Expanding the proteomics toolbox with intelligent data acquisition and genomic locus protein mapping

Christopher McGann

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Devin K. Schweppe, Chair
Brian J. Beliveau
Michael J. MacCoss

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Expanding the proteomics toolbox with intelligent data acquisition and genomic locus protein mapping

Christopher McGann

Chair of supervisory committee:

Devin K. Schweppe

Department of Genome Sciences

Mass spectrometry-based proteomics has emerged as a cornerstone technology for understanding biological systems, yet significant computational and methodological challenges still exist. This dissertation aims to improve three important areas of need in proteomics: intelligent data acquisition methodologies, peptide identification efficiency, and scalability of methods for characterizing DNA-protein interactions. Chapter 2 addresses the computational demands of modern proteomics by implementing fragment ion indexing in the widely-used Comet search algorithm. Chapter 3 introduces real-time spectral library search (RTLS), an intelligent data acquisition method that leverages whole-proteome spectral libraries to guide instrument decision-making during acquisition. Finally, Chapter 4 presents DNA O-MAP, a scalable method for characterizing locus-specific chromatin interactions that overcomes limitations of existing approaches.

ACKNOWLEDGEMENTS

I'd like to start my acknowledgements with my advisor, Dr. Devin Schweppe. Devin has been an incredibly supportive mentor and has made graduate school a rewarding experience. He's been invaluable to my development both scientifically and professionally and his easy-going unflappability helped transform all the setbacks over the years to opportunities for growth. I joined his lab because I believed it would be a place to both do cutting-edge research as well as a great environment for training and 5 years later I believe that only more strongly. I'm leaving graduate school with immense pride in being the first graduate student of the Schweppe lab.

I'd also like to acknowledge the other members of my thesis committee: Mike MacCoss, Brian Beliveau, David Shechner, and Andrew Stergachis. I feel fortunate to have had the guidance of such exceptional scientists and their help was invaluable. Many projects I contributed to were supervised by Brian Beliveau and his keen problem solving taught me much. To Jimmy Eng and Mike Hoopmann, thank you for all you've taught me and letting me barge into your office to pick your brain on a regular basis.

I owe much to my pre-graduate school mentors as well. Dr. Punit Shah introduced me both to the wonderful world of proteomics and the idea that science could be a fun, rewarding experience. I've yet to take your constant advice of work smart, not hard but I'll get there some day. The proteomics team at Neon Therapeutics was crucial for my development. Jenn Abelin, Daniel Rothenberg, and John Salud were terrific colleagues who inspire me to this day. I owe a special debt of gratitude to Scott Goulding and Terri Addona, they believed in me far beyond what I believed in myself and I would not be in graduate school today without their support and encouragement.

I'd like to thank all the members of the Schweppe Lab: Rose Fields, Valerie Lynch, Meagan Gadzuk-Shea, Sahar Attar, Chelsea Lin, Katarina Vlajic, Erik Bergstrom, Jacob Cogan, Mauri Butzke, Conor Herlihy, Catherine Sniezek, and Kyle Brandt. I could not have asked for a better collection of colleagues and I'm proud to have worked with you all. The proteomics community in Seattle is tremendously supportive and brilliant, the members of the MacCoss, Villen, and Riley labs often provided me both advice and camaraderie. A special shoutout to Nick Riley, Emmajay Sutherland, and Haley Schramm. My collaborations have defined my

experience here and I'm grateful to everyone from UW and Thermo Fisher that I worked with, with special thanks to Yuzhen Liu and Will Barshop. Yuzhen and I had a fantastic journey together through the world of O-MAP and Will's shared enthusiasm of making everything real-time kept me going at times. My friend Taylor Real has been an invaluable part of my grad school experience. I maintained a desk full of knick-knacks and art thanks to her creative and generous spirit.

To my friends on the east coast, namely Charlie Garcia and Gramoz Kondacki, thank you for your dependability and all the memories. Thank you to my amazing family for their love and support. Michelle Sumner always made sure I knew she was thinking of me and is an amazing sister to my mother. The world has known few people as caring as Mary McGann and I am forever indebted to her. Thanks to my younger brother, Dan McGann, I'd never be able to do this without him, he motivates me in a multitude of ways, reminds me how to have fun and relax. He always keeps me grounded and reminds me that no matter what issues I'm having in the lab or with my work it's not nearly as important as basketball. Thank you to my wonderful girlfriend, Dr. Lia Serrano. She is both a caring, supportive, reliable partner as well as a brilliant scientist in her own right. We've gone on a parallel journey through graduate school and her level-headedness, drive, and patience both made the successes more satisfying and reinforced me when things were tough.

Thank you to my stepmother, Sarah for all her unwavering support as I've navigated graduate school, I'm lucky to have her as family. My father is my single greatest source of inspiration and I wouldn't have attempted or succeeded in this endeavor without his assistance. I can't find the words to express how much he has taught me, any success I come by is owed to him. Lastly I'd like to thank my mother, Fran McGann. There is nothing I'm more certain of than her love and pride in me. She provided a foundation that let me weather any storm, I am eternally grateful to her.

DEDICATION

To my parents

TABLE OF CONTENTS

INTRODUCTION..... 1
1.1 Quantitative proteomics..... 1
1.3 Peptide identification with database searching..... 3
1.3 Mass spectrometry acquisition methods..... 4
1.4 Intelligent data acquisition..... 7
1.7 Locus proteomics..... 9
1.6 Organization of this dissertation..... 10
1.6 References..... 10
COMET FRAGMENT-ION INDEXING FOR ENHANCED PEPTIDE SEQUENCING..17
2.1 Introduction..... 17
2.2 Methods..... 18
2.3 Results and Discussion..... 22
2.4 Conclusions..... 30
2.5 References..... 31
REAL-TIME SPECTRAL LIBRARY MATCHING FOR SAMPLE MULTIPLEXED QUANTITATIVE PROTEOMICS..... 36
3.1 Introduction..... 36
3.2 Methods..... 39
3.3 Results and Discussion..... 45
3.4 Conclusions..... 69
3.5 References..... 69
DNA O-MAP UNCOVERS THE MOLECULAR NEIGHBORHOODS ASSOCIATED WITH SPECIFIC GENOMIC LOCI..... 78
4.1 Introduction..... 78
4.2 Methods..... 81
4.3 Results and Discussion..... 93
4.4 Conclusions..... 114
4.5 References..... 114
CONCLUSIONS AND FUTURE DIRECTIONS..... 124
5.1 Conclusions..... 124
5.2 Future directions..... 125

TABLE OF FIGURES

Figure 1-1: Overview of a TMT experimental workflow……2

Figure 1-2: Candidate selection in database searching……4

Figure 1-3: Example of effective MS3 rate……7

Figure 1-4: Real-time search diagram……..9

Figure 2-1: Overview of fragment indexing in Comet……..23

Figure 2-2: Evaluating Comet-FI in various workflows……26

Figure 2-3: Open modification search timing……27

Figure 2-4: Real time search……..29

Figure 2-5: Open modification RTS……..30

Figure 3-1: Overview of RTLS workflow……47

Figure 3-2: Grid search of cosine score weights……50

Figure 3-3: Score comparisons for different proteomics search methods ……51

Figure 3-4: RTLS experimental design……..52

Figure 3-5: Venn diagrams showing the overlap between search results……..54

Figure 3-6: Comparison between empirical and predicted spectral libraries……55

Figure 3-7: Quantifying peptides and proteins from human and yeast in HyPro……..56

Figure 3-8: RTLS with and without FAIMS……..58

Figure 3-9: Comparing IDA methods……59

Figure 3-10: RTLS run with HCD and CID collision energies ……60

Figure 3-11: RTLS methods comparison on fractionated, whole-proteome samples……61

Figure 3-12: Volcano plots from belinostat experiment showing all three cell lines with the different acquisition methods……63

Figure 3-13: Comparison of RTLS methods run on a Thermo Scientific Orbitrap Eclipse Tribrid or a Orbitrap Ascend Tribrid……..64

Figure 3-14: Chimeric spectra and RTLS……..65

Figure 3-15: Percentage of chimeric MS2 spectra…….67

Figure 3-16: SPS match percentage……...…68

Figure 4-1: Overview of DNA O-MAP workflow and label-free quantitative proteomics analysis of telomeres……95

Figure 4-2: DNA O-MAP reveals distinct features of the sub-proteomes at peri-centromeric alpha satellites, telomeres, and the mitochondrial genome……..100

Figure 4-3: Predicted genome-wide binding profile of the pan-alpha probe……...102

Figure 4-4: Replicate analysis of multi-target DNA O-MAP proteomics experiment……104

Figure 4-5: DNA O-MAP efficiently labels single-copy chromatin loop anchors……..106

Figure 4-6: Relative quantitation for the multi-target DNA O-MAP proteomics experiment compared to no-probe control and mtDNA datasets……108

Figure 4-7: Comparison of histone proteins between telomere and pan-alpha probes……109

Figure 4-8: DNA O-MAP biotin purification sequencing of chr3 left, chr3 right, chr10 non-loop anchors, and no-primary-probe control……111

Figure 4-9: DNA O-MAP biotin purification sequencing of multiplexed targeting of chr3 left, chr10 right, chr19 right anchors, and no-primary-probe control in duplicates……....113

Chapter 1

INTRODUCTION

1.1 Quantitative proteomics

The proteome, the collection of proteins in any given organism or cell, is involved in nearly every biological function and therefore of great interest to researchers. Proteomes by nature are highly dynamic, responding to metabolic state, environmental factors, and oncogenic signaling among other stimuli, resulting in a measurable molecular phenotype. As such, systems-level protein quantification provides a path forward to attain deep mechanistic insights and disease characterization.

However, quantifying the proteome is a challenging task due to the complexity and dynamic range of proteins. The approximately 20,000 human genes are theorized to produce over one million proteoforms when taking into account post-translational modifications and variants.[1] Additionally, the dynamic range within a cell can span seven orders of magnitude, with some proteins having only a handful of copies per cell.[2] This can be even more dramatic in certain samples like tissue wherein the proteome has been observed to span eight orders of magnitude or in biofluids like plasma that spans ten orders of magnitude.[3,4]

Mass spectrometry has become the primary analytical tool for this purpose, using the mass-to-charge (m/z) ratio and intensities of peptide ions to identify and quantify proteins, respectively. “Bottom-up” proteomics has been the dominant technique, which generally employs the strategy of subjecting whole proteins to enzymatic digestion, resulting in smaller, more mass spectrometry compatible peptides.[5] Peptide samples are commonly separated by reversed-phase liquid chromatography to reduce the complexity of the ions as they enter the mass spectrometer for analysis. After analysis, peptide level measurements are mapped back to

their originating proteins to generate protein- level quantitation. To improve experimental throughput and data completeness, researchers commonly multiplex proteomic samples by labeling them with an isobaric chemical barcode.[6–8] "Reporter" ions will release from labelled peptides during collision induced fragmentation and can be used to provide relative quantification across samples (Figure 1-1).[9–11] Thus, each fragment ion scan provides both peptide identification and relative quantification information. This technique improves upon both resolution and selectivity limitations for which MS1-based quantification can be vulnerable with increasingly higher plexes. While multiple versions of tags have been developed and successfully implemented, tandem mass tags (TMT/TMTpro) are the most commonly used. Newer versions of these tags allow for up to 35 samples to be quantified simultaneously.[12] "Bottom-up" mass spectrometry based proteomics studies have been used to interrogate many important and varied topics such as mammalian development, aging, drug discovery, and host-pathogen interactions.[13–16]

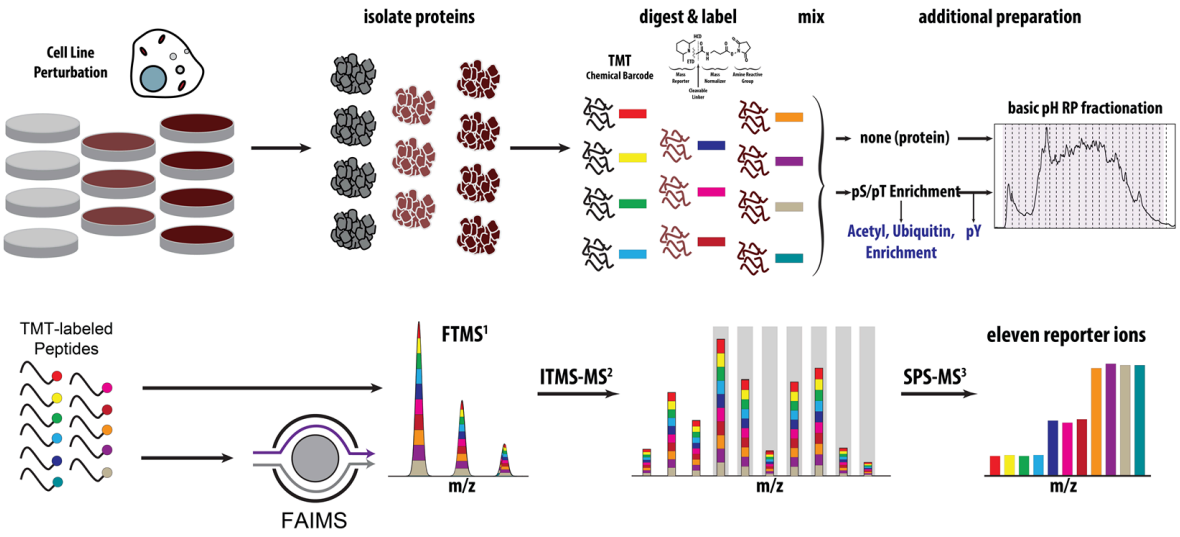


Figure 1-1. Overview of a TMT experimental workflow. Starting with cell lines or tissue samples, proteins are isolated before being digested and being labelled with a unique isobaric

tag. From there, samples can undergo optional enrichment for lower abundant peptide species such as phosphopeptides, acetylated peptides, or ubiquitinylated peptides. TMT-labeled peptides will appear as a single peak in the MS1 precursor peak and MS2 fragment ion peaks but will produce reporter ions that can be used to calculate relative quantitation between conditions.

1.2 Peptide identification with database searching

Peptide sequencing through the interpretation of mass spectra is the backbone of any mass spectrometry-based proteomics workflow. Multiple methods of peptide-spectrum assignment exist, yet they are largely built on the predictability of peptide fragmentation during specified dissociation techniques. When undergoing collision-induced dissociation (CID), peptides typically cleave at the amide bond, resulting in what is referred to as N-terminal "b" and C-terminal "y" ions. While spectra can be manually sequenced, modern mass spectrometers are capable of collecting hundreds of scans per second which necessitates a high-throughput and accurate approach for sequencing peptides. This led to the creation and popularization of "database" searching.[17,18]

In database searching, a protein database, commonly the human proteome, is digested *in silico* to generate a list of all potential peptides. For each peptide in this list, all theoretical fragment ions can be calculated. From there, for each experimentally derived spectrum collected, the best peptide assignment can be calculated by using a score based upon the presence of those theoretical fragment ions. Traditionally, to reduce the number of required comparisons as well as remove impossible matches, the scored candidate peptides are first filtered from the list by their precursor mass (Figure 1-2). [19,20] .

Scoring in database searching has two goals: one, to identify which peptide is most likely to give rise to a specific spectrum and two, to compare the quality of one peptide-spectrum

match to another. Results are commonly filtered to a specified false discovery rate using target-decoy competition.[21,22] Here, database peptides or "target" peptides have their sequences shuffled or reversed to create "decoy" peptides, which act as known false positives. Under the assumption that an incorrect match is equally likely to be from a decoy peptide as an incorrectly assigned target peptide, the number of decoys can be used to estimate the false discoveries returned by the algorithm. Due to the difficult nature of this task and the fact that many primary scoring algorithms are not well calibrated, classifiers such as support vector machines or discriminative models can incorporate myriad spectral quality metrics and features to ultimately estimate well-calibrated false discovery rates [23,24] In order to keep pace with the latest instrumentation and sample types, a major goal of database searching development has been on improving the speed of these searches.[25,26]

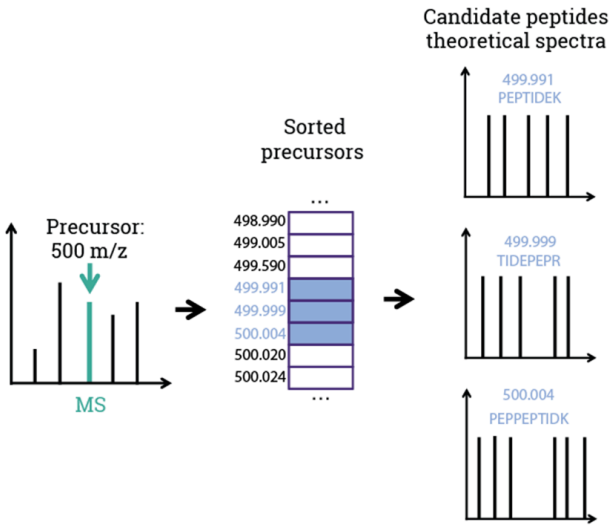


Figure 1-2. Candidate selection in database searching. Example of a precursor m/z being used to select potential peptides from a precursor-indexed database. Only the candidates within the narrow tolerance are scored against the experimental spectrum.

1.3 Mass spectrometry acquisition methods

Proteomics experiments can vary widely in goal, sample composition, and instrumentation availability. Therefore, there are multiple ways to acquire data on a mass spectrometer based upon experimental needs and goals. The three major acquisition methodologies are data-dependent acquisition, data-independent acquisition, and targeted acquisition methods.

Targeted acquisition methods are built around an *a priori* list of mass-to-charge values corresponding to specific peptides of interest.[27] This can be done at the MS1 level but is more commonly done at the MS2 level, isolating and quantifying peptide fragment ions.[28] Targeted methods have unparalleled sensitivity yet are limited in scope as monitoring too many targets can cause the cycle time to grow long enough to negatively affect quantitation. Modern techniques have leveraged new algorithms and instrumentation to push the number of targets per experiment into the thousands.[29] Despite the advantages of targeted methods, they are incompatible with discovery based goals.

Data-independent acquisition (DIA) methods cycle across the desired m/z range with set quadrupole isolation windows, subjecting all precursor ions within the window to fragmentation before collecting an MS/MS scan.[30,31] The unbiased inclusion of precursors for fragmentation provides sensitivity and multiple scans can be used to build curves of fragment-ion intensity values for accurate quantitation.[32–34] The consistency in the precursor isolation regime along with independence from precursor measurement greatly reduces missing values , especially in large studies. However, the MS/MS spectra collected in DIA methods are highly chimeric, i.e. composed of multiple peptides, which increases spectral complexity, complicating identification. Samples multiplexed with isobaric barcodes are not compatible with DIA methods as chimericity

produces reporter ion intensities that are averages of the peptides present rather than individual values. While DIA methods are compatible with short gradient times, their throughput is still ultimately limited by the difficulty associated with using them in a multiplexed setting.

Data-dependent acquisition (DDA) methods aim to isolate individual precursors with narrow isolation windows and collect MS/MS spectra from one peptide at a time.[35] Ions detected in an MS survey scan are selected for isolation and sequencing based upon their inferred charge state, isotopic distribution, intensity, etc., with the prevailing logic being to choose the N most intense ions that pass the specified filters.[36] By selecting high-quality precursors with narrow isolation windows, DDA methods result in spectra that are easier to interpret. Reliance on selection from a survey scan however, can limit sensitivity and introduce stochasticity to the set of peptides that are analyzed in a given run.[37,38] While DDA can quantify proteins using extracted precursor ion chromatograms , it is also compatible with TMT-based quantification.

TMT multiplexing requires high resolution scans to resolve the 2.9 mDa mass difference between mass tags and fully use their multiplexing capabilities.[39] Consequently, TMT samples are primarily analyzed with Orbitrap containing instruments. While DDA minimizes the precursor co-isolation that leads to chimeric spectra, it does not eliminate it and this leads to an effect known as ratio compression.[40] Given co-isolated peptides will produce a reporter ion profile that is roughly the ion current weighted average of their component parts, extreme ratios between reporters are often suppressed by more common, less differentially expressed values, leading to undermeasuring the true fold change differences between conditions. Synchronous precursor selection MS/MS/MS (SPS-MS3)-based methods improve quantitative accuracy by selecting intense MS/MS fragments for additional fragmentation.[41,42]Similar to label-free DDA

methods, TMT multiplexed proteomics suffers from stochasticity in precursor selection leading to issues with missing values.

As acquisition methods continue to develop, the existing branches will coalesce as the researchers look to combine the strengths of all. This can be seen in narrow-window DIA or wide-window DDA methods that start to bridge the gap between DDA and DIA methods.[43–45] Future methods will focus on maximizing the efficiency of the instrument while producing complete, accurately quantified results.

1.4 Intelligent data acquisition

In order to improve the weaknesses of the traditional acquisition methods, "intelligent" data acquisition (IDA) methods seek to dynamically control the instrument in parallel to acquisition.[46,47] Similar to the classic top N selection from DDA, these methods make real-time decisions on collected scans to determine how future scans will be collected. These methods are often custom-made and controlled through application program interfaces (APIs). Intelligent acquisition has been used extensively in targeted proteomics, with either dynamically using retention time information to decide what targets are included or monitoring for synthetic and/or isotopically-heavy peptides to trigger a specific scan event.[48–51] IDA methods have also been beneficial with isobaric-labelled samples, specifically real-time database searching.[52,53] While SPS-MS3 methods are required to achieve the most accurate data, they can be inefficient as the MS3 scan is comparatively long (~86-200ms), and ~50% of them are unused, being dependent on a MS2 scans that did not lead to a confident identification(Figure 1-2).

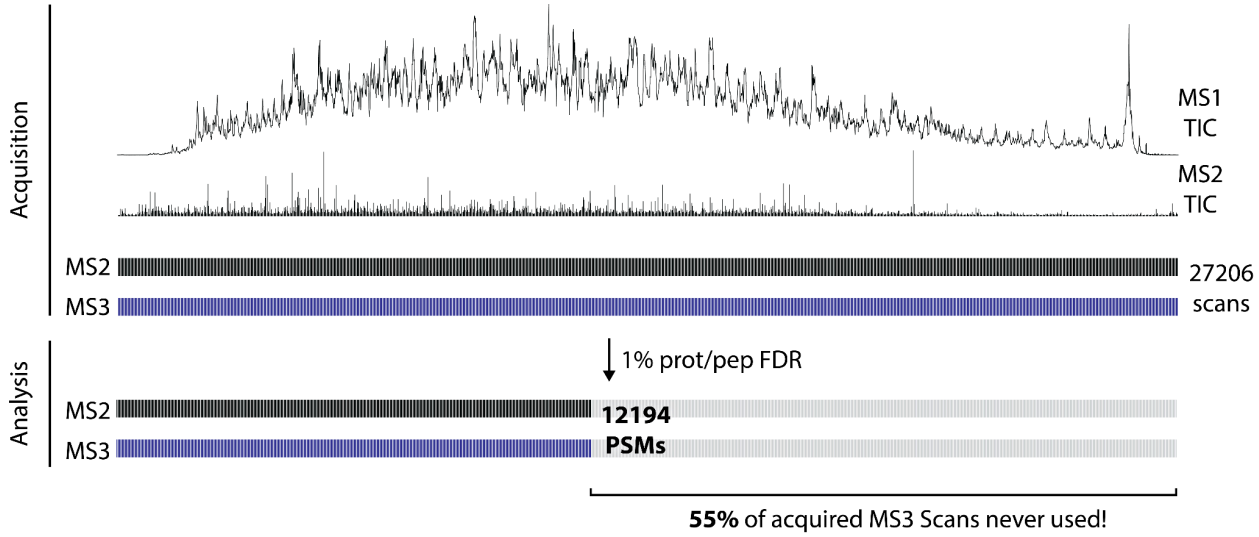


Figure 1-3. Example of effective MS3 rate. In this example raw file, over 27,000 MS2 and MS3 scans were collected. After searching and filtering results to 1% FDR, only 12,000 spectra led to a confident identification, resulting in 55% of all MS3s being unnecessary.

Real-time search (RTS) performs a database search against each MS/MS spectrum as it is acquired, identifying and scoring the best peptide match.[54] An MS3 is only queued if the given PSM passes a score cutoff, saving time on attempting to quantify low-quality identifications unlikely to end up in our final dataset. Additionally, SPS ion selection can be modified to only select fragments that match our putative peptide sequence and that have a TMT-tag attached, rather than prioritizing intensity. More informed SPS ion selection improves quantitative accuracy by further reducing interference. RTS also sets the foundation for secondary filters such as preventing MS3 scans from being performed on proteins that are already confidently quantified.[55] RTS can dramatically improve efficiency in practice, reducing total analysis time by 50% in fractionated samples. Intelligent data acquisition methods present a path forward to continuously improve how proteomics data is collected.

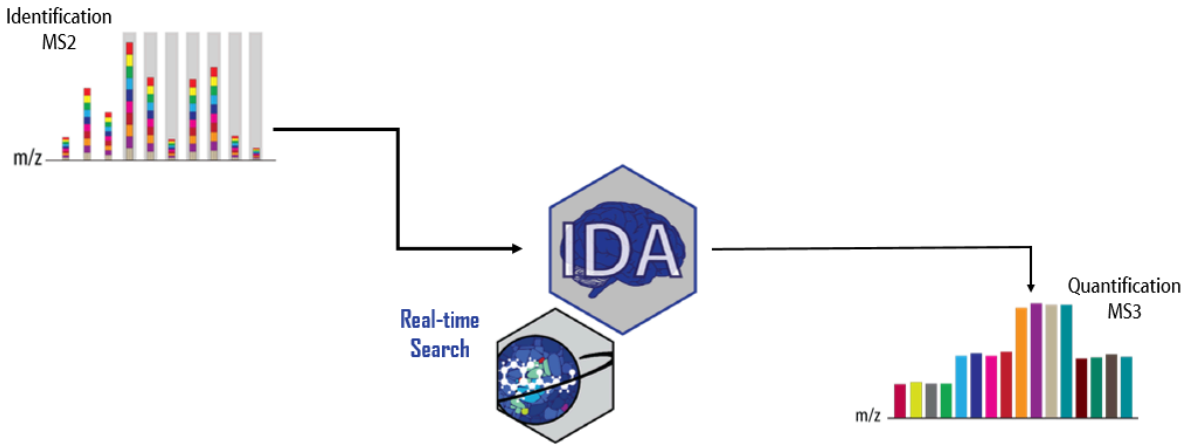


Figure 1-4. Real-time search. RTS operates in between the identification and quantification steps. It acts both as a pass/fail filter for triggering an MS3 as well as which fragments are selected for the MS3 scan.

1.7 Locus proteomics

An important yet challenging subset of proteins to study are those associated with DNA. Proteins govern many important DNA related-functions such as gene regulation, transcription, and DNA repair yet there exists only a few ways to isolate and study these proteins in a high-throughput manner.[56] The methods that do exist operate on a protein-centric basis, cross-linking DNA to protein then using an antibody to pull-down DNA sequences interacting with the protein of interest.[57,58] While powerful, these methods do not allow the interrogation of the local chromatin composition of a specific genomic loci. Current strategies for loci-centric proteomic methods utilize a CRISPR-directed proximity biotinylation approach.[59–61] Catalytically inactive Cas9 (dCas9) is fused to either a peroxidase or biotin ligase, and targeted to a genomic loci with guide RNAs. From there, proximity biotinylation can be induced with hydrogen peroxide and biotin phenol, tagging nearby proteins for affinity purification and

analysis by mass spectrometry. These methods are significant advances yet the high cell input requirements and the need for stably expressed cell lines hinder their throughput. In order to investigate the quantitative protein dynamics of local chromatin in a comprehensive manner, more flexible tools are needed.

1.6 Organization of this dissertation

This dissertation will center around three projects aimed at adding to the repertoire of proteomics technologies. The following chapter describes the improvement of Comet, the popular, free, and open-source database search tool. By implementing a fragment-ion indexed approach, the search times for all sample types, especially large, difficult search spaces are decreased dramatically. Chapter three, describes the development of a new intelligent data acquisition method, real-time spectral library searching. Building spectral libraries both from machine learning models and existing data, scans are quickly and accurately searched to improve multiplexed analysis. Chapter 4 details the development of DNA O-MAP, a novel genomic locus proteomics method that shows improved flexibility and sensitivity to existing methods. Finally, chapter 5 will cover overall conclusions and future directions of my graduate work.

1.6 References

1. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
2. Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13**, 723–726 (2013).
3. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
4. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human

tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).

5. Shuken, S. R. An introduction to mass spectrometry-based proteomics. *J. Proteome Res.* **22**, 2151–2171 (2023).
6. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020).
7. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined tandem mass tag (SL-TMT) protocol: An efficient strategy for quantitative (phospho)proteome profiling using tandem mass tag-synchronous precursor selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).
8. 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).
9. Choe, L. *et al.* 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics* **7**, 3651–3660 (2007).
10. Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).
11. Virreira Winter, S. *et al.* EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15**, 527–530 (2018).
12. Zuniga, N. R. *et al.* Achieving a 35-plex tandem mass tag reagent set through deuterium incorporation. *J. Proteome Res.* **23**, 5153–5165 (2024).
13. Garge, R. K. *et al.* The proteomic landscape and temporal dynamics of mammalian gastruloid development. *bioRxivorg* (2024) doi:10.1101/2024.09.05.609098.
14. Keele, G. R. *et al.* Global and tissue-specific aging effects on murine proteomes. *Cell Rep.* **42**, 112715 (2023).
15. Mitchell, D. C. *et al.* A proteome-wide atlas of drug mechanism of action. *Nat. Biotechnol.* **41**, 1–13 (2023).

16. Schweppe, D. K. *et al.* Host-Microbe Protein Interactions during Bacterial Infection. *Chem. Biol.* **22**, 1521–1530 (2015).
17. 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).
18. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
19. Fenyö, D. & Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75**, 768–774 (2003).
20. Eng, J. K., Fischer, B., Grossmann, J. & Maccoss, M. J. A fast SEQUEST cross correlation algorithm. *J. Proteome Res.* **7**, 4598–4602 (2008).
21. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
22. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
23. Du, X. *et al.* Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications. *J. Proteome Res.* **7**, 2195–2203 (2008).
24. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).
25. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
26. Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and

Quantification at Scale. *J. Proteome Res.* **22**, 3652–3659 (2023).

27. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).
28. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11**, 1475–1488 (2012).
29. Plubell, D. L. *et al.* Development of highly multiplex targeted proteomics assays in biofluids using the Stellar mass spectrometer. *bioRxivorg* (2024) doi:10.1101/2024.06.04.597431.
30. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Mol. Cell. Proteomics* **19**, 1088–1103 (2020).
31. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).
32. Searle, B. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).
33. Ting, Y. S. *et al.* PECAN: Library free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017).
34. Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* **39**, 229–244 (2020).
35. Stahl, D. C., Swiderek, K. M., Davis, M. T. & Lee, T. D. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J. Am. Soc. Mass Spectrom.* **7**, 532–540 (1996).
36. Kalli, A., Smith, G. T., Sweredoski, M. J. & Hess, S. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers. *J. Proteome Res.* **12**, 3071–3086 (2013).

37. O'Connell, J. D., Paulo, J. A., O'Brien, J. J. & Gygi, S. P. Proteome-wide evaluation of two common protein quantification methods. *J. Proteome Res.* **17**, 1934–1942 (2018).
38. Dowell, J. A., Wright, L. J., Armstrong, E. A. & Denu, J. M. Benchmarking quantitative performance in label-free proteomics. *ACS Omega* **6**, 2494–2504 (2021).
39. He, Y. *et al.* TMT-based multiplexed (chemo)proteomics on the Orbitrap Astral mass spectrometer. *Mol. Cell. Proteomics* **24**, 100968 (2025).
40. Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* **12**, 3586–3598 (2013).
41. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).
42. 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 (2014).
43. Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-023-02099-7.
44. Heil, L. R. *et al.* Evaluating the performance of the Astral mass analyzer for quantitative proteomics using data-independent acquisition. *J. Proteome Res.* **22**, 3290–3300 (2023).
45. Frejno, M. *et al.* Unifying the analysis of bottom-up proteomics data with CHIMERYS. *Nat. Methods* **22**, 1017–1027 (2025).
46. Bailey, D. J. *et al.* Instant spectral assignment for advanced decision tree-driven mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8411–8416 (2012).
47. Bailey, D. J., McDevitt, M. T., Westphall, M. S., Pagliarini, D. J. & Coon, J. J. Intelligent data acquisition blends targeted and discovery methods. *J. Proteome Res.* **13**, 2152–2161 (2014).
48. Yu, Q. *et al.* Sample multiplexing-based targeted pathway proteomics with real-time

analytics reveals the impact of genetic variation on protein expression. *Nat. Commun.* **14**, 555 (2023).

