Advancing technologies for the study of proteoforms and protein-nucleic acid interactomes

by

Katherine B. Henke

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 12/17/2021

The dissertation is approved by the following members of the final oral committee:

Ying Ge, Professor, Cell and Regenerative Biology & Chemistry
Lingjun Li, Professor, Pharmaceutical Sciences & Chemistry
Douglas G. McNeel, Professor, Medicine
Lloyd M. Smith, Professor, Chemistry

ACKNOWLEDGEMENTS

I have been incredibly fortunate to have had the support, guidance, and encouragement of many people throughout my education. I would like to take a moment to recognize these people and wholeheartedly thank them for all they have done for me.

First, I would like to thank Professor Lloyd Smith for being a truly fantastic advisor throughout my graduate career. As a scientist, Lloyd enjoys working on the cutting edge, developing technologies that enable us to ask and answer new questions and drive scientific progress. He has instilled this mindset and love of technology development in me and, importantly, has also taught me to be patient with myself and with the scientific process, as failure is an inevitable part of doing things that have not been done before. As a mentor, Lloyd is thoughtful, attentive, kind, encouraging, and understanding. I always looked forward to my bi-weekly meetings with Lloyd because I knew that they were not only an opportunity for me to share my research and hear his feedback and advice, but also a chance for me to hear his "free associations", which inevitably introduced me to new topics, ideas, or areas of research. Even when my research wasn't progressing as I'd hoped, I always felt better walking out of a meeting with Lloyd than I felt walking in, and I think that is a mark of a truly great advisor. Lloyd has always fostered a collaborative, creative, fun environment where everyone's ideas are welcome, and it has been an absolute privilege to work in his group.

I would also like to thank the Smith group's three staff scientists—Drs. Michael Shortreed, Brian Frey, and Mark Scalf. Michael Shortreed is an exceptionally creative scientist who is unafraid to explore new ideas and encourages students to do the same. He is also an incredibly encouraging and selfless mentor, who genuinely cares for his students' well-being and is invested in helping them to succeed. I am grateful for the kindness and support he has shown me. Brian Frey is not only a wealth of knowledge about seemingly all things related to science, but he is also an excellent teacher. There have been times when I felt I was stumbling around in the dark, trying to grasp a concept, and a quick

conversation with Brian gave me clarity and understanding. I strive to become the kind of logical, clear thinker that Brian is. Finally, Mark Scalf is a skilled researcher, a patient mentor, and an awesome lab mate. Mark has an immense amount of experience with the practical aspects of performing proteomics experiments, and he is the first person I turn to for help when my proteins won't solubilize, when I'm not sure which sample prep method to use, or when I can't figure out the source of that darn contamination. He is also the only person I know who, after sample prep, can look at what appears to be an empty Eppendorf tube and somehow know that $1/16^{th}$ is the correct proportion to inject for good mass spec data. Mark creates a fun and friendly atmosphere in the lab by just being himself, and I am grateful for Mark's patience, positivity, and willingness to teach me throughout my graduate career.

I would also like to thank Professors Lingjun Li, Ying Ge, and Doug McNeel for taking time out of their busy schedules to serve on my thesis committee. Professor Li has been a member of my committee since my second-year TBO exam, and I am grateful for her support, kindness, and thoughtful questions. Professor Ge taught me a great deal about proteomics both in her class and through her group's papers, and I admire her passion for science and bright personality. Professor McNeel helped expose me to the world of immunopeptidomics, and through our collaboration I had the opportunity to expand my horizons and work in an area of research that was new and exciting to me. I appreciate his willingness to explain biological concepts to me as well as his collaborative spirit. I would also like to thank Professors Eric Streiter and Jim Weisshaar who have now left UW, but who served on my TBO and RP committees and asked thoughtful questions and provided helpful advice. Finally, I would like to thank Professors Samantha Glazier and Janel Owens for giving me a taste for research during my undergraduate years. Both of these women are kind, patient mentors and excellent role models for female scientists.

None of the work in this dissertation would have been possible without the efforts, ideas, and feedback of the students and postdocs of the Smith group, both past and present. I would like to thank all of them for being such bright scientists, good people, and wonderful groupmates. I am grateful to

Dr. Yunxiang (Sean) Dai for teaching me about HyCCAPP and for letting me follow him around and patiently answering my questions during my first year. Sean was also my office mate for multiple years and my primary collaborator on Chapter 3 of this dissertation, and I enjoyed chatting with him and bouncing ideas off one another. Dr. Rachel Knoener, Dr. Michele Spiniello, and Isabella Whitworth were excellent collaborators on the HyPR-MS team, and I am grateful to each of them for their thoughtful questions, insight, and, most of all, kindness. I would like to thank Rachel Miller and Drs. Leah Schaffer, Rob Millikin, Zach Rolfs, Anthony Cesnik, Stefan Solntsev, and Lei Lu for being fun lab mates and brilliant programmers, and for patiently taking the time to help me become a better user of their proteomics software. Lastly, I am grateful to Dr. Matt Holden and Maisie Steinbrink, John Pavek, Sam Markovich, Samantha Shrum, Austin Carr, Kyndalanne Pike, and Yuling Dai for being excellent researchers and groupmates. There's not one bad apple in the bunch—it has been a pleasure working with you all.

The UW Chemistry department is extremely fortunate to have so many wonderful staff and faculty members working to keep the research enterprise moving forward, and I am thankful to each of them for the work they do behind the scenes. The past few years have been very difficult for the University and the Chemistry department in particular, with the flood, pandemic, and building construction imposing restrictions and causing multiple prolonged building closures. While these challenges were, at times, extremely frustrating, I know that everything would have been infinitely worse without the hard work, leadership, and advocacy of people like Chancellor Rebecca Blank, Chemistry Chairs and Professors Judith Burstyn and Clark Landis, Professors Bob McMahon and John Moore, as well as Jeff Nielsen, Matt Sanders, and Pat Egan. I hope you know how much you are appreciated, despite the griping and angst you have no doubt had to endure. I am also grateful to Sue Martin Zernicke, Liv West, Cheri Stephens, Bruce Goldade, and Mike Bradley for their work to support me and my research, always with kindness and humor. Additionally, I am grateful to the Genomic Sciences Training Program for providing financial support and opportunities for me to grow as a

scientist, as well as to the American taxpayers, who ultimately supported all of the research in this dissertation.

My family has been a constant source of love and encouragement throughout my entire life, and I am more grateful to them than I can possibly articulate. My mom, Deb, is my biggest cheerleader, my confidant, and my friend. She is the first person I call whether I have good news, bad news, or just need to talk, and her endless support and belief in me has played a huge role in getting me to this point. My dad, Clint, is an excellent listener and a fountain of wisdom and good advice. More than that, he is a truly kind, compassionate, and selfless person, and I hope he knows that I look up to him in so many ways. I have also been blessed with two outstanding step-parents—Darryl and Margaret—who have shown me nothing but unconditional love and support since the day I entered their lives. Additionally, I have been fortunate to have four wonderful grandparents—Cliff, Nancy, Howard, and Eileen—who have been huge influences on me throughout my entire life. Although we lost three of them during my graduate school years, I know how lucky I have been to have such incredible pillars of love and support.

Last, but definitely not least, I would like to thank my wonderful husband, Austin. Austin, who recently received his Ph.D. from the UW Department of Chemistry, has been in my life since 2013 and has been by my side in Madison since my second year of graduate school. I feel so fortunate to have shared this experience with him, as he understands better than anyone the highs and lows of life as a graduate student. I am grateful every day for his love, support, encouragement, perspective, and goofy humor, and for the much-needed balance he provides in my life. Austin, thank you for being a fantastic partner and my very best friend. I can't wait to see what the future has in store for us.

AUTHOR CONTRIBUTIONS

The work presented in this dissertation was conducted in the Department of Chemistry at the University of Wisconsin-Madison under the supervision of Professor Lloyd M. Smith. Funding for the work presented herein was provided by the National Human Genome Research Institute (NHGRI), the National Institute of General Medical Sciences (NIGMS), the National Cancer Institute (NCI), and the National Heart, Lung, and Blood Institute (NHLBI). Three chapters in this dissertation have been published elsewhere, while a fourth chapter describes the ongoing development of a technology capable of identifying proteoforms bound to target RNA transcripts.

I have had the pleasure of collaborating with many creative, thoughtful scientists throughout my Ph.D. studies. Below are descriptions of my contributions (as K. E. Buxton and K. B. Henke) to the work presented in this dissertation, as well as the contributions of my respected colleagues.

Chapter 2 was adapted from a publication in the *Journal of Proteome Research*. It was reprinted with permission from:

Buxton, K. E.; Kennedy-Darling, J.; Shortreed, M. R.; Zaidan, N. Z.; Olivier, M.; Scalf, M.; Sridharan, R.; Smith, L. M. Elucidating protein-DNA interactions in human alphoid chromatin via hybridization capture and mass spectrometry. *J. Proteome Res.* **2017**, *16* (9), 3433-3442. https://doi.org/10.1021/acs.jproteome.7b00448.

K.E.B., J.K.D., M.O., and L.M.S. conceived of HyCCAPP experiments. K.E.B. and J.K.D. developed experimental aspects of HyCCAPP, and K.E.B. performed HyCCAPP experiments. M.S. and K.E.B. analyzed captured proteins via mass spectrometry. K.E.B. and M.R.S. analyzed mass spectrometric data. K.E.B, N.Z.Z., and R.S. provided biological insight into proteins identified as enriched at alpha satellite DNA via HyCCAPP. K.E.B., N.Z.Z., and R.S. conceived of ChIP validation experiments, and K.E.B. executed these experiments. K.E.B. and L.M.S. prepared the manuscript. All authors edited and reviewed the manuscript.

Chapter 3 was adapted from a publication in the *Journal of Proteome Research*. It was reprinted with permission from:

Dai, Y.;[#] Buxton, K. E.;[#] Schaffer, L. V.; Miller, R. M.; Millikin, R. J.; Scalf, M.; Frey, B. L.; Shortreed, M. R.; Smith, L. M. Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. *J. Proteome*

Res. **2019**, *18* (10), 3671-3680. https://doi.org/10.1021/acs.jproteome.9b00339.

[#]Y.D. and K.E.B. contributed equally to this work. L.M.S., M.R.S., B.L.F., M.S., K.E.B., and Y.D. conceived the use of NeuCode intact-mass and top-down proteomics for human proteoform analysis. K.E.B. and R.M.M. prepared samples for analysis. M.S. facilitated LC/MS analysis. K.E.B. and Y.D. processed raw data. R.M.M., Y.D., K.E.B., R.J.M. and B.L.F. built G-PTM-D databases from bottom-up data. Y.D. and L.V.S. used Proteoform Suite to integrate proteomic data and identified proteoforms and proteoform families. Y.D. and K.E.B. wrote the manuscript. All authors edited the manuscript.

Chapter 4 was adapted from a publication in *Post-Transcriptional Gene Regulation*, 3rd ed., part of the Methods in Molecular Biology book series. It was reproduced with permission from Springer Nature.

Henke, K. B.; Miller, R. M.; Knoener, R. A.; Scalf, M.; Spiniello, M.; Smith, L. M. Identifying protein interactomes of target RNAs using HyPR-MS. In *Post-Transcriptional Gene Regulation*, 3rd ed., Methods in Molecular Biology, vol. 2404; Dassi, E., Ed.; Humana Press: New York, NY, 2022; pp 219-244. https://doi.org/10.1007/978-1-0716-1851-6_12.

K.B.H. and R.M.M. wrote the manuscript. K.B.H. and R.A.K. made the figures. All authors edited the manuscript.

Chapter 5 describes the ongoing development of a technology capable of identifying proteoforms bound to target RNA transcripts.

Henke, K. B.; Miller, R. M.; Scalf, M.; Smith, L. M. Identifying proteoforms bound to target RNAs with HyPR-MS: current status and future directions.

K.B.H., R.M.M., and L.M.S. conceived of top-down HyPR-MS experiments. K.B.H. and R.M.M. performed cell culture. R.M.M. performed reverse cross-linking experiments and K.B.H. analyzed these data. K.B.H. and R.M.M. conceived of using suspension trapping (S-Trap) for top-down proteomics sample preparation. M.S. provided guidance on proteomics sample preparation and LC/MS analysis and performed instrument maintenance. K.B.H. performed all other experiments and data analysis. K.B.H. wrote the chapter.

TABLE OF CONTENTS

ABSTRACT

Proteins are critical actors within the cell, enabling complex biological processes essential for cell survival, development, and homeostasis. Generally, proteins function via their interactions with other biomolecules, such as nucleic acids, making knowledge of these interactions and the players involved essential to our understanding of even basic biological function. This dissertation describes the advancement of technologies for the identification of proteins interacting with target nucleic acid sequences (e.g., genomic loci or RNA transcripts), as well as the development of approaches for better understanding the diversity of proteins expressed in human cells. **Chapter 2** presents the application of HyCCAPP (Hybridization Capture of Chromatin-Associated Proteins for Proteomics), a technology developed in the Smith lab for the study of protein-DNA interactions in a locus-specific manner, to identify the protein interactome of human centromeric alpha satellite DNA. We identified 90 proteins as enriched in alphoid chromatin, and this list included many known centromere-binding proteins in addition to multiple novel alpha satellite-binding proteins. This work represents the first application of the HyCCAPP technology in mammalian cells and is the first DNA-centric examination of human protein-alpha satellite interactions. In **Chapter 4**, we present a detailed and comprehensive guide for the application of HyPR-MS (Hybridization Purification of RNA-protein complexes followed by Mass Spectrometry), a technology developed in the Smith lab for the identification of proteins interacting with target RNA transcripts. It is our hope that the practical advice provided in this chapter will enable the widespread utilization of this technology. In **Chapter 3**, we explored how different types of

proteomics data could be integrated to maximize proteoform identifications from a human cell line. Proteoforms are the specific molecular forms of proteins expressed in the cell, accounting for genetic variation, alternative splicing, and post-translational modifications. Through the integration of intact-mass, top-down, and bottom-up proteomics data, we were able to identify ~1,200 proteoforms representing 484 genes from the human Jurkat cell line. Finally, in **Chapter 5**, we present the current status of our work to combine proteoform analysis with HyPR-MS to enable the first ever study of the proteoforms bound to a target RNA transcript.

CHAPTER 1

INTRODUCTION

1.1 PROTEIN-NUCLEIC ACID INTERACTIONS

Proteins are critical actors in cellular biology, enabling complex processes essential for cell survival, development, and homeostasis. Generally, proteins function via their interactions with other biomolecules, making knowledge of these interactions and the players involved essential to our understanding of even basic biological function. A primary focus of this dissertation is the development and application of technologies to study the interactions of proteins with nucleic acids (i.e., genomic DNA and cellular RNA transcripts).

Protein-DNA interactions are crucial to numerous cellular processes, including but not limited to DNA replication, DNA repair, dictation of chromatin structure, and gene expression.[1–6] In addition

to the various roles that these interactions play in healthy cells, anomalous protein-DNA interactions are implicated in numerous disease mechanisms, including neurodegenerative diseases[7–11] and cancer.[12–14] Similarly, protein-RNA interactions are integral to multiple aspects of cellular function, including RNA transcription, translation, splicing, localization, and degradation,[15–21] and disrupted or aberrant protein-RNA interactions are responsible for numerous pathological states.[22–25]

1.2 TECHNOLOGIES FOR STUDYING PROTEIN-NUCLEIC ACID INTERACTIONS

Recognition of the importance of protein-nucleic acid interactions has led to the development of numerous technologies for their study and elucidation. These technologies can be broadly classified as either protein occupancy methods, targeted protein-centric methods, or targeted locus-/transcript-centric methods, and we will briefly discuss each of these categories in turn.

We first turn our attention towards tools meant to characterize the global degree of protein occupancy across the entire genome/transcriptome. In the realm of DNA-protein interactions, techniques such as DNase hypersensitivity (DHS) mapping[26,27] and formaldehyde-assisted isolation of regulatory elements (FAIRE)[28] are used to identify regions of "open" chromatin. The impetus for such technologies is that open regions of chromatin are more accessible to transcriptional machinery, and therefore these experiments may reveal sites of regulatory elements, including enhancers, promoters, insulators, and silencers.[26,28] Cognate tools exist to assess the degree of protein occupancy across the transcriptome. For example, orthogonal organic phase separation (OOPS)[29] and phenol toluol

extraction (PTex)[30] are recently-published protocols that enable the separation of "free" RNA from protein-bound RNA in UV-cross-linked samples. The RNA in both populations can be identified via sequencing, providing a snapshot of protein occupancy across the transcriptome. Additionally, these protocols separate "free" protein from RNA-protein complexes, and therefore mass spectrometry can be used to identify the RNA-bound proteome. While tools which reveal the degree of protein occupancy across the genome/transcriptome can be useful, these techniques generally do not provide extensive insight into specific protein-DNA/protein-RNA interactions. In other words, we know which regions of the genome/transcriptome are densely populated by or depleted of proteins, but we typically do not know which proteins are interacting with which genomic loci/RNA transcripts. This level of detail is necessary for identifying the players involved in specific cellular processes.

Targeted protein-centric technologies can provide more detailed information about specific protein-nucleic acid interactions. These technologies operate by first selecting a specific protein to study and then attempting to elucidate its interactions with genomic loci/RNA transcripts. Typically, an antibody is used to pull down the protein of interest from in vivo cross-linked cell lysate, and associated DNA/RNA is identified via quantitative PCR (qPCR) or high-throughput sequencing. When used to study protein-DNA interactions, this process is known as chromatin immunoprecipitation (ChIP), and modifications to this basic workflow have been developed to address a variety of research questions.[31–38] When used to study protein-RNA interactions, this process is

known as cross-linking immunoprecipitation (CLIP) or RNA immunoprecipitation (RIP) and, again, multiple variants of this general procedure have been published.[39–46]

ChIP and CLIP are powerful tools, particularly if one's study is aimed at investigating a single protein and understanding where in the genome/transcriptome that protein binds. These technologies have several advantages, including (1) they can provide insight into in vivo protein-nucleic acid interactions, which are essentially impossible to accurately mimic in vitro, (2) they can provide specific information about the identities of both the protein and the DNA/RNA sequence involved in a particular interaction, and (3) they require relatively small amounts of starting material, since the captured DNA/RNA fragments can be PCR-amplified prior to analysis. However, despite these strengths, none of these targeted protein-centric technologies allow for the characterization of the entire collection of proteins interacting at a specific genomic locus/RNA transcript. For that, we turn to our final category of methods to study protein-nucleic acid interactions.

Proteins often bind to DNA/RNA in a combinatorial and context-specific fashion to accomplish critical cellular functions.[17,47–49] Thus, a thorough understanding of the biology occurring at a genomic locus/RNA transcript of interest requires knowledge of the complete ensemble of proteins interacting at that locus/transcript. Targeted locus-/transcript-centric methods for studying protein-nucleic acid interactions can provide this information by selectively isolating a locus/transcript of interest and then identifying its associated proteins, typically by mass spectrometry. However, there are a number of important technical challenges to consider when using captured protein as a readout

of the experiment. Unlike ChIP/CLIP, whose readout is DNA/RNA which may be PCR-amplified prior to analysis, it is not possible to amplify protein, and therefore experimental conclusions must be drawn based solely on the amount of protein captured in the experiment. This problem is compounded by the fact that protein expression levels within the cell vary by many orders of magnitude, and some proteins of interest may be present only in very low quantities.[50] Thus, these locus-/transcript-centric tools are limited by (1) the detection limit of the mass spectrometer used for protein identification and (2) the amount of background protein "noise" present in an experiment, which may prevent interesting protein interactors from being detected.

Despite these significant challenges, a number of locus-/transcript-centric technologies for studying protein-nucleic acid interactions have been developed.[39,51–53] Gauchier et al. recently reviewed DNA-centric tools for studying protein-DNA interactions,[52] including but not limited to ChAP-MS,[54–56] PICh,[57–61] and several tools relying on nuclease-dead Cas9 for recognition of the target locus.[62–66] Similarly, RNA-centric methods for studying protein-RNA interactions, such as CHART-MS,[67] ChIRP-MS,[68] and RAP-MS,[69] were recently reviewed by Gräwe et al.[53]

In this work, we describe two nucleic acid-centric technologies developed in the Smith lab for studying DNA-/RNA-protein interactions—HyCCAPP (DNA) and HyPR-MS (RNA) (Figure 1.1). Both technologies utilize sequence-specific oligonucleotide(s) to selectively hybridize to target DNA/RNA sequence(s) in in vivo formaldehyde-cross-linked cell lysate. A biotin tag on the capture oligonucleotide(s) is then used to isolate the hybridized DNA-/RNA-protein complexes on

streptavidin-coated magnetic beads. The target DNA-/RNA-protein complexes are then eluted from the beads and the proteins are analyzed via mass spectrometry. By comparing the proteins identified in the target capture sample to those identified in control samples, one can identify the in vivo protein interactome of the target genomic locus/RNA transcript.

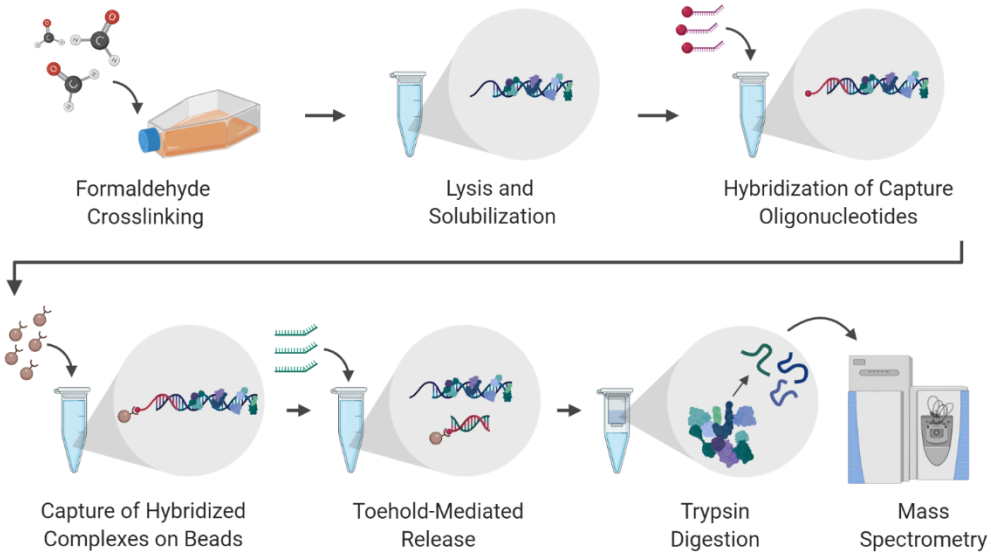


Figure 1.1 Overview of the HyCCAPP technology. Cells are cross-linked with formaldehyde in vivo to covalently fix protein-DNA interactions. Cells are then lysed and chromatin is solubilized via sonication. Sequence-specific, biotinylated capture oligonucleotides (magenta) are then introduced to the lysate and hybridize to the target DNA locus. Hybridized complexes are captured on streptavidin-coated magnetic beads and the beads are washed to remove nonspecific interactors. Release oligonucleotides (green) are then added and target DNA-protein complexes are released from the beads via toehold-mediated strand displacement.[70] The proteins are then digested with trypsin prior to mass spectrometric analysis to identify the protein interactome of the target DNA locus. Note that the procedure for HyPR-MS is similar to the procedure depicted here, except the target nucleic acid for HyPR-MS is RNA, not DNA. (This figure was created with BioRender.com and was adapted from Knoener et al.[71] under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/))

HyCCAPP (Hybridization Capture of Chromatin-Associated Proteins for Proteomics) was initially reported in 2014 and applied for the identification of proteins interacting with three multi-copy loci and one single-copy locus in *S. cerevisiae*.[72] It has since been used to identify proteins associated with the ENO2 and GAL1 promotor regions of *S. cerevisiae* grown using glucose versus galactose as the carbon source[73] and multiplexed using a toehold-mediated release strategy[70] to identify proteins interacting with four genomic loci in *S. cerevisiae* grown under normal and salt-stressed conditions.[74] In **Chapter 2** of this dissertation, we present the application of HyCCAPP to study the protein interactome of human alpha satellite DNA, a repetitive class of DNA found in human centromeres.[75] This work represents the first application of the HyCCAPP technology in mammalian cells and is the first DNA-centric examination of human protein-alpha satellite interactions.

HyPR-MS (Hybridization Purification of RNA-protein complexes followed by Mass Spectrometry) was first reported in 2017 and applied for the identification of proteins interacting with HIV-1 RNA in HIV-infected cells.[76] It has since been applied to identify host proteins interacting with different HIV-1 RNA splice variants,[71] to identify the protein interactomes of three human long noncoding RNAs (MALAT1, NEAT1, and NORAD),[77] and to identify proteins interacting with c-Myc mRNA in human cells.[78] In **Chapter 4** of this dissertation we present a detailed and comprehensive guide for the application of HyPR-MS,[79] discussing experimental considerations and providing practical advice to enable the widespread utilization of this technology.

1.3 PROTEOFORMS AND THEIR IMPORTANCE FOR BIOLOGICAL FUNCTION

Thus far, we have discussed proteins as the molecules which drive cellular function, either independently or through their interactions with other biomolecules. The complex processes required for cellular homeostasis are made possible by the diversity of proteins expressed by the cell, and it is therefore prudent to examine the source(s) of this diversity. The human genome encodes ~20,000 protein-coding genes, but the size of the human proteome extends beyond this number when we consider the heterogeneity introduced by post-translational modifications (PTMs) and sequence variations due to single-nucleotide polymorphisms and alternative splicing.[80] In recognition of the fact that even a single gene can give rise to a wide array of distinct protein molecules, the term "proteoform" was introduced to describe a particular molecular form of a protein, with a specified amino acid sequence and localized set of PTMs.[81] Additionally, the term "proteoform family" was introduced to describe the set of proteoforms derived from the same gene.[82]

Even proteoforms belonging to the same family can exhibit vastly different biological functions. A classic example is that of histone proteoforms, where the specific combination of PTMs on a histone protein can serve to recruit other proteins and influence the open/closed nature of chromatin, thereby helping to regulate processes such as gene expression.[83] Another example is that of Elk-1 proteoforms. Elk-1 is a transcription factor with multiple phosphorylation sites, and early phosphorylation of some of these sites promotes the interaction of Elk-1 with the Mediator

transcriptional coactivator complex, resulting in transcriptional activation.[84] However, as additional phosphorylation sites are progressively modified, recruitment of the Mediator complex is inhibited and transcriptional activation is limited.[84] Thus, these different Elk-1 proteoforms have opposite effects on transcriptional activity, despite the fact that their parent genes and base amino acid sequences are the same. These examples make it clear that a complete understanding of cellular biology requires analysis of the proteome on the proteoform level.

1.4 MASS SPECTROMETRY-BASED PROTEOFORM ANALYSIS

Mass spectrometry (MS) is a powerful and high-throughput tool for the analysis of proteins in complex samples. The most commonly used approach for performing MS-based proteomics is the so-called "bottom-up" or "shotgun" method.[85,86] In a standard bottom-up proteomics workflow, proteins in a sample are digested with a protease and the resultant peptides are analyzed via liquid chromatography-tandem MS (LC-MS/MS) (Figure 1.2, top). The peptides are separated via online LC and converted to gas-phase ions, which are then introduced into the mass spectrometer. The mass-to-charge (m/z) ratios of the intact peptide ions are then measured by the mass analyzer, and certain species (typically the most abundant) are selected for fragmentation and the m/z ratios of the fragment ions are measured. The precursor peptide masses and associated fragmentation spectra are then searched against an in silico-digested protein database for identification, and identified peptides are used to infer the presence of proteins in the sample.

Bottom-up proteomics is a robust and widely used approach, and its key feature is that proteins are digested into peptides prior to analysis. The reason for this is that peptides are easier to analyze than their parent proteins, as peptides are typically easily solubilized and separated.[87] Additionally, peptides generated using the standard trypsin protease tend to ionize and fragment well and fall within a mass range that is suitable for mass spectrometric analysis.[86,87] Unfortunately, bottom-up proteomics does not allow for proteoform identification. This is because the proteolytic digestion inherent to bottom-up proteomics destroys the molecular context of the intact proteoform, making it impossible to decipher the parent proteoform of an identified peptide (Figure 1.2, top). Thus, additional strategies are required for MS-based proteoform analysis.

Top-down proteomics[88–92] is an alternative to bottom-up proteomics wherein intact proteoforms are analyzed directly, without the intermediary step of protease digestion (Figure 1.2, middle). Similar to bottom-up proteomics, the *m/z* ratios of precursor (in this case, intact proteoform) ions are measured and certain species are selected for fragmentation. The resultant precursor masses and fragmentation spectra are then searched against a proteoform database and, because the relationship between amino acid sequence and PTMs is preserved, top-down proteomics is capable of proteoform identification. However, top-down mass spectrometry experiments do face analytical challenges, including the low abundance of many proteoforms, low signal-to-noise ratios for large proteoforms,[93] insufficient proteoform fragmentation for PTM localization, complex data analysis, and coelution of proteoforms in standard LC-MS/MS experiments.[88]

An additional approach for MS-based analysis of proteoforms is "intact-mass" proteomics (Figure 1.2, bottom). In this approach, as in top-down proteomics, intact proteoform molecules are analyzed directly without protease digestion. However, unlike top-down proteomics, proteoforms are not fragmented and therefore must be identified based on their intact mass alone or by their intact mass coupled with some other piece of information (e.g., the number of lysine residues[82]). A key benefit of foregoing fragmentation in intact-mass experiments is that more MS acquisition time can be spent collecting intact proteoform measurements. Additionally, because only a limited number of proteoforms can be fragmented over the time-scale of an LC-MS/MS run, many proteoforms go unidentified in top-down proteomics experiments either because they were not selected for fragmentation or because they generated low-quality fragmentation spectra. Intact-mass proteomics offers the opportunity to identify such proteoforms. However, key limitations of intact-mass proteomics are that, in general, PTMs cannot be localized to specific residues and proteoforms with the same mass cannot be distinguished from one another.

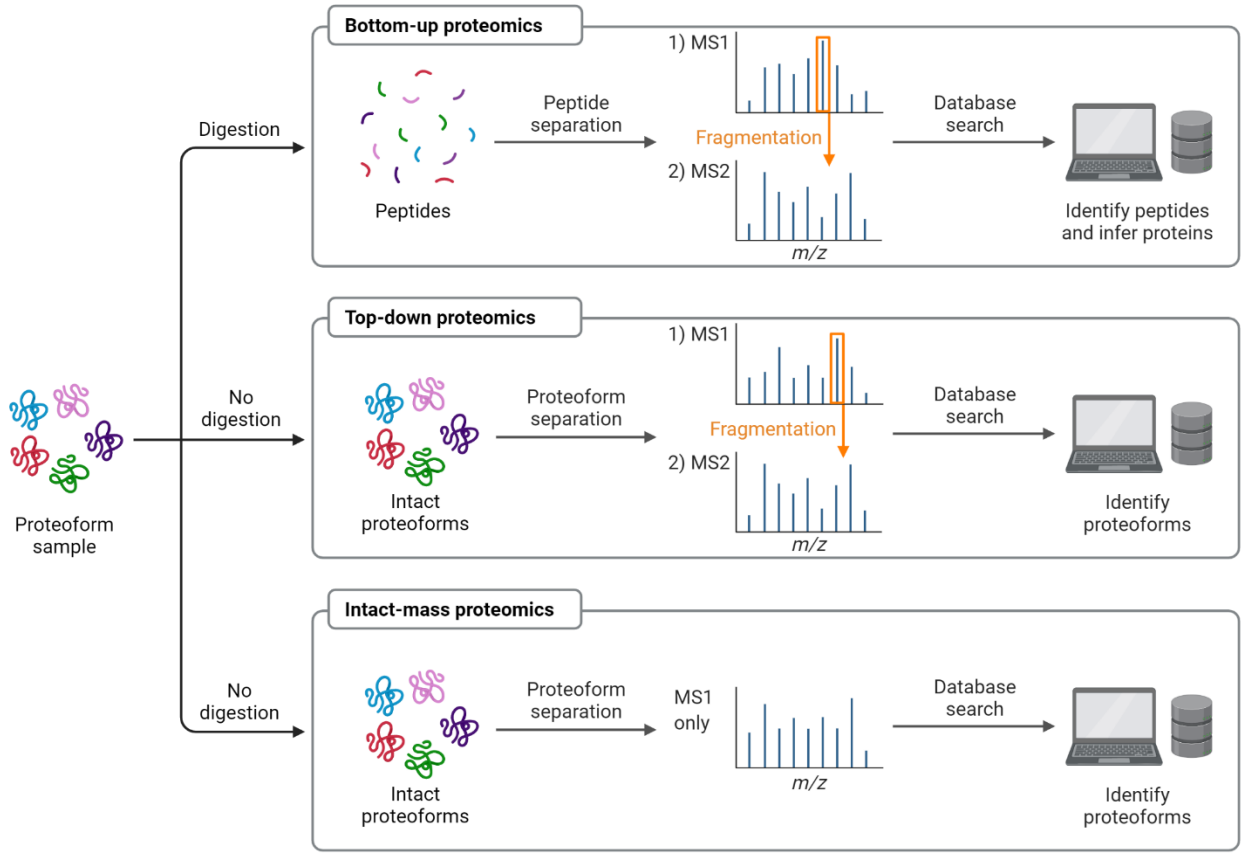


Figure 1.2 Schematics of bottom-up, top-down, and intact-mass proteomics workflows. (Figure adapted from "Types of Proteomics Workflows (Bottom-up, Middle-down and Top-down)", by BioRender.com (2021). Retrieved from https://app.biorender.com/biorender-templates)

Bottom-up, top-down, and intact-mass proteomics approaches all have their own strengths and limitations. Fortunately, it is possible to integrate these strategies to leverage the advantages of each, as has been demonstrated in a variety of studies.[94–101] In **Chapter 3** of this dissertation, we present the integration of bottom-up, top-down, and intact-mass proteomics data to maximize proteoform identifications from the human Jurkat cell line.[102] Finally, in recognition of the importance of both proteoforms and protein-RNA interactions, we present the current status of our work to combine the

HyPR-MS strategy described above with top-down proteomics in **Chapter 5**. This work serves as the foundation for the first ever analysis of proteoforms bound to specific target RNA(s).

1.5 REFERENCES

(1) Burgers, P. M. J.; Kunkel, T. A. Eukaryotic DNA replication fork. *Annu. Rev. Biochem.* **2017**, *86*, 417–438.

(2) Chatterjee, N.; Walker, G. C. Mechanisms of DNA damage, repair and mutagenesis. *Environ. Mol. Mutagen.* **2017**, *58* (5), 235–263.

(3) Olins, D. E.; Olins, A. L. Chromatin history: our view from the bridge. *Nat. Rev. Mol. Cell Biol.* **2003**, *4* (10), 809–814.

(4) Wolffe, A. P.; Guschin, D. Review: chromatin structural features and targets that regulate transcription. *J. Struct. Biol.* **2000**, *129* (2–3), 102–122.

(5) Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T. The human transcription factors. *Cell* **2018**, *172* (4), 650–665.

(6) Vaquerizas, J. M.; Kummerfeld, S. K.; Teichmann, S. A.; Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **2009**, *10* (4), 252–263.

(7) Benn, C. L.; Sun, T.; Sadri-Vakili, G.; McFarland, K. N.; DiRocco, D. P.; Yohrling, G. J.; Clark, T. W.; Bouzou, B.; Cha, J.-H. J. Huntingtin modulates transcription, occupies gene promoters in vivo, and binds directly to DNA in a polyglutamine-dependent manner. *J. Neurosci.* **2008**, *28* (42), 10720–10733.

(8) Jiang, W.; Han, Y.; Zhou, R.; Zhang, L.; Liu, C. DNA is a template for accelerating the aggregation of copper, zinc superoxide dismutase. *Biochemistry* **2007**, *46* (20), 5911–5923.

(9) Jiménez, J. S. Protein-DNA interaction at the origin of neurological diseases: a hypothesis. *J. Alzheimer's Dis.* **2010**, *22* (2), 375–391.

(10) Camero, S.; Benítez, M. J.; Jiménez, J. S. Anomalous protein-DNA interactions behind neurological disorders. *Adv. Protein Chem. Struct. Biol.* **2013**, *91*, 37–63.

(11) Cordeiro, Y.; Macedo, B.; Silva, J. L.; Gomes, M. P. B. Pathological implications of nucleic acid interactions with proteins associated with neurodegenerative diseases. *Biophys. Rev.* **2014**, *6* (1),

97–110.

(12) Freed-Pastor, W. A.; Prives, C. Mutant p53: one name, many proteins. *Genes Dev.* **2012**, *26* (12), 1268–1286.

(13) Alvarado-Ortiz, E.; de la Cruz-López, K. G.; Becerril-Rico, J.; Sarabia-Sánchez, M. A.; Ortiz-Sánchez, E.; García-Carrancá, A. Mutant p53 gain-of-function: role in cancer development, progression, and therapeutic approaches. *Front. Cell Dev. Biol.* **2021**, *8*, 607670.

(14) Shiroma, Y.; Takahashi, R.; Yamamoto, Y.; Tahara, H. Targeting DNA binding proteins for cancer therapy. *Cancer Sci.* **2020**, *111* (4), 1058–1064.

(15) Moore, M. J. From birth to death: the complex lives of eukaryotic mRNAs. *Science* **2005**, *309* (5740), 1514–1518.

(16) Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **2008**, *582* (14), 1977–1986.

(17) Mitchell, S. F.; Parker, R. Principles and properties of eukaryotic mRNPs. *Mol. Cell* **2014**, *54* (4), 547–558.

(18) Re, A.; Joshi, T.; Kulberkyte, E.; Morris, Q.; Workman, C. T. RNA-protein interactions: an overview. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, Methods in Molecular Biology, vol. 1097; Gorodkin, J., Ruzzo, W. L., Eds.; Humana Press: Totowa, NJ, 2014; pp 491–521.

(19) Matera, A. G.; Terns, R. M.; Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (3), 209–220.

(20) Mayr, C. Regulation by 3'-untranslated regions. *Annu. Rev. Genet.* **2017**, *51*, 171–194.

(21) Marchese, F. P.; Raimondi, I.; Huarte, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* **2017**, *18* (1), 206.

(22) Allerson, C. R.; Cazzola, M.; Rouault, T. A. Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. *J. Biol. Chem.* **1999**, *274* (37), 26439–26447.

(23) Lukong, K. E.; Chang, K.; Khandjian, E. W.; Richard, S. RNA-binding proteins in human genetic disease. *Trends Genet.* **2008**, *24* (8), 416–425.

(24) Corbett, A. H. Post-transcriptional regulation of gene expression and human disease. *Curr. Opin. Cell Biol.* **2018**, *52*, 96–104.

(25) Conlon, E. G.; Manley, J. L. RNA-binding proteins in neurodegeneration: mechanisms in

aggregate. *Genes Dev.* **2017**, *31* (15), 1509–1528.

(26) Thurman, R. E.; Rynes, E.; Humbert, R.; Vierstra, J.; Maurano, M. T.; Haugen, E.; Sheffield, N. C.; Stergachis, A. B.; Wang, H.; Vernot, B.; et al. The accessible chromatin landscape of the human genome. *Nature* **2012**, *489* (7414), 75–82.

(27) Vierstra, J.; Stamatoyannopoulos, J. A. Genomic footprinting. *Nat. Methods* **2016**, *13* (3), 213–221.

(28) Giresi, P. G.; Kim, J.; McDaniell, R. M.; Iyer, V. R.; Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **2007**, *17* (6), 877–885.

(29) Queiroz, R. M. L.; Smith, T.; Villanueva, E.; Marti-Solano, M.; Monti, M.; Pizzinga, M.; Mirea, D.-M.; Ramakrishna, M.; Harvey, R. F.; Dezi, V.; et al. Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **2019**, *37* (2), 169–178.

(30) Urdaneta, E. C.; Vieira-Vieira, C. H.; Hick, T.; Wessels, H.-H.; Figini, D.; Moschall, R.; Medenbach, J.; Ohler, U.; Granneman, S.; Selbach, M.; et al. Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nat. Commun.* **2019**, *10* (1), 990.

(31) Mukhopadhyay, A.; Deplancke, B.; Walhout, A. J. M.; Tissenbaum, H. A. Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in Caenorhabditis elegans. *Nat. Protoc.* **2008**, *3* (4), 698–709.

(32) Buck, M. J.; Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **2004**, *83* (3), 349–360.

(33) Barski, A.; Cuddapah, S.; Cui, K.; Roh, T.-Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell* **2007**, *129* (4), 823–837.

(34) Adli, M.; Zhu, J.; Bernstein, B. E. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat. Methods* **2010**, *7* (8), 615–618.

(35) Lambert, J.-P.; Mitchell, L.; Rudner, A.; Baetz, K.; Figeys, D. A novel proteomics approach for the discovery of chromatin-associated protein networks. *Mol. Cell. Proteomics* **2009**, *8* (4), 870–882.

(36) Lambert, J.-P.; Fillingham, J.; Siahbazi, M.; Greenblatt, J.; Baetz, K.; Figeys, D. Defining the budding yeast chromatin-associated interactome. *Mol. Syst. Biol.* **2010**, *6* (1), 448.

(37) Wang, C. I.; Alekseyenko, A. A.; LeRoy, G.; Elia, A. E. H.; Gorchakov, A. A.; Britton, L.-M. P.; Elledge, S. J.; Kharchenko, P. V; Garcia, B. A.; Kuroda, M. I. Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila. *Nat. Struct. Mol. Biol.* **2013**, *20* (2), 202–209.

(38) Guillen-Ahlers, H.; Shortreed, M. R.; Smith, L. M.; Olivier, M. Advanced methods for the analysis of chromatin-associated proteins. *Physiol. Genomics* **2014**, *46* (13), 441–447.

(39) Ramanathan, M.; Porter, D. F.; Khavari, P. A. Methods to study RNA-protein interactions. *Nat. Methods* **2019**, *16* (3), 225–234.

(40) Ule, J.; Jensen, K. B.; Ruggiu, M.; Mele, A.; Ule, A.; Darnell, R. B. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **2003**, *302* (5648), 1212–1215.

(41) Licatalosi, D. D.; Mele, A.; Fak, J. J.; Ule, J.; Kayikci, M.; Chi, S. W.; Clark, T. A.; Schweitzer, A. C.; Blume, J. E.; Wang, X.; et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **2008**, *456* (7221), 464–469.

(42) Hafner, M.; Landthaler, M.; Burger, L.; Khorshid, M.; Hausser, J.; Berninger, P.; Rothballer, A.; Ascano Jr., M.; Jungkamp, A.-C.; Munschauer, M.; et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **2010**, *141* (1), 129–141.

(43) König, J.; Zarnack, K.; Rot, G.; Curk, T.; Kayikci, M.; Zupan, B.; Turner, D. J.; Luscombe, N. M.; Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **2010**, *17* (7), 909–915.

(44) Van Nostrand, E. L.; Pratt, G. A.; Shishkin, A. A.; Gelboin-Burkhart, C.; Fang, M. Y.; Sundararaman, B.; Blue, S. M.; Nguyen, T. B.; Surka, C.; Elkins, K.; et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **2016**, *13* (6), 508–514.

(45) Kim, B.; Kim, V. N. fCLIP-seq for transcriptomic footprinting of dsRNA-binding proteins: lessons from DROSHA. *Methods* **2019**, *152*, 3–11.

(46) Gagliardi, M.; Matarazzo, M. R. RIP: RNA immunoprecipitation. In *Polycomb Group Proteins: Methods and Protocols*, Methods in Molecular Biology, vol. 1480; Lanzuolo, C., Bodega, B., Eds.; Humana Press: New York, NY, 2016; pp 73–86.

(47) Gerstein, M. B.; Kundaje, A.; Hariharan, M.; Landt, S. G.; Yan, K.-K.; Cheng, C.; Mu, X. J.; Khurana, E.; Rozowsky, J.; Alexander, R.; et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **2012**, *489* (7414), 91–100.

(48) Siggers, T.; Gordân, R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.* **2014**, *42* (4), 2099–2111.

(49) Gerstberger, S.; Hafner, M.; Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **2014**, *15* (12), 829–845.

(50) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **2011**, *7* (1), 549.

(51) Vermeulen, M.; Déjardin, J. Locus-specific chromatin isolation. *Nat. Rev. Mol. Cell Biol.* **2020**, *21* (5), 249–250.

(52) Gauchier, M.; van Mierlo, G.; Vermeulen, M.; Déjardin, J. Purification and enrichment of specific chromatin loci. *Nat. Methods* **2020**, *17* (4), 380–389.

(53) Gräwe, C.; Stelloo, S.; van Hout, F. A. H.; Vermeulen, M. RNA-centric methods: toward the interactome of specific RNA transcripts. *Trends Biotechnol.* **2021**, *39* (9), 890–900.

(54) Byrum, S. D.; Raman, A.; Taverna, S. D.; Tackett, A. J. ChAP-MS: a method for identification of proteins and histone posttranslational modifications at a single genomic locus. *Cell Rep.* **2012**, *2* (1), 198–205.

(55) Byrum, S. D.; Taverna, S. D.; Tackett, A. J. Purification of a specific native genomic locus for proteomic analysis. *Nucleic Acids Res.* **2013**, *41* (20), e195.

(56) Waldrip, Z. J.; Byrum, S. D.; Storey, A. J.; Gao, J.; Byrