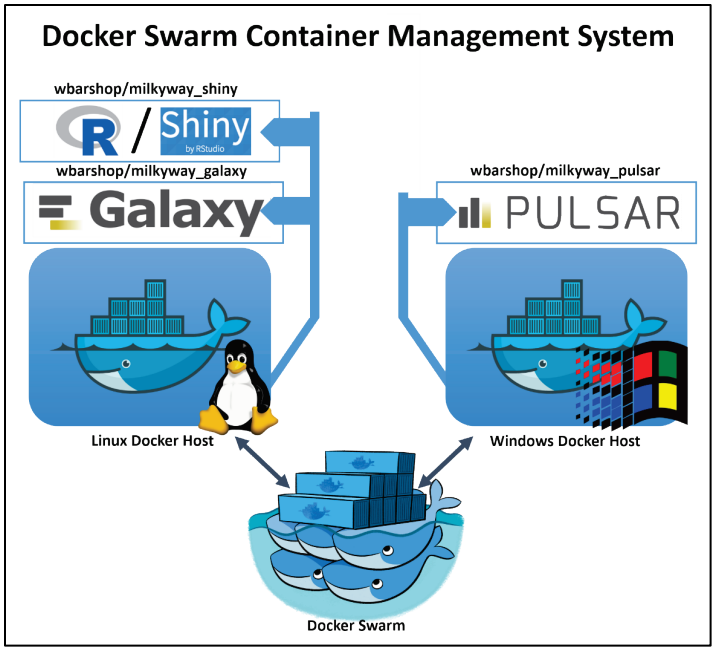


**Figure 4-1.**
The default deployment architecture of MilkyWay is built atop the Docker Swarm container orchestration utility, which handles the networking an intercommunication of containers. This diagram shows the three core containers for the platform which contain the Galaxy server, R/Shiny front-end, and Pulsar server.

## *File Upload and Experimental Topology Definition*

Users are provided with an R/Shiny web-app interface runs as a container within the Docker Swarm, through which they may upload new data for analysis, and explore completed analyses. The initial upload of data into MilkyWay is done through either the individual experiment upload tool, in which a set of *either* DDA or DIA files are uploaded, and file metadata annotated to allow for comparisons between biological conditions to be determined (**Figure 4-2**). Alternatively, we provide a specialized tool to handle the simultaneous upload of combined DDA and DIA acquisition sets, in which the DDA is used for generation of identifications and the DIA for quantitative signal extraction. Once file upload has been carried out to the R/Shiny container, the files are subsequently transferred to the Galaxy instance as a new history. Metadata for the

analysis such as sample origin and a user defined analysis name defined during file upload are embedded in the new history as tags through the Galaxy BioBlend API[83]. Once experimental data is available within Galaxy, users submit the proper tool execution workflow directly from the Galaxy web interface.

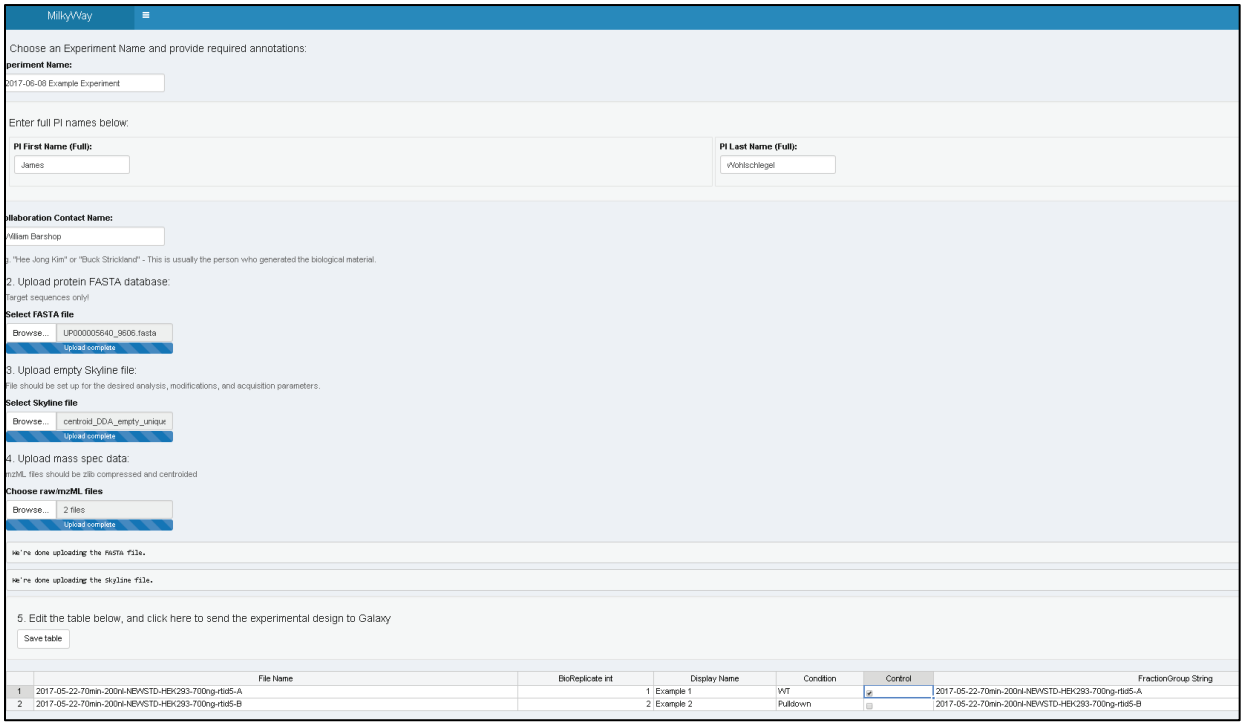


**Figure 4-2.**

A screenshot of the MilkyWay data upload utility. For each analysis, a user must define an experiment name, the laboratory (PI) name, and a submitting user contact name (collaboration contact) which are embedded in the new Galaxy history as tags. A proteome FASTA file, a configured Skyline document (omitted if only spectral counts are desired), and the raw mass spec data are uploaded and transferred to Galaxy, along with the experimental topology data defined in the populated table (bottom).

## *Data Organization and Exploratory Analysis*

Completed workflow results are read and organized into a browsable format after parsing the Galaxy user history and reading history tags to organize the analyses into a hierarchical format for browsing (**Figure 4-3**). When an analysis output package is loaded, the ".Rdata" file is

transferred into the R/Shiny instance and provides the necessary data to render exploratory plots and analytical outputs tables described below.

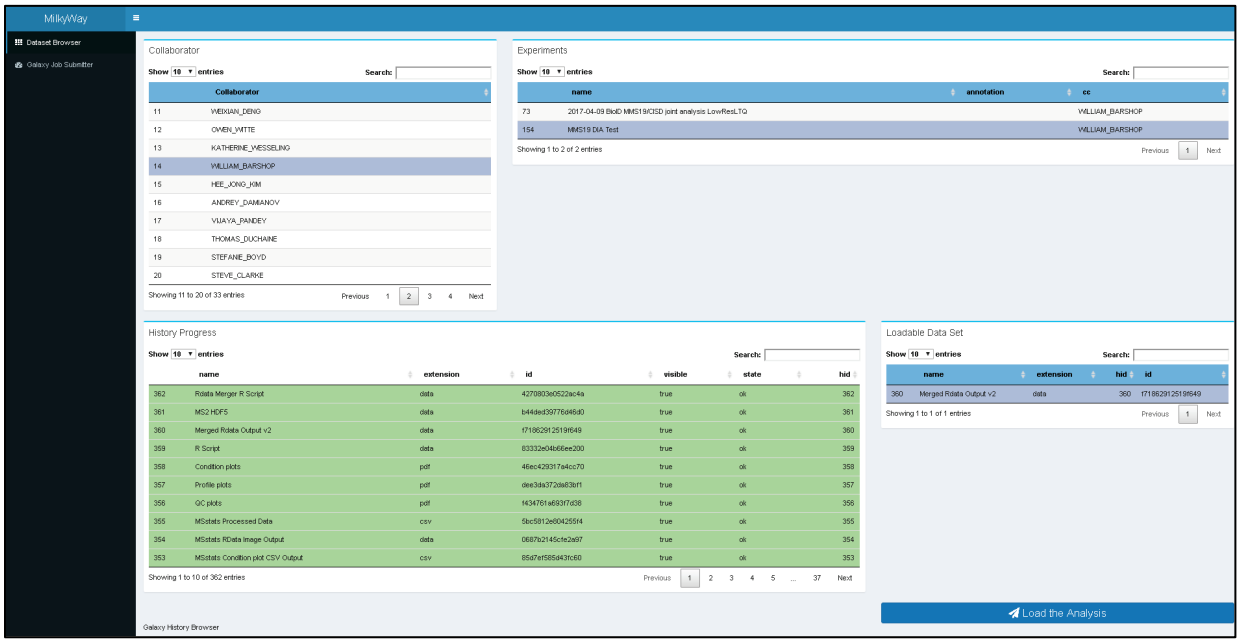


**Figure 4-3.**

A screenshot of the MilkyWay analysis browser within the R/Shiny web-application. This interface allows for the organization and loading of completed analyses from the Galaxy instance which have been summarized through the MilkyWay Rdata packaging tool. Analyses are sorted by collaborator (PI) name, and experiment name within those groups. For each experiment, a view of the history progress from Galaxy is provided, along with a filtered list of available .Rdata output packages for loading.

Upon the loading of an analysis output within the application, users are able to explore a variety of visualizations illustrating the comparative topology, identification counts, and identification mass accuracy distributions (**Figure 4-4**). These plots are rendered in real time, and the filter thresholds may be adjusted per the user's wishes. Users are also provided with plots showing mProphet chromatographic peak picking target-decoy competition, and a protein and peptide identification ROC curves (**Figure 4-5A**). MSstats comparisons yield fold-change estimates and differential abundance probabilities which can be explored via the interactive volcano plot (**Figure**

**4-5B**). Violin plots for peptide level intensities, and protein level intensities are available, along with embedded copy of the lorikeet interactive spectral annotation browser[84] (**Figure 4-5C-D**).


**Figure 4-4.**

An example analysis depicts the sample comparison topology diagram (left), identification delta mass distributions (top right) and confident PSM count per run (bottom right). These visual outputs are a first-stop check to ensure that portions of the computational analysis and machine acquisition have been performed without substantive error, and to quality-check each data acquisition by the count of confident PSM identifications.

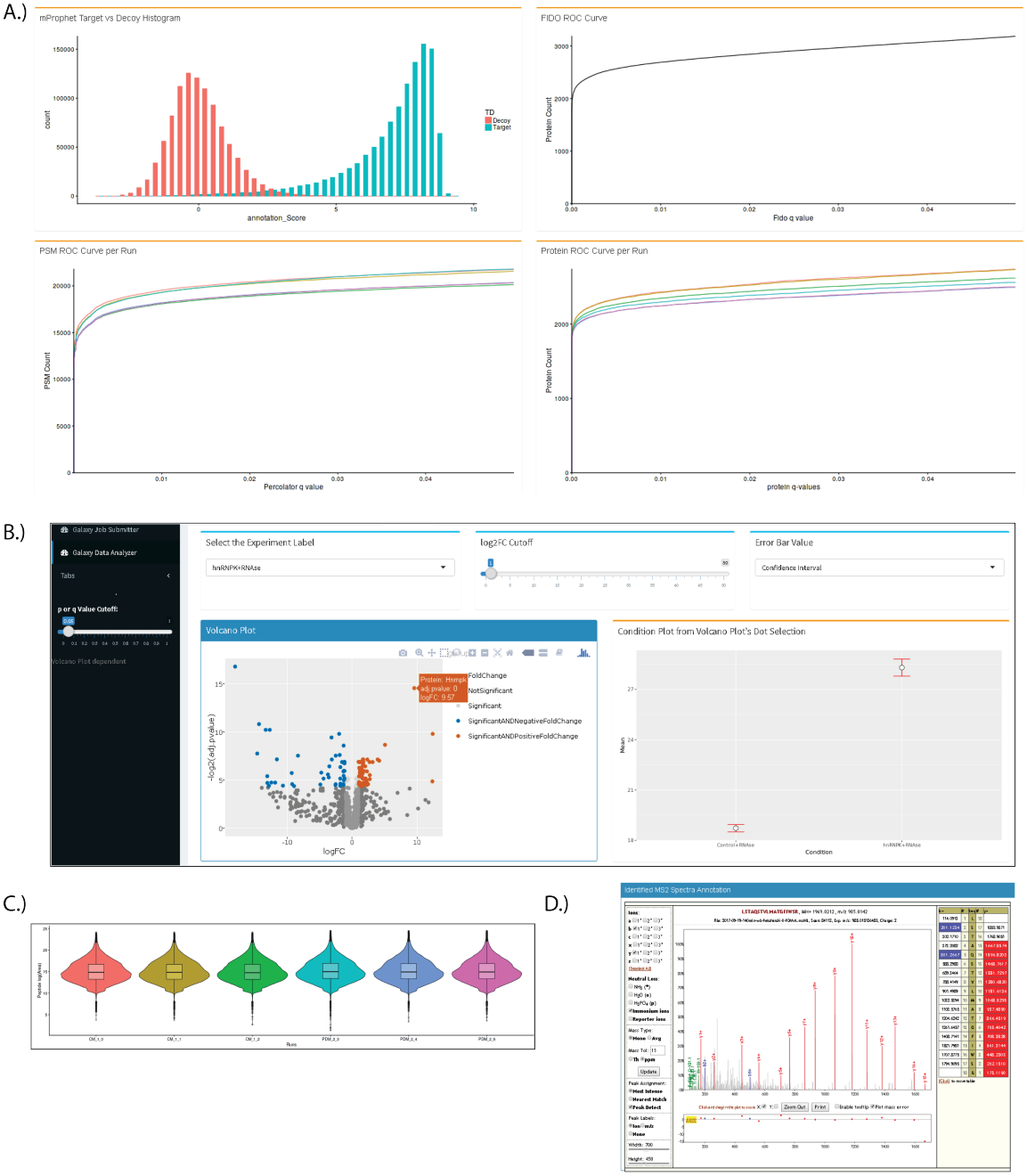


**Figure 4-5.**

A subset of interactive plots available through the MilkyWay R/Shiny interface for data exploration.

A.) Top left: mProphet model target-decoy competition score distribution. Top right: Analysis-wide joint FIDO protein ROC curve. Bottom left: individual run PSM-level ROC curves. Bottom right: individual run protein-level ROC curves.

B.) A MilkyWay screenshot displaying the interactive volcano plot, and a linked protein intensity boxplot. Values are generated by MSstats for these plots, and are additionally available in a table view.

C.) A run level peptide intensity violin plot generated from MilkyWay. Violin plots are available for both peptide and protein level intensity distributions and provide a useful means of sanity checking normalization and run quality.
D.) Example lorikeet rendering of an identified fragmentation spectrum. Users may filter and sort through identification tables before selecting spectra to be interrogated by the lorikeet annotation utility.

## *Conclusions*

The MilkyWay platform is a dockerized Galaxy flavor for comparative label-free proteomic data analysis which runs atop the Docker Swarm container orchestration utility. With an auxiliary R/Shiny web-application interface, MilkyWay facilitates rapid and powerful data exploration and export with interactive plots and filterable tables. As the rate of data generation from instrumentation increases, the utility of automation of computational analysis becomes a necessity. Simultaneously, the automation of informatic workflows naturally allows for the careful recording of parameters and settings used for the various analytical substeps to enhance the reproducility of analytical results. MilkyWay serves to accelerate the analysis of proteomic datasets through a powerful set of common workflows deployable on a mixed Windows/Linux Docker Swarm.

# Chapter 5: Conclusions

Experiments involving the use of bottom-up proteomic mass spectrometry have become a persistent presence in cell biology and biochemistry labs. Many of the existing analytical methodologies have reached mature implementation with broad adoption and efficacy. Still, in this age of advanced sample preparation, data acquisition, and informatics, there remain many avenues for improvement. In this thesis, I have described work to improve or explore these aspects of the bottom-up proteomics experimental ecosystem.

The use of isobaric tags for relative quantitation and multiplexing continues to be a highly popular methodology. The method, however, typically retains the use of semi-stochastic analyte selection and fragmentation by Data Dependent Acquisition (DDA). The Sequential Windowed Acquisition of Reporter Masses (SWARM) method is a Data Independent Acquisition (DIA)-like cycle in which the interrogated precursor mass range is segmented and sequentially isolated and fragmented, with mass and intensity data only scanned for the reporter ion mass range. This type of pre-scanning allows for decision making in how to allocate acquisition time, which we leverage to bias machine acquisition toward high fold-change analytes. By biasing machine acquisition, we are able to narrow the instrument focus toward analytes displaying quantitative characteristics indicative of biological interest. The SWARM+PRM implementation appears effective, and should be considered a proof-of-concept of the SWARM paradigm, and indicative of potential for further exploration of real-time SWARM acquisition methods.

The contamination of affinity enrichment samples caused by proteolysis of the affinity matrix during digestive elution has been a disruptive reality, especially when using enrichment systems for which alternate methods of elution are inefficient. This is a common situation when performing AP-MS experiments using streptavidin to enrich biotinylated protein targets. While some workflows replace streptavidin with mutants of lower biotin binding affinity to aid elution,

we have explored a method for protecting the immobilized streptavidin from tryptic proteolysis. The method appears highly effective at reducing streptavidin contamination, along with solving chromatographic warping caused by streptavidin peptide elution. Additionally, the modified streptavidin appears to retain biotin binding capabilities. We extended the method to an immunoglobulin cultivated against the hemagglutinin affinity enrichment tag. That antibody showed a similar pattern of protection from tryptic digestion while retaining binding. While we focus on applications of digestion protection in bottom-up proteomics, we expect that the method may be broadly applicable to many preparative workflows.

Lastly, the expansion of available bioinformatic utilities for proteomic mass spectrometry has been of great benefit to proteomics researchers, but many of the tools lack easy interoperability or integration into automated workflows. We present MilkyWay, a docker based Galaxy flavor which integrates a tapestry of bioinformatic algorithms together into a set of coherent analytical workflows. These workflows facilitate the rapid and reproducible analysis and exploration of DDA, DIA, and mixed acquisition datasets. The R/Shiny companion web-app enables effective organization and retrieval of analysis results for many projects, and provides powerful interactive plots for the interpretation of result outputs. The default deployment of MilkyWay is designed to run on the Docker Swarm container orchestrator, and requires both a Windows and Linux node. MilkyWay is available on GitHub (http://github.com/wohllab/milkyway_compose).

# References


1 Turriziani, B., von Kriegsheim, A. & Pennington, S. R. Protein-Protein Interaction Detection Via Mass Spectrometry-Based Proteomics. *Adv Exp Med Biol* **919**, 383-396, doi:10.1007/978-3-319-41448-5_18 (2016).

2 Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu Rev Anal Chem (Palo Alto Calif)* **8**, 61-80, doi:10.1146/annurev-anchem-071114-040325 (2015).

3 Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71 (1989).

4 Annesley, T. M. Ion suppression in mass spectrometry. *Clin Chem* **49**, 1041-1044 (2003).

5 Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun* **445**, 683-693, doi:10.1016/j.bbrc.2014.02.041 (2014).

6 Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R., 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* **113**, 2343-2394, doi:10.1021/cr3003533 (2013).

7 Smith, L. M., Kelleher, N. L. & Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nature methods* **10**, 186-187, doi:10.1038/nmeth.2369 (2013).

8 Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* **4**, 1419-1440, doi:10.1074/mcp.R500012-MCP200 (2005).

9 Mikesh, L. M. *et al.* The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta* **1764**, 1811-1822, doi:10.1016/j.bbapap.2006.10.003 (2006).

10 Cleland, T. P. *et al.* High-Throughput Analysis of Intact Human Proteins Using UVPD and HCD on an Orbitrap Mass Spectrometer. *Journal of proteome research* **16**, 2072-2079, doi:10.1021/acs.jproteome.7b00043 (2017).

11 Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of proteome research* **11**, 5479-5491, doi:10.1021/pr3007045 (2012).

12 Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976-989 (1994).

13 Kim, S., Gupta, N. & Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of proteome research* **7**, 3354-3363, doi:10.1021/pr8001244 (2008).

14 Kim, S. *et al.* The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & cellular proteomics : MCP* **9**, 2840-2852, doi:10.1074/mcp.M110.003731 (2010).

15 Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications* **5**, 5277, doi:10.1038/ncomms6277 (2014).

16 Serang, O., MacCoss, M. J. & Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research* **9**, 5346-5357, doi:10.1021/pr100594k (2010).

17 Lundgren, D. H., Hwang, S. I., Wu, L. & Han, D. K. Role of spectral counting in quantitative proteomics. *Expert Rev Proteomics* **7**, 39-53, doi:10.1586/epr.09.69 (2010).

18 Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *Journal of proteome research* **5**, 2339-2347, doi:10.1021/pr060161n (2006).

19 Sturm, M. *et al.* OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* **9**, 163, doi:10.1186/1471-2105-9-163 (2008).

20 Rost, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods* **13**, 741-748, doi:10.1038/nmeth.3959 (2016).

21 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

22 MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968, doi:10.1093/bioinformatics/btq054 (2010).

23 Weisser, H. & Choudhary, J. S. Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *Journal of proteome research* **16**, 2964-2974, doi:10.1021/acs.jproteome.7b00248 (2017).

24 Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods* **1**, 39-45, doi:10.1038/nmeth705 (2004).

25 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & cellular proteomics : MCP* **11**, O111 016717, doi:10.1074/mcp.O111.016717 (2012).

26 Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature communications* **8**, 291, doi:10.1038/s41467-017-00249-5 (2017).

27 Tsou, C. C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* **12**, 258-264, 257 p following 264, doi:10.1038/nmeth.3255 (2015).

28 Teo, G. *et al.* mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *Journal of proteomics* **129**, 108-120, doi:10.1016/j.jprot.2015.09.013 (2015).

29 Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP* **1**, 376-386 (2002).

30 Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904 (2003).

31 Erickson, B. K. *et al.* A Strategy to Combine Sample Multiplexing with Targeted Proteomics Assays for High-Throughput Protein Signature Characterization. *Mol Cell* **65**, 361-370, doi:10.1016/j.molcel.2016.12.005 (2017).

32 Gingras, A. C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol* **8**, 645-654, doi:10.1038/nrm2208 (2007).

33 Lam, S. S. *et al.* Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nature methods* **12**, 51-54, doi:10.1038/nmeth.3179 (2015).

34 Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of cell biology* **196**, 801-810, doi:10.1083/jcb.201112098 (2012).

35 Chaiet, L. & Wolf, F. J. The Properties of Streptavidin, a Biotin-Binding Protein Produced by Streptomycetes. *Arch Biochem Biophys* **106**, 1-5 (1964).

36 Cheah, J. S. & Yamada, S. A simple elution strategy for biotinylated proteins bound to streptavidin conjugated beads using excess biotin and heat. *Biochem Biophys Res Commun* **493**, 1522-1527, doi:10.1016/j.bbrc.2017.09.168 (2017).

37 McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **86**, 7150-7158, doi:10.1021/ac502040v (2014).

38 Pascovici, D., Handler, D. C., Wu, J. X. & Haynes, P. A. Multiple testing corrections in quantitative proteomics: A useful but blunt tool. *Proteomics* **16**, 2448-2453, doi:10.1002/pmic.201600044 (2016).

39 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10**, 1794-1805, doi:10.1021/pr101065j (2011).

40 Trachsel, C. *et al.* rawDiag: An R Package Supporting Rational LC-MS Method Optimization for Bottom-up Proteomics. *Journal of proteome research* **17**, 2908-2914, doi:10.1021/acs.jproteome.8b00173 (2018).

41 Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534-2536, doi:10.1093/bioinformatics/btn323 (2008).

42 Gatto, L. & Lilley, K. S. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **28**, 288-289, doi:10.1093/bioinformatics/btr645 (2012).

43 Liao, H. *et al.* Quantitative proteomic analysis of cellular protein modulation upon inhibition of the NEDD8-activating enzyme by MLN4924. *Molecular & cellular proteomics : MCP* **10**, M111 009183, doi:10.1074/mcp.M111.009183 (2011).

44 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic acids research* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

45 Schilling, B. *et al.* Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Molecular & cellular proteomics : MCP* **11**, 202-214, doi:10.1074/mcp.M112.017707 (2012).

46 Soucy, T. A. *et al.* An inhibitor of NEDD8-activating enzyme as a new approach to treat cancer. *Nature* **458**, 732-736, doi:10.1038/nature07884 (2009).

47 Aberle, H., Bauer, A., Stappert, J., Kispert, A. & Kemler, R. beta-catenin is a target for the ubiquitin-proteasome pathway. *EMBO J* **16**, 3797-3804, doi:10.1093/emboj/16.13.3797 (1997).

48 Rybak, J. N., Scheurer, S. B., Neri, D. & Elia, G. Purification of biotinylated proteins on streptavidin resin: a protocol for quantitative elution. *Proteomics* **4**, 2296-2299, doi:10.1002/pmic.200300780 (2004).

49 Fukuyama, H. *et al.* On-bead tryptic proteolysis: an attractive procedure for LC-MS/MS analysis of the Drosophila caspase 8 protein complex during immune response against bacteria. *Journal of proteomics* **75**, 4610-4619, doi:10.1016/j.jprot.2012.03.003 (2012).

50 Vashisht, A. A., Yu, C. C., Sharma, T., Ro, K. & Wohlschlegel, J. A. The Association of the Xeroderma Pigmentosum Group D DNA Helicase (XPD) with Transcription Factor IIH Is Regulated by the Cytosolic Iron-Sulfur Cluster Assembly Pathway. *J Biol Chem* **290**, 14218-14225, doi:10.1074/jbc.M115.650762 (2015).

51 Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Current protocols in bioinformatics* **46**, 13 24 11-19, doi:10.1002/0471250953.bi1324s46 (2014).

52 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).

53 McIlwain, S. *et al.* Crux: rapid open source protein tandem mass spectrometry analysis. *Journal of proteome research* **13**, 4488-4491, doi:10.1021/pr500741y (2014).

54 The, M., MacCoss, M. J., Noble, W. S. & Kall, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry* **27**, 1719-1727, doi:10.1007/s13361-016-1460-7 (2016).

55 Reiter, L. *et al.* mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature methods* **8**, 430-435, doi:10.1038/nmeth.1584 (2011).

56 Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524-2526, doi:10.1093/bioinformatics/btu305 (2014).

57 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

58 Oliphant, T. E. Python for Scientific Computing. *Computing in Science & Engineering* **9**, 10-20, doi:10.1109/MCSE.2007.58 (2007).

59 Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic acids research* **45**, D1100-D1106, doi:10.1093/nar/gkw936 (2017).

60 Jarnuczak, A. F. & Vizcaino, J. A. Using the PRIDE Database and ProteomeXchange for Submitting and Accessing Public Proteomics Datasets. *Current protocols in bioinformatics* **59**, 13 31 11-13 31 12, doi:10.1002/cpbi.30 (2017).

61 Chumsae, C. *et al.* Arginine modifications by methylglyoxal: discovery in a recombinant monoclonal antibody and contribution to acidic species. *Anal Chem* **85**, 11401-11409, doi:10.1021/ac402384y (2013).

62 Schiapparelli, L. M. *et al.* Direct detection of biotinylated proteins by mass spectrometry. *Journal of proteome research* **13**, 3966-3978, doi:10.1021/pr5002862 (2014).

63 O'Sullivan, V. J. *et al.* Development of a tetrameric streptavidin mutein with reversible biotin binding capability: engineering a mobile loop as an exit door for biotin. *PloS one* **7**, e35203, doi:10.1371/journal.pone.0035203 (2012).

64 Nagaraj, 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. *Molecular & cellular proteomics : MCP* **11**, M111 013722, doi:10.1074/mcp.M111.013722 (2012).

65 Kearney, P. & Thibault, P. Bioinformatics meets proteomics--bridging the gap between mass spectrometry data analysis and cell biology. *J Bioinform Comput Biol* **1**, 183-200 (2003).

66 Chandramouli, K. & Qian, P. Y. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics* **2009**, doi:10.4061/2009/239204 (2009).

67 Martens, L. *et al.* mzML--a community standard for mass spectrometry data. *Molecular & cellular proteomics : MCP* **10**, R110 000133, doi:10.1074/mcp.R110.000133 (2011).

68 Griss, J. *et al.* The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & cellular proteomics : MCP* **13**, 2765-2775, doi:10.1074/mcp.O113.036681 (2014).

69 Kuenzi, B. M. *et al.* APOSTL: An Interactive Galaxy Pipeline for Reproducible Analysis of Affinity Proteomics Data. *Journal of proteome research* **15**, 4747-4754, doi:10.1021/acs.jproteome.6b00660 (2016).

70 Sheynkman, G. M. *et al.* Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC genomics* **15**, 703, doi:10.1186/1471-2164-15-703 (2014).

71 Liu, G. *et al.* ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nature biotechnology* **28**, 1015-1017, doi:10.1038/nbt1010-1015 (2010).

72 Boekel, J. *et al.* Multi-omic data analysis using Galaxy. *Nature biotechnology* **33**, 137-139, doi:10.1038/nbt.3134 (2015).

73 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research* **46**, W537-W544, doi:10.1093/nar/gky379 (2018).

74 Wang, J. *et al.* MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nature methods* **12**, 1106-1108, doi:10.1038/nmeth.3655 (2015).

75 Frewen, B. & MacCoss, M. J. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Current protocols in bioinformatics* **Chapter 13**, Unit 13 17, doi:10.1002/0471250953.bi1307s20 (2007).

76 Granholm, V. *et al.* Fast and accurate database searches with MS-GF+Percolator. *Journal of proteome research* **13**, 890-897, doi:10.1021/pr400937n (2014).

77 Teo, G. *et al.* SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *Journal of proteomics* **100**, 37-43, doi:10.1016/j.jprot.2013.10.023 (2014).

78 Teo, G. *et al.* SAINTq: Scoring protein-protein interactions in affinity purification - mass spectrometry experiments with fragment or peptide intensity data. *Proteomics* **16**, 2238-2245, doi:10.1002/pmic.201500499 (2016).

79 Fermin, D., Avtonomov, D., Choi, H. & Nesvizhskii, A. I. LuciPHOr2: site localization of generic post-translational modifications from tandem mass spectrometry data. *Bioinformatics* **31**, 1141-1143, doi:10.1093/bioinformatics/btu788 (2015).

80 Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research* **10**, 5354-5362, doi:10.1021/pr200611n (2011).

81 Kohlbacher, O. *et al.* TOPP--the OpenMS proteomics pipeline. *Bioinformatics* **23**, e191-197, doi:10.1093/bioinformatics/btl299 (2007).

82 da Veiga Leprevost, F. *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **33**, 2580-2582, doi:10.1093/bioinformatics/btx192 (2017).

83 Sloggett, C., Goonasekera, N. & Afgan, E. BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics* **29**, 1685-1686, doi:10.1093/bioinformatics/btt199 (2013).

84 Sharma, V., Eng, J. K., Maccoss, M. J. & Riffle, M. A mass spectrometry proteomics data management platform. *Molecular & cellular proteomics : MCP* **11**, 824-831, doi:10.1074/mcp.O111.015149 (2012).