49. Remes, P. M., Yip, P. & MacCoss, M. J. Highly multiplex targeted proteomics enabled by real-time chromatographic alignment. *Anal. Chem.* **92**, 11809–11817 (2020).
50. Pollock, S. B. *et al.* Sensitive and quantitative detection of MHC-I displayed neoepitopes using a semiautomated workflow and TOMAHAQ mass spectrometry. *Mol. Cell. Proteomics* **20**, 100108 (2021).
51. Wichmann, C. *et al.* MaxQuant.Live enables global targeting of more than 25,000 peptides. *Mol. Cell. Proteomics* **18**, 982–994 (2019).
52. Erickson, B. K. *et al.* Active instrument engagement combined with a real-time database search for improved performance of sample multiplexing workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).
53. Schweppe, D. K. *et al.* Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).
54. Pelletier, A. R. *et al.* MealTime-MS: A machine learning-guided real-time mass spectrometry analysis for protein identification and efficient dynamic exclusion. *J. Am. Soc. Mass Spectrom.* **31**, 1459–1472 (2020).
55. Budayeva, H. G., Ma, T. P., Wang, S., Choi, M. & Rose, C. M. Increasing the throughput and reproducibility of activity-based proteome profiling studies with hyperplexing and intelligent data acquisition. *J. Proteome Res.* **23**, 2934–2947 (2024).
56. Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).
57. Reimer, J. J. & Turck, F. Genome-wide mapping of protein-DNA interaction by chromatin immunoprecipitation and DNA microarray hybridization (ChIP-chip). Part A: ChIP-chip molecular methods. *Methods Mol. Biol.* **631**, 139–160 (2010).

58. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

59. Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat. Methods* **15**, 437–439 (2018).

60. Gao, X. D. *et al.* C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9-APEX2. *Nat. Methods* **15**, 433–436 (2018).

61. Liu, X. *et al.* In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* **170**, 1028–1043.e19 (2017).

Chapter 2

COMET FRAGMENT-ION INDEXING FOR ENHANCED PEPTIDE SEQUENCING

This chapter is adapted from:McGann, C. D., Bergstrom, E. J., Sharma, V., Heil, L. R., Yu, Q., Eng, J. K., & Schweppe, D. K. Comet fragment-ion indexing for enhanced peptide sequencing. J. Proteome Res., 2025.

2.1 Introduction

Mass spectrometry is capable of producing thousands of tandem mass spectra per minute.[1] Database searching has remained the most common method of identifying peptides from these acquired spectra. Scoring each experimental spectrum against a database of in-silico generated theoretical peptides has proven to be a robust and sensitive approach that has improved over time.[2] Descending from the original database search algorithm, Comet, has remained one of the most popular search tools for over a decade.[3–5] Comet's free and open-source nature has led to its inclusion in a myriad of proteomics workflows and projects[6–15], establishing itself as a hallmark of the field. Comet has been both a critical tool for individual users and a key component of several widely-used proteomics platforms, such as quantms[16], Galaxy[17], and OpenMS.[7] As the field of mass spectrometry proteomics has expanded, so has the application of database searching. With this expansion comes the need for tools to handle the latest developments, driving ever increasing numbers of samples and spectra that must be analyzed. Techniques such as open-modification searching, non-specific immunopeptidomic searches, proteogenomic searching, and real-time searching are exciting but impose significant

computational demands. Alongside the ongoing advancements in instrumentation, this creates a pertinent need to improve the efficiency of database peptide spectral matching.

Fragment ion indexing has been a major advancement in speeding up the search process.[18] Originally implemented by MSFragger, this inverted index style approach involves in-silico digestion of proteins into a database of theoretical fragment ions, allowing the filtering of candidate peptides based on the presence of these ions (in addition to filtering based on the precursor tolerance). Fragment ion indexing significantly reduces spectral scoring time and has been effectively applied to tasks such as open-modification searches.[19] This strategy has been adopted by other tools as well, achieving impressive search speeds.[20] In this study, we adapt a fragment ion indexed approach for the widely-used, open-source Comet search algorithm, significantly accelerating its search capabilities (Figure 2-1A). This enhancement enables Comet to keep pace with new applications and modern instrumentation, maintaining its relevance as a key tool in proteomics research.

2.2 Methods

Fragment ion indexing in Comet

MS/MS spectra are pre-processed by representing peaks in an 1D array, binning masses based on the fragment bin tolerance/offset parameters. To generate the fragment ion index, Comet first generates a peptide index file, denoted with an ".idx" file name extension, based on search/digest parameters (Figure 2-1A). To perform a fragment ion index search, the user specifies the peptide index file instead of a standard FASTA as the query database. The peptide index contains peptide sequences, their unmodified masses, file positions of each parent protein within the original protein sequence FASTA database, and combinatorial bitmasks representing potential variable modification positions. At search time, a user controlled setting can direct

Comet to calculate fragment ion m/z (y3-yn, b3-bn) only for theoretical peptides with precursor m/z's detected in the input spectrum and bins fragments by the same tolerance used to pre-process spectra. The fragment ion index is segmented per search thread. And to mitigate over-populating any specific index entry, the index is further divided into a large number of precursor m/z bins. Binned fragments are then queried against the fragment ion index and the number of matched ions for any potential peptide hits are incremented. Candidate peptides are filtered based on the number of matched ions before undergoing full XCorr[21] scoring and E-value[22] calculation. Output files contain the same content as a non-indexed search and are compatible with the same downstream workflows.

Analysis of publicly available data

To evaluate the performance of Comet's newly implemented indexing strategy, publicly available datasets (PXD016766[34], PXD046453[25], PXD013649[36], PXD029860[35], and PXD019853[37]) retrieved from ProteomeXchange[38] were searched with and without a fragment ion index. Searches were run on a desktop computer (Intel i9-11900) using 12 threads and 64 GB memory on Windows Subsystem for Linux. Peptide-spectrum matches (PSMs) were exported in the Percolator input file format and re-scored using mokapot[39], a Python implementation of the widely used Percolator approach for target-decoy discrimination.[40] For false discovery rate estimation, mokapot [39] was used to aggregate multiple features into a single score, using a Support Vector Machine classifier, to increase the sensitivity of peptide detections. If applicable, protein inference was also performed using mokapot. PSM, peptide, and protein identifications using each search mode were evaluated using Comet's XCorr score, E-value score, and the

mokapot post-processing score.[39] In the Philosopher[41] open search, the recommended workflow was used.

Data collection for real-time search analyses

Data for some real-time search benchmarks was collected in house. Human+yeast and human phosphoproteomics samples were prepared as previously described.[42–44] For the human+yeast sample, peptides were eluted over 60 min gradients running from 96% Buffer A (5% acetonitrile, 0.125% formic acid) and 4% buffer B (95% acetonitrile, 0.125% formic acid) to 30% buffer B. Sample eluate was electrosprayed (2700 V) into a Thermo Scientific Orbitrap Eclipse mass spectrometer for analysis. High field asymmetric waveform ion mobility spectrometry (FAIMS) was set at "standard" resolution, 4.6 L/min gas flow, and 3 CVs: −40/–60/–80 were used. MS1 scans were conducted at 120,000 resolving power with a 50 ms max injection time, and the AGC target set to 100%. Peaks from the MS1 scans were filtered by intensity (minimum intensity $>5 \times 10^3$), charge state ($2 \leq z \leq 6$), and detection of a monoisotopic mass (monoisotopic precursor selection, MIPS). Dynamic exclusion was used with a duration of 60 s, repeat count of 1, mass tolerance of 10 ppm, and the "exclude isotopes" option checked. For each MS1, 8 data-dependent MS/MS scans were collected. MS/MS scans were conducted in the linear ion trap with the "rapid" scan rate, 50 ms max injection time, AGC target set to 200%, CID collision energy of 35% with 10 ms activation time, and 0.5 m/z isolation window. SPS ions were set to 10 and MS3 scans were performed at a resolving power of 50,000, with an HCD collision energy of 45%, AGC of 200%, with a maximum injection time of 200 ms.

Phosphopeptides were eluted over a 120 min gradient running from 94% Buffer A (5% acetonitrile, 0.125% formic acid) and 4% buffer B (80% acetonitrile, 0.125% formic acid) to 28% buffer B. Sample eluate was electrosprayed (2700 V) into a Thermo Scientific Orbitrap

Ascend mass spectrometer for analysis. High field asymmetric waveform ion mobility spectrometry (FAIMS) was set at "standard" resolution, 4.6 L/min gas flow, and 3 CVs: −40/–60/–80 were used. MS1 scans were conducted at 120,000 resolving power with a 251 ms max injection time, and the AGC target set to 100%. Peaks from the MS1 scans were filtered by intensity (minimum intensity $>5 \times 10^3$), charge state ($2 \leq z \leq 6$), and detection of a monoisotopic mass (monoisotopic precursor selection, MIPS). Dynamic exclusion was used, with a duration of 90 s, repeat count of 1, mass tolerance of 10 ppm, and the "exclude isotopes" option checked. For each MS1, 8 data-dependent MS/MS scans were collected. MS/MS scans were conducted in the linear ion trap with the "rapid" scan rate, 75 ms max injection time, AGC target set to 250%, HCD normalized collision energy of 32% and 0.5 m/z isolation window. Phosphoproteomic real-time search database settings included a precursor tolerance of 20 ppm, max 2 missed cleavages, 7-50 amino acids in length, oxidation on M and phosphorylation on S/T/Y as variable modifications, maximum 3 variable mods per peptide. Immunopeptidomic real-time search database settings included a precursor tolerance of 20 ppm and a non-enzymatic digest of 8 to 15 amino acids long. Real-time search results were simulated using the publicly available real-time search option for Comet through in-house platform, Orbiter[15].

All searches were performed against a target-decoy database where the decoy sequences were composed of reversing each protein sequence and appending those reverse sequences to the target sequence database. After ranking the Comet search results by either the XCorr, E-value, or mokapot[39] score FDR was calculated using target and decoy peptides as $FDR = \frac{n_{decoys}}{n_{targets}}$ when using XCorr and E-value or $FDR = \frac{n_{decoys}+1}{n_{targets}}$ when using mokapot.

Code Availability

Fragment ion indexed Comet is freely available under an Apache 2.0 license with version v2024.02.0 and newer on Github, https://github.com/UWPR/Comet. Classic search can be run the same as previous versions. A fragment ion index search is invoked by specifying a peptide index file instead of a FASTA file for the search database. Precursor search tolerance parameters were updated to allow asymmetric windows. Minimum number of matched fragment ions for reporting and scoring can be set with the "fragindex_min_ions_report" and "fragindex_min_ions_score" parameters, respectively.

2.3 Results and Discussion

The main cross correlation (XCorr) scoring function, which is well-established in the field[2,3,21], was maintained moving from Comet to Comet-FI. As such, comparison of XCorr values for peptide-spectral matches (PSMs) from the same analytical run for Comet and Comet-FI resulted in near-identical values (Figure 2-1B). Extending this to a twelve fraction TMT-labeled dataset, we detect a very similar number of filtered PSMs for both linear ion trap and Orbitrap spectra at a 1% FDR as calculated by mokapot. Notably, using Comet-FI, these searches finished 1.7-fold faster for the ITMS2 data and 2.3-fold faster for the HRMS2 data with no significant change in total filtered PSMs (Fig. 2-1C, D).

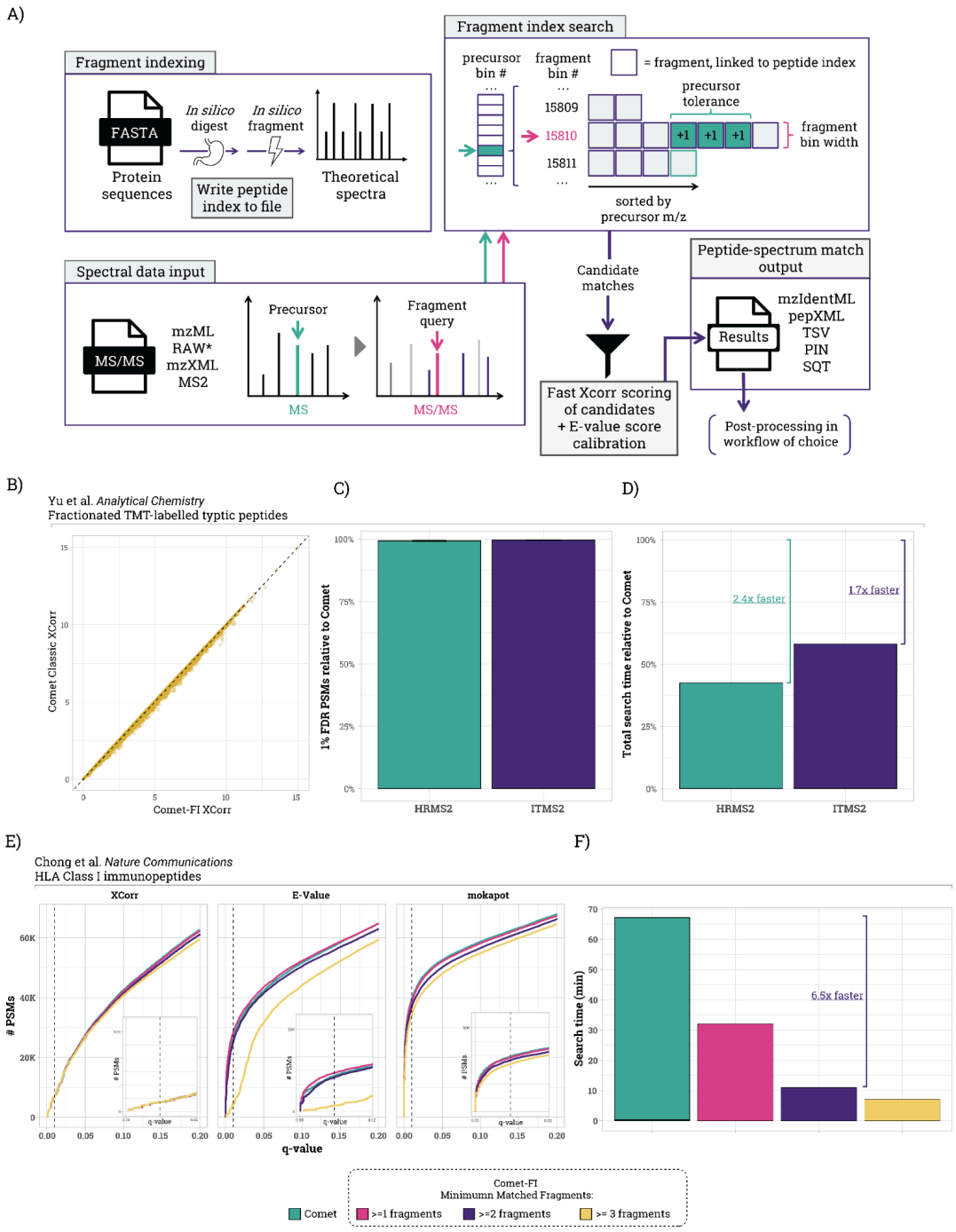


Figure 2-1. Overview of fragment indexing in Comet. A) Graphical workflow describing the new indexing strategy integrated into Comet. B) Comparison of XCorr values between Comet and Comet-FI for the same peptide spectral match. C) Number of FDR filtered peptide spectral matches for TMT-labeled fractionated data using Comet-FI as compared to baseline with Comet. D)Total search time relative to Comet baseline for same dataset E) PSMs at different q-values as

determined by XCorr, E-value, or mokapot scores for HLA Class I data. The dotted line indicates a q-value of 0.01 (or 1%). Inset plots zoom in on the q-value range from 0 to 0.02. Classic search is precursor indexed and has no minimum matched fragments. F) Search times for the raw files shown in E.

Comet calculates a calibrated score called the expectation value (E-value) which is better for target-decoy discrimination and FDR estimation. The E-value is calculated based on fitting a linear regression to the survival function of the XCorr score distribution for a given spectrum[22]. Fragment ion indexing allows for candidate peptides with matched fragment counts below a threshold to not be scored. We've observed that despite the low chance that a peptide with 1 to 2 matched fragments is a correct identification, skipping these peptides alters the XCorr score distributions and can affect the discriminating power of the E-value (Fig. 2-1E). Peptidomic studies often require searches unconstrained by enzyme specificity, where all possible unique peptides within a given length range (e.g., 8-15 amino acids) are considered. The large pool of candidate peptides provides a good dataset for evaluating the relationship between the minimum number of fragment ions considered, the E-values reported, and the total sensitivity for PSM detection. Benchmarking on HLA-I data with challengingly large search spaces due to non-enzymatic digestions, we evaluated how changing the minimum number of fragment ions required for scoring affects the number of peptide spectral matches across the range of q-values using XCorr, E-value, or mokapot's discriminant score to rank PSMs. Using E-value alone with two or more fragment ions severely reduced the number of PSMs passing FDR filters. (Fig. 2-1E). As such, it is inadvisable to use an E-value in isolation for FDR determination. However, when used with programs like mokapot or Percolator, there is very little difference between 0, 1,

and 2 matched ions. Using the default parameters with a minimum of two matched fragments, Comet-FI reduced the runtime of HLA peptide searches by 6.5-fold (Fig. 2-1F).

The latest generation of mass spectrometers can acquire spectra at 300 Hz, producing data at a rate that makes post-hoc peptide identification one of the rate limiting steps in the complete analysis workflow[23]. To test Comet-FI on these modern instruments with a large and challenging search space, we analyzed data from Guzman et al. 2024 (PXD046453)[24,25] collected with a wide 2 Th precursor isolation window DDA method on the Orbitrap Astral. Based on the isolation widths, we used Comet-FI with a precursor tolerance of ±1 Th (vs the default Comet tolerance of ± 20 ppm). At 1% FDR, we observed a 3.6% boost in identified PSMs for Comet-FI compared to Comet (Fig. 2-2A). More importantly, Comet-FI was 3.43-fold faster than Comet when searching these Astral data (Comet 155.6 min; Comet-FI 45.3 min) (Fig. 2B). Comet-FI's enhanced search speed means that post hoc searching can now keep up with the increasing speed of modern, high-speed instruments.

Searches with a very wide precursor tolerance (hundreds of daltons) have been used to identify genetic and chemical variants of canonical peptide sequences [19,26,27]. Open searching dramatically increases the candidate peptide sequence space, necessitating improved search speeds [18,20]. Thus, while Comet has been used for open search pipelines[28], the compute time was not practical on common laboratory workstations. To evaluate the utility of Comet-FI for open searching, we used the Philosopher software pipeline which includes both Comet and the closed-source fragment indexing search engine, MSFragger[18]. Comet-FI for open searching was able to generate data consistent with MSFragger's established workflow (Fig. 2-2C). Importantly, Comet-FI performed open-searching ~50-times faster than Comet. Search times were previously reported to be over 151-times slower than MSFragger. While MSFragger's

closed-source search algorithm was still faster than Comet-FI, the difference in speeds for open-searching was reduced to 3.8-fold (Fig. 2-3, average of 21 min for MSFragger and 80 min per file for Comet-FI). Thus, Comet-FI's open-source fragment ion indexing implementation can readily be used for open searching and easily integrated anywhere that Comet has classically been used before. We note that MSFragger has been actively developed since 2017. Thus, ongoing Comet-FI code development, optimizations, and improvements are on track to offer a premier, open-source alternative to other fragment-ion indexing approaches.

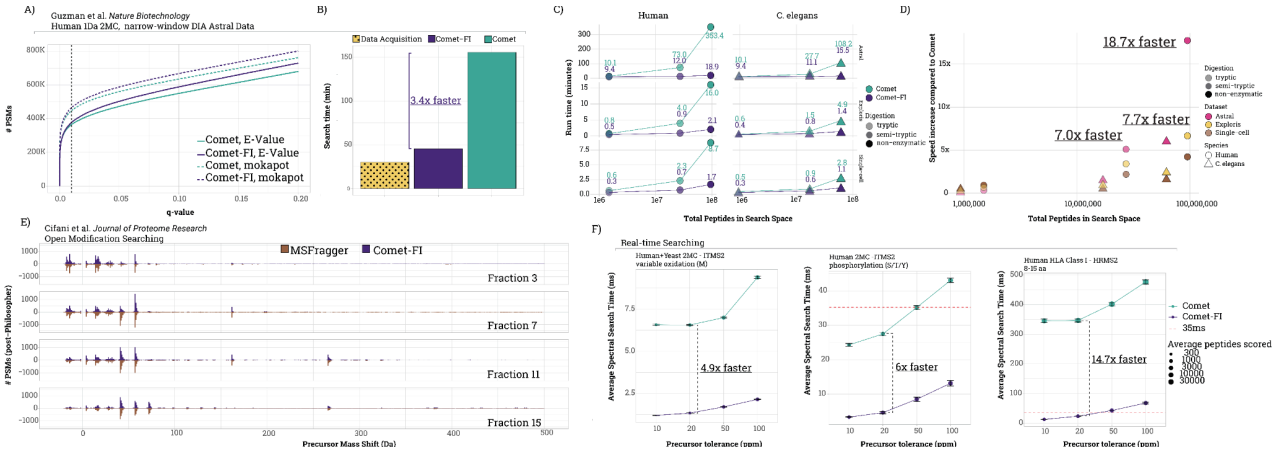


Figure 2-2. Evaluating Comet-FI in various workflows. A) Peptide spectral matches at different q-values using either E-value or mokapot score for Comet and Comet-FI using narrow window DIA data collected on an Astral. B) Total search times for the same data using Comet and Comet-FI with the instrument acquisition time also plotted. C) Relationship between run time and total peptides in search space of different datasets using tryptic, semi-tryptic, and non-enzymatic digestion settings (see Table S1 for file names). D) The relative increase of single file search speed increases compared to classic Comet from the data portrayed in C. E) Precursor mass shift on peptide spectral matches from running Philosopher with Comet-FI or MSFragger. -1.5 to 3.5 Da are excluded from the plot for readability. Fractions 3, 7, 11, and 15 are plotted. F)

Real-time search time comparisons for various data types and search parameters as described above each plot. Red dotted line at 35 ms is the standard MS2 fill time as part of an RTS-MS3 based workflow[34]. Data in F were collected in this study (left and middle panels) or Stopfer et al. (PXD029860, right panel)[35].

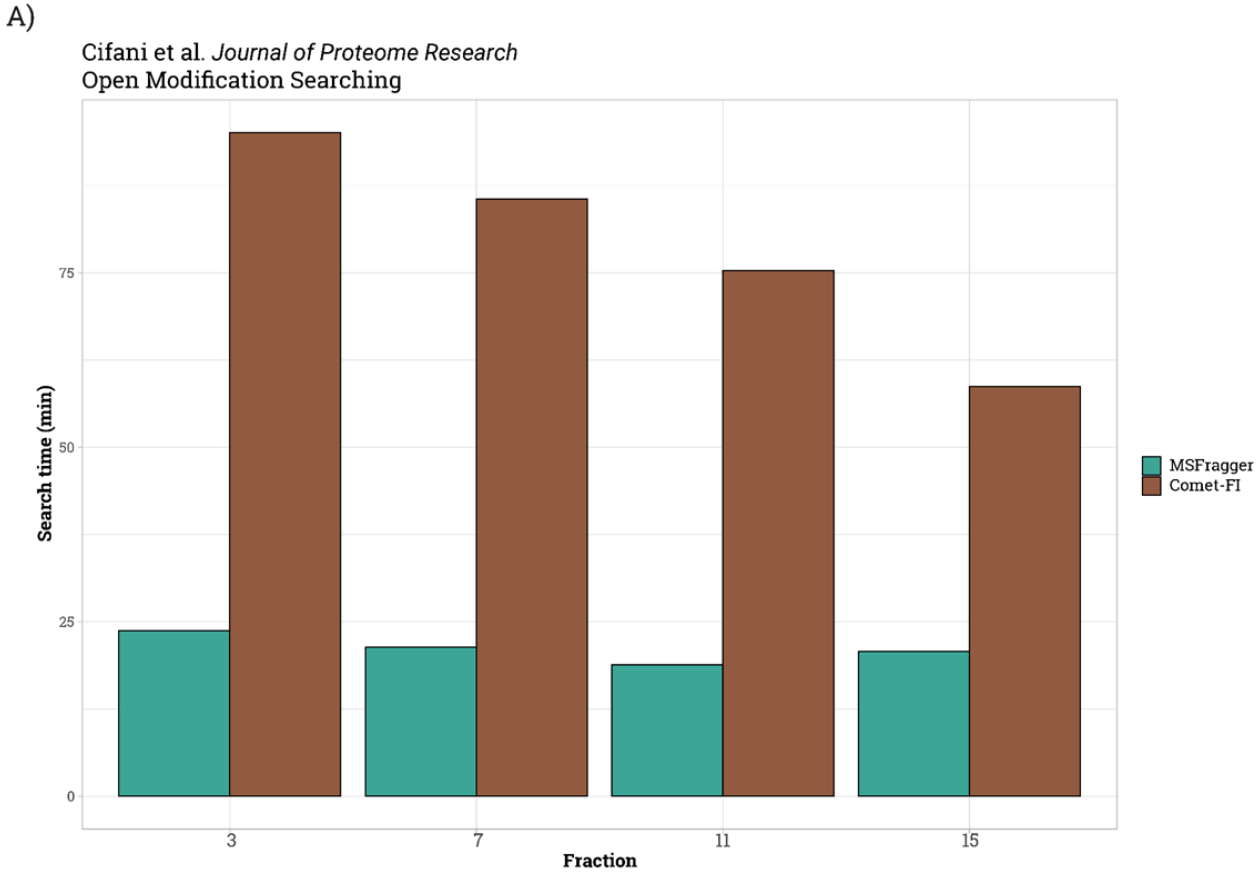


Figure 2-3. Open modification search timing. A) Search time in minutes from running open search on fractions from Cifani et al. with MSFragger and Comet-FI.

Real-time searching (RTS), i.e., simultaneous identification of peptides with instrument acquisition, allows mass spectrometric methods to use peptide sequence information to inform spectral acquisition [15,29–32] with applications to isobarically multiplexed samples (TMT-labeled) and single-cell proteomics[33]. Owing to the open-source codebase, Comet was the first

full-featured algorithm used for TMT-based RTS methods and the first RTS integrated into vendor instrument control software[34]. While Comet's classic implementation worked well for canonical proteomics applications (single species database, unmodified peptide sequences, minimal missed cleavages, sub-50 ppm precursor tolerances, etc.), for applications requiring larger sequence search spaces, the spectral search time often exceeded the MS2 acquisition time (35 ms). Therefore, it could not be perfectly parallelized with instrument acquisition. This was particularly true for searches with multiple missed cleavages, wide precursor windows, variable modifications (e.g. phosphoproteomics), and immunopeptidomics searches for HLA-bound peptides. Comet-FI integrated in the Orbiter real-time search platform[15] reduced the median single-spectrum search times for RTS by 4.9-fold for tryptic peptide searches (Fig. 2-2D,E, 2-4), 6-fold for phosphopeptide searches (Fig. 2-2F), and 14.7-fold for HLA Class I peptide searches (Fig. 2-2G) at 20 ppm precursor tolerance. Single spectrum search times that are commensurate with MS2 acquisition rates for both unmodified and modified searches, as well as HLA-I derived peptides, are now possible with Comet-FI for RTS, opening the door for biologically-aware instrument control.

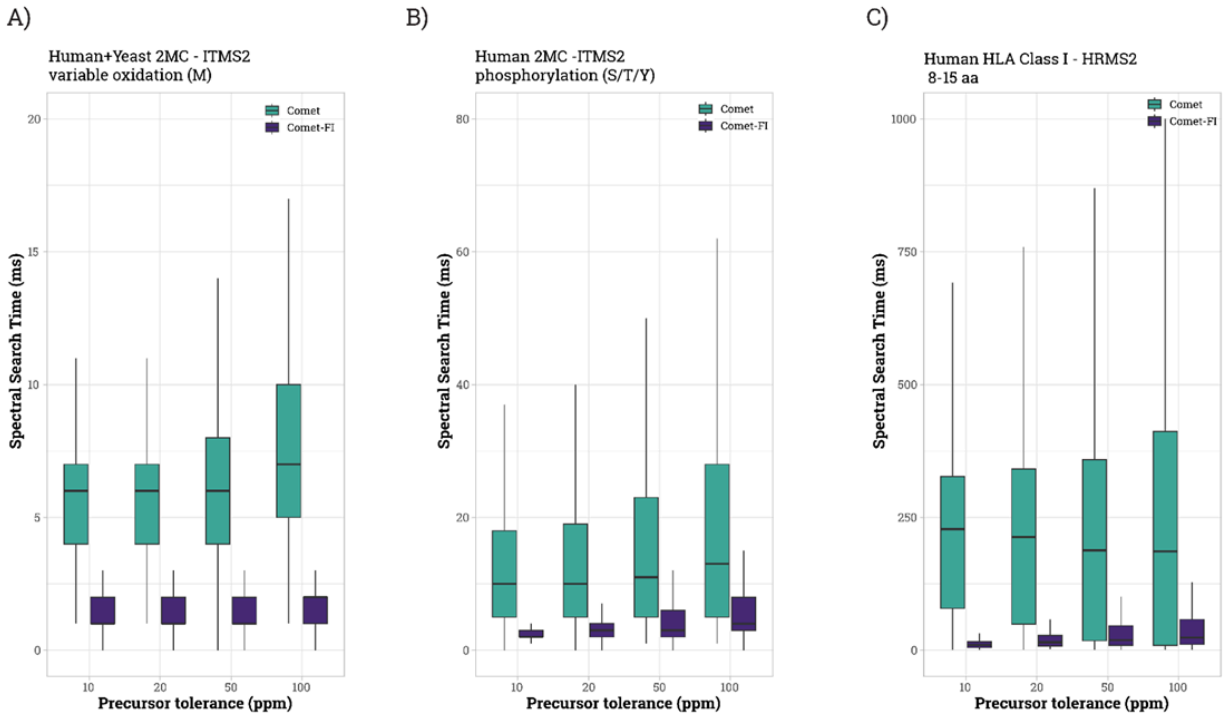


Figure 2-4. **Real time search**. A-C) Boxplots of RTS spectral search time for different search parameters.

To further illustrate the relationship between diverse search spaces due to database size and enzymatic constraints, we queried one file of the TMT-labeled dataset against yeast and human databases using three different enzyme constraints (fully tryptic, semi-tryptic, and non-enzymatic). For this analysis, we compared both the total number of peptides as the total search space and the number of PSMs that were eventually scored by the XCorr calculation. As shown in Figure 2-3, the search space grows significantly as the enzyme constraint is relaxed. This analysis shows that constraints applied during fragment-ion indexing (e.g., number of required fragments) greatly reduces the number of PSMs that are eventually scored compared to classic Comet. In the case of the human database searches, there is a ~400 to ~800-fold reduction in PSMs that are scored by XCorr when fragment-ion indexing is applied. The result of

the reduced scoring burden was an 18-fold increase in search speeds for non-enzymatic searches with Comet-FI compared to classic Comet (Fig. 2-2).

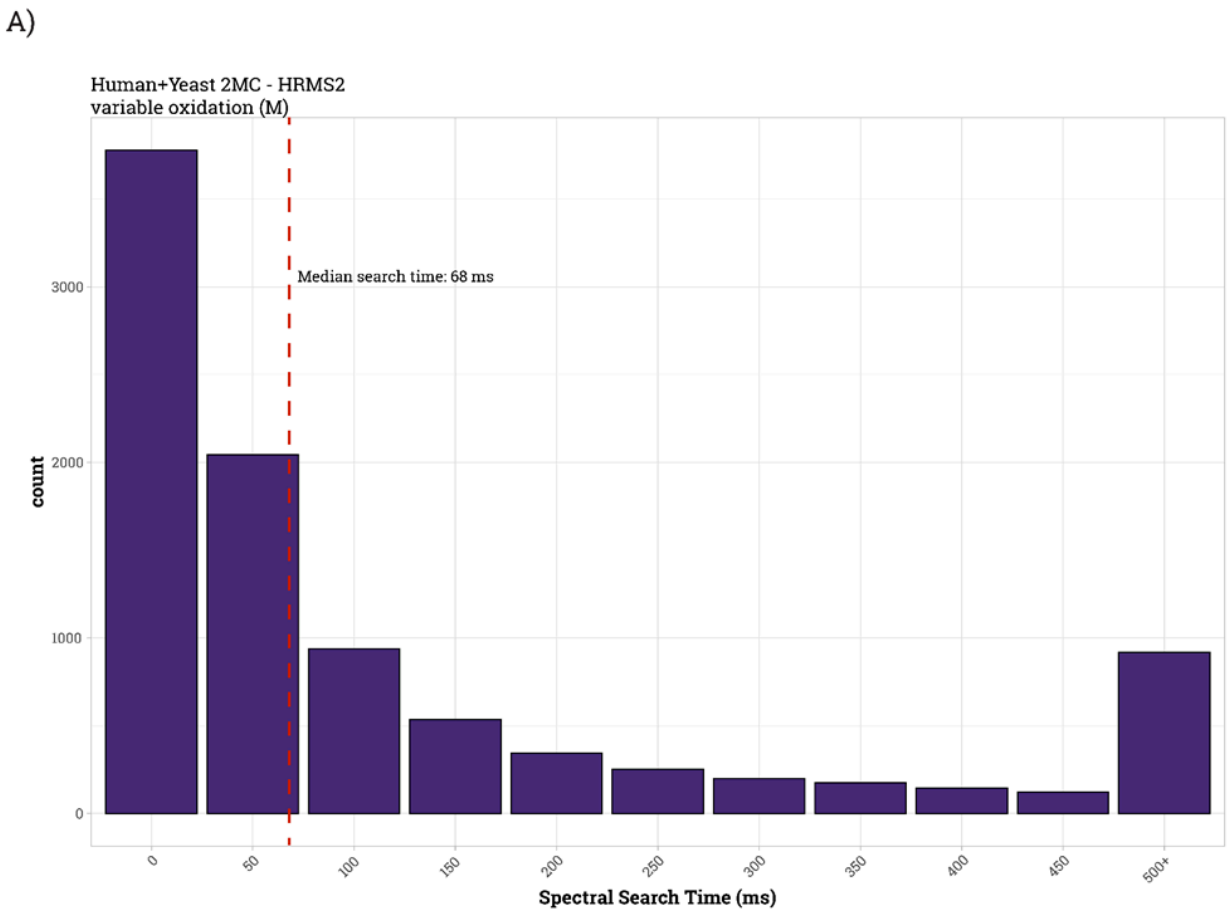


Figure 2-5. Open Modification RTS. Histogram of search times for open modification (-150 to +500 Da) real-time searching. Red dotted line is at the median search time.

2.4 Conclusions

The implementation of fragment ion indexing in the open-source Comet search algorithm significantly enhances post hoc and real-time spectral searching for proteomics analyses across data and instrument types. Comet-FI accelerates search times, enables practical open searching, and increases the real-time analysis capabilities with adaptable and extensible code based on

well-established scoring and spectral matching. Comet's open-source code, ecosystem of developers, and track record of diverse implementations will ensure that the enhanced performance and versatility of Comet-FI will provide ongoing benefit for the proteomics research community. These advancements position the Comet search engine to keep pace with the ongoing evolution of the proteomics field.

2.5 References

1. Stewart, H. I. *et al.* Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis. *Anal. Chem.* **95**, 15656–15664 (2023).
2. 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).
3. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
4. Eng, J. K., Hoopmann, M. R. & Jahan, T. A. A deeper look into Comet—implementation and features. *Journal of the* (2015).
5. Tabb, D. L. The SEQUEST family tree. *J. Am. Soc. Mass Spectrom.* **26**, 1814–1819 (2015).
6. Deutsch, E. W., Mendoza, L., Shteynberg, D. & Farrah, T. A guided tour of the Trans‐Proteomic Pipeline. (2010).
7. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
8. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11**, 996–999 (2011).

9. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33**, 22–24 (2015).
10. Orsburn, B. C. Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **9**, 15 (2021).
11. McIlwain, S. *et al.* Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **13**, 4488–4491 (2014).
12. Carvalho, P. C. *et al.* Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. *Nat. Protoc.* **11**, 102–117 (2016).
13. Liu, G. *et al.* ProHits: integrated software for mass spectrometry–based interaction proteomics. *Nat. Biotechnol.* **28**, 1015–1017 (2010).
14. Winkler, R. MASSyPup--an “out of the box” solution for the analysis of mass spectrometry data. *J. Mass Spectrom.* **49**, 37–42 (2014).
15. Schweppe, D. K. *et al.* Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).
16. Dai, C. *et al.* Quantms: A cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data. *Nat. Methods* **21**, 1603–1607 (2024).
17. Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res.* **52**, W83–W94 (2024).
18. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
19. Yu, F. *et al.* Identification of modified peptides using localization-aware open search. *Nat. Commun.* **11**, 4065 (2020).

20. Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *J. Proteome Res.* **22**, 3652–3659 (2023).
21. Eng, J. K., Fischer, B., Grossmann, J. & Maccoss, M. J. A fast SEQUEST cross correlation algorithm. *J. Proteome Res.* **7**, 4598–4602 (2008).
22. Fenyö, D. & Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75**, 768–774 (2003).
23. Peters-Clarke, T. M., Coon, J. J. & Riley, N. M. Instrumentation at the leading edge of proteomics. *Anal. Chem.* **96**, 7976–8010 (2024).
24. Matzinger, M. *et al.* Micropillar arrays, wide window acquisition and AI-based data analysis improve comprehensiveness in multiple proteomic applications. *Nat. Commun.* **15**, 1019 (2024).
25. Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-023-02099-7.
26. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **33**, 743–749 (2015).
27. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
28. Bagwan, N. *et al.* Comprehensive quantification of the modified proteome reveals oxidative heart damage in mitochondrial heteroplasmy. *Cell Rep.* **23**, 3685-3697.e4 (2018).
29. Bailey, D. J., McDevitt, M. T., Westphall, M. S., Pagliarini, D. J. & Coon, J. J. Intelligent

data acquisition blends targeted and discovery methods. *J. Proteome Res.* **13**, 2152–2161 (2014).

30. Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J. & Mann, M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol. Cell. Proteomics* **11**, M111.013185 (2012).
31. Erickson, B. K. *et al.* Active instrument engagement combined with a real-time database search for improved performance of sample multiplexing workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).
32. McGann, C. D. *et al.* Real-time spectral library matching for sample multiplexed quantitative proteomics. *J. Proteome Res.* **22**, 2836–2846 (2023).
33. Furtwängler, B. *et al.* Real-time search-assisted acquisition on a tribrid mass spectrometer improves coverage in multiplexed single-cell proteomics. *Mol. Cell. Proteomics* **21**, 100219 (2022).
34. Yu, Q. *et al.* Benchmarking the Orbitrap Tribrid Eclipse for Next Generation Multiplexed Proteomics. *Anal. Chem.* **92**, 6478–6485 (2020).
35. Stopfer, L. E. *et al.* MEK inhibition enhances presentation of targetable MHC-I tumor antigens in mutant melanomas. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2208900119 (2022).
36. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
37. Cifani, P. *et al.* Discovery of Protein Modifications Using Differential Tandem Mass Spectrometry Proteomics. *J. Proteome Res.* **20**, 1835–1848 (2021).
38. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).

39. Fondrie, W. E. & Noble, W. S. mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *J. Proteome Res.* **20**, 1966–1971 (2021).

40. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

41. da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870 (2020).

42. Navarrete-Perea, J., Gygi, S. P. & Paulo, J. A. HYpro16: A two-proteome mixture to assess interference in isobaric tag-based sample multiplexing experiments. *J. Am. Soc. Mass Spectrom.* **32**, 247–254 (2021).

43. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined tandem mass tag (SL-TMT) protocol: An efficient strategy for quantitative (phospho)proteome profiling using tandem mass tag-synchronous precursor selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).

44. Paulo, J. A., Navarrete-Perea, J. & Gygi, S. P. Multiplexed proteome profiling of carbon source perturbations in two yeast species with SL-SP3-TMT. *J. Proteomics* **210**, 103531 (2020).

Chapter 3

REAL-TIME SPECTRAL LIBRARY MATCHING FOR SAMPLE MULTIPLEXED QUANTITATIVE PROTEOMICS

This chapter is adapted from: McGann, C. D., Barshop, W. D., Canterbury, J. D., Lin, C., Gabriel, W., Huang, J., Bergen, D., Zabrouskov, V., Melani, R.D., Wilhelm, M., McAlister G.C., & Schweppe, D. K. (2023). Real-time spectral library matching for sample multiplexed quantitative proteomics. Journal of proteome research, 22(9), 2836-2846

3.1 Introduction

Nearly every data-dependent analysis suffers from stochastic precursor selection effects. These effects reduce run-to-run coverage of the proteome, increase the number of missing values, and limit quantitative access to the proteome[1]. This problem is particularly challenging for methods that require longer scan or fill times to reach necessary resolutions or sensitivity. For example, the most accurate methods for sample multiplexed quantitative proteomics require tertiary quantitative scans - synchronous precursor selection (SPS)-MS3 scans[2,3]. Acquisition of these scans leads to multifold increases in the time necessary to quantify proteomics samples at sufficient depth (~8000 proteins).

To combat these challenges, intelligent data acquisition (IDA) methods have been employed to improve both the throughput and quantitative accuracy of sample multiplexed proteomics workflows[4,5]. Initially IDA methods were largely focused on optimizing fragmentation schemes for precursors in real-time with instrument acquisition or adjusting acquisition settings based on retention time features[4-6]. Since then, these efforts have expanded to improve targeted[7,8] and discovery-based proteomic workflows[9]. IDA strategies rely on

interpretation and utilization of raw spectral data within a few milliseconds of scan acquisition. This data can then be used to trigger additional scans, optimize scan parameters, or filter low utility spectra[10].

In previous work, we showed that real-time database searching (RTS) could double instrument acquisition efficiency for whole proteome profiling[10]. Comet-based[11] scoring enabled real-time peptide spectral matching that could be used to inform the instrument when to collect an SPS-MS3 scan and to target b- and y-ions for SPS selection. Leveraging the increased flexibility, users have improved efficiency and sensitivity in a range of applications from drug development to aging to immunology[12,13]. For example, RTS was recently applied to chemical proteomics studies to identify cysteine reactive compounds as well as diverse mechanisms of action[14,15]. In the cysteine study, the integration of TMTpro and RTS led to a 42-fold increase in the sample throughput for reactive cysteine quantification[14]. In addition, RTS-based decision making was used to improve the sensitivity for the detection of single cell proteomes[13].

While RTS is a powerful means to improve instrument acquisition efficiency, we know from post hoc analysis of proteomics datasets that additional information can be used to increase the sensitivity of peptide detection[16-19]. Outside of RTS-based spectral matching, intelligent data acquisition methods have been used for tasks such as retention time shifting-correction in targeted or DIA methods and heavy-labelled peptide triggered methods[7,8,20-24]. Recently, new tools for Real-Time spectral Library Search (RTLS) were developed to enable the real-time characterization of metabolites[25]. Several groups have now used spectral libraries of predicted or empirical spectra to identify fragmentation signatures to triage useful spectra or to select fragmentation schemes for further analysis[26,27]. This has now been applied to identify metabolites, improve structural characterization of lipids, and enrich for crosslinked peptide

species[25-27]. Though these methods have largely focused on spectral libraries of less than 200 individual spectra, they evidenced how IDA methods, and RTLS in particular, can be applied to a diverse array of biological samples.

Building from the efforts of real-time methods for quantitative proteomics and now RTLS, we wanted to explore the use of spectral library peptide matching for proteome-wide quantitative methods[28]. Early work with spectral library searching for proteomics relied on the construction of empirically derived spectra to generate libraries using well established workflows such as SpectraST to confidently match peptides based on common score metrics (dot product, cosine score, spectral similarity)[29,30]. Recent advances in deep learning have now contributed multiple pipelines for the in silico prediction of peptide spectra[16,31]. Algorithms such as Prosit enable users to predict peptide spectra for whole proteomes (2.6 million peptide spectra for human cells) and have recently been extended to incorporate isobaric labelled samples[32,33]. These predicted spectral libraries can then be used to efficiently score new empirical spectra or combined with database searching algorithms to re-score spectra for improved sensitivity[34-36]. Spectral library searching has been shown to be a sensitive and accurate way to identify peptides, especially those of complex spectra such as data-independent acquisition experiments[37-41]. Lessons learned from using spectral libraries in data-independent acquisition experiments have recently been leveraged to improve spectral library searches for data-dependent acquisition methods as well[36]. These latest spectral library search algorithms show that while using even a predicted library, there is a sensitivity gain compared to cutting-edge database search methods[15]. Yet, as noted these algorithms generate libraries of millions of individual spectra which can be challenging to process and compare in real-time with instrument acquisition.

Here, we sought to develop and implement RTLS for whole proteome, sample multiplexed quantitative proteomics. Previous work established that RTLS could be used for small molecule spectral matching with spectral libraries of a single small molecule, and to improve analysis of cross-linked peptides with a two-spectrum library of diagnostic ions[42]. To enable whole proteome RTLS, we needed to enable RTLS to process full proteome libraries of up to 4 million spectra in a few milliseconds to allow for real time processing with spectral acquisition. To accomplish this, we (1) developed a flexible library processing workflow for both predicted and experimental libraries from common proteomics resources, (2) optimized the online scoring for whole proteome sample multiplexed analyses, (3) determined optimal parameters for improved instrument efficiency and quantitative accuracy, and (4) benchmarked the RTLS workflow compared to traditional acquisition methods using established standards and complex proteomic samples. Using our optimized workflow and methods, RTLS increased instrument acquisition efficiency 2-fold, and consistently outperformed RTS for high-resolution spectral matching. In addition, RTLS improved quantitative accuracy for chimeric spectra from whole proteome single shot samples and enabled fast and efficient identification of post-translationally modified peptides. Thus, RTLS has proved to be a useful addition to the IDA toolkit with great potential for sample multiplexed quantitative proteomics and future development.

3.2 Methods

Sample collection and preparation.

Human cell lines were grown to confluence in DMEM containing 10% fetal bovine serum and 1% streptomycin/puromycin. Cells were harvested by manual scraping and washed twice with PBS. Cells were syringe lysed in lysis buffer (8M urea, 50mM EPPS pH 8.5, 150mM

NaCl, and Roche protease inhibitor tablet) and the resulting lysates were cleared via centrifugation.

Saccharomyces cerevisiae (BY4742) was grown YPD cultures to an OD600 of 0.8 then washed twice with PBS, pelleted, and stored at −80 °C until use. Cells were resuspended in lysis buffer (8 M urea, 50 mM EPPS pH 8.5, 150 mM NaCl, Roche protease inhibitor tablet) and lysed by bead beating. After lysis and bead removal, the lysate was centrifuged to remove cellular debris and the supernatant was collected for use.

A two proteome (human and yeast) HyPro standard labeled with TMTpro was prepared as previously described[43]. In brief, HCT116 cells were prepared according to the SL-TMT protocol[44] and labeled at 1:1 across all channels. S. cerevisiae (BY4716) was prepared similarly but with ratios of 0:1:1:1:2:2:2:4:4:4:8:8:8:10:10:10 across the 16 TMTpro reporter ion channels. The HCT116 and S. cerevisiae were combined so the final sample was 90% human peptides and 10% yeast peptides (w/w). In the small molecule perturbation studies, A549, H292, or PSC1 cells were treated with 10 μM of a given HDAC inhibitor for 24 hours (belinostat, abexinostat, CUDC-101, vorinostat).

Mass spectrometry data acquisition methods and analysis.

Samples were resuspended in 5% acetonitrile/2% formic acid prior to being loaded onto an in-house pulled C18 (Thermo Accucore, 2.6 Å, 150 μm) 30 cm column. Peptides were eluted over 30, 60, 90, 120, or 180 minute gradients running from 96% Buffer A (5% acetonitrile, 0.125% formic acid) and 4% buffer B (95% acetonitrile, 0.125% formic acid) to 30% buffer B. Sample eluate was electrosprayed (2,700 V) into a Thermo Scientific Orbitrap Eclipse or Orbitrap Ascend mass spectrometer for analysis. The scan procedure for MS1 scans (Orbitrap scan at 120,000 resolving power, 50 ms max injection time, and AGC target set to 100%) and

MS2 scans (linear ion trap, “rapid” scan rate, 50 ms max injection time, AGC target set to 200%, CID collision energy of 35% with 10 ms activation time, and 0.5 m/z isolation width) was constant for all analyses with a gradient length above 30 minutes. For 30 minute gradient methods, the ion trap scan rate was set to “turbo” and the maximum injection time lowered to 11ms. Peaks from the MS1 scans were filtered by intensity (minimum intensity > 5e3), charge state ($2 \leq z \leq 6$), and detection of a monoisotopic mass (monoisotopic precursor selection, MIPS). Dynamic exclusion was used, with a duration of 90s, repeat count of 1, mass tolerance of 10ppm, and with the “exclude isotopes” option checked. High field asymmetric waveform ion mobility spectrometry (FAIMS) was set at “standard” resolution, 4.6 L/minute gas flow and 3 CVs: -40/-60/-80. Each CV value was set to top N mode with number of dependent scans set to 6. For MS3 scans, SPS ions were set to 10, MS1 Isolation window was 2 m/z and MS2 isolation window was 3 m/z. MS3 scans were performed at a resolving power of either 45,000 (Ascend) or 50,000 (Eclipse) with an HCD collision energy of 45%.

Real-time spectral library searching.

Predicted spectral libraries were generated using Prosit-TMT (https://www.proteomicsdb.org/prosit/) and output to '.msp’ files. MSP files were converted to an RTLS/mzVault compatible format (‘.db’ extension) based on an indexed SQLite database structure. These files can be consumed by the RTLS filter in the instrument’s method editor. Conversion was done using the in-house developed DBKey R Shiny application (https://github.com/SchweppeLab/DBKey). Unless otherwise noted, libraries were generated with no missed cleavages, fixed TMTpro at n-term and lysine and variable oxidation on methionines. For each library candidate within the specified precursor tolerance, a cosine

similarity score was calculated against the acquired spectra, with the cosine similarity score defined as such (Equation 1):

$$\frac{\sum_{matched} I_L^a mz_L^b \; I_U^a mz_U^b}{\sqrt{\sum_{all} I_L^a mz_L^{b^2}} \sqrt{\sum_{all} I_U^a mz_U^{b^2}}} \qquad \text{Eq. 1}$$

where L is library peaks, U are acquired spectral peaks, I is intensity of peak, mz is m/z of peak, a and b are the weight factors. Unless otherwise specified, the minimum cosine score threshold used was set to 20 to approximate the triggering rate of RTS-based methods. We note that the thresholds used in this work are specific to the underlying spectral libraries that were used. This is due to factors such as: the number of peaks allowed in each spectra of the library, the minimum basepeak intensity used to filter out 'noise', the origin of the libraries (empirical or predicted).

Real-time spectral searching and analysis was done using Orbitrap Eclipse instrument control software version 4.0, unmodified from the released version except where noted. Unless otherwise specified, the minimum cosine score threshold used was 20. Precursor tolerance was set to 10 ppm and isotope correction was set to 0/1. TMT-SPS mode was selected, such that only matched fragments with a TMTpro tag were selected for SPS-MS3 scans. For methods targeting yeast in HyPro, peptides derived from human proteins were rejected from triggering further SPS-MS3 scans, using the keyword promote/reject feature. Spectral libraries built with SpectraST 5.0 were processed through the Trans-Proteomic Pipeline[45]. Raw files were converted to mzXMLs with msconvert and subsequently searched with Comet using settings described in the data analysis section[46]. Results were then processed using PeptideProphet before SpectraST

was used to build a consensus library with default settings. Decoy library entries were generated via precursor swap [47-49].

Real-time database searching and concatenate methods.

Real-time search was done using the instrument control software, Tune version 4.0, unmodified from the release version except where noted. For the fractionated HDAC-treated cell line experiment, a human database was used with fixed Cys(Cam) and variable Met(Ox), one missed cleavage, with TMT mode enabled, a minimum XCorr of 1.0, a minimum dCn of 0.05, precursor tolerance of 10 ppm and FDR/protein closeout disabled. Unless otherwise noted for HyPro runs, a concatenated human-yeast database (a single FASTA file starting with the human proteome then the yeast proteome) was used with the same modifications, "TMT mode" enabled, a minimum XCorr of 1.4, a minimum dCn of 0.05, and precursor tolerance of 10 ppm.

For sequential RTS/RTLS analysis, a single CV of -50 was used. MS2 nodes are set up sequentially with a real-time filter after each scan, the first filter is set to "trigger-only" and the filer set to pass along every MS2 (Cosine Score less than 100/minimum XCorr of 0) and the second filter set to block any further scans (Cosine Score/XCorr greater than 100). For FTMS2 analysis in this experiment, a resolving power of 15,000, a max fill time of 54ms and an AGC target of 200% were used. All raw data are available through accession PXD039855.

Post hoc data analysis.

Raw files were searched using the Proteome Discoverer 3.1 software. Unless otherwise noted, Comet and Sequest searches were performed against databases downloaded from Uniprot, with 2 missed cleavages and a 20 ppm precursor mass tolerance. For ITMS2-MS3 methods, a 0.6 Dalton fragment tolerance was used, and for FTMS2 methods, a 10 ppm fragment tolerance was used. Charge state was restricted to 2-6 and peptide length was limited to 7-30 amino acids to

comply with INFERYS. All searches were performed with variable methionine oxidation (+15.99491), static cysteine carboxyamidomethylation (+57.02146), and static TMTpro modifications on lysine and the peptide N-termini (+304.207126). MSPepSearch was run with an INFERYS predicted library with 1 missed cleavage, fragment charges 2 to 4, variable methionine oxidation (+15.99491), static cysteine carboxyamidomethylation (+57.02146), and static TMTpro modifications on lysine and the peptide N-termini (+304.207126). MSPepSearch precursor tolerance was 10 ppm and the fragment tolerance was 0.6 Dalton for ITMS2-MS3 methods and 10 ppm for FTMS2 methods. Peptide spectral matches and spectrum spectral matches were filtered to a peptide and protein false discovery rate (FDR) of less than 1%[50]. Quantification was done by selecting the centroid with the highest signal-to-noise ratio within a 0.003 Dalton tolerance of the reporter ion's theoretical m/z. Unless otherwise noted, peptides identified using FTMS2 or SPS-MS3 methods were considered quantified if the sum of the reporter ions' signal-to-noise ("s:n")ratios was greater than 100 and precursor isolation specificity was greater than 0.5. Peptides identified using RTS or RTLS methods were considered quantified if the sum of the reporter ions' signal-to-noise ratios was greater than 100, regardless of precursor isolation specificity. Chimeric spectra analysis was performed using CHIMERYS under the same settings as above, using the "inferys_2.1_fragmentation" prediction model. Quantitative accuracy was assessed with the HYPER interference-free index (IFI, Equation 2), a modified version of the original TKO interference free index where the empty channels in the HyPro standard serve as substitute for the empty KO channels [51,52].

$$IFI = 1 - \frac{s{:}n\ KO}{s{:}n\ non\text{-}KO} \qquad \text{Eq. 2}$$

Statistical analyses and plotting was done using the R project for statistical computing[53]. When comparing real-time and post hoc data, true positives were considered as those PSM/SSMs

that were confidently identified post hoc ('ground truth') and in real time ('test') (Equations 3 & 4).

$$sensitivity = \frac{True\ Positive_{Passed\ RT\ \&\ Post\ hoc}}{True\ Positive_{Passed\ RT\ \&\ Post\ hoc} + False\ Negative_{Passed\ Post\ hoc\ only}} \quad \text{Eq. 3}$$

$$specificity = \frac{True\ Negative_{Failed\ RT\ \&\ Post\ hoc}}{True\ Negative_{Failed\ RT\ \&\ Post\ hoc} + False\ Positive_{Failed\ Post\ hoc\ only}} \quad \text{Eq. 4}$$

Important considerations of RTLS analyses

As noted above, the spectral library used can have strong effects on the underlying score distributions for RTLS matching and therefore future users should make sure to adjust the score thresholds accordingly. Additionally, post hoc analysis of peptides quantified from RTLS triggered MS3 scans should be carefully analyzed. We have pointed out that database searching with rescoring based on spectral libraries is a powerful means to improve sensitivity, but users should also consider key metrics of successful quantitation, such as the number of SPS ions (selected by RTLS) derived from the matched peptide (selected by post hoc analysis). Ideally this fraction should be close to 100%, otherwise inaccurate quantitation may have occurred.

3.3 Results and Discussion

Spectral library searching has been shown previously to increase the sensitivity of peptide detection in quantitative proteomics by leveraging either acquired spectra or predicted fragment ion intensities[41,54]. With a diverse array of scoring metrics, analysis pipelines, and applications, spectral library searching has proved to be a robust method for a wide array of quantitative proteomics methods, though it is predominantly used for label-free quantitation, and data-independent acquisition methods (DIA)[55]. To enable spectral library searching for real-time decision making we had to (1) determine a method and memory compatible means to store

full-proteome spectral libraries (2) enable conversion of diverse spectral library formats into this common format, and (3) optimize scoring functions for performant real-time decision making (Figure 3-1A).

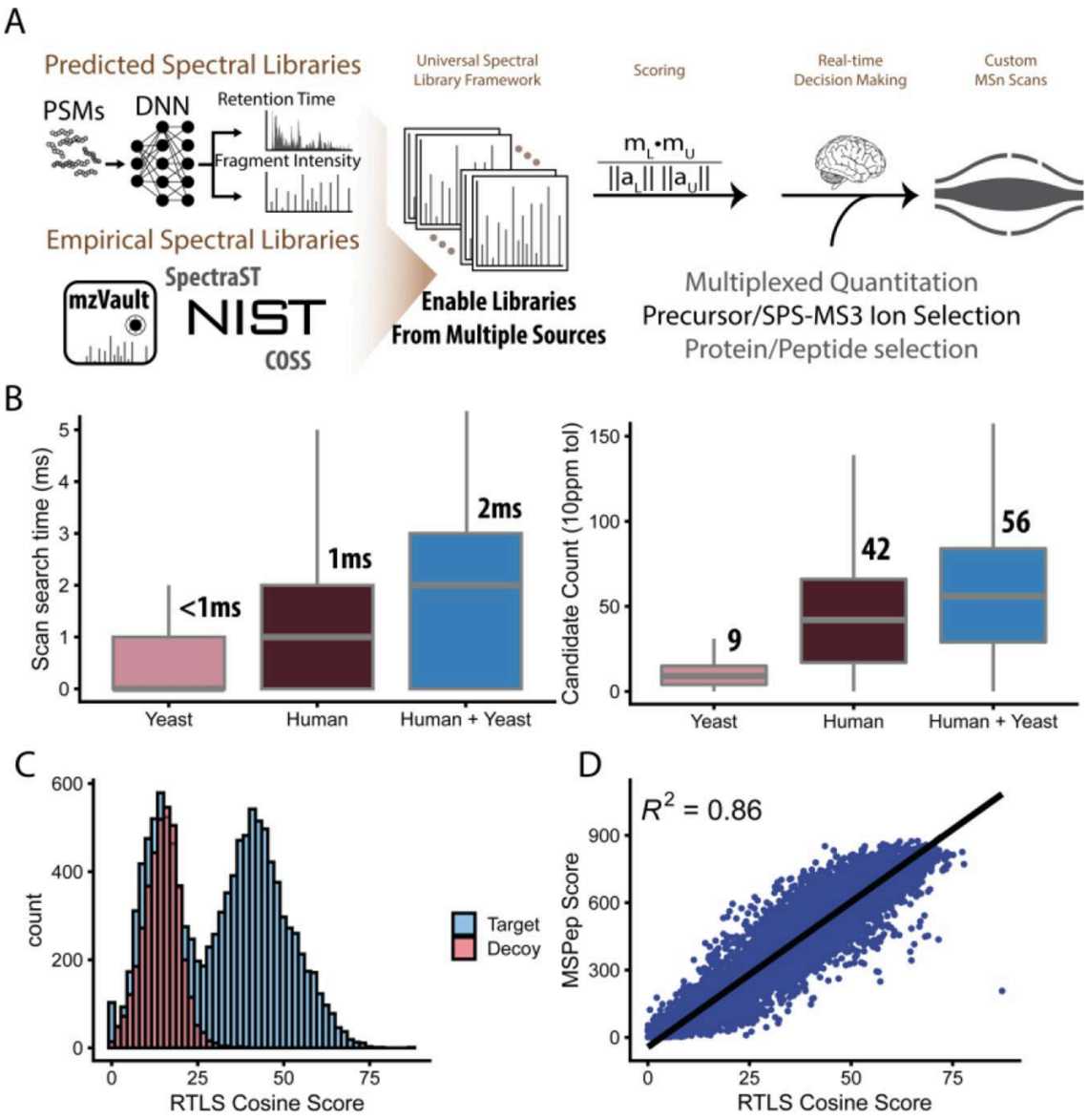


Figure 3-1. Overview of RTLS workflow. (A) The DBKey-RTLS workflow enables the use of spectral libraries from repositories, deep neural network (DNN) predictions, and empirical data. Spectral library matching and scoring are then used to define subsequent scans (SPS-MS3). (B) (Left) Boxplots of the search time of Prosit-TMT libraries for yeast, human, and concatenated human-yeast. (Right) Library candidates scored per scan for the same libraries. (C) Target-decoy separation for a yeast sample labeled with TMTpro. (D) Scatterplot comparing RTLS cosine score to that of an established offline spectral library search, MSPepSearch.

Development of a unified tool for spectral library processing for RTLS

We began by building a library processing tool, called DBKey, that can take in spectral libraries from both empirical and predicted spectral sources including repositories such as NIST, SpectraST, and Prosit. DBKey then converts these input file types to a single, RTLS-compatible data type for real-time processing. Our common data type is also compatible with mzVault and is based on an indexed SQLite data structure that can be held in memory during instrument acquisition and stored for repeated use ('.db' file extension). By integrating DBKey into a Docker image, we can process these files efficiently in a Docker environment for workstation or cloud extensible deployment.

Using DBkey, we built libraries from both predicted libraries and publicly available datasets to evaluate their compatibility with a multiplexed RTLS-MS3 workflow. Prosit-TMT[32,33] was used to make whole-proteome predicted libraries from S. cerevisiae and human FASTA files (497,718 and 2,225,832 spectra respectively) (Table S1). The total size on disk of these databases was 0.41GB for yeast, 2.95GB for human, and 3.23GB for the concatenated human-yeast databases (Table S1). Queries to the database generally require less than 1ms for single proteome spectral libraries even when the library consists of millions of individual spectra (Figures 3-1B).

Table S1: Description of spectral libraries used in study

Species	Missed Cleavage	Variable Mods	Source	Entries	Size (GB)
S. cerevisiae	0	Oxidation (M)	Prosit	497718	0.411
S. cerevisiae	0	Oxidation (M)	PXD02482	66145	0.271
Human	0	Oxidation (M)	Prosit	2255832	2.950
Human	1	Oxidation (M)	Prosit	4853167	6.930
Human+S. cerevisiae	0	Oxidation (M)	Prosit	2753550	3.230

Table 3-1. RTLS Libraries used in this study**.**

Optimization of RTLS cosine score weights

To enable rapid scoring of spectral-spectral matching, we chose to use the weighted, or 'modified', cosine score because it is very fact to calculate and has been used and tested extensively[30,56-58]. In particular we used the weighted cosine score which has been shown to be more sensitive than the unweighted cosine score for larger spectral libraries[40,56]. While cosine score weights have been optimized previously for label-free proteomes they had never been tuned for TMTpro-labelled peptides[59]. We performed a parameter sweep for the scoring based on Equation 1 and found that the weights a = 0.4 (applied to the intensity) and b = 0.9 (applied to the m/z) provided the highest sensitivity for peptide spectral matching of TMTpro labeled peptides (Figure 3-2). This was in-line with previous work that found down-weighting dominant peaks and up-weighting larger fragment ions improves sensitivity[30]. PSMs scored with our optimized cosine score discriminate target and decoys well and show strong agreement with Comet database search results (Figure 3-3).

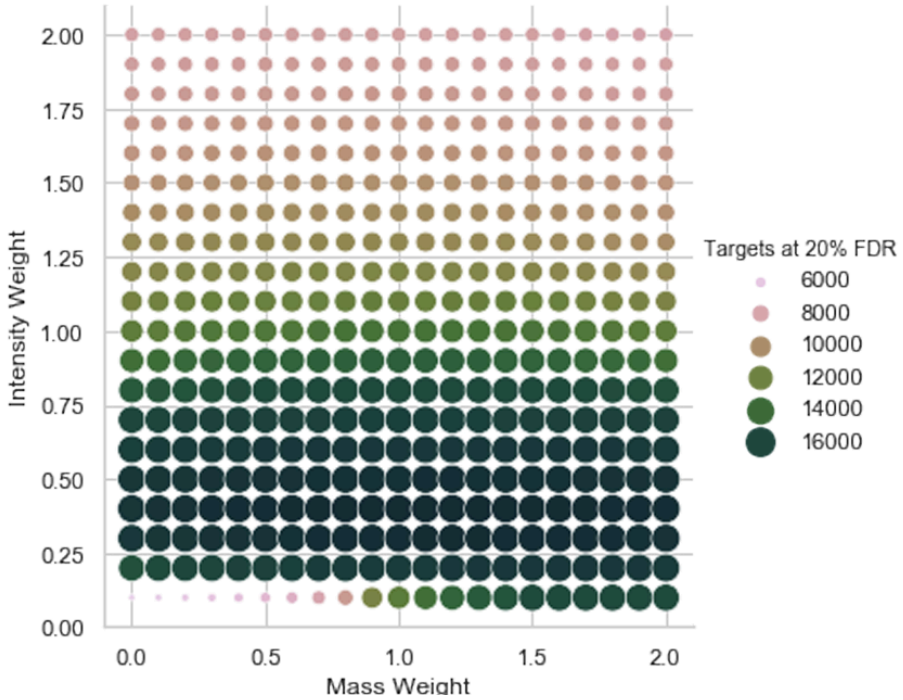


Figure 3-2. Parameter sweep at different cosine score weights run on human and yeast TMTpro labelled peptides, measured at 20% FDR. At an intensity weight of 1 and a mass weight of 1, the score will equal the unweighted cosine score.

Using the optimized score weights, we sought to establish RTLS' potential for efficiently matching spectra in real time with instrument acquisition. Ideally, spectral matching should occur fast enough to enable parallelization scan acquisition. Using optimized indexing from the .db files, we were able to match experimental spectra to library spectra from predicted and empirical libraries with median search times of 1 ms for single proteome databases (human or yeast) and 2ms for a concatenated human-yeast library (Figure 3-1B). Importantly, even as the number of candidate spectra considered for each search increased, the relative search time remained well below our target of 35ms. RTLS maintained accurate discrimination of target and decoy peptides

using the weighted cosine score (Figure 3-1C) and these scores correlated well with post hoc spectral match scoring from MSPepSearch (Figure 3-1D, Figure 3-3C).

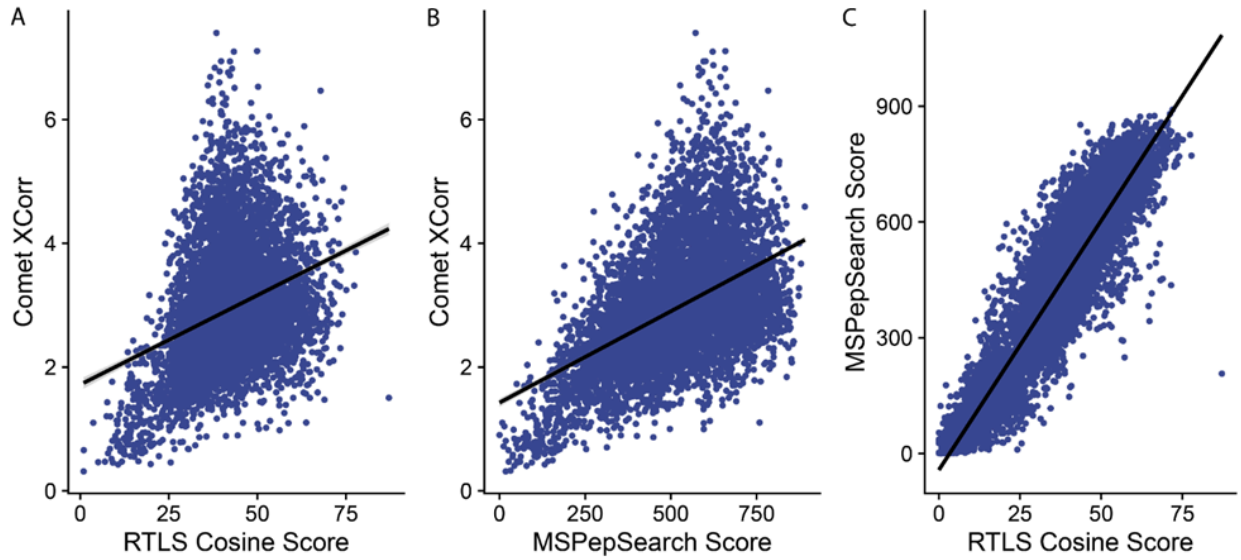


Figure 3-3. Score comparisons for different proteomics search methods. A) Scatter plot of RTLS cosine score and Comet XCorr. B) Scatter plot of Comet XCorr and MSPepSearch Score. C) Scatter plot of MSPepSearch and RTLS Cosine Score.

Evaluation of post hoc search methods for RTLS quantified peptides

While we performed our initial post hoc analysis with MSPepSearch, we wanted to evaluate multiple search pipelines to determine the optimal post-RTLS methods for sensitive detection of labeled peptides. To this end, we searched a 12-fraction whole-proteome multiplexed sample set collected with RTLS methods with three different search workflows: Comet (canonical database searching), MSPepSearch (spectral library searching), and Sequest with INFERYS[36] rescoring – a database search rescored on spectral similarity to predicted

fragment ion intensities (Figure 3-4). Settings were kept as similar as possible across informatics workflows (see methods). Though we observed a high degree of overlap between search methods, Sequest+INFERYS returned the highest number of confidently identified PSMs (Figure 3-5). Due to the improved detection sensitivity for PSMs and the incorporation of aspects of database and library searching with Sequest+INFERYS, we proceeded to use this pipeline for post hoc analysis of RTLS acquired data.

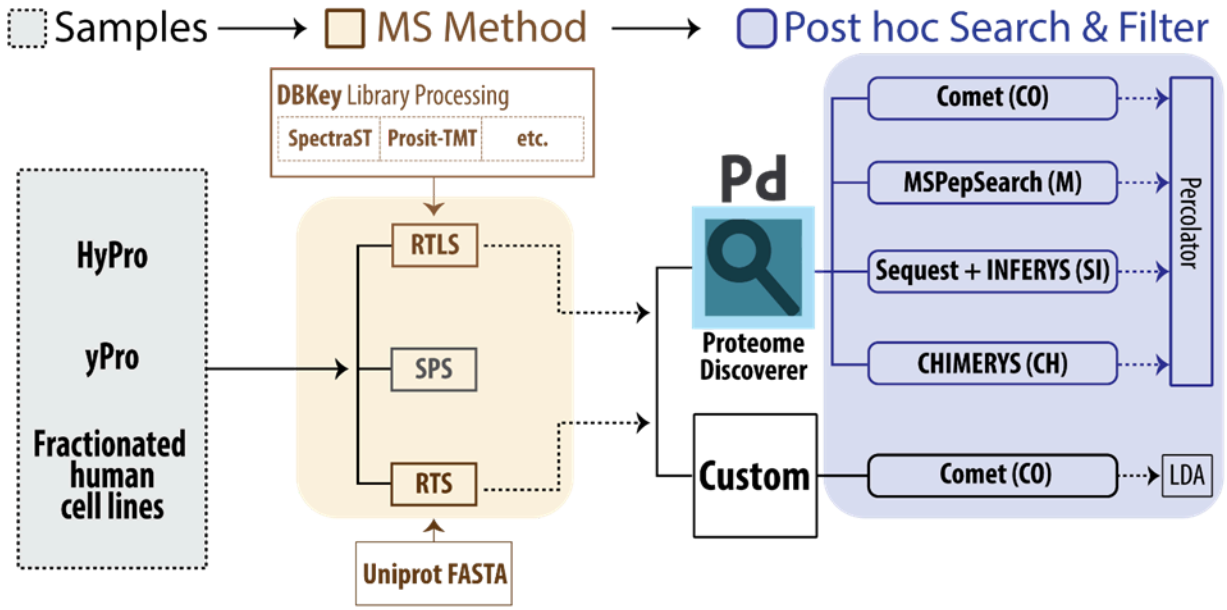


Figure 3-4. A) Diagram showing the samples (HyPro, yPro, and fractionated human cell lines), methods (SPS, RTS, RTLS) and data analysis (Comet, MsPepSearch, Sequest+INFERYS, CHIMERYS) performed in this study.

Spectral library searching is a highly flexible approach, but can be influenced by library spectra sourcing, peptide fragmentation, spectral quality and spectral purity during instrument acquisition[60]. To test this, we measured peptide detection sensitivity and quantitative accuracy

across a panel of acquisition methods comparing: (1) predicted or empirically derived spectral libraries, (2) with or without FAIMS, and (3) using both CID and HCD MS2 fragmentation (Figure 3-6-8). First, we generated an empirical spectral library from fractionated yeast samples labeled with TMTpro (PXD014546)[61] and assembled a library using SpectraST (66,415 spectra). We compared this empirical yeast library to a predicted library built using Prosit-TMT. Using these two libraries we found that RTLS could efficiently match spectra and trigger SPS-MS3 scans for both empirical and predicted libraries. We observed the cosine scores distribution skewed higher in the empirical spectra results, most likely due to incorporation of non b/y fragments, yet the number of quantified peptides and the concordance with offline searching was lower (Figure 3-6). Thus, when using empirical libraries, it is important to determine score distributions to establish an optimal score threshold for a given filter-library pairing. Overall, we observed that the larger predicted library led to 22.4% more quantified peptides compared to the empirically derived library (Figure 3-6) while only 5.9% of the peptides were missing from the empirical library. While scoring for an empirical library would likely be improved using data generated on the same instrument and optimized library construction, the promising results, robustness, and flexibility of predicted libraries, we primarily used Prosit-derived libraries for this work.

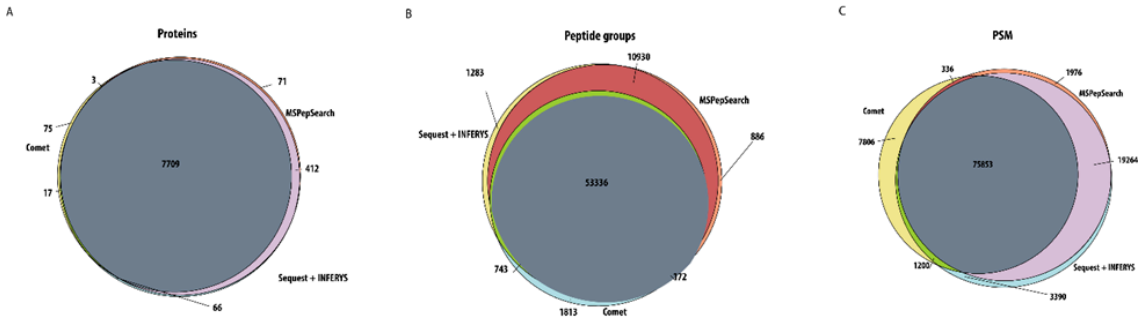


Figure 3-5. Venn diagrams showing the overlap between Sequest + INFERYS, MSPepSearch and Comet in Proteome Discoverer at protein (A), peptide group (B), and PSM (C) level for fractionated whole proteomics data.

Second, when comparing RTLS methods with or without using FAIMS, we observed a broader score distribution, higher median score (20.7 to 26), and better target/decoy discrimination (Figure 3-7). We believe this is due to FAIMS ability to reduce precursor co-isolation[52] thereby generating spectra with fewer interfering peptide fragments and higher spectral match scores. For this reason, we proceeded with FAIMS for further experiments. Third, peptide fragmentation methods used during acquisition can greatly influence library-spectral matching[16]. We examined both CID and HCD RTLS workflows for use in multiplexed method. For these analyses, we tested if predicted library spectra based on a specific fragmentation type and energy would affect the sensitivity of RTLS peptide detection (Figure 3-7). Due to the similar score distributions and small (3.8%) difference in peptides detected we chose to use CID peptide fragmentation as it is the most similar to canonical SPS-MS3 workflows (Figure 3-8).

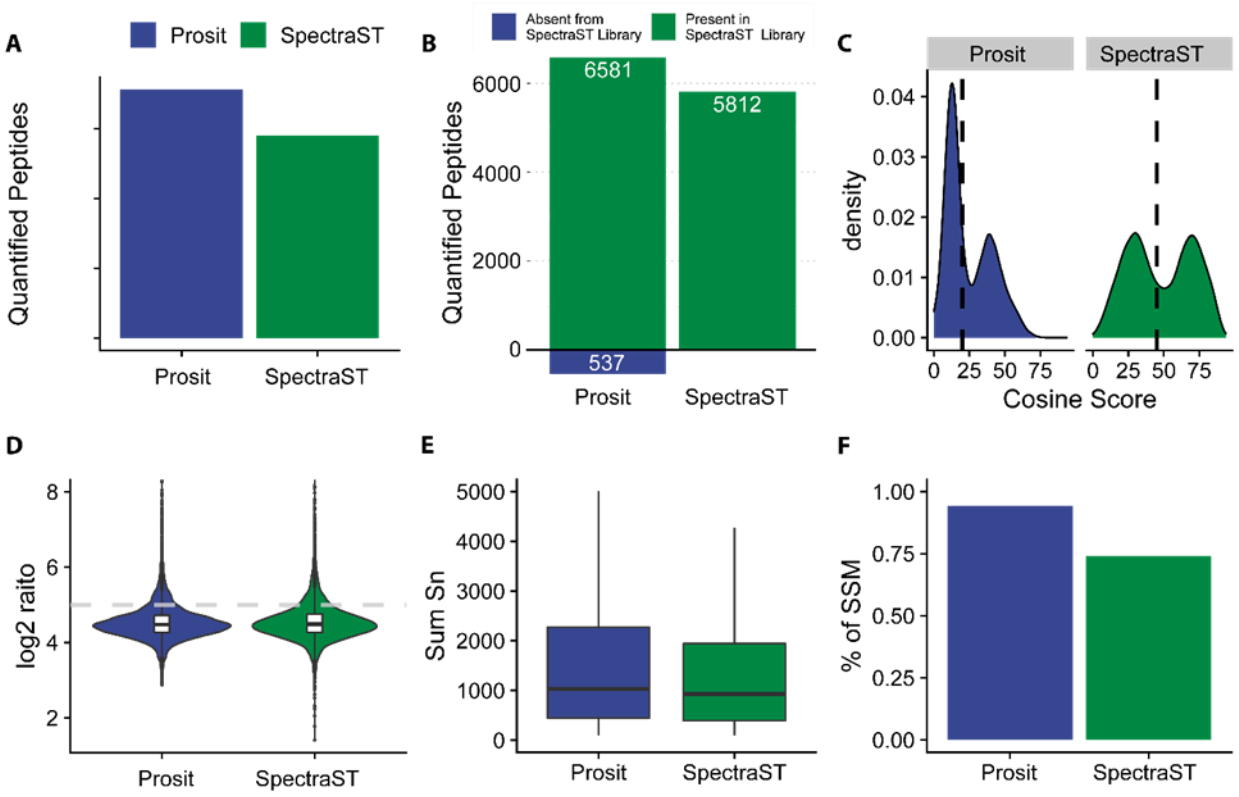


Figure 3-6. Comparison between empirical and predicted spectral libraries. SpectraST libraries were built from fractionated DDA data with default settings. A) Total quantified peptides using either SpectraST or Prosit libraries. B) Same as A, but for the Prosit method highlighting whether the peptides were present in the empirical SpectraST library. C) RTLS cosine score distribution. The dotted lines represent the approximate score thresholds for each of these libraries (see Experimental Procedures). D) Violin plot showing quantitative accuracy based on the *l*pg2(133*nn*, 134*cc*, 134*nn* / 127*nn*, 127*cc*, 128*nn*). Gray dotted line represents the median ratio when injecting only the yeast proteome. E) Comparison of PSM level reporter ion summed signal-to-noise (Sn). F) Percent of matches that agree with Comet database search.

Having established an optimized set of methods parameters we began by comparing RTLS to traditional SPS methods for whole proteome single-shot methods (Figure 3-7A). Running 180 minute gradients with HyPro we confirmed that RTLS resulted in more selective triggering and an increase in quantified peptides and proteins of 19.4% and 8.4%, respectively (Figure 3-7B). We then moved to the challenging task of sub-proteome analysis, quantifying only a subset of our sample, in this case the lower-abundant yeast proteins in our standard samples (10% of the peptides by mass Figure 3-7A, inset). When comparing SPS-MS3 to RTLS methods of equal length (180 minute gradients) RTLS increased the number of unique quantified peptides and proteins by 23.1% and 36.5%, respectively. Strikingly, when we compared RTLS methods targeting only yeast proteins and using a shortened gradient time (90 minute), the differences in total quantified yeast peptides and proteins was no significant between SPS-MS3 at 180 minute gradients and RTLS at 90 minute gradients (Figure 3-7B). When we compared the proportion of MS3 spectra triggered during instrument acquisition that were used to obtain our final set of quantified peptides and proteins, we found that RTLS increased MS3 usage from 11.6% to 59.8% and maintained high quantitative accuracy (Figure 3-7C, D). We observed that increasing the gradient length resulted in a slight decrease in the percentage of useful MS3 scans. We believe this is a function of sampling more low quality MS2 spectra as the gradient length increases which would in part explain why the percentages for either the yeast-only or human-yeast library methods remain consistent (Figure 3-7C).

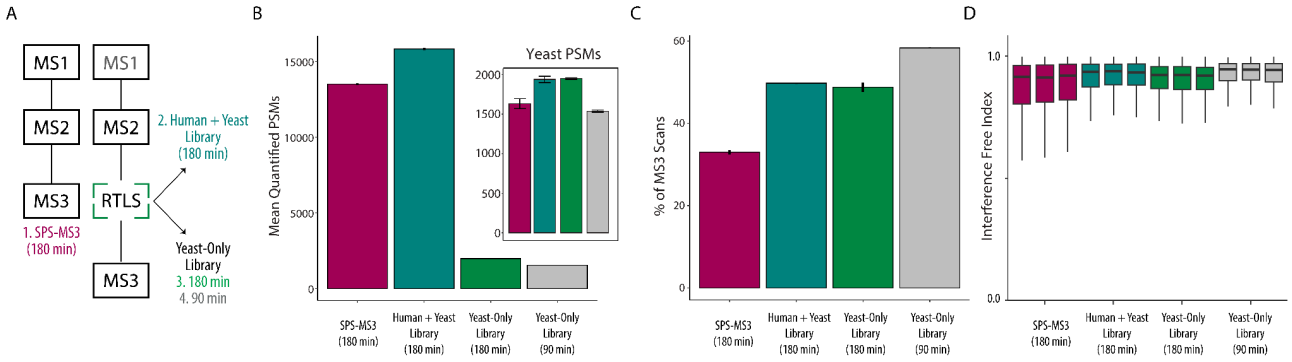


Figure 3-7. Quantifying peptides and proteins from human and yeast in HyPro. (A) Depiction of the four methods used for the comparison of SPS-MS3 and RTLS methods. (B) Total quantified PSMs from 1 µg HyPro standard for each of the four methods highlighted in A for 3 replicates. Inset: Mean quantified yeast PSMs for each method. Bars are the average number of quantified PSMs for triplicate HyPro injections; error bars represent standard error of the mean. (C) Percent of total acquired MS3 scans that led to quantified peptides for the SPS-MS3 and Human + Yeast Library data as well as the percentage of total acquired MS3 scans that led to quantified yeast peptides for the yeast-only library data for 3 replicates. Bars represent the average number of quantified PSMs for three replicate injections; error bars represent standard error of the mean. (D) Box plot of the HyPro IFI for quantified yeast proteins for each of 3 replicates per method

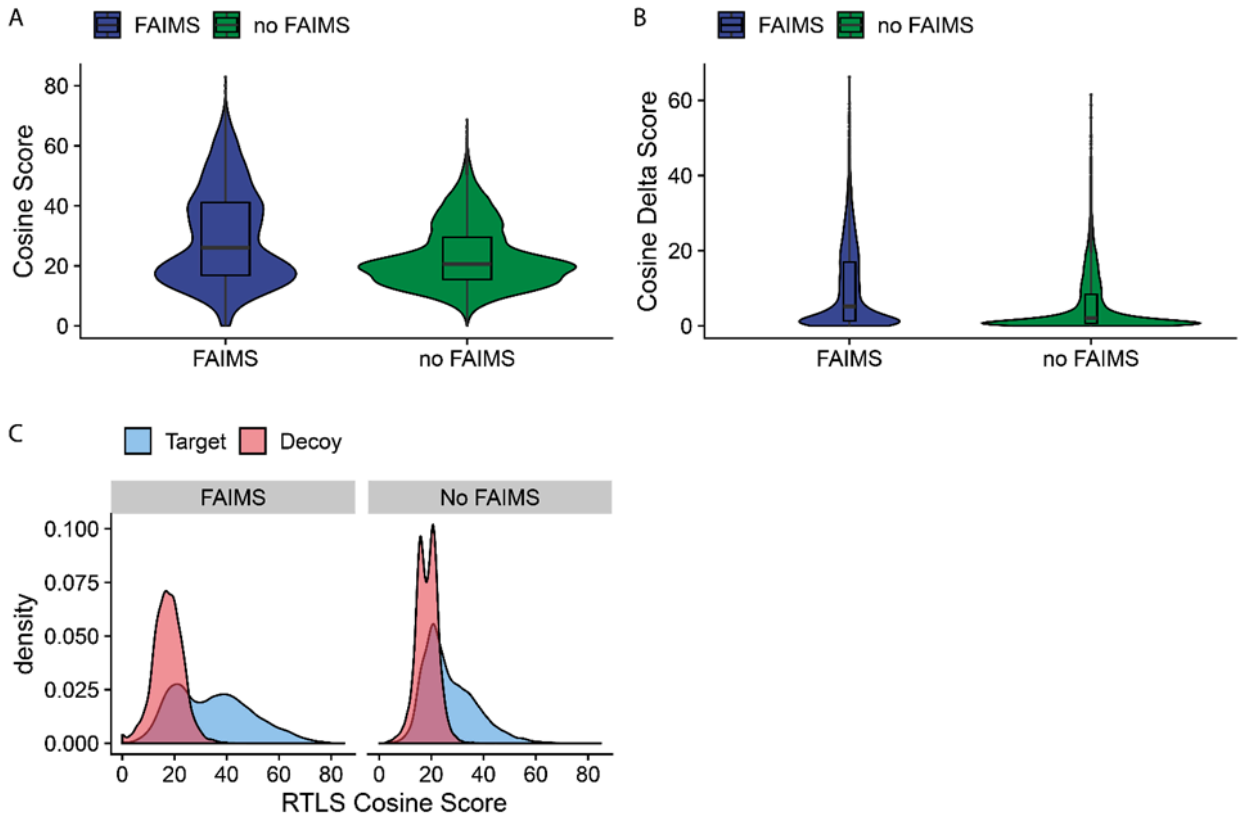


Figure 3-8. RTLS with and without FAIMS resulting in 8631 total quantified peptides without FAIMS and 8551 total quantified peptides with FAIMS. A) Violin plots showing cosine score distribution. B) Violin plots showing delta cosine score distribution. C) Target/decoy discrimination for FAIMS and no FAIMS.

We next sought to investigate whether one or the other of these methods was better suited to sample multiplexed proteomics analysis. To do this, we ran a series of single-shot methods with settings matched as closely as possible between RTS and RTLS so as to generate similar rates of triggering quantitative MS3 scans (Figure 3-9A). For this work, we chose an RTLS score threshold initially because a weighted cosine score filter of 20 resulted in a similar number of total quantified proteins and peptides for ITMS2 RTLS-based methods compared to RTS

methods (Experimental Procedures, Figure 3-11A). We observed a slight but consistent increase in the number of quantified peptides when using the RTLS methods for FTMS2 analyses (Figure 3-9A). To determine why RTLS consistently generated more quantified peptides we ran a sequential real-time method to directly compare RTS and RTLS on the same set of precursors in the same analytical run. In this method, the MS2 level was branched so that each precursor produced two MS2 spectra. Each of these subsequent MS2 scans were then analyzed in real time by either RTS or RTLS to generate a matched set of filtering and quantification events. We then processed this data through the common Sequest+INFERYS pipeline which combines database searching and library rescoring. With the matched set of RTS and RTLS triggering events, we compared XCorr (Comet-based RTS) and weighted cosine score (from RTLS) for their ability to classify target and decoy peptides (Figure 3-9B). In replicate analyses, we found that RTLS was more sensitive at low FDR thresholds (1-5%) at detecting confirmed peptide spectral matches from the post hoc analysis.

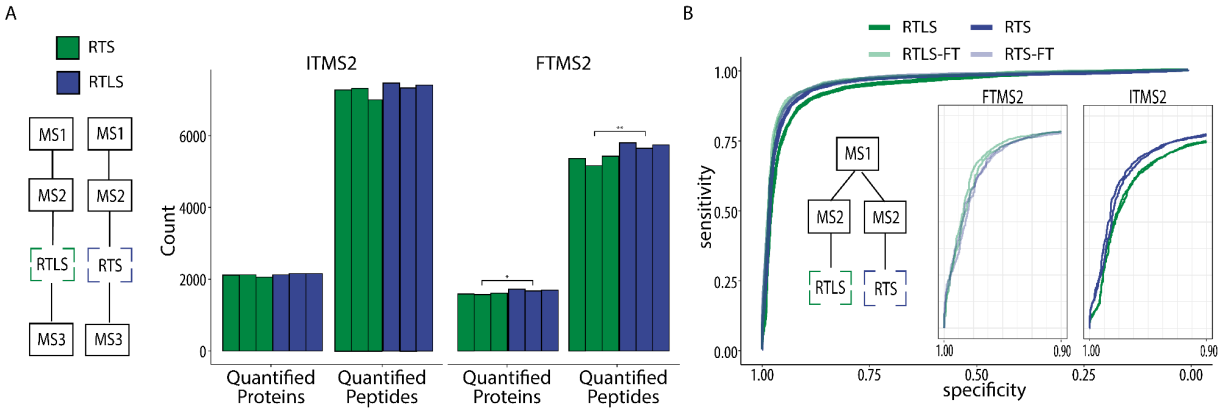


Figure 3-9. Comparing IDA methods. (A) 60 min HyPro runs (n = 3) comparing RTS and RTLS done with MS/MS from both the ion trap (IT) and Orbitrap (FT). MS2s, MS3s, quantified peptides, and quantified proteins plotted. For FTMS methods: *, p-valueproteins ~0.0001; **,

p-valuepeptides = 0.0192. (B) Receiver-operator characteristic plots of both ion trap and Orbitrap MS2 spectral matching with either RTS or RTLS based on XCorr (RTS) or cosine score (RTLS). Sensitivity and selectivity were calculated based on the comparison between RTS and RTLS spectral matching compared to post hoc peptide spectral matching ("ground truth").

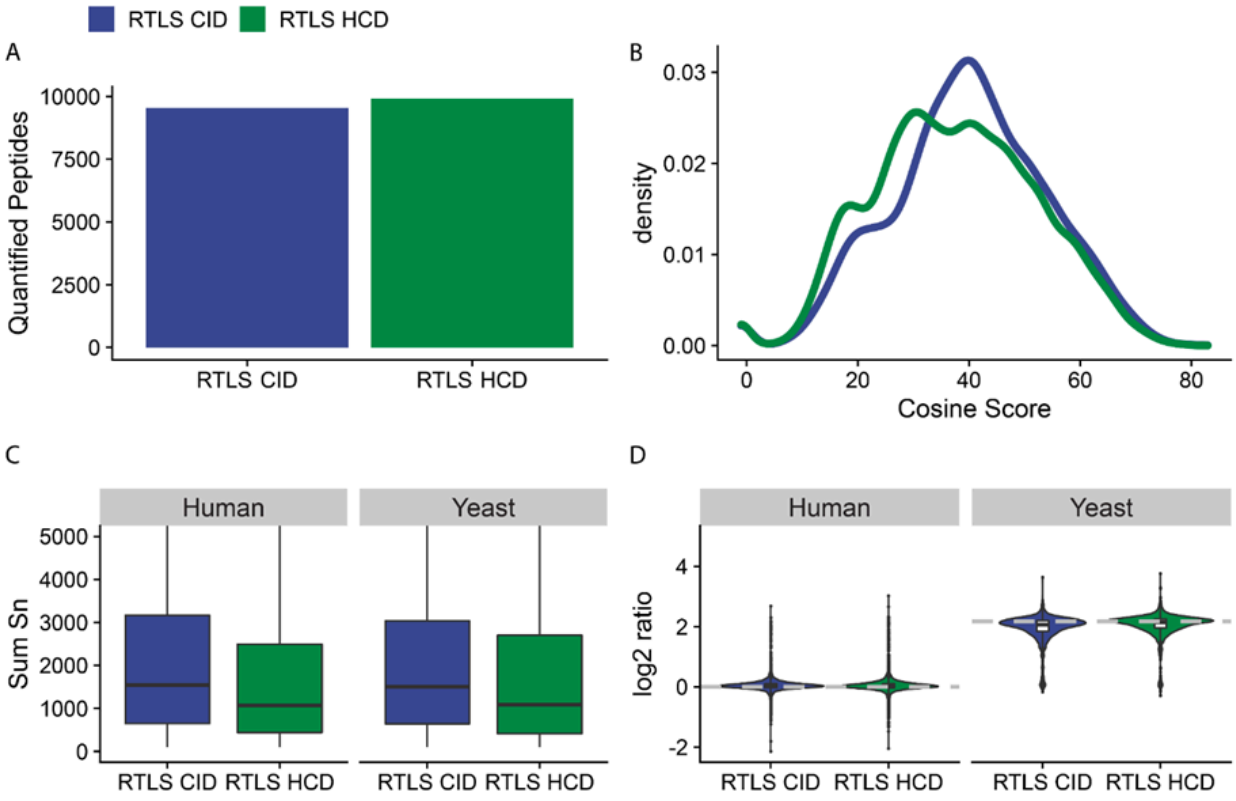


Figure 3-10. RTLS run with HCD and CID collision energies. Prosit was used to generate CE specific library for HCD spectra. A) Comparison of quantified peptides in 120 minute run. B) Distribution of cosine scores between predicted library and acquired spectra. C) Boxplot comparing PSM reporter ion signal of the two proteomes found in HyPro. D) Violin plot showing quantitative accuracy by looking at the . Gray dotted line represents the median ratio when injected as single proteome

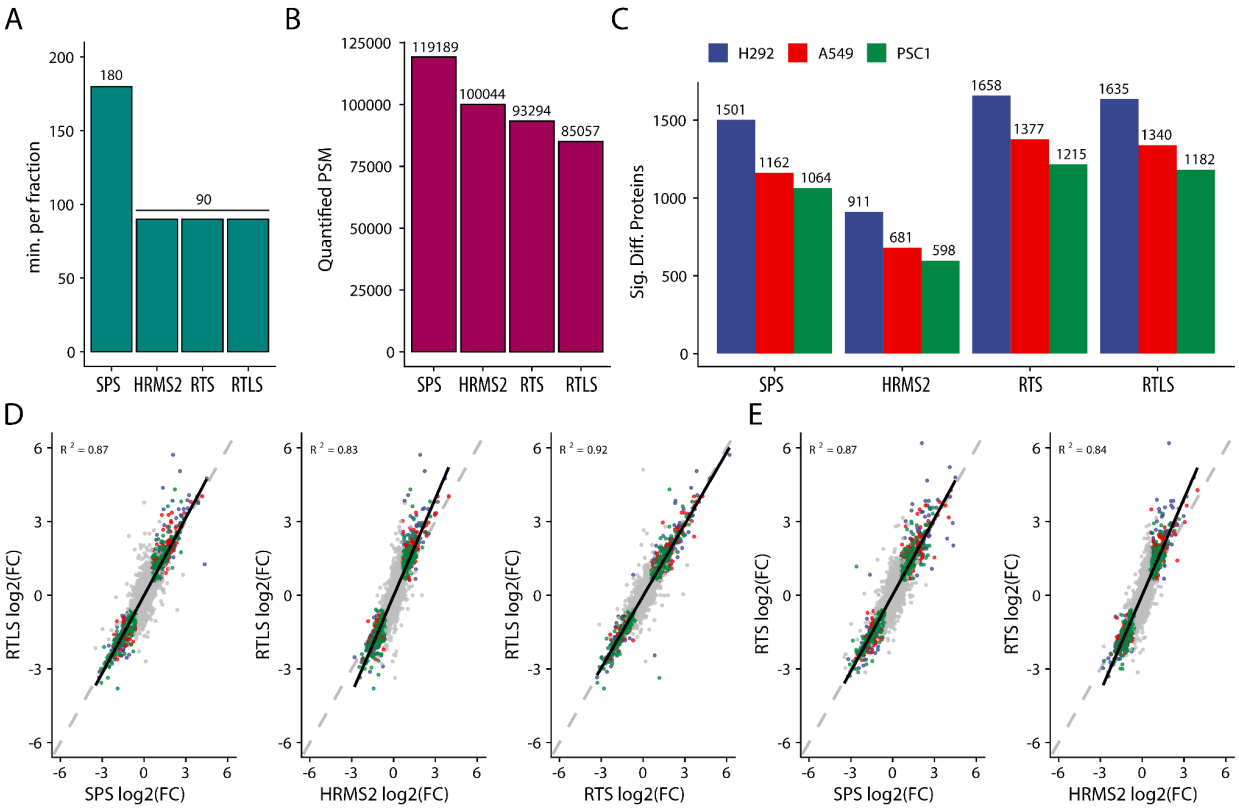


Figure 3-11. RTLS methods comparison on fractionated, whole-proteome samples. A) Total gradient time used per fraction in minutes. B) Total number of quantified peptide spectral matches for each method type. C) Total number of significant differentially abundant proteins (FC > 1.5 and q-value < 0.05) for H292, A549, and PSC1 cell lines. D) Scatter plots comparing the $\log_2$ fold changes of common significantly differential proteins found when comparing RTLS to SPS-MS3, HRMS2, and RTS as well as their respective the coefficients of determination (R^2). E) Same as D for the comparison of RTS versus SPS-MS3 and HRMS2 methods.

Deep, quantitative proteome analysis of belinostat treated cells with RTLS

While single-shot methods are valuable, one of the most common uses for multiplexed proteomics is for the analysis of fractionated proteomes, quantifying differential abundance across conditions. To test RTLS methods across fractionated samples, we treated three human

cell lines (A549, H292, PSC1) with the histone deactylase inhibitor belinostat, using four different acquisition methods: SPS-MS3, HRMS2, RTS, and RTLS. Perturbation with belinostat treatment alters chromatin state and leads to pleotropic remodeling of the proteome after sustained treatments. In keeping with previous analyses, we compared belinostat treatment responses with SPS samples run as twelve 180 minute runs while the HRMS2, RTS, and RTLS methods were collected using half of the gradient time (90 min per fraction, Figure 3-11A). In total, the canonical SPS-MS3 (180 min per fraction) and HRMS2 (90 min per fraction) methods quantified 119,189 and 100,044 PSMs, respectively (Figure 3-11B). The two real-time methods, RTS and RTLS (90 min per fraction), generated fewer total quantified PSMs of 93,294 and 85,057, respectively (Figure 3-11B). Interestingly, though RTLS methods for single shot whole proteome analyses performed similarly to RTS in terms of quantified peptides (Figure 3-9A), for the large, fractionated sample comparison RTLS was more conservative and quantified 91% of the peptides obtained with RTS (Figure 3-11B). However, even though both RTS and RTLS quantified fewer total PSMs, we observed more proteins with significant fold-change differences with RTS and RTLS compared to SPS-MS3 or HRMS2 methods (Figure 3-11C). In totally, RTS and RTLS methods quantified 11.8-14.4% more significant differentially abundant proteins (q-value < 0.05, fold-change > 1.5) across all three cell lines when compared to the SPS-MS3 method (Figure 3-11A). Comparing the $\log_2$ fold-changes of shared significantly quantified proteins, IDA methods have higher absolute fold-changes due to SPS ion selection of b- and y-ions leading to improved quantitative accuracy (Figure 3-11D-E, Figure 3-12).

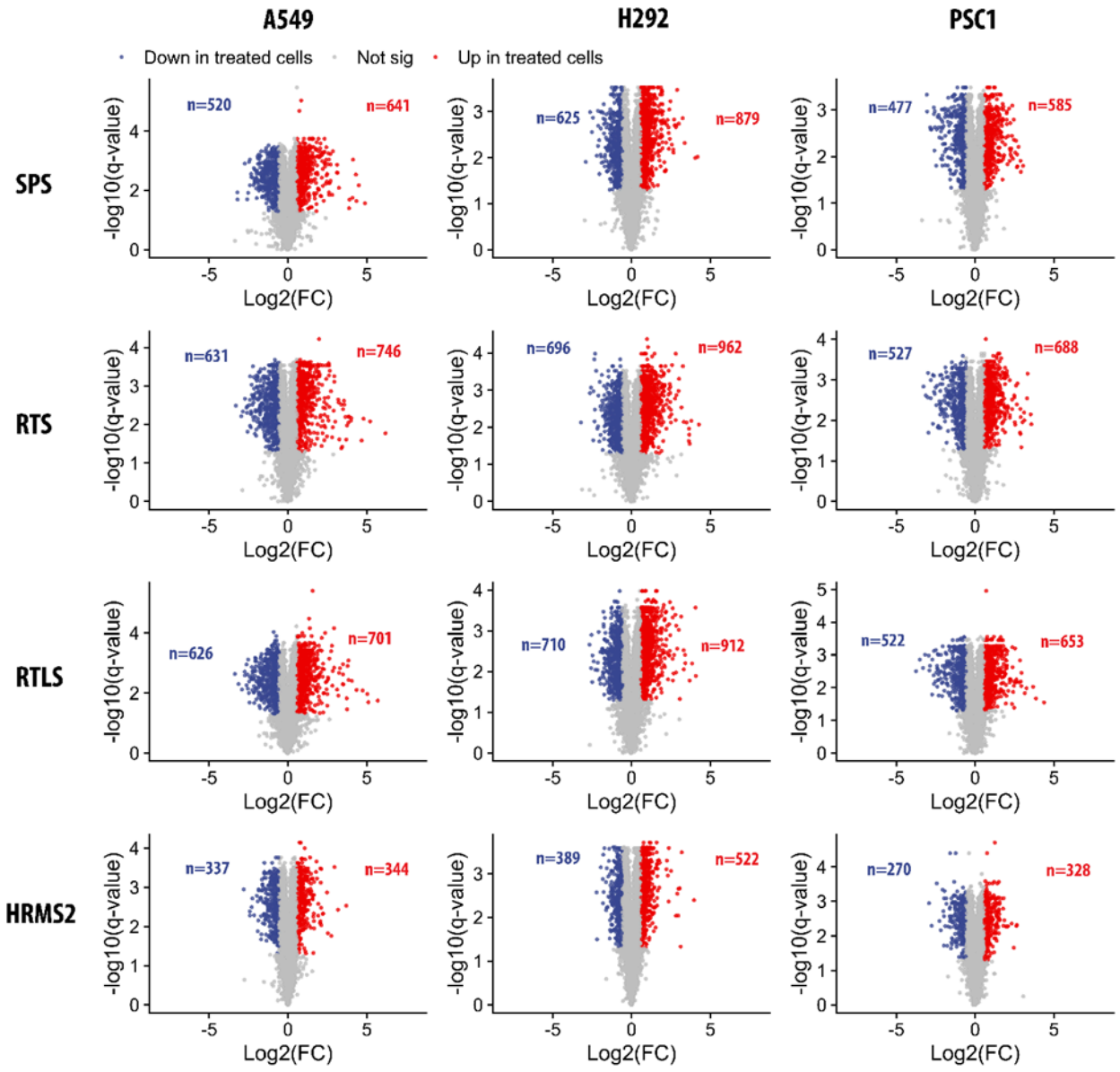


Figure 3-12. Volcano plots from belinostat experiment showing all three cell lines with the different acquisition methods. Proteins significantly more abundant than control are in red and less abundant in blue.

Extension of RTLS methods to new instrumentation

Using our study of cellular responses to belinostat we compared the utility of RTLS when using new instrumentation for sample multiplexed quantitative proteomics workflows. In

particular, we wanted to determine if improved quantitation of peptides when running RTLS methods. Therefore, we compared the reporter ion sensitivity for matched RTLS methods and matched LC systems to quantify the proteomes of A549 cells treated with three HDAC inhibitors (abexinostat, CUDC-101, vorinostat). From fractionated analysis of short gradient runs (60 minute) on both an Orbitrap Eclipse and Orbitrap Ascend using RTLS, we found that the Orbitrap Ascend increased the detected reporter ion signal-to-noise by 127% for peptides and 145% for proteins when running RTLS methods (Figure 3-13).

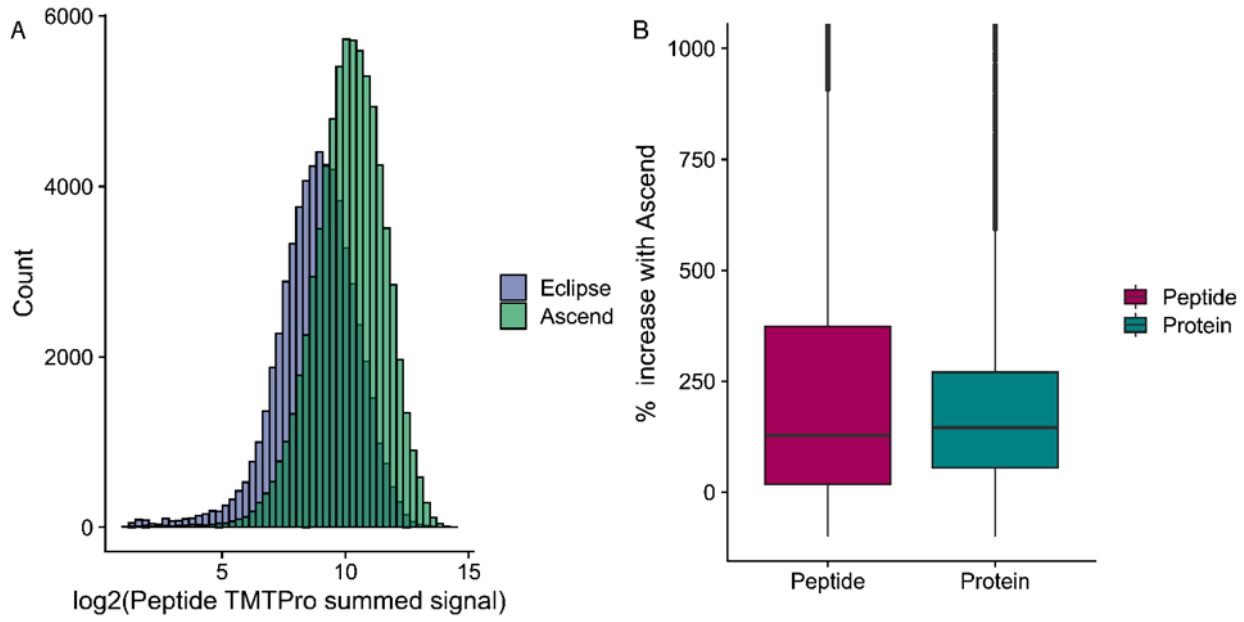


Figure 3-13. Comparison of RTLS methods run on a Thermo Scientific Orbitrap Eclipse Tribrid or a Orbitrap Ascend Tribrid. A) Distributions of peptide level TMTpro signal for the two instruments on fractioned human cell lines. B) Boxplot showing percent signal increase for shared peptides and proteins.

Deconvolution and simultaneous triggering of multiple MS3 scans from chimeric spectra using RTLS

Including the work above, most multiplexed proteomics methods are designed to minimize or mitigate precursor co-isolation, as failure to do so leads to chimeric spectra, which impair accurate quantitation. However, co-isolation remains a general challenge in complex sample analyses. Interestingly, recent work with library searching of label-free samples illustrated that purposefully generating chimeric spectra can increase identifications[62]. We hypothesized that if RTLS can correctly identify multiple precursors in a chimeric MS2 spectra, we could potentially trigger multiple separate MS3 scans from the same MS2 that would lead to an increase in sensitivity and quantified peptides across the run. To this end, we tested methods where the MS2 isolation width was increased from 0.4 Th to 2.0 Th and enabled the multiple precursor search in RTLS. This option allows the search engine to consider multiple precursors within the isolation window when performing the search on a single MS2 spectrum.

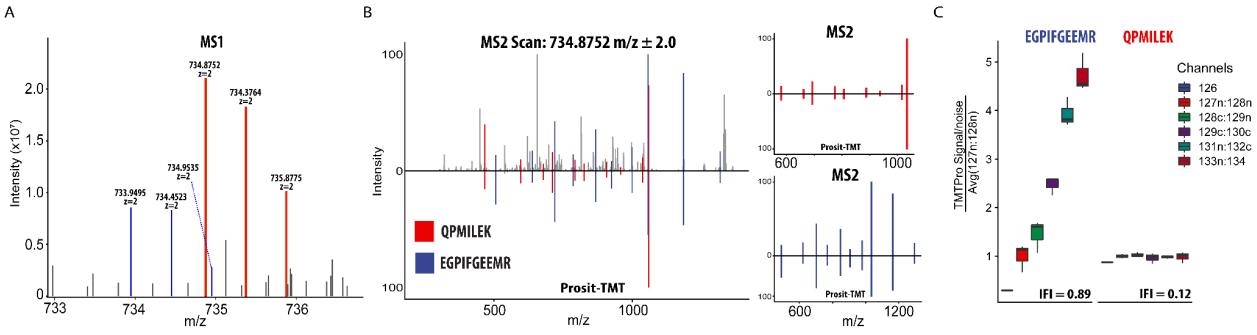


Figure 3-14. Chimeric spectra and RTLS. (A) Acquired MS1 scan with overlapping precursor isotopic envelopes. The red peak for QPMILEK comes from the human EFTU (P49411) and the blue peak for EGPIFGEEMR comes from the yeast EF2 (P32324). (B) ddMS2 triggered from the MS1 in A with matched RTLS fragments colored for both precursors (right) Mirror plots of

the fragments matched by RTLS and the library entry from a concatenated Prosit library of human and yeast peptides. (C) Hypro interference-free indexes (IFI) of the two MS3 spectra that were triggered from the single MS2 spectrum in panel B. Channels are grouped by their expected ratios (see Materials and Methods).

In developing multi-precursor RTLS methods, we found that canonical post hoc searching could not efficiently identify peptides from chimeric spectra. To address this, we performed post hoc searching using CHIMERYS, a newly developed search engine focused on deconvolution of chimeric spectra[36,63]. CHIMERYS successfully validated hundreds of chimeric spectra from our multiple precursor methods. Importantly, we found that RTLS could also correctly identify multiple precursors from a single MS2 scan. As a proof of principle, CHIMERYS identified 77.9% of all PSMs to be derived from chimeric spectrum when using a wide 2.4 Da isolation width (Figure 3-15). In 43.6% of the chimeric spectra, RTLS returned at least one of the CHIMERYS validated peptides as the top precursor match. There was concordance between RTLS and CHIMERYS on multiple PSMs within a spectrum across 13.5% of the validated chimeras. Due to the differing known quantitative profiles of the human and yeast peptides in our HyPro standard we were able to validate detection of chimeric peptide spectra from a single MS2 scan.

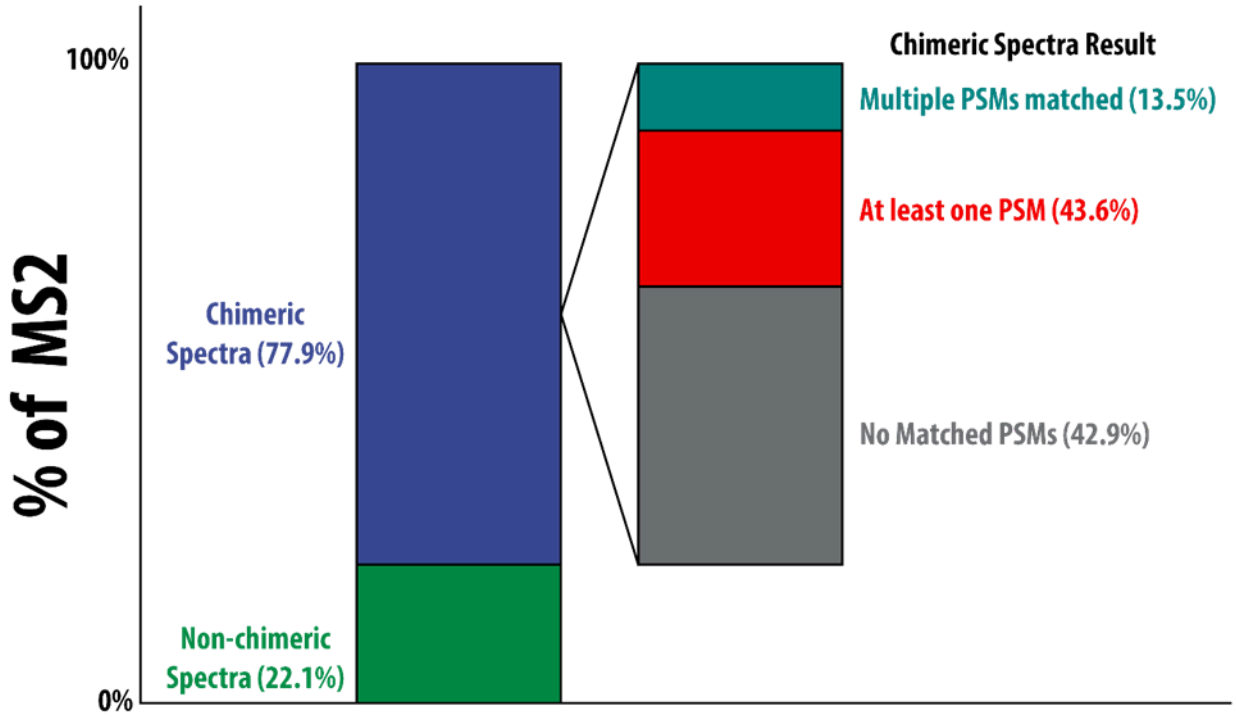


Figure 3-15. Percentage of chimeric MS2 spectra. Spectra were collected from a 120 minute HyPro analysis using 2.4 Da isolation widths. Of the total MS2 spectra, 77.9% were determined by CHIMERYS to be chimeric spectra. Of the chimeric spectra, 13.5% were observed to match to multiple PSMs.

We found that RTLS could properly assign both a human and yeast peptide to a chimeric MS2 spectra and then trigger MS3-based quantification consistent with the known concentrations across TMTpro channels (Figure 3-14C). We calculated interference free indices (IFI) for both the human peptide (IFI = 0.12) and yeast peptide (IFI = 0.89) and observed distinct quantification profiles across TMTpro channels that were consistent with the coisolation and fragmentation of two peptides (Figure 3-14C). In general, we also observed that RTLS more accurately picked the 10 SPS ions selected for quantifying MS3 scans. We found that 94.9% of quantified peptides analyzed with RTLS methods had 0/10 or 1/10 unmapped SPS ions, and this

outperformed RTS and SPS-MS3 methods which had 91.9% and 47.2% of quantified peptides with 0/10 or 1/10 unmapped SPS ions, respectively (Figure 3-16). This suggests high concordance of the real-time triggering peptide identification and the post hoc identified peptide. Future improvements in methods for chimeric spectra generation and detection sensitivity are still needed to implement this method. Despite these considerations, chimeric spectra triggering with RTLS serves as the first step towards addressing chimeric spectra isolation in sample multiplexed proteomics and potentially leveraging wider isolation width methods for isobaric multiplexed samples.

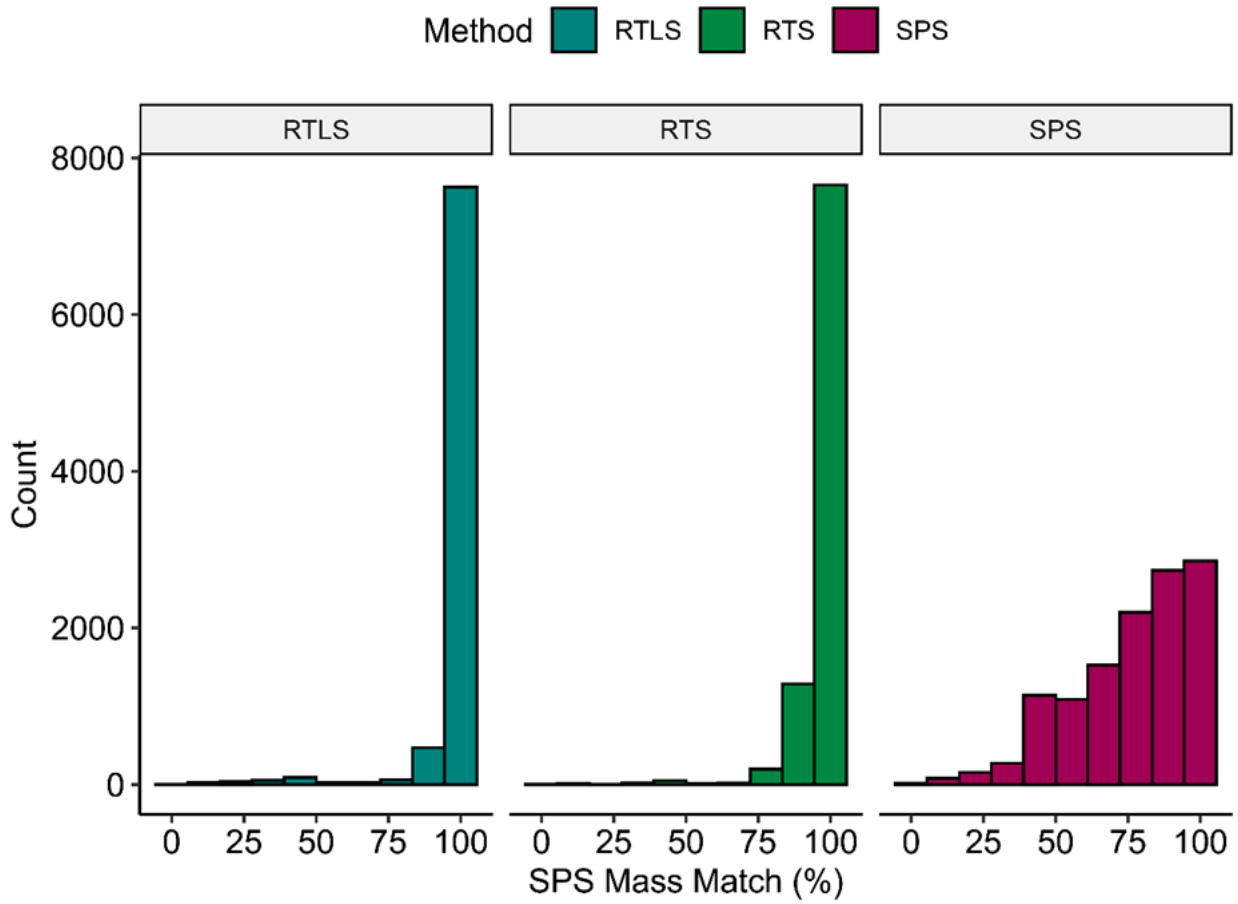


Figure 3-16. The percentage of SPS ions selected by the acquisition methods that are annotated as fragment ions of the post hoc identified peptide.

3.4 Conclusions

We have reported the development and use of RTLS, a modular, integrated intelligent data acquisition strategy based on spectral library searching in real time for sample multiplexed proteomics. The use of RTLS methods resulted in improved instrument efficiency and increased the number of quantified proteins and peptides when compared to traditional methods. Establishing RTLS for multiplexed proteomics lays the groundwork for future work utilizing library searches in areas where library searching can potentially make a large impact. This includes studies of post-translational modifications, where including modifications in the match scoring can lead to large search spaces. As we have shown, RTLS will also be useful in chimeric spectra deconvolution, and in combination with other IDA methods, such as RTS, for highly selective and adaptive instrument methods. In addition to the core methods, we present optimized RTLS scoring weights for sample multiplexed analyses and demonstrate the utility of integrating RTLS and FAIMS for improved discrimination of low confidence peptides. These optimizations can be further improved upon using new instrumentation and the addition of spectral library close out procedures to increase the sensitivity of detection of quantified peptides and proteins. Finally, we demonstrate that RTLS is capable of triggering multiple, quantitatively distinct MS3 spectra from the same MS2 spectrum. Together these findings highlight how new IDA methods can be used to improve sample multiplexed quantitative proteomics methods.

3.5 References

1. Liu, H., Sadygov, R. G. & Yates, J. R., 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193-4201, doi:10.1021/ac0498563 (2004).

2. 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).
3. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**, 937-940, doi:10.1038/nmeth.1714 (2011).
4. Bailey, D. J., McDevitt, M. T., Westphall, M. S., Pagliarini, D. J. & Coon, J. J. Intelligent data acquisition blends targeted and discovery methods. *J Proteome Res* **13**, 2152-2161, doi:10.1021/pr401278j (2014).
5. Graumann, J., Scheltema, R. A., Zhang, Y., Cox, J. & Mann, M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol Cell Proteomics* **11**, M111 013185, doi:10.1074/mcp.M111.013185 (2012).
6. Bailey, D. J. *et al.* Instant spectral assignment for advanced decision tree-driven mass spectrometry. *Proc Natl Acad Sci U S A* **109**, 8411-8416, doi:10.1073/pnas.1205292109 (2012).
7. Yu, Q. *et al.* Sample multiplexing for targeted pathway proteomics in aging mice. *Proc Natl Acad Sci U S A* **117**, 9723-9732, doi:10.1073/pnas.1919410117 (2020).
8. Rose, C. M. *et al.* TomahaqCompanion: A Tool for the Creation and Analysis of Isobaric Label Based Multiplexed Targeted Assays. *J Proteome Res* **18**, 594-605, doi:10.1021/acs.jproteome.8b00767 (2019).
9. Erickson, B. K. *et al.* Active Instrument Engagement Combined with a Real-Time Database Search for Improved Performance of Sample Multiplexing Workflows. *J Proteome Res* **18**, 1299-1306, doi:10.1021/acs.jproteome.8b00899 (2019).

10. Schweppe, D. K. *et al.* Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J Proteome Res* **19**, 2026-2034, doi:10.1021/acs.jproteome.9b00860 (2020).
11. 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, doi:10.1016/1044-0305(94)80016-2 (1994).
12. Keele, G. R. *et al.* Global and tissue-specific aging effects on murine proteomes. *bioRxiv*, 2022.2005.2017.492125, doi:10.1101/2022.05.17.492125 (2022).
13. Furtwangler, B. *et al.* Real-Time Search-Assisted Acquisition on a Tribrid Mass Spectrometer Improves Coverage in Multiplexed Single-Cell Proteomics. *Mol Cell Proteomics* **21**, 100219, doi:10.1016/j.mcpro.2022.100219 (2022).
14. Kuljanin, M. *et al.* Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nat Biotechnol* **39**, 630-641, doi:10.1038/s41587-020-00778-3 (2021).
15. Mitchell, D. C. *et al.* A proteome-wide atlas of drug mechanism of action. *Nat Biotechnol*, doi:10.1038/s41587-022-01539-0 (2023).
16. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **16**, 509-518, doi:10.1038/s41592-019-0426-7 (2019).
17. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: Boosting Peptide Identifications by Chimeric Spectra Identification and Retention Time Prediction. *J Proteome Res* **17**, 2581-2589, doi:10.1021/acs.jproteome.7b00836 (2018).
18. AS, C. S., Bouwmeester, R., Martens, L. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search

engine scoring functions. *Bioinformatics* **35**, 5243-5248, doi:10.1093/bioinformatics/btz383 (2019).

19. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* **20**, e1900334, doi:10.1002/pmic.201900334 (2020).
20. van Bentum, M. & Selbach, M. An Introduction to Advanced Targeted Acquisition Methods. *Mol Cell Proteomics* **20**, 100165, doi:10.1016/j.mcpro.2021.100165 (2021).
21. Plank, M. J. Modern Data Acquisition Approaches in Proteomics Based on Dynamic Instrument Control. *J Proteome Res* **21**, 1209-1217, doi:10.1021/acs.jproteome.2c00096 (2022).
22. Remes, P. M., Yip, P. & MacCoss, M. J. Highly Multiplex Targeted Proteomics Enabled by Real-Time Chromatographic Alignment. *Anal Chem* **92**, 11809-11817, doi:10.1021/acs.analchem.0c02075 (2020).
23. Stopfer, L. E. *et al.* High-Density, Targeted Monitoring of Tyrosine Phosphorylation Reveals Activated Signaling Networks in Human Tumors. *Cancer Res* **81**, 2495-2509, doi:10.1158/0008-5472.CAN-20-3804 (2021).
24. Yu, Q. *et al.* Sample multiplexing-based targeted pathway proteomics with real-time analytics reveals the impact of genetic variation on protein expression. *Nat Commun* **14**, 555, doi:10.1038/s41467-023-36269-7 (2023).
25. Bills, B. *et al.* Novel Real-Time Library Search Driven Data Acquisition Strategy for Identification and Characterization of Metabolites. *Anal Chem* **94**, 3749-3755, doi:10.1021/acs.analchem.1c04336 (2022).

26. Brademan, D. R. *et al.* Improved Structural Characterization of Glycerophospholipids and Sphingomyelins with Real-Time Library Searching. *Anal Chem* **95**, 7813-7821, doi:10.1021/acs.analchem.2c04633 (2023).
27. Ruwolt, M. *et al.* Real-Time Library Search Increases Cross-Link Identification Depth across All Levels of Sample Complexity. *Anal Chem* **95**, 5248-5255, doi:10.1021/acs.analchem.2c05141 (2023).
28. Yates, J. R., 3rd, Morgan, S. F., Gatlin, C. L., Griffin, P. R. & Eng, J. K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal Chem* **70**, 3557-3565, doi:10.1021/ac980122y (1998).
29. Lam, H. Building and searching tandem mass spectral libraries for peptide identification. *Mol Cell Proteomics* **10**, R111 008565, doi:10.1074/mcp.R111.008565 (2011).
30. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom* **5**, 859-866, doi:10.1016/1044-0305(94)87009-8 (1994).
31. Degroeve, S., Maddelein, D. & Martens, L. MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res* **43**, W326-330, doi:10.1093/nar/gkv542 (2015).
32. Gabriel, W. *et al.* Prosit-TMT: Deep Learning Boosts Identification of TMT-Labeled Peptides. *Anal Chem* **94**, 7181-7190, doi:10.1021/acs.analchem.1c05435 (2022).
33. Gabriel, W., Giurcoiu, V., Lautenbacher, L. & Wilhelm, M. Predicting fragment intensities and retention time of iTRAQ- and TMTPro-labeled peptides with Prosit-TMT. *Proteomics* **22**, e2100257, doi:10.1002/pmic.202100257 (2022).

34. Yang, K. L. *et al.* MSBooster: Improving Peptide Identification Rates using Deep Learning-Based Features. *bioRxiv*, 2022.2010.2019.512904, doi:10.1101/2022.10.19.512904 (2022).
35. Declercq, A. *et al.* MS(2)Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol Cell Proteomics* **21**, 100266, doi:10.1016/j.mcpro.2022.100266 (2022).
36. Zolg, D. P. *et al.* INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. *Rapid Commun Mass Spectrom*, e9128, doi:10.1002/rcm.9128 (2021).
37. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat Commun* **11**, 146, doi:10.1038/s41467-019-13866-z (2020).
38. Shiferaw, G. A. *et al.* Sensitive and Specific Spectral Library Searching with CompOmics Spectral Library Searching Tool and Percolator. *J Proteome Res* **21**, 1365-1370, doi:10.1021/acs.jproteome.2c00075 (2022).
39. Dorl, S., Winkler, S., Mechtler, K. & Dorfer, V. MS Ana: Improving Sensitivity in Peptide Identification with Spectral Library Search. *J Proteome Res* **22**, 462-470, doi:10.1021/acs.jproteome.2c00658 (2023).
40. Cranney, C. W. & Meyer, J. G. CsoDIAq Software for Direct Infusion Shotgun Proteome Analysis. *Anal Chem* **93**, 12312-12319, doi:10.1021/acs.analchem.1c02021 (2021).
41. Searle, B. C., Shannon, A. E. & Wilburn, D. B. Scribe: Next Generation Library Searching for DDA Experiments. *J Proteome Res* **22**, 482-490, doi:10.1021/acs.jproteome.2c00672 (2023).

42. Ruwolt, M. *et al.* Real-time library search increases cross-link identification depth across all levels of sample complexity. *bioRxiv*, 2022.2011.2016.516769, doi:10.1101/2022.11.16.516769 (2022).

43. Navarrete-Perea, J., Gygi, S. P. & Paulo, J. A. HYpro16: A Two-Proteome Mixture to Assess Interference in Isobaric Tag-Based Sample Multiplexing Experiments. *J Am Soc Mass Spectrom* **32**, 247-254, doi:10.1021/jasms.0c00299 (2021).

44. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined Tandem Mass Tag (SL-TMT) Protocol: An Efficient Strategy for Quantitative (Phospho)proteome Profiling Using Tandem Mass Tag-Synchronous Precursor Selection-MS3. *J Proteome Res* **17**, 2226-2236, doi:10.1021/acs.jproteome.8b00217 (2018).

45. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl* **9**, 745-754, doi:10.1002/prca.201400164 (2015).

46. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods Mol Biol* **1550**, 339-368, doi:10.1007/978-1-4939-6747-6_23 (2017).

47. Cheng, C. Y., Tsai, C. F., Chen, Y. J., Sung, T. Y. & Hsu, W. L. Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications. *J Proteome Res* **12**, 2305-2310, doi:10.1021/pr301039b (2013).

48. Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* **13 Suppl 16**, S1, doi:10.1186/1471-2105-13-S16-S1 (2012).

49. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667, doi:10.1002/pmic.200600625 (2007).

50. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* **604**, 55-71, doi:10.1007/978-1-60761-444-9_5 (2010).

51. Paulo, J. A., O'Connell, J. D. & Gygi, S. P. A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J Am Soc Mass Spectrom* **27**, 1620-1625, doi:10.1007/s13361-016-1434-9 (2016).

52. Schweppe, D. K. *et al.* Characterization and Optimization of Multiplexed Quantitative Analyses Using High-Field Asymmetric-Waveform Ion Mobility Mass Spectrometry. *Anal Chem* **91**, 4010-4016, doi:10.1021/acs.analchem.8b05399 (2019).

53. Team, R. C. R: A language and environment for statistical computing. (2013).

54. Zhang, X., Li, Y., Shao, W. & Lam, H. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* **11**, 1075-1085, doi:10.1002/pmic.201000492 (2011).

55. Shen, J. Q. *et al.* Spectral Library Search Improves Assignment of TMT Labeled MS/MS Spectra. *Journal of Proteome Research* **17**, 3325-3331, doi:10.1021/acs.jproteome.8b00594 (2018).

56. Huber, F. *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol* **17**, e1008724, doi:10.1371/journal.pcbi.1008724 (2021).

57. Koo, I., Kim, S. & Zhang, X. Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry. *J Chromatogr A* **1298**, 132-138, doi:10.1016/j.chroma.2013.05.021 (2013).

58. Schollee, J. E. *et al.* Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. *J Am Soc Mass Spectrom* **28**, 2692-2704, doi:10.1007/s13361-017-1797-6 (2017).

59. Yen, C. Y., Houel, S., Ahn, N. G. & Old, W. M. Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol Cell Proteomics* **10**, M111 007666, doi:10.1074/mcp.M111.007666 (2011).

60. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem* **78**, 5678-5684, doi:10.1021/ac060279n (2006).

61. Paulo, J. A., Navarrete-Perea, J. & Gygi, S. P. Multiplexed proteome profiling of carbon source perturbations in two yeast species with SL-SP3-TMT. *J Proteomics* **210**, 103531, doi:10.1016/j.jprot.2019.103531 (2020).

62. Mayer, R. L. *et al.* Wide Window Acquisition and AI-based data analysis to reach deep proteome coverage for a wide sample range, including single cell proteomic inputs. *bioRxiv*, 2022.2009.2001.506203, doi:10.1101/2022.09.01.506203 (2022).

63. Frejno, M. *et al.* CHIMERYS: An AI-Driven Leap Forward in Peptide Identification. (Annual Meeting of the American Society for Mass Spectrometry, 2021).

Chapter 4

DNA O-MAP UNCOVERS THE MOLECULAR NEIGHBORHOODS ASSOCIATED WITH SPECIFIC GENOMIC LOCI

This chapter is adapted from: Liu, Y., McGann, C. D., Krebs, M., Perkins Jr, T. A., Fields, R., Camplisson, C. K.,Nwizugbo, D.Z., Hsu, C., Avanessian, S.C., Tsue, A.F. and Kania, E.E., Shechner D.M., Beliveau B.J., Schweppe, D. K. (2024). DNA O-MAP uncovers the molecular neighborhoods associated with specific genomic loci. eLife, 13

4.1 Introduction

Eukaryotic cells store their genetic material in the form of chromatin, a DNA-protein complex. The function of a eukaryotic DNA locus is executed through the cooperation between its nucleotide sequence and the hundreds of protein factors assembled around it. DNA-protein interactions thus play a fundamental role in regulating both the genome's structure and message storing functions[1]. Therefore, developing methods to decipher DNA-protein interactions in cells has been a focus of technology development efforts for decades[2]. For instance, chromatin immunoprecipitation followed by sequencing (ChIP-seq[3]), which has emerged as a core technology for epigenomics[4], surveys the genome-wide binding profile of a target DNA-associated protein. ChIP-seq and related technologies (e.g., DamID[5], CUT&Tag[6]) have produced an abundance of high-quality datasets that enabled the establishment of database consortia such as ENCODE[7,8] and IHEC[9], and significantly accelerated chromatin state annotation efforts[10,11]. Such methods, which profile DNA-protein interactions through a protein-centric lens, require the *a priori* knowledge of which protein(s) to target and rely on the

availability of suitable reagents such as antibodies or genetically engineered cell lines. By targeting a single protein at a time, these methods also inherently ignore the context of protein complexes or transient interactions that may be present at a given locus.

In addition to methods that profile the DNA bound by specific proteins, efforts have been dedicated to addressing the inverse problem—identifying the full collection of proteins assembled on a given DNA locus[12–15]. Such methods include the foundational proteomics of isolated chromatin segment (PICh) technology, which uses a biotinylated oligonucleotide (oligo) probe to affinity label specific genomic DNA intervals via *in situ* hybridization (ISH)[16]. To enhance the stability of probe-chromatin interactions throughout the purification workflow, PICh utilizes oligos containing locked nucleic acid residues[17], which are highly efficient as hybridization probes against repetitive DNA targets but cost-prohibitive to use to target non-repetitive intervals that require dozens to hundreds of probes to produce visible signal[18]. As noted in follow-up work, PICh was effective for repeat sequences but would require significant additional work to extend to more complex genomic sequences[19]. Additionally, even with the increased stability gained from the use of locked nucleic acid probes, the probe-chromatin hybrids can be difficult to maintain when coupled with stringent purification washes[19], limiting detection sensitivity. As a consequence, an input of one trillion cells was required for one purification and successful identification of proteins interacting with telomeres[16].

To reach a higher degree of enrichment, which is critical for lower abundance DNA targets, an alternative strategy is to directly biotinylate the proteins that occupy a target DNA locus. This biotinylation can be achieved via targeted proximity labeling using promiscuous biotin ligases[20,21] or the engineered ascorbate peroxidase (APEX/APEX2) enzymes[22,23]. Since the development of APEX, several methods including C-BERST[12] and GLoPro[13], have combined

APEX with CRISPR genome targeting to endow it with locus specificity. This involves fusing APEX to a catalytically dead RNA-guided nuclease, Cas9 (dCas9) and directing the fusion enzyme to a specific locus of interest by single guide RNAs (sgRNAs). The locus-docked dCas9-APEX biotinylates the neighboring proteins on electrophilic amino acid side chains, such as tyrosine, enabling protein purification and subsequent identification by mass spectrometry (MS). In the case of GLoPro, APEX-based proximity labeling enhanced protein detection sensitivity, reducing the input required for each replicate analysis to ~300 million cells—a 10-fold reduction in cell input compared to PICh, which used 3 billion cells. Nevertheless, a notable limitation of CRISPR-guided proximity labeling is requiring the introduction of the fusion dCas9-APEX enzyme and sgRNAs into a suitable host cell line. Since a successful locus purification canonically requires tens to hundreds of millions of cells, if not more, most current methods aim to create stable cell lines for this purpose. These requirements limit the use of previous locus proteomics methods since efficient and well-tolerated gene delivery remains a major challenge and considerable effort in primary cells[24]. In addition, the labeling reagents necessary for APEX-based proximity labeling—hydrogen peroxide and biotin phenoxyl radicals—are toxic to cells and living organisms, limiting the use of CRISPR-based proximity labeling to cell lines amenable to genetic engineering. Owing to the large numbers of cells required and the need to maximize sensitivity, previous methods often compared only 1–2 biological replicates[12,13]. In some cases, this was limited by the use of stable-isotope-based quantification methods that can only multiplex up to three samples per analysis[12]. Thus, an unmet need exists for extensible methods capable of scaling and profiling multiple genomic loci. Moreover, these methods would ideally be capable of scaling and multiplexing comparisons

between multiple local proteomes or one local proteome in response to multiple stimuli or perturbations.

We address these pressing technical limitations by introducing DNA O-MAP, a locus purification method that uses oligo-based ISH probes to recruit peroxidase activity to specific DNA intervals. DNA O-MAP builds on our previously introduced RNA O-MAP[25] and pSABER[26] techniques, which target peroxidase activity to specific RNAs and RNAs/DNA intervals for purification or visualization, respectively. Here, we describe a cost-effective and scalable bulk hybridization and biotinylation workflows capable of processing millions of cells in parallel in just a few days, and demonstrate the recovered material is compatible with sample multiplexed proteomics[27]. We benchmark the specificity of our approach by recovering telomere-specific DNA binding proteins after targeting telomeric DNA. We further showcase the scalability and sample multiplexing capacity of DNA O-MAP by distinguishing the DNA-associated proteomes around human pericentromeric alpha-satellite repeats, telomeres, and mitochondrial genomes in quadruplicates using tandem mass tags[27]. Finally, we establish that DNA O-MAP can be used to capture functionally relevant DNA-DNA interactions, read out by DNA sequencing, from intervals as small as 20 kilobases. We anticipate that the flexible targeting, scalable protocol, and robust labeling capabilities provided by DNA O-MAP will lead to its adoption as a platform technology for uncovering locus-specific chromatin interactions.

4.2 Methods

Cell culture and fixation

Colorectal cancer HCT-116 cells were grown in ATCC-formulated McCoy's 5A Medium Modified (ATCC 30-2007) supplemented with 10% fetal bovine serum and 100 U/ml

Penicillin-Streptomycin at 37°C in a humidified atmosphere of 5% CO2. For each purification, 20 million HCT-116 cells were seeded into one T-500 flask (Thermo Scientific 132867) to culture for 36-48 hours to reach 90–120 million cells. Before collection, cells were briefly rinsed once with Dulbecco's phosphate buffered saline (DPBS) and then incubated with 25 ml of TrypLE Express Enzyme (Gibco 12604-021) at 37°C for two minutes or until loosely attached. The cell suspension was collected into two 50 ml conical tubes and the T-500 flask was rinsed with DPBS. The wash was combined with the cell suspension and centrifuged at 300 G for 5 minutes. After a DPBS wash to remove remaining TrypLE, cells were fixed in 4% paraformaldehyde (wt/vol) (Electron Microscopy Sciences 15710) in PBS in suspension at room temperature for 10 minutes with rotation, followed by 125 mM Glycine quenching for 5 minutes at room temperature with rotation and 15 minutes on ice. Fixed cells were collected by centrifugation at 350G for 5 minutes, and stored in fresh DPBS at 4°C until liquid-phase hybridization. Fixed cells were used within 3-5 days.

Primary oligo probes

Primary oligos targeting the human alpha satellite repeat and telomere were purchased as individually column-synthesized DNA oligos from Integrated DNA Technologies. Probe sets targeting mtDNA (chrM:1-16,569), chr3 left anchor (chr3:187,729,712-187,749,712), chr3 right anchor (chr3:188,939,711-188,964,711), chr10 non-looping anchor (chr10:123,187,984-123,207,984), chr10 right anchor (chr10:123,957,984-123,977,984), and chr19 right anchor (chr19:33,750,000-33,775,000) were designed using PaintSHOP[68] and ordered in oPool format from Integrated DNA Technologies. More than 300 primary oligos were designed to cover each single-copy DNA interval to ensure a sufficient number of probes at the

locus for FISH. The sequences of the oligo and oligo pools used are listed in Supplementary Dataset 1.

Primer exchange reaction (PER)

To extend primary oligos with PER concatemers, reactions were set up as previously described[71] in 100 ul-volume containing 10 mM MgSO4, 300 uM dATP/dCTP/dTTP mix, 100 nM Clean.G hairpin, 80 U/ml Bst DNA Polymerase, Large Fragment (NEB M0275L), 1 uM hairpin, and 1 uM primary oligos in PBS. To verify the length of primary oligos, the reactions were assessed with denaturing polyacrylamide gel electrophoresis. Primary oligos extended to 300-500 nucleotides were used in hybridizations downstream. Unpurified reactions were dehydrated using vacuum concentrators and stored dry at −20°C until hybridization.

In-solution hybridization and biotinylation of cell pellets

Oligo hybridizations were performed on cells in solution for the cost-effectiveness of primary and secondary oligos. Fixed cells were split into 6e7 cell aliquots in 1.5 ml microcentrifuge tubes. All washes and buffer exchanges were performed as follows: centrifuging at 350G for 3.5 minutes or until pelleted, pouring away used buffers from the pellets, adding new buffers, and gentle shaking or low speed vortexing to dislodge cell pellets into tiny clusters or cell suspensions for incubations or washes. Cells in fresh wash buffer were rotated on a low speed nutator for 5 minutes.

Cells were rinsed once with fresh phosphate buffer saline (PBS), and permeabilized in PBS-0.5% TritonX-100 (Sigma T8787) for 10 minutes with nutation. After a PBS-0.1% Tween20 (PBS-T) (Sigma T2287) wash, permeabilized cells were incubated in 0.1 N hydrochloric acid (HCl) for 5 minutes. After a PBS-T wash to remove acid, cells were incubated in PBS-T-0.5% hydrogen

peroxide to block endogenous peroxidases. After a 2X saline sodium citrate-0.1% Tween20 (2X SSC-T) wash to remove acid, cells were incubated in 2X SSC-T-50% formamide for 20 minutes at 60°C on a Thermomixer C dry block (Eppendorf 2231001005). Cells were exchanged into primary hybridization buffer (Hyb1) comprising 2X SSC-T, 50% (vol/vol) formamide, 10% (wt/vol) dextran sulfate, 0.4 μg/ul RNAse A, and ~1 μM extended primary oligos (resuspended dry, unpurified PER reactions). The cell-Hyb1 mixture was distributed into PCR strip tubes at 1e7-1.5e7 cells in 100 μL volumes. The cells were denatured and primary oligos were hybridized to the genome in the PCR strip tubes in a thermocycler using the cycling protocol: 78°C 3 minutes, 37°C ∞ incubating overnight for more than 18 hours.

The next day, cells were rinsed with 60°C 2X SSC-T into 1.5 ml microcentrifuge tubes, followed by two 2X SSC-T buffer exchanges to remove residual Hyb1. Cell pellets were then washed in 1 ml 2X SSC-T at 60°C, followed by two two-minute washes in 2X SSC-T at room temperature. Fully washed cell pellets were exchanged into 1 ml PBS, and then exchanged into 100 nM secondary HRP oligo that map to the PER concatemer sequence on the primary oligo (custom synthesis by Integrated DNA Technologies or Bio-Synthesis Inc) in PBS. Secondary hybridization was performed at 37°C with nutation for one hour. Cell pellets underwent three 5-minute washes in 1 ml PBS-T at 37°C with nutation. Fully washed cells were incubated in 5 uM desthiobiotin tyramide (Iris Biotech LS-1660) and 1 mM hydrogen peroxide in PBS-T for 5 minutes at room temperature with nutation. To quench the HRP activity, biotinylated cells were washed twice in 10 mM sodium ascorbate and 10 mM sodium azide in PBS-T for 5 minutes at room temperature with nutation. Quenched cells were washed with PBS to remove residual sodium azide. After sampling cells for quality control, the cell pellets were stored dry in −80°C until chromatin solubilization and affinity purification.

Microscopy-based quality control assays for hybridization and biotinylation

We routinely sample cells along the workflow of preparing AP-MS or NGS samples to monitor the locus specificity of primary oligo hybridization. To assess the quality of primary oligo hybridization, we sampled roughly 5% of fully washed cells from primary hybridization to a new 1.5 ml tube. Cells were incubated with 400 nM fluorescent oligos in PBS at 37°C for an hour with nutation. Hybridized cells underwent three washes in 1 ml PBS-T at 37°C with nutation to remove unbound fluorescent oligos. Washed cells were immobilized on glass slides with Slowfade Gold Antifade Mountant with DAPI (Thermo Fisher S36938) and coverslips for confocal imaging of FISH signal.

We assessed the quality of biotinylation specificity for all samples entering the proteomics or genomics workflow. Roughly 5% of fully quenched cells were sampled into a new 1.5 ml tube and incubated with 0.5-1 μg/ml Alexa Fluor 647-streptavidin (Thermo Fisher S32357) in PBS-T, 1% bovine serum albumin at 37°C for 30 minutes with nutation. Stained cells underwent four washes in 1 ml PBS-T at 37°C with nutation to remove unbound Alexa Fluor 647-streptavidin conjugate. Washed cells were immobilized on glass slides with Slowfade Gold Antifade Mountant with DAPI and coverslips for confocal imaging of Alexa-Fluor 647-streptavidin signals.

Confocal microscopy

Confocal imaging was performed using a Yokogawa CSU-W1 SoRa spinning disc confocal device attached to a Nikon ECLIPSE Ti2 microscope. Excitation light was emitted at 30% of maximal intensity from 405 nm, 488 nm, 561 nm, or 640 nm lasers housed inside of a Nikon LU-NF laser unit. Laser excitation was delivered via a single-mode optical fiber into the

CSU-W1 SoRa unit. Excitation light was directed through a microlens array disk and a SoRa spinning disk containing 50 um pinholes to the rear aperture of a 100x N.A. 1.49 Apo TIRF oil immersion objective lens by a prism in the base of Ti2. Emission light was collected by the same objective and directed by a prism in the base of Ti2 back into the SoRA unit, where it was relayed by a 1x lens (conventional imaging) or 2.8x lens (super-resolution imaging) through the pinhole disk and then directed to the emission path by a quad-band dichroic mirror (Semrock Di01-T405/488/568/647-13X15X0.5). Emission light was then spectrally filtered by one of four single-bandpass filters (DAPI:Chroma ET455/50M; ATTO488: Chroma ET525/36M; ATTO565:Chroma ET605/50M; Alexa Fluor 647: Chroma ET705/72M) and focused by a 1x relay lens onto an Andor Sona 4.2B-11 camera with a physical pixel size of 11 um, resulting in an effective resolution of 110 nm (conventional), or 39.3 nm (super-resolution). The Sona was operated in 16-bit mode with rolling shutter readout and exposure times of 70-300 ms.

FISH-biotinylation co-localization experiment

Fixed cells were split into 5e6 cell aliquots in 1.5 ml microcentrifuge tubes. Primary hybridization and washes were performed similarly to described in the in-solution hybridization and biotinylation of cell pellets with fewer cells. Fully washed cell pellets were exchanged into a secondary co-hybridization buffer containing 30 nM of fluorescent oligos and 100 nM of HRP-oligos in PBS, instead of solely HRP-oligos, for simultaneous hybridization of both species. After washes and biotinylation, the pellets were stained with 0.5-1 μg/ml Alexa-Fluor 647-streptavidin. Cells were immobilized on glass slides with Slowfade Gold Antifade Mountant with DAPI and coverslips for confocal imaging of both FISH and Alexa-Fluor 647-streptavidin signals.

Affinity Purification and sample preparation for proteomics

Biotinylated cell pellets were removed from −80°C to thaw at room temperature. Each cell pellet was resuspended in roughly 0.9 ml of lysis buffer consisting of 1% SDS and 200 mM EPPS with protease inhibitors (Roche 11836170001). The cell mixture was boiled at 95°C for 30 minutes. The boiled cell mixture was sonicated at 4°C using a Covaris LE-220 focused ultrasonicator with the following protocol: 300W peak incident power, 50% duty factor, 200 cycles per burst, with a treatment time of 420 seconds in 1-ml milliTUBEs with AFA fiber (Covaris 520135). The sonicated cell mixture was boiled for a second time at 95°C for 30 minutes. The boiled lysates were cleared by centrifuging at 21130 G for 30 minutes in an Eppendorf 5424 Microcentrifuge at room temperature. The supernatants were transferred to a fresh 1.5-ml tube. To prevent any remnants of cell debris, the supernatants were cleared for a second time by centrifuging at 21130 G for 30 minutes and the supernatants were transferred to a fresh 1.5-ml tube. The supernatants were stored in −80°C until protein quantification.

The cleared cell lysates were quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher 23225). Pierce Streptavidin Magnetic Beads (Thermo Fisher 88817) were washed using 1% SDS, 200 mM EPPS lysis buffer three times before use. From each labeled cell pellet, 2.17 milligrams of protein was used to couple with 500 μg of streptavidin beads in a Protein Lo-Bind tube (Eppendorf EP022431081). The lysates were incubated with the bead slurry for one hour at room temperature with nutation allowing biotinylated proteins to bind. The coupled beads were collected and separated from the flow-through using a magnetic rack (Sergi Lab Supplies 1005a). After the flow-through was removed, the beads underwent the following washes: 2% SDS with 20 mM EPPS twice, 0.1 M Na2CO3, 2 M urea, and 1 M KCl with 20 mM EPPS twice. All washes were performed as follows: after immobilizing the beads on a magnetic rack for 5

minutes, the supernatant was removed, and the beads were resuspended in the new wash buffer and incubated for 5 minutes with nutation. Finally, the beads were rinsed once with 20 mM EPPS to remove the excess salt.

The washed streptavidin beads were resuspended in 50 µl of 5 mM TCEP, 200 mM EPPS, pH 8.5 for a 20-minute on-bead protein reduction. The proteins were alkylated on-bead using 10 mM iodoacetamide for one hour in the dark. Then DTT was added to the final concentration of 5 mM to quench the alkylation for 15 minutes. The beads were rinsed twice with 200 mM EPPS for on-bead digest. Assuming 20 µg of eluate protein, 200 ng LysC (Wako) was added to the beads in a 50-ul volume and incubated for 16 hours with vortexing. The next day, 200 ng of trypsin (Promega V5113) was added to the beads and incubated for six hours at 37°C at 200 rpm. After digestion, the peptide-containing supernatant was collected in a fresh 0.5-ml Protein Lo-Bind tube. The beads were rinsed once with 100 µl 50% acetonitrile, 5% formic acid and the wash was combined with the peptides. Peptides were desalted via the stop and go extraction (StageTip)[72] method and dried in a vacuum concentrator.

For label free telomere-enriched samples, one sample consisted of HCT-116-Rad21-mAID cells[73]. For samples intended to be multiplexed, dried, desalted peptides were reconstituted in 4 µl of 200 mM EPPS, pH 8.5. The peptides were labeled using 25 µg of TMTpro 16plex Label Reagents (Thermo Fisher A44520) at 33.3% acetonitrile for one hour at room temperature. The labeling reaction was quenched with the addition of 1 µl of 5% hydroxylamine and incubated at room temperature for 15 minutes. The pooled sample was acidified using formic acid and peptides were desalted using a StageTip cartridge. Peptides were eluted in 70% acetonitrile, 1% formic acid and dried by vacuum centrifugation

Mass Spectrometry Data Acquisition Methods and Analysis

Samples were resuspended in 5% acetonitrile/2% formic acid prior to being loaded onto an in-house pulled C18 (Thermo Accucore, 2.6 Å, 150 μm) 30 cm column. Peptides were eluted over 180 min gradients running from 96% Buffer A (5% acetonitrile, 0.125% formic acid) and 4% buffer B (95% acetonitrile, 0.125% formic acid) to 30% buffer B. Sample eluate was electrosprayed (2700 V) into a Thermo Scientific Orbitrap Eclipse mass spectrometer for analysis. High field asymmetric waveform ion mobility spectrometry (FAIMS) was set at "standard" resolution, 4.6 L/min gas flow, and 3 CVs: −40/–60/–80 were used. MS1 scans were conducted at 120,000 resolving power with a 50 ms max injection time, and the AGC target set to 100%. Peaks from the MS1 scans were filtered by intensity (minimum intensity >5 × 103), charge state ($2 \leq z \leq 6$), and detection of a monoisotopic mass (monoisotopic precursor selection, MIPS). Dynamic exclusion was used, with a duration of 90 s, repeat count of 1, mass tolerance of 10 ppm, and the "exclude isotopes" option checked. For each MS1, 8 data-dependent MS/MS scans were collected. MS/MS scans were conducted in the linear ion trap with the "rapid" scan rate, 50 ms max injection time, AGC target set to 200%, CID collision energy of 35% with 10 ms activation time, and 0.5 m/z isolation window. For TMTPro labelled samples, an MS3 scan was also included in the method. Unless otherwise noted in the methods, the real-time search filter was enabled[43]. Using a human fasta downloaded from Uniprot, fixed modifications for the TMTpro mass (+304.207146) were added to n-terminal residues and lysines. Carbamidomethly (+57.021464) was added for cysteines. Oxidation (+15.9949) was added as a variable modification on methionines. Missed cleavages were set to maximum of 1. "TMT mode" was enabled and thresholds of 1 and 0.05 for Xcorr and dCn respectively were used as minimums to trigger SPS-MS3 scans. SPS ions were set to 10 and MS3 scans were performed at a resolving

power of 50,000, with an HCD collision energy of 45%, AGC of 200%, with a maximum injection time of 200 ms.

Label-free mass spectrometry data was analyzed with MSFragger[74] search algorithm searched against a full human protein database with forward and reverse protein sequences. Fixed modifications included Carbamidomethyl (+57.021464) on cysteines. Variable modifications included were Oxidation (+15.9949) on methionine and formylation (+27.994915) on lysines. Peptides up to 2 missed cleavages were included. Peptide spectral matches and proteins were filtered to a 1% false discovery rate using Percolator[75].

Multiplexed raw mass spectrometry data was analyzed using the Comet[76] search algorithm, searched against a full human protein database with forward and reverse protein sequences (Uniprot 10/2020). Precursor monoisotopic peaks were estimated using the Monocle package. Fixed modifications included TMTpro (+304.207146) on n-terminal residues and lysines and Carbamidomethyl (+57.021464) on cysteines. Variable modifications included were Oxidation (+15.9949) on methionine and formylation (+27.994915) on lysines. Peptides up to 2 missed cleavages were included. Peptide spectral matches and proteins were filtered to a 1% false discovery rate using the rules of parsimony and protein picking. Protein quantification was done using signal-to-noise estimates of reporter ions. Samples were column normalized for total protein concentration. After filtering for contaminants, we performed a two-sided t-test comparing each O-MAP condition using Benjamini-Hochberg adjusted p values (i.e. q-values). Log2 fold changes of the mean of the biological replicates were also calculated for each biological condition. Human Protein Atlas[54] subcellular locations were downloaded and the "main location" was assigned to each protein with a supported or enhanced reliability level. SAINT scores and interaction false discovery rates were calculated with the SAINTexpress

software[77,78]. Significant hits were those with a SAINT calculated FDR less than 1%[79]. BioPlex interaction networks were accessed through the online BioPlex Explorer[80] (https://bioplex.hms.harvard.edu/). Networks were imaged using Cytoscape 3.10.02[81]. Protein complex members were accessed through CORUM[82]. Gene set enrichment analysis was performed with clusterProfiler[83] and fgsea[84] packages.

Preparation of soluble chromatin for affinity purification followed by next generation sequencing

For confirmation of single-copy O-MAP labeling, loop anchor-biotinylated pellets of 10-20 million cells were removed from −80°C to thaw at room temperature. Each cell pellet was resuspended in an SDS lysis buffer consisting of 1% SDS and 200 mM EPPS with protease inhibitors. The cell mixture was sonicated at 4°C using a Covaris LE-220 focused ultrasonicator with the following protocol: 300W peak incident power, 15% duty factor, 200 cycles per burst, with a treatment time of 20-30 minutes in 130-µl microTUBEs with AFA fiber (Covaris 520077). After the samples had returned to room temperature, the sheared fixed chromatin was transferred to fresh 1.5-ml Protein Lo-Bind tubes and centrifuged at 21130 G for 10 minutes to pellet cellular debris. The supernatants were transferred to a new set of tubes. The cleared chromatin samples were quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher 23225). Next, 50 µl of sheared chromatin was sampled for reverse crosslinking, DNA extraction, and gel electrophoresis to verify that a significant amount of DNA had been sheared to <700 base pairs. A sample of 10 µg sheared chromatin was reserved and stored at −20°C as immunoprecipitation input. 200 µg of chromatin was used to couple with 200 µg of streptavidin beads for one hour in a Protein Lo-Bind tube at room temperature with nutation. The coupled beads were collected and separated from the flow-through using a magnetic rack. After the flow-through was removed, the beads underwent the following washes:

- 2% SDS with 20 mM EPPS
- 2% SDS with 20 mM EPPS
- High Salt Buffer containing 500 mM NaCl, 1 mM EDTA, 50 mM of HEPES pH7.5, 0.1% sodium deoxycholate, and 1% TritonX-100
- LiCl Buffer containing 250 mM LiCl, 1 mM EDTA, 10 mM Tris-HCl pH 8.0, and 0.5% of IGEPAL CA-630
- TE Buffer with 10 mM Tris and 1 mM EDTA
- TE Buffer with 10 mM Tris and 1 mM EDTA

The washes were performed as follows: briefly spin and immobilize the beads on a magnetic rack, pipette out the supernatant as much as possible, resuspend the beads in 0.8 ml of wash buffer, and incubate for 5 minutes with nutation. The washed beads were resuspended in 300 ul of reverse crosslinking buffer containing 300 mM NaCl, 300 mM Tris-HCl pH 8.0, and 1 mM EDTA. Both the eluate beads and the input chromatin were incubated at 65°C for 16 hours for reverse crosslinking. The next day, 4 ul of 20 mg/ml proteinase K (Roche 3115836001) was added to the eluates and inputs and incubated at 50°C for 2 hours to cleave away proteins. The DNA was isolated from the mixture using phenol chloroform extraction followed by ethanol precipitation. Before sequencing library generation, the precipitated DNA was further purified using SPRI beads. The purified DNA was used to generate next-generation sequencing libraries using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB E7645S) and NEBNext Multiplex Oligos for Illumina Index Primers Set 1 and 3 (NEB E7335S, E7710S) and PCR-amplified for 15 cycles. The sequencing libraries were quantified using the Qubit 4 fluorometer and library sizes were quantified using the D1000 ScreenTape assay (Agilent 5067-5582) on the TapeStation 4200 automated electrophoresis platform.

DNA sequencing and data analysis

The libraries were mixed and sequenced pair-ended at 50-bp read length on an Illumina NextSeq 2000 sequencer to depths of 14.1-351.8 million reads per eluate sample and 3.14-16.45 millions reads per input sample using the NextSeq 1000/2000 P2 Reagents (100 Cycles) kit (Illumina 20046811). Reads were demultiplexed and adapters were removed using Cutadapt[85]. Trimmed reads were mapped to the reference genome (GRCh38) using Bowtie2 version 2.5.3 with the parameter −X 1000 keeping reads with a MAPQ>=30[86]. Duplicate reads were removed using Picard 3.1.1[87]. Eluate reads were normalized to input reads using DeepTools[88] bamCompare with the following parameters: –binSize 20 –normalizeUsing BPM –smoothLength 60 – extendReads 150. Normalized data were visualized using Coolbox 0.3.9[89].

4.3 Results and Discussion

Design of DNA O-MAP

DNA O-MAP is a molecular profiling methodology that combines the targeting flexibility of oligo-based (ISH) with the ability of horseradish peroxidase (HRP) to catalyze the localized deposition of small biomolecules at sites where it is bound. DNA O-MAP works by recruiting a 'secondary' HRP-conjugated oligo to sites where the primary ISH probes are bound. HRP-mediated deposition of biotin at specific genomic sites then enables the pull-down and purification of chromatin associated proteins and DNA from *trans*-interacting genomic loci. As in RNA O-MAP, the specificity of ISH and/or biotinylation can be assessed by microscopy using a small sample of cells immobilized on solid support before the cell pellets enter affinity purification downstream. Importantly, the HRP-conjugated oligo is available via several

commercial sources, allowing researchers without the expertise to perform their own conjugations to utilize DNA O-MAP.

DNA O-MAP deploys a scalable in-solution hybridization-biotinylation workflow

During the development of DNA O-MAP, it became clear that performing *in situ* hybridization on samples adhered to solid substrates such as microscope slides or well plates would create significant scaling challenges, both in terms of reagent costs and sample processing time. We addressed these challenges by developing a suspension-based hybridization workflow for cost-efficient genomic labeling (Figure 4-1A). We began with adherent cells grown on multi-layer flasks, each yielding 90-120 million cells, and subsequently released and fixed (4% PFA) in order to be compatible with DNA ISH. Samples can be processed in parallel, thereby increasing the number of samples that could be handled in parallel by one experimentalist. Critically, this approach reduces reagent costs by ~1,000-fold relative to conventional ISH protocols performed on solid substrates, making the labeling of millions or more cells with oligo-based ISH probe sets, including those targeting non-repetitive DNA, cost-feasible.

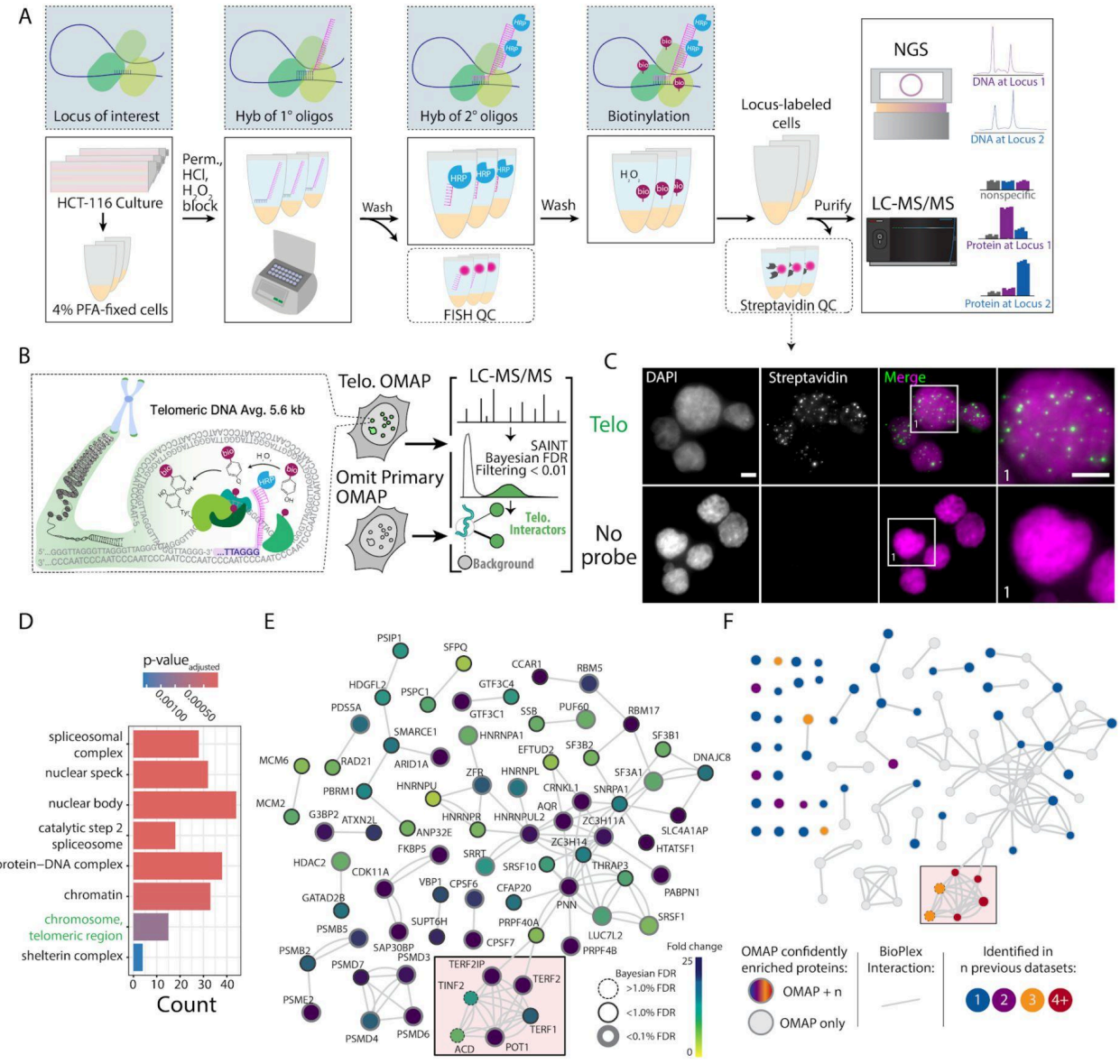


Figure 4-1. Overview of DNA O-MAP workflow and label-free quantitative proteomics analysis of telomeres. A) Schematic of DNA O-MAP. B) Overview of telomere targeted DNA O-MAP experiment. C) Fluorescent microscopy data showing the observed patterns of DNA (DAPI, left) and *in situ* biotinylation detected by staining with fluorescent streptavidin conjugates (middle, left)l. D) Significant gene sets identified by the Gene Set Enrichment Analysis of the proteins enriched by the telomere probe. E) DNA O-MAP telomeric proteins

mapped onto the BioPlex interaction network[28,29]. The red box highlights shelterin complex proteins. Nodes are colored by the fold-enrichment compared to a no-primary-probe control shown in C, excluding unconnected nodes. F) Telomeric proteins observed in five previous datasets (PICh, C-BERST, CAPLOCUS, CAPTURE, BioID) superimposed onto Figure 4-1E, colored by the number of prior datasets where the protein was present and including unconnected nodes. Scale bars, 5 μm.

DNA O-MAP reveals the organization of the telomeric proteome

To demonstrate that O-MAP can successfully purify proteins from small genomic viewpoints, we selected human telomeres for initial testing (Figure 4-1B). Mammalian telomeres are several kilobases of tandemly repeated arrays of 5'-TTAGGG-3' hexamers with terminal 3' single-stranded overhangs at the ends of chromosomes[30]. Telomeric DNA is specifically bound by a proteinaceous cap that protects the natural chromosome ends from being recognized as damaged DNA—the shelterin complex[31,32]. Shelterin is a six-subunit complex, which is comprised of the telomeric repeat-binding factor 1 (TERF1), telomeric repeat-binding factor 2 (TERF2), protection of telomeres protein 1 (POT1), adrenocortical dysplasia protein homolog (ACD), TERF2-interacting protein 1 (TERF2IP), and TERF1-interacting nuclear factor 2 (TINF2). Due to the unique telomeric sequence and characteristic DNA structure, the shelterin proteins accumulate exclusively at the ends of the chromosomes. Accordingly, this well-defined set of proteins has been widely accepted as goalposts for a successful locus-specific enrichment experiment[12,13,16]. In the near-diploid HCT-116 cells, telomeres have an average length of 5.6 kb and their cumulative length approximates 0.017% (~500kb) of the human genome[33]. Compared to other repetitive elements in the human genome, telomeres are relatively short in HCT-116 cells

and thus serve as a rigorous test case for DNA viewpoints of around 500 kb in aggregate across the genome.

We performed a DNA O-MAP experiment in which we either targeted telomeric DNA or omitted the primary hybridization probe (negative control). We purified biotinylated proteins from <60 million cells in three technical replicates followed by imaging of biotinylation and identification of proteins using label-free quantitative proteomics. By streptavidin staining, the punctate fluorescence pattern of biotin-labeled biomolecules closely mimicked telomere FISH, whereas we did not observe patterns of these puncta in the negative control samples (Figure 4-1C). From our label-free proteomics analysis, we identified 163 proteins as significantly enriched at telomeres. As expected, gene set enrichment analysis[34] identified significant enrichment of telomeric chromosomal components, chromatin, and protein-DNA complexes (Figure 4-1D-E). Importantly, we identified all six shelterin proteins in the telomere sample and these proteins were completely absent from the control samples. Of the six shelterin proteins, four (TERF1, TERF2, TERF2IP, POT1) passed stringent false-discovery rate control while ACD and TINF2 did not due to low spectral intensity. To benchmark DNA O-MAP, we compared the full set of telomeric proteins to proteins observed in five established telomeric datasets (PICh, C-BERST, CAPLOCUS, CAPTURE, BioID)[12,14,16,35,36] (Figure 4-1F). We then overlaid each called interactor on direct protein interaction data and found that DNA O-MAP enabled greater coverage of known protein interactors, even those not previously identified as enriched at telomeres by other methods. In addition to shelterins, we identified multiple heterogeneous nuclear ribonucleoproteins (hnRNPs) previously annotated as telomere-associated, including HNRNPA1 and HNRNPU. HNRNPA1 has been demonstrated to displace replication protein A (RPA) and directly interact with single-stranded telomeric DNA to regulate telomerase

activity[37–39]. In addition, HNRNPU belongs to the telomerase-associated proteome[40] where it binds the telomeric G-quadruplex to prevent RPA from recognizing chromosome ends[41]. Taken together, this data supports the effectiveness of DNA O-MAP for sensitively and selectively isolating loci-specific proteomes.

DNA O-MAP enables multiplexed detection of locus proteomes

We next evaluated the utility of DNA O-MAP to quantitatively delineate locus-specific proteomes. We integrated sample multiplexing quantitative[27,43,44] proteomics downstream of DNA O-MAP to enable spectral quantification of all samples simultaneously (Figure 4-2A). In our experimental design, we selected three well-characterized DNA loci with distinct protein occupants in the human genome: 1) telomeres, 2) peri-centromeric alpha satellite repeats; 3) the mitochondrial genome (Figure 4-2B). Centromeres are epigenetically defined chromosomal loci where kinetochore proteins assemble for spindle microtubule attachment to ensure equal chromosome segregation during cell division[45,46]. Human centromeres are located within the AT-rich alpha satellite repeats, which are higher-order repeats composed of 171-base-pair monomeric units[47,48]. Due to the sequence independence of centromeres, we utilized a previously described probe[26,49] that targets a subset of alpha satellite repeats to represent centromeres, hereafter denoted as the 'Pan Alpha Sat.' probe. The predicted genome-wide binding profile[50] of the pan-alpha probe closely overlaps with centromeres (Figure 4-3). Mitochondria are intracellular organelles of eukaryotic cells with their own genome (mtDNA). The mtDNA is a circular double-stranded DNA molecule of about 16.6 kb, located in the mitochondrial matrix associated with the inner membrane[51,52]. To demonstrate the locus-specificity of biotinylation using the new oligo/oligo pools, we performed DNA O-MAP in human HCT-116 cells with a co-hybridization of both fluorescent oligos and HRP oligos in order to observe fluorescent *in situ*

hybridization (FISH) and *in situ* biotinylation signals in the same cell. Biotinylation patterns of the pan-alpha, telomere, and mtDNA probes showed strong concordance with FISH (Figure 4-2C). To quantify the local proteomes corresponding to each of these biotinylated patterns, we prepared replicate (n=4) samples for each probe and control. After *in situ* HRP-mediated labeling, we performed thermal reversal of fixation of cells prior to lysis, enrichment of biotinylated proteins[53], tryptic digestion, and labeling with isobaric TMTpro barcodes[27]. We note that artificial lysine alkylation due to cellular fixation with PFA may affect TMTpro labeling of protein, thus we tracked artificial lysine modifications during mass spectrometric analysis to ensure minimal effects of alkylation on protein quantification (1.38% of lysines were alkylated).

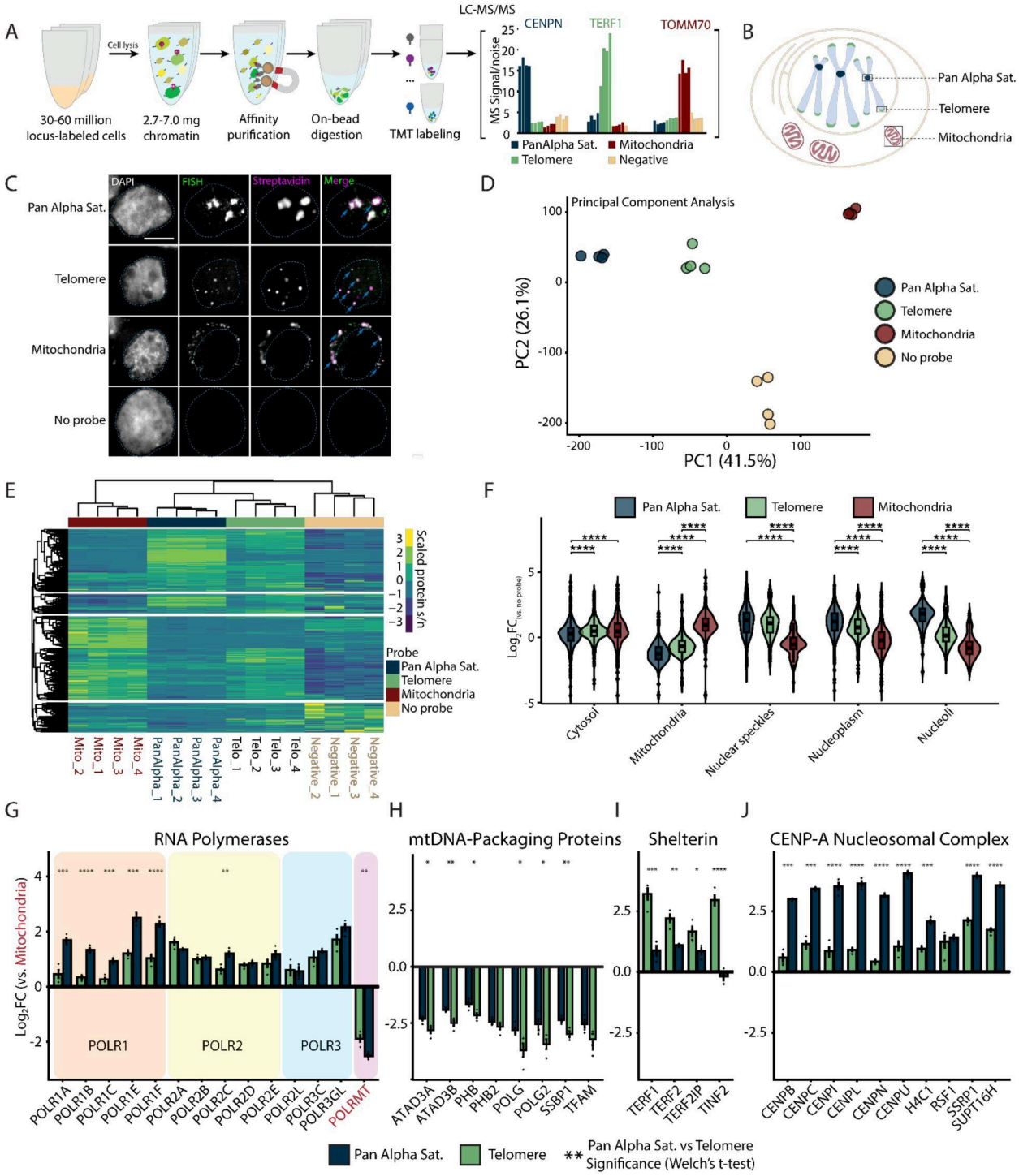


Figure 4-2. DNA O-MAP reveals distinct features of the sub-proteomes at peri-centromeric alpha satellites, telomeres, and the mitochondrial genome. A) Workflow of DNA O-MAP integrated with sample multiplexing quantitative proteomics B) Schematic of the three DNA loci

examined in the TMT16plex experiment: peri-centromeric alpha satellites, telomeres, and mitochondrial genomes. C) Co-localization of DNA FISH and the streptavidin staining of the proteins biotinylated by DNA O-MAP targeting the peri-centromeric alpha satellites, telomeres, and mitochondrial genomes. Scale bar: 5 μm. D) Principal component analysis of scaled intensities of proteins enriched by the pan-alpha probe, telomere probe, mitochondrial genome oligo pool, and no-primary-probe control. E) Unsupervised hierarchical clustering of scaled intensities of proteins enriched by the pan-alpha probe, telomere probe, mitochondrial genome oligo pool, and no-primary-probe control. F) Log2 fold change of proteins compared to no-primary-probe control, grouped by HPA subcellular location. Significance calculated based on Welch's t-test for pairwise comparisons (****: p-value <0.0001). G–J) Log2 fold change of proteins compared to mitochondrial probe enriched proteins for the RNA Polymerases (G), mtDNA nucleoid packaging proteins[42] (H), Shelterin (I), and CENP-A nucleosomal complexes (J). Significance calculated based on Welch's t-test for pairwise comparisons (p-value: *<0.05, **<0.01, ***<0.001, ****<0.0001).

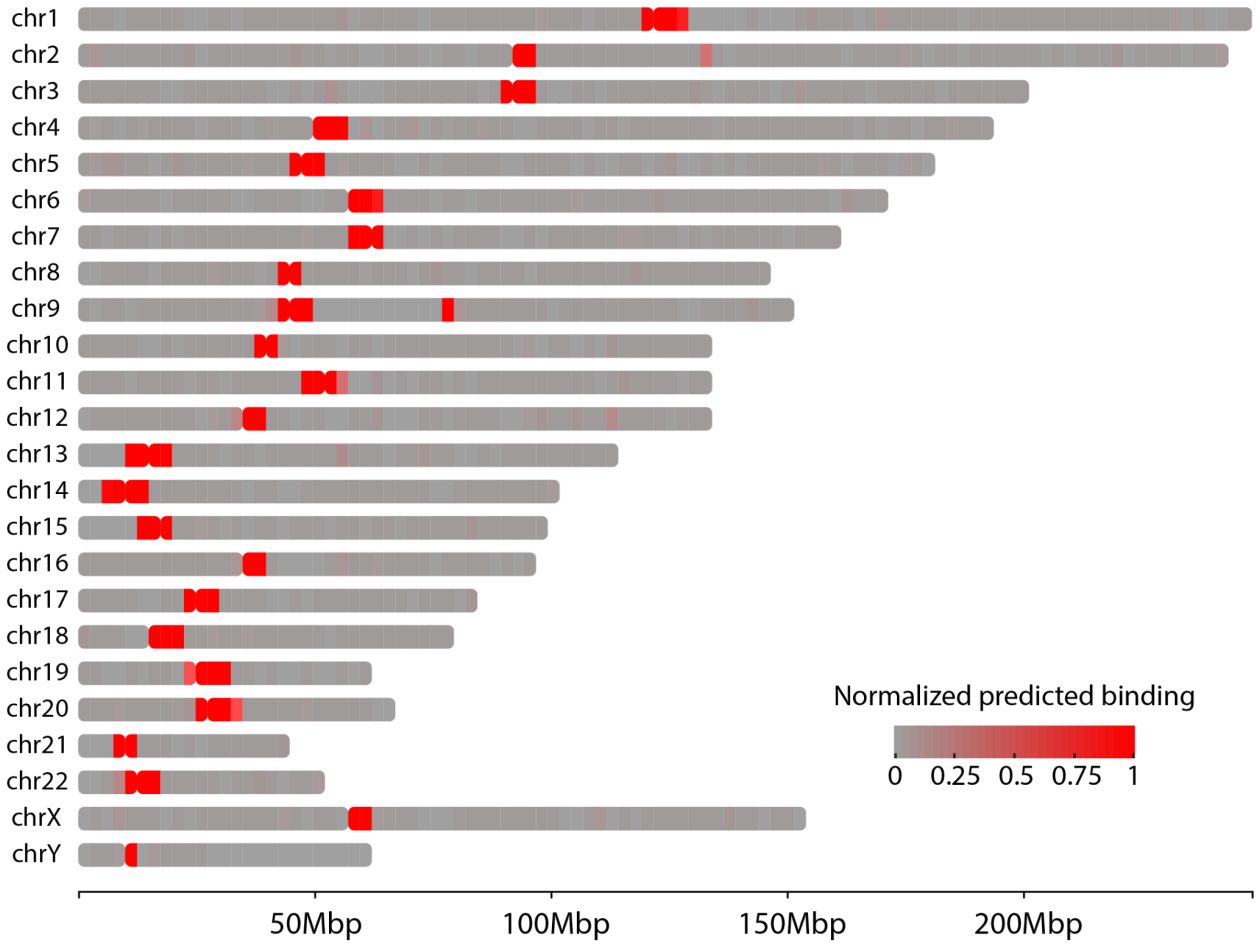


Figure 4-3. Predicted genome-wide binding profile of the pan-alpha probe. The intensity of red indicates the amount of predicted probe binding.

In total we quantified 3,055 proteins across all four conditions (Figure 4-2D–E). We observed consistent proteome enrichment by principal component analysis and correlation analyses, with tight clustering of replicates (Figure 4-2D–E**,** S2). Based on Human Protein Atlas annotations[54], we observed significant enrichment of mitochondrial proteins with the mtDNA-probe proteomes and proteins from nuclear locations such as nuclear speckles, nucleoplasm, and nucleoli enriched by the telomere and pan-alpha probes (Figure 4-2F**,** 4-4). Notably, the pan-alpha probe enriched proteins from the nucleoli, consistent with the known nucleoli-centromere associations[55]; chromosomal passenger complex member AURKB, consistent with the centromeric localization of AURKB in early mitosis to ensure faithful

chromosome segregation[62,63] and the localization of chromosomal passenger complex members to pericentromeric heterochromatin[56,57]. We also observed pericentromeric enrichment of spindle and chromosomal segregation associated proteins TPX2[58] and KIF20A[59] (Figure 4-4, 4-5). Next, we explored the enrichment of several multi-unit protein complexes across the examined loci. To dissect the differences between enriched proteomes for each probe, we chose a subset of proteins of interest and measured the fold change of the two nuclear targets compared to mitochondria. RNA Polymerase I,II,III subunits were all higher in the nuclear probes than mitochondria, however in contrast to RNA Polymerase II and III, POLR1 proteins are significantly enriched in pan-alpha compared to telomere (Figure 4-2G). This enrichment is likely due to clustering of centromeres around nucleoli[60,61], the location of ribosomal RNA synthesis by RNA Polymerase I. Conversely, mitochondrial RNA Polymerase POLRMT abundance was significantly lower in the nuclear probe proteomes compared to the mitochondrial probe proteome (log2 Pan-Alpha Sat./Mito.= −2.51; log2 Telomere/Mito.= −1.88). Similarly, we observed enrichment of mtDNA-packaging nucleoid components[42] with the mtDNA probes (TFAM, SSBP1, POLG, POLRMT, Lon, ATAD3A/B, and PHB/PHB2; Figure 4- 2G–H). As above, we observed consistent enrichment of shelterin components at telomeres (Figure 4-2I). We also observed CENP-A nucleosomal complexes enriched in the pan-alpha proteomes (Figure 4-2J). Histones were enriched with our nuclear probes and a subset (H2A1C, H2AX, and H4C1) were significantly enriched by the pan-alpha probe compared to the telomere probe (Figure 4-7). We also observed enrichment of catenins CTNNB1 and CTNND1 at telomeres (Figure 4-6). The transcription factor CTNNB1 has been observed at the transcriptional start site of h*TERT* where it regulates h*TERT* expression[64]. The h*TERT* gene is located in the subtelomeric region of chromosome 5 (chr5:1,253,167-1,295,068) and expressed in HCT-116 cells[65]. Collectively, these results demonstrate the sensitivity and subcompartment specificity of DNA O-MAP and highlight how coupling quantitative proteomics with DNA O-MAP can distinguish differential compartment components even for ubiquitous chromatin constituents like histones.

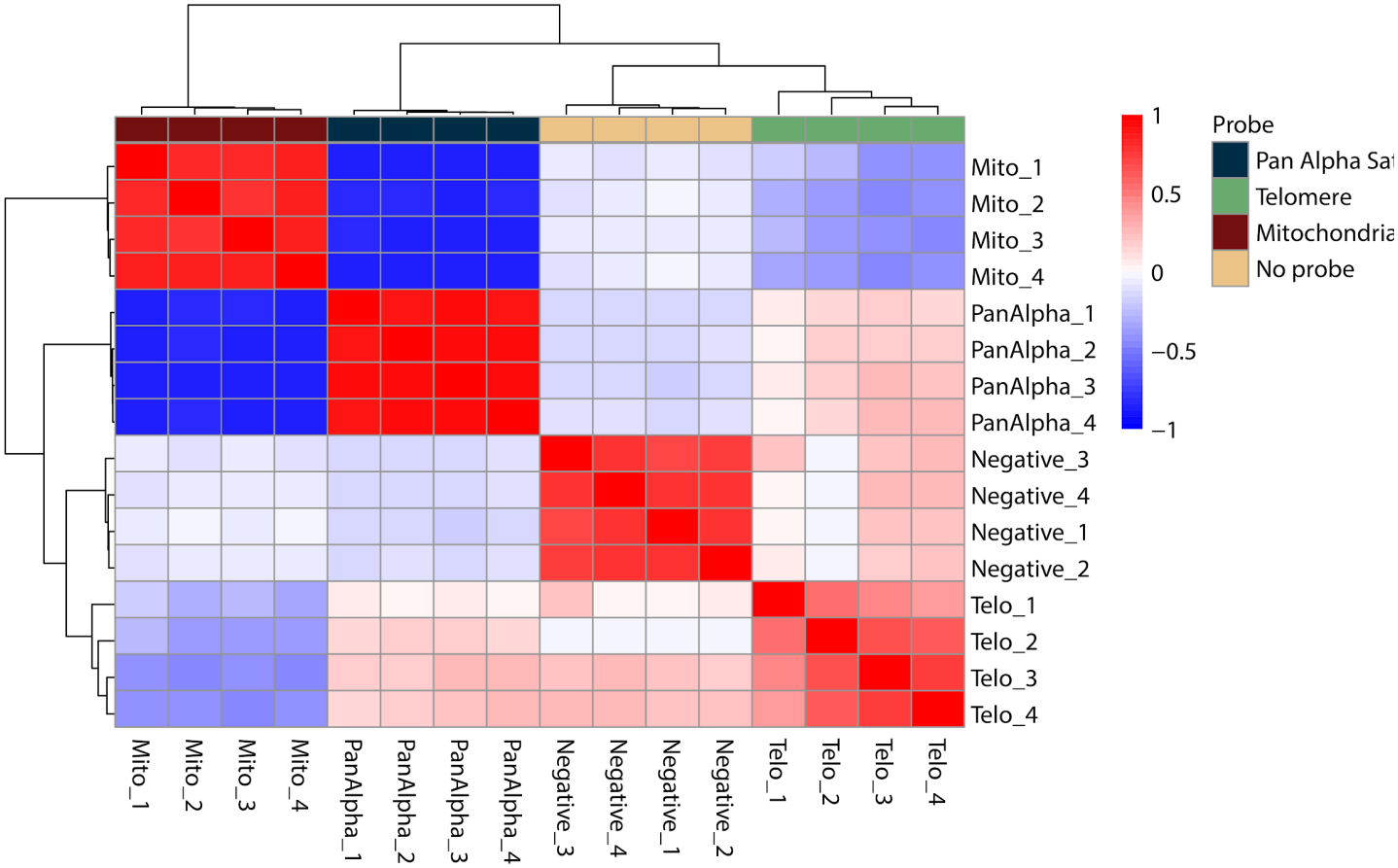


Figure 4-4. Replicate analysis of multi-target DNA O-MAP proteomics experiment. A) Pearson correlation coefficient of the raw protein intensity values for each replicate of the multiplex with hierarchical clustering on the rows and columns.

DNA O-MAP can uncover DNA-DNA interactions from non-repetitive DNA loci

Beyond repetitive regions in the human genome, we explored whether DNA O-MAP can recover material from small, single-copy DNA intervals. To this end, we designed an experiment in which we performed *in situ* biotinylation followed by chromatin extraction, affinity purification, and sequencing (Figure 4-5A). The human genome is folded into thousands of chromatin loops where two loci on the same chromosome are tethered to each other (Figure

4-5B). The anchors of the loops are bound by the insulator protein CTCF. The ring-shaped cohesin protein complex is thought to to often stall at CTCF-bound sites while dynamically moving along the genome, creating contact domains of preferential DNA-DNA interaction[66]. In HCT-116 cells, these contacts between chromatin loop anchors have been captured genome-wide with *in situ* Hi-C[67]. Normally present in two copies per genome, these 20–25 kb loop anchor intervals are considerably less abundant than telomeres.

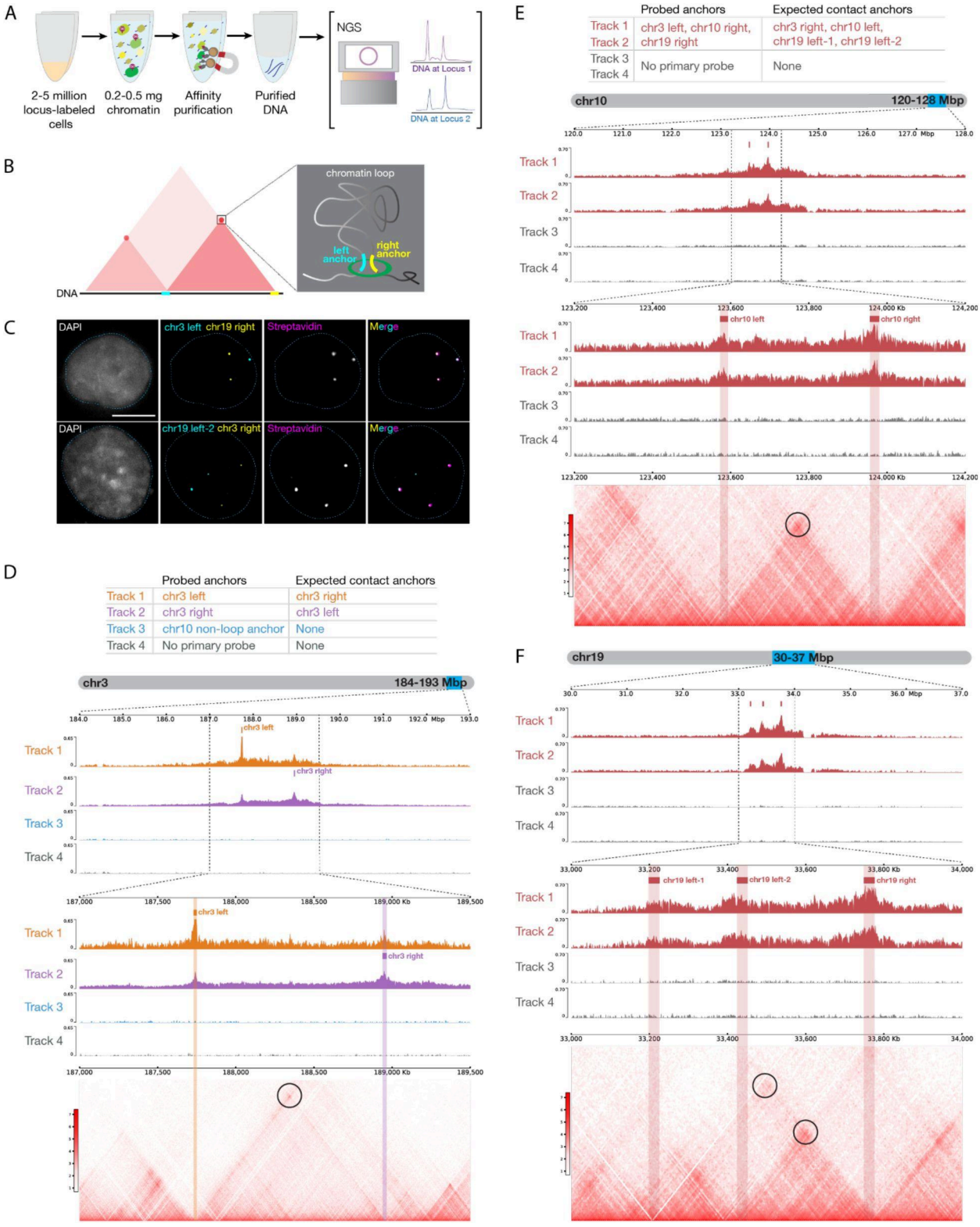


Figure 4-5. DNA O-MAP efficiently labels single-copy chromatin loop anchors. A) Workflow of DNA O-MAP integrated with biotin purification sequencing B) Schematic of a pair

of chromatin loop anchors on a hypothetical Hi-C map and 3-dimensional space C) DNA FISH and the streptavidin staining of the proteins biotinylated by DNA O-MAP targeting anchors of chromatin loops on chromosome 3 and chromosome 19 D) Table listing the three anchors (Track 1-3) and no-primary-probe control (Track 4) biotinylated by DNA O-MAP and their expected anchors in contact in each track (top). Desthiobiotin purification sequencing signals across the 9-Mb region on chromosome 3 corresponding to the chr3 chromatin loop (middle). Desthiobiotin purification sequencing signals and pairwise contact map at 5-kb resolution across the 2.5-Mb region on chromosome 3 corresponding to the chr3 chromatin loop. Black circle on the contact map indicates the presence of a loop. (bottom). E) Table listing the three chromatin loop anchors (Track 1-2) and no-primary-probe controls (Track 3-4) biotinylated by DNA O-MAP in duplicates and their expected anchors in contact in each track (top). Desthiobiotin purification sequencing signals across the 8-Mb region on chromosome 10 corresponding to the chr10 chromatin loop targeted (middle). Desthiobiotin purification sequencing signals and pairwise contact map at 5-kb resolution across the 1-Mb region on chromosome 10 corresponding to the chr10 chromatin loop. Black circle on the contact map indicates the presence of a loop. (bottom). F) Desthiobiotin purification sequencing signals across the 7-Mb region on chromosome 19 corresponding to the chr19 chromatin loops targeted (top). Desthiobiotin purification sequencing signals and pairwise contact map at 5-kb resolution across the 1-Mb region on chromosome 19 corresponding to the chr19 chromatin loops. Black circles on the contact map indicate the presence of loops (bottom).

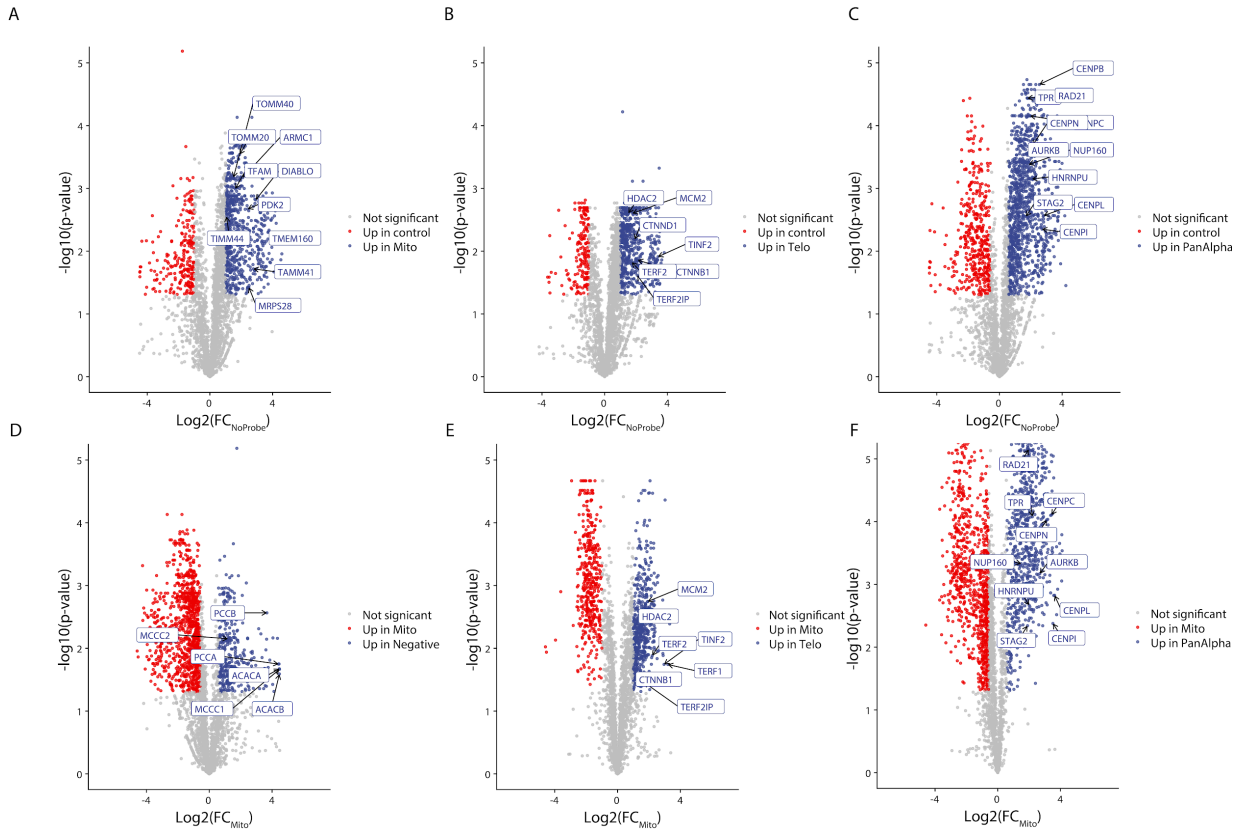


Figure 4-6. Relative quantitation for the multi-target DNA O-MAP proteomics experiment compared to no-probe control and mtDNA datasets. Volcano plots from multiplexed proteomics experiments with proteins of interest highlighted. A-C) Fold-changes and significance calculated compared to no probe. D-F) Fold-changes and significance calculated compared to mtDNA probe.

We first evaluated whether DNA O-MAP can specifically biotinylate loop anchors with microscopy by a co-hybridization of both fluorescent oligos and HRP oligos at four anchors: chr3 left (chr3:187,729,712-187,749,712), chr3 right (chr3:188,939,711-188,964,711), chr19 left-2 (chr19:33,425,000-33,450,000), and chr19 right (chr19:33,750,000-33,775,000). DNA O-MAP specifically biotinylated the biomolecules proximal to these small DNA intervals, as

observed in the co-localizing patterns of FISH and streptavidin staining in the same cells (Figure 4-5C). We next evaluated whether DNA O-MAP could recover the DNA interactions originally discovered by Hi-C. We targeted a pair of intervals with high contact frequency—chr3 left and chr3 right anchors, one non-looping interval (chr10:123,187,984-123,207,984), and no-primary-probe control. We performed DNA O-MAP to biotinylate these DNA intervals, subjected the labeled cells to chromatin solubilization and desthiobiotin purification, and sequenced the eluate DNA. As expected, all three probed DNA intervals were highly enriched compared with other genomic regions, indicating efficient purification of the loci (Figures 4-5D**,** S5A). Furthermore, chr3 left and chr3 right anchors reciprocally recovered each other, indicating that DNA O-MAP was able to recover known DNA interactions mediated by proteins. In contrast, the non-looping chr10 anchor did not enrich any other peak except itself (Figure 4-9B). Lastly, in the cells that received no primary oligos, no pronounced enrichment was observed genome wide (Figure 4-9B).

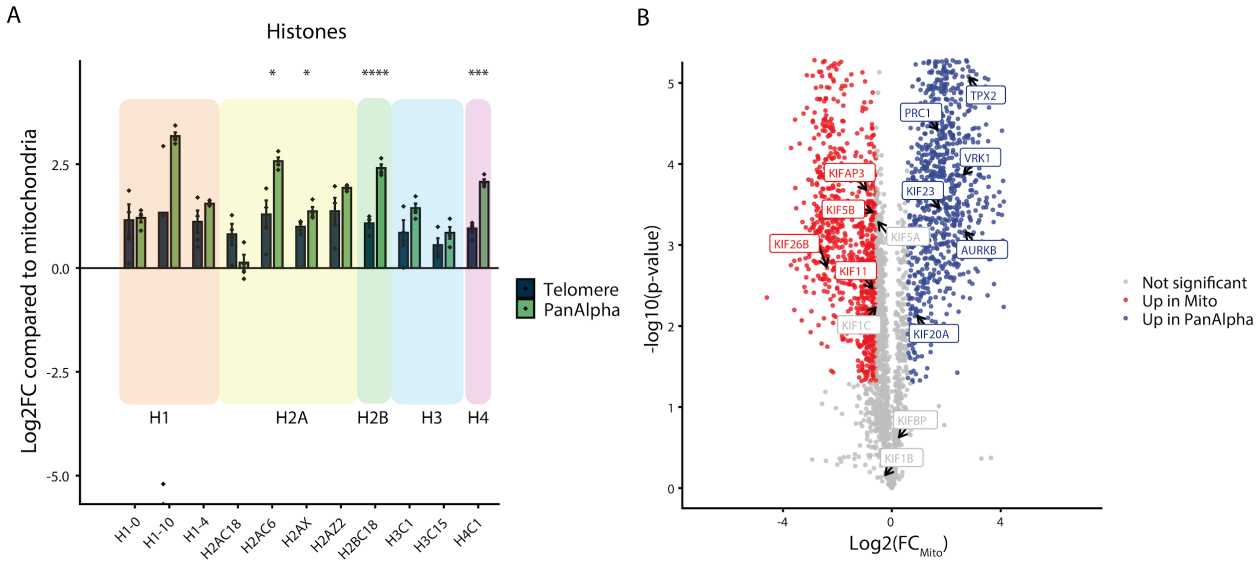


Figure 4-7. Comparison of histone proteins between telomere and pan-alpha probes. A) Log2 fold change of proteins compared to mitochondrial probe enriched histone complex

proteins. Significance calculated based on Welch's t-test for pairwise comparisons (p-value: *<0.05, **<0.01, ***<0.001, ****<0.0001). B) Volcano plot comparing the fold change of pan-alpha to the mtDNA probe with spindle proteins highlighted.

To examine the multiplexability and reproducibility of DNA O-MAP, we simultaneously targeted three chromatin loop anchors: chr3 left, chr10 right (chr10:123,957,984-123,977,984), and chr19 right anchors in duplicates and subjected the cell pellets to purification and DNA sequencing. All three targeted anchors, chr3 left, chr10 right, and chr19 right anchors were successfully enriched (Figures 4-5E–F, 4-10A), whereas no pronounced enrichment was observed in the no-primary-probe controls genome-wide (Figure 4-9B). Furthermore, chr10 left (contacting chr10 right), chr19 left-1, and chr19 left-2 (both contacting chr19 right) were also efficiently recovered, accurately matching the Hi-C contact maps and the signals from two replicates was consistent (Figure 4-5E–F). These imaging and genomics data demonstrate that DNA O-MAP is capable of labeling small, single-copy DNA intervals with high specificity.

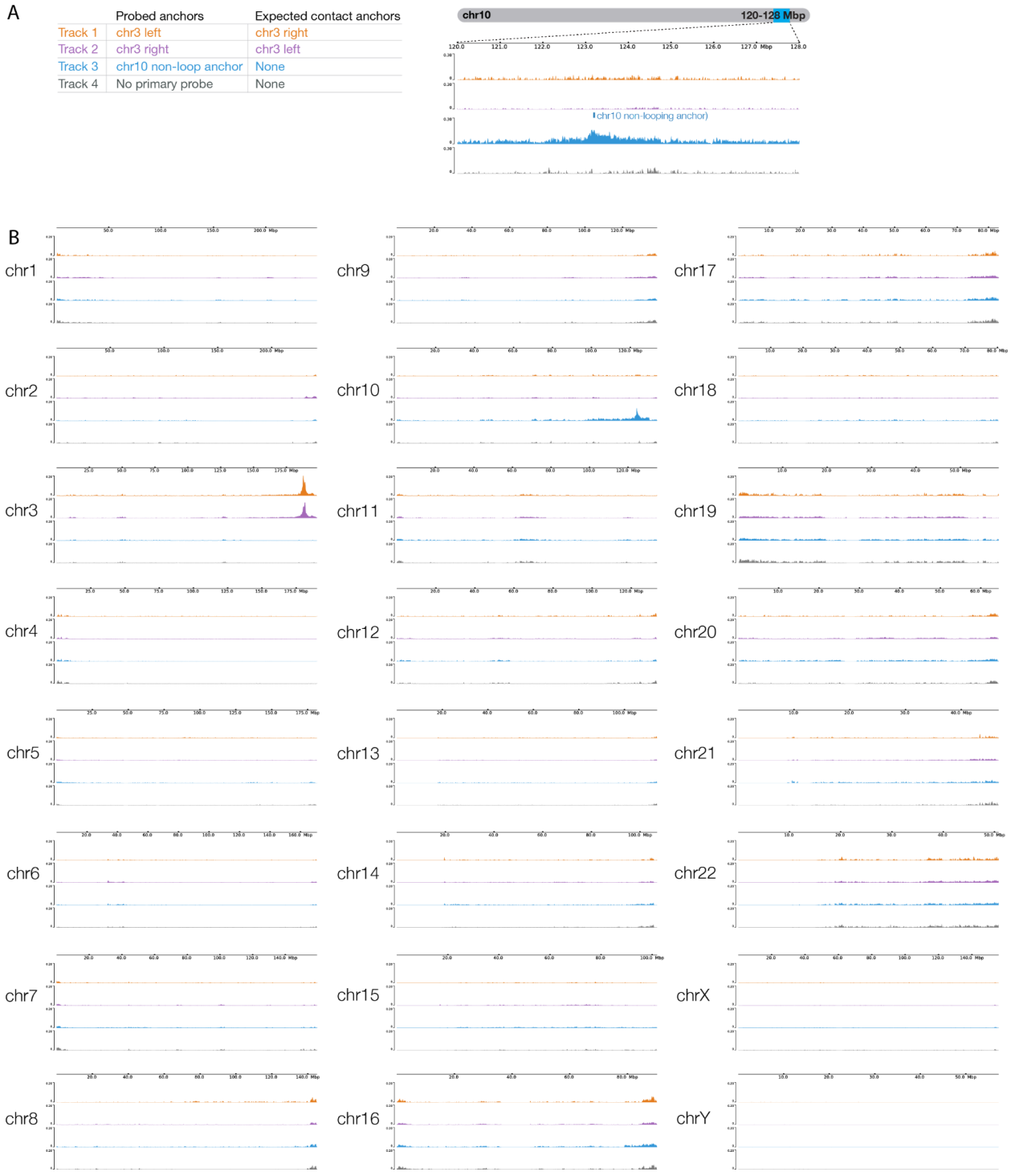
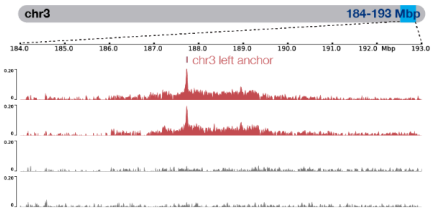


Figure 4-8. DNA O-MAP biotin purification sequencing of chr3 left, chr3 right, chr10 non-loop anchors, and no-primary-probe control. A) Table listing the three anchors (Track 1-3) and no-primary-probe control (Track 4) biotinylated by DNA O-MAP and their expected contact anchors (left). Biotin purification sequencing signals across the 8-Mb region on

chromosome 10 corresponding to the chr10 non-loop anchor targeted (right). B) Biotin purification sequencing signals across every chromosome in the genome for this experiment.

A

	Probed anchors	Expected contact anchors
Track 1 Track 2	chr3 left, chr10 right, chr19 right	chr3 right, chr10 left, chr19 left-1, chr19 left-2
Track 3 Track 4	No primary probe	None



B

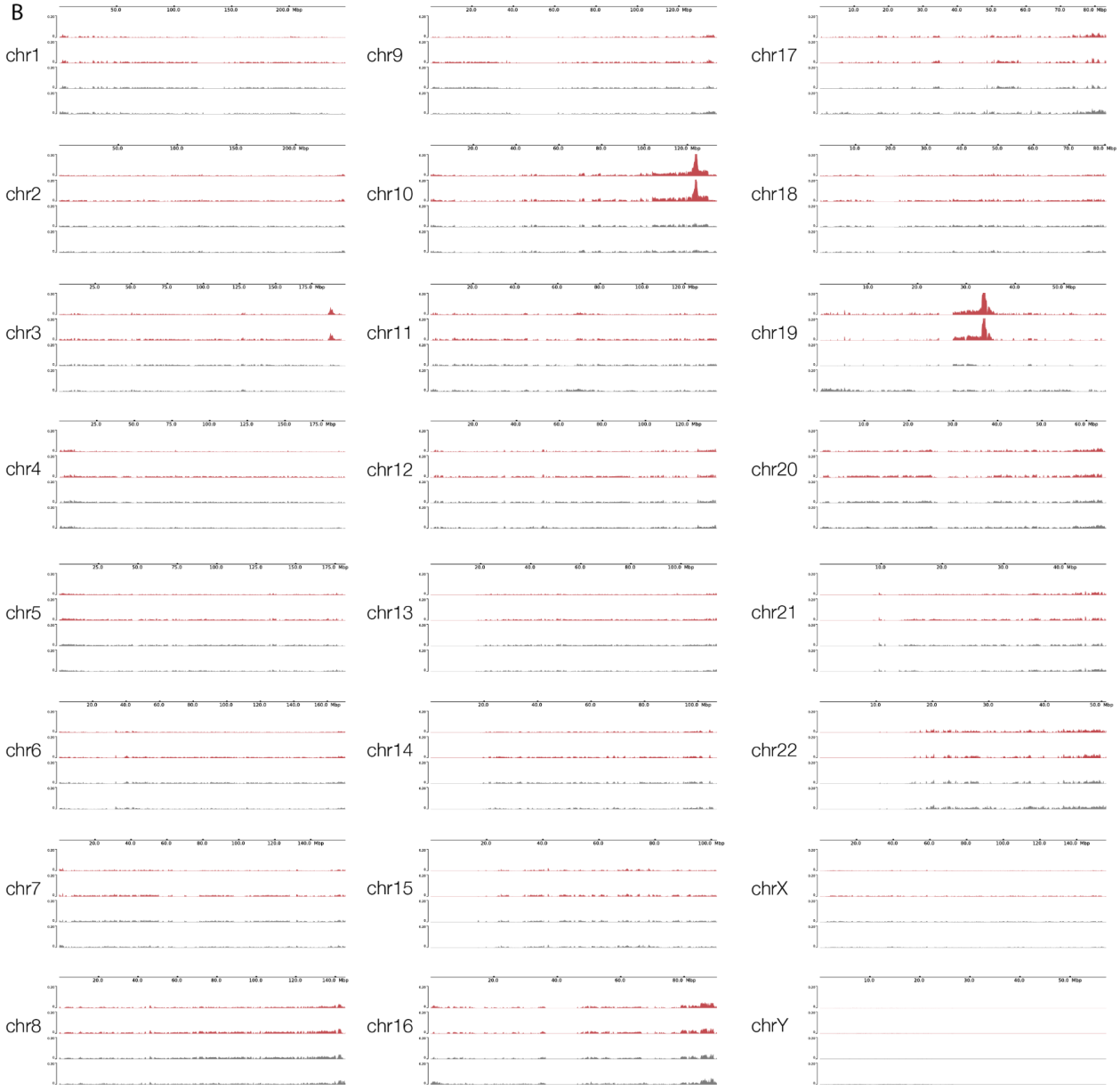


Figure 4-9. DNA O-MAP biotin purification sequencing of multiplexed targeting of chr3 left, chr10 right, chr19 right anchors, and no-primary-probe control in duplicates. A) Table listing the three anchors (Track 1-2) and no-primary-probe control (Track 3-4) biotinylated by DNA O-MAP and their expected contact anchors (left). Biotin purification sequencing signals across the 9-Mb region on chromosome 3 corresponding chr3 left anchor targeted in Track 1-2 (right). B) Biotin purification sequencing signals across every chromosome in the genome for this experiment.

Discussion

By combining the versatility of hybridization-based genome targeting with robustness of proximity biotinylation, DNA O-MAP offers a scalable approach to study DNA-associated proteomes through a locus specific lens. The liquid-phase hybridization-biotinylation workflow allows for efficient processing of samples and is compatible with both proteomic and genomic readouts. Integration with multiplexed quantitative proteomics enables simultaneous analysis of multiple loci or conditions, increasing data completeness and throughput. Label-free analysis of the telomeres shows strong concordance of labeling with in-situ hybridization and recapitulates previous similar proteomic datasets. Our tri-locus experiment was able to differentiate proteins with a quantitative profile suggesting general nuclear location from those specifically associated with telomeres and peri-centromeres. DNA O-MAP's ability to target single-copy loci, as evidenced by the chromatin loop anchor experiments, opens up possibilities for studying protein-mediated DNA interactions at a finer resolution than previously possible.

O-MAP has now been shown to be a highly flexible technology for the exploration of biomolecular interactions with RNAs[25] and DNA loci. Using oligos to target the DNA locus,

DNA O-MAP can be theoretically adapted for use in any sample types amenable to *in situ* hybridization, including cultured cells, tissue sections, and primary tissue samples[26,50,68]. As the purification tag is decoupled from the probe oligos, labeled chromatin fragments can undergo stringent washes to achieve efficient purification with minimal background. Moreover, without the need to genetically modify the biological system at hand, the probes in this dataset alone could be used to explore telomeric remodeling in cancer cells[36], spindle-associated proteome dynamics at the pericentromere[69], and molecular drivers of hetero- or euchromatin formation[70] at nearly any locus in the human genome (O-MAP probes can feasibly cover >99% of the human genome)[50,68].

4.4 Conclusions

While this work has laid the foundation for generalized and extensible locus proteomics, further work will be required to achieve the sensitivity required for small, single copy locus proteomics. By taking a comparative quantitative approach, we remove the need to pre-define the local context of probe localization, but experimental design is critical and novel interactors likely need further validation to confirm their co-localization at a given locus (e.g., with imaging/FISH). With developments in automation and instrument sensitivity, DNA O-MAP has the potential to expand to locus specific post-translational modifications and be used for large-scale chromatin perturbation screens. We anticipate that DNA-OMAP will have broad utility for research questions seeking to understand the intricate relationships between DNA sequence, chromatin structure, and cellular function.

4.5 References

1. Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of

interphase chromosomes. *Cell* **152**, 1270–1284 (2013).

2. Jerkovic, I. & Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **22**, 511–528 (2021).
3. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
4. Ho, J. W. K., Alekseyenko, A. A., Kuroda, M. I. & Park, P. J. Genome-wide mapping of protein-DNA interactions by ChIP-seq. in *Tag-Based Next Generation Sequencing* 139–151 (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2012).
5. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
6. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
7. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
8. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
9. Bujold, D. *et al.* The International Human Epigenome Consortium Data Portal. *Cell Syst* **3**, 496–499.e2 (2016).
10. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
11. Hoffman, M. M., Buske, O., Bilmes, J. A. & Noble, W. S. Segway: a dynamic Bayesian network method for segmenting genomic data. *Invert. Neurosci.* (2009).
12. Gao, X. D. *et al.* C-BERST: defining subnuclear proteomic landscapes at genomic elements

with dCas9–APEX2. *Nat. Methods* **15**, 433–436 (2018).

13. Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat. Methods* **15**, 437–439 (2018).
14. Qiu, W. *et al.* Determination of local chromatin interactions using a combined CRISPR and peroxidase APEX2 system. *Nucleic Acids Res.* **47**, e52 (2019).
15. Ugur, E., Bartoschek, M. D. & Leonhardt, H. Locus-Specific Chromatin Proteome Revealed by Mass Spectrometry-Based CasID. *Methods Mol. Biol.* **2175**, 109–121 (2020).
16. Déjardin, J. & Kingston, R. E. Purification of proteins associated with specific genomic Loci. *Cell* **136**, 175–186 (2009).
17. Silahtaroglu, A. N., Tommerup, N. & Vissing, H. FISHing with locked nucleic acids (LNA): evaluation of different LNA/DNA mixmers. *Mol. Cell. Probes* **17**, 165–169 (2003).
18. Beliveau, B. J. *et al.* Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21301–21306 (2012).
19. Ide, S. & Dejardin, J. End-targeting proteomics of isolated chromatin segments of a mammalian ribosomal RNA gene promoter. *Nat. Commun.* **6**, 6674 (2015).
20. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196**, 801–810 (2012).
21. Cho, K. F. *et al.* Proximity labeling in mammalian cells with TurboID and split-TurboID. *Nat. Protoc.* **15**, 3971–3999 (2020).
22. Lam, S. S. *et al.* Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods* **12**, 51–54 (2015).
23. Martell, J. D. *et al.* Engineered ascorbate peroxidase as a genetically encoded reporter for

electron microscopy. *Nat. Biotechnol.* **30**, 1143–1148 (2012).

24. Mangeot, P. E. *et al.* Genome editing in primary cells and in vivo using viral-derived Nanoblades loaded with Cas9-sgRNA ribonucleoproteins. *Nat. Commun.* **10**, 45 (2019).
25. Tsue, A. F. *et al.* Oligonucleotide-directed proximity-interactome mapping (O-MAP): A unified method for discovering RNA-interacting proteins, transcripts and genomic loci in situ. *bioRxiv* (2023) doi:10.1101/2023.01.19.524825.
26. Attar, S. *et al.* Programmable peroxidase-assisted signal amplification enables flexible detection of nucleic acid targets in cellular and histopathological specimens. *bioRxiv* (2023) doi:10.1101/2023.01.30.526264.
27. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020).
28. Huttlin, E. L. *et al.* Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040.e28 (2021).
29. Schweppe, D. K., Huttlin, E. L., Harper, J. W. & Gygi, S. P. BioPlex Display: An Interactive Suite for Large-Scale AP-MS Protein-Protein Interaction Data. *J. Proteome Res.* **17**, 722–726 (2018).
30. Chakravarti, D., LaBella, K. A. & DePinho, R. A. Telomeres: history, health, and hallmarks of aging. *Cell* **184**, 306–322 (2021).
31. Sfeir, A. & de Lange, T. Removal of shelterin reveals the telomere end-protection problem. *Science* **336**, 593–597 (2012).
32. de Lange, T. Shelterin-Mediated Telomere Protection. *Annu. Rev. Genet.* **52**, 223–247 (2018).
33. Myung, K. *et al.* Regulation of telomere length and suppression of genomic instability in

human somatic cells by Ku86. *Mol. Cell. Biol.* **24**, 5050–5059 (2004).

34. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
35. Liu, X. *et al.* In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* **170**, 1028–1043.e19 (2017).
36. Garcia-Exposito, L. *et al.* Proteomic Profiling Reveals a Specific Role for Translesion DNA Polymerase η in the Alternative Lengthening of Telomeres. *Cell Rep.* **17**, 1858–1871 (2016).
37. LaBranche, H. *et al.* Telomere elongation by hnRNP A1 and a derivative that interacts with telomeric repeats and telomerase. *Nat. Genet.* **19**, 199–202 (1998).
38. Zhang, Q.-S., Manche, L., Xu, R.-M. & Krainer, A. R. hnRNP A1 associates with telomere ends and stimulates telomerase activity. *RNA* **12**, 1116–1128 (2006).
39. Flynn, R. L. *et al.* TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. *Nature* **471**, 532–536 (2011).
40. Fu, D. & Collins, K. Purification of human telomerase complexes identifies factors involved in telomerase biogenesis and telomere length regulation. *Mol. Cell* **28**, 773–785 (2007).
41. Izumi, H. & Funa, K. Telomere Function and the G-Quadruplex Formation are Regulated by hnRNP U. *Cells* **8**, (2019).
42. Matilainen, O., Quirós, P. M. & Auwerx, J. Mitochondria and Epigenetics - Crosstalk in Homeostasis and Stress. *Trends Cell Biol.* **27**, 453–463 (2017).
43. Schweppe, D. K. *et al.* Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **19**, 2026–2034 (2020).
44. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined tandem mass tag

(SL-TMT) protocol: An efficient strategy for quantitative (phospho)proteome profiling using tandem mass tag-synchronous precursor selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).

45. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).
46. Talbert, P. B. & Henikoff, S. The genetics and epigenetics of satellite centromeres. *Genome Res.* **32**, 608–615 (2022).
47. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).
48. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
49. Deng, Z. & Beliveau, B. J. An open source 16-channel fluidics system for automating sequential fluorescent in situ hybridization (FISH)-based imaging. *HardwareX* **12**, e00343 (2022).
50. Aguilar, R. *et al.* Tigerfish designs oligonucleotide-based in situ hybridization probes targeting intervals of highly repetitive DNA at the scale of genomes. *Nat. Commun.* **15**, 1027 (2024).
51. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
52. Rackham, O. & Filipovska, A. Organization and expression of the mammalian mitochondrial genome. *Nat. Rev. Genet.* **23**, 606–623 (2022).
53. Paek, J. *et al.* Multidimensional Tracking of GPCR Signaling via Peroxidase-Catalyzed Proximity Labeling. *Cell* **169**, 338–349.e11 (2017).

54. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).

55. Bersaglieri, C. *et al.* Genome-wide maps of nucleolus interactions reveal distinct layers of repressive chromatin domains. *Nat. Commun.* **13**, 1483 (2022).

56. Rangasamy, D., Berven, L., Ridgway, P. & Tremethick, D. J. Pericentric heterochromatin becomes enriched with H2A.Z during early mammalian development. *EMBO J.* **22**, 1599–1607 (2003).

57. Ono, T., Fang, Y., Spector, D. L. & Hirano, T. Spatial and temporal regulation of Condensins I and II in mitotic chromosome assembly in human cells. *Mol. Biol. Cell* **15**, 3296–3308 (2004).

58. Kufer, T. A. *et al.* Human TPX2 is required for targeting Aurora-A kinase to the spindle. *J. Cell Biol.* **158**, 617–623 (2002).

59. Khongkow, P. *et al.* Paclitaxel targets FOXM1 to regulate KIF20A in mitotic catastrophe and breast cancer paclitaxel resistance. *Oncogene* **35**, 990–1002 (2016).

60. Politz, J. C. R., Scalzo, D. & Groudine, M. Something silent this way forms: the functional organization of the repressive nuclear compartment. *Annu. Rev. Cell Dev. Biol.* **29**, 241–270 (2013).

61. Rodrigues, A. *et al.* Nucleoli and the nucleoli–centromere association are dynamic during normal development and in cancer. *MBoC* **34**, br5 (2023).

62. Liang, C. *et al.* Centromere-localized Aurora B kinase is required for the fidelity of chromosome segregation. *J. Cell Biol.* **219**, (2020).

63. Broad, A. J., DeLuca, K. F. & DeLuca, J. G. Aurora B kinase is recruited to multiple discrete kinetochore and centromere regions in human cells. *J. Cell Biol.* **219**, (2020).

64. Hoffmeyer, K. *et al.* Wnt/β-catenin signaling regulates telomerase in stem cells and cancer

cells. *Science* **336**, 1549–1554 (2012).

65. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
66. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
67. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
68. Hershberg, E. A. *et al.* PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. *Nat. Methods* **18**, 937–944 (2021).
69. Santos-Barriopedro, I., van Mierlo, G. & Vermeulen, M. Off-the-shelf proximity biotinylation for interaction proteomics. *Nat. Commun.* **12**, 5015 (2021).
70. Iglesias, N. *et al.* Native Chromatin Proteomics Reveals a Role for Specific Nucleoporins in Heterochromatin Organization and Maintenance. *Mol. Cell* **77**, 51–66.e8 (2020).
71. Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable autonomous synthesis of single-stranded DNA. *Nat. Chem.* **10**, 155–164 (2018).
72. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).
73. Natsume, T., Kiyomitsu, T., Saga, Y. & Kanemaki, M. T. Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Rep.* **15**, 210–218 (2016).
74. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).

75. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

76. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

77. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* **8**, 70–73 (2011).

78. Teo, G. *et al.* SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J. Proteomics* **100**, 37–43 (2014).

79. Choi, H. *et al.* Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr. Protoc. Bioinformatics* **Chapter 8**, 8.15.1–8.15.23 (2012).

80. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).

81. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

82. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **36**, D646–50 (2008).

83. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

84. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

85. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

EMBnet.journal **17**, 10–12 (2011).

86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Picard. https://broadinstitute.github.io/picard/.
88. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
89. Xu, W. *et al.* CoolBox: a flexible toolkit for visual analysis of genomics data. *BMC Bioinformatics* **22**, 489 (2021).
90. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).

Chapter 5
CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Conclusions

This work developed new tools across the full range of proteomics workflows, from sample preparation, to mass spectrometry data acquisition, to downstream analysis. All three areas are important to consider in order to best take advantage of new instrumentation and biotechnologies.

Chapter 2 detailed that as mass spectrometry capabilities expand to produce thousands of spectra per minute and new applications impose significant computational burdens, existing database search tools struggle to maintain necessary speed. By adapting the fragment ion indexing approach into Comet's open-source framework, we significantly accelerated search capabilities while maintaining the algorithm's robustness and accessibility. This enhancement ensures Comet remains relevant for emerging applications and modern instrumentation demands.

Chapter 3 focused on intelligent data acquisition methods. Traditional data-dependent acquisition suffers from stochastic precursor selection, leading to missing values and reduced quantitative accuracy, particularly for sample-multiplexed methods. By implementing RTLS with optimized library processing workflows capable of searching millions of spectra in milliseconds, we achieved a 2-fold increase in instrument acquisition efficiency while outperforming existing real-time database searching approaches in some situations. RTLS demonstrated the first instance of chimeric spectra in a multiplexed sample and enabled efficient identification of post-translationally modified peptides, establishing itself as a valuable addition to the data acquisition toolkit.

Chapter 4 focused on the development of DNA O-MAP. Current methods for identifying proteins assembled at specific genomic loci either require a litany of upfront work in generating constructs, stably expressing them in cells, and growing cells in the billions per replicate. DNA O-MAP employs oligo-based in situ hybridization probes to recruit peroxidase activity to specific DNA locations , enabling cost-effective processing with a fraction of the cells. We demonstrate DNA O-MAP's specificity by recovering telomere-binding proteins and showcase its intracompartment specificity by distinguishing proteomes around pericentromeric repeats, telomeres, and mitochondrial genomes. Additionally, DNA O-MAP captures functionally relevant DNA-DNA interactions from intervals as small as 20 kilobases, establishing the possibility of single copy loci sensitivity.

Collectively, this work advances the proteomics field by addressing fundamental limitations in data acquisition intelligence, computational efficiency, and chromatin proteomics methodology. These developments enable more comprehensive, accurate, and scalable approaches to understanding protein function and regulation across diverse biological systems.

5.2 Future directions

Proteomics, and science generally, is driven forward by the development of new tools. New developments in instrumentation, machine learning, and other bioanalytical methods all create opportunities to advance the field. The work presented in this thesis contributes to open-source and accessible science by speeding up an important bioinformatic tool to match current computational demands. Science builds upon itself and supporting open software like Comet allows others to utilize it as a foundation for their own work rather than reinvent the wheel. Continuing to support Comet and similar tools is crucial to the field.

Modern mass spectrometers can measure billions of ions within an hour. Continuing to make acquisition methods "smarter" to most efficiently utilize all this information will be important in allowing proteomics from achieving the depth and throughput necessary to advance biology. Real-time library search offers a new method that can be used by itself but also combined with other IDA tools to further improve sensitivity and throughput, especially for analytically challenging samples. In order to routinely perform studies with sample counts in the tens or hundreds of thousands, multiplexing strategies will need to be implemented as short gradients will not be enough. Thinking on how intelligent acquisition can continue to improve upon isobaric or other multiplexing strategies will be critical.

Profiling how proteins at specific genomic loci respond to different stimuli has the potential to unveil both mechanisms of disease and identify potential therapeutic avenues. Previously, these types of experiments were time and resource intensive enough that doing a single loci was arduous and dozens were impossible. DNA O-MAP is the first step into making these experiments approachable and scalable. Improvements and new variations are already underway and chromatin-wide or multi-drug chemoproteomic screens are now feasible. Pushing sensitivity even further to profile post-translational modifications in these experiments will prove an important development as well as an opportunity to harness new intelligent data acquisition methods. This thesis ultimately sought to develop new tools that will lead to new discoveries and new ideas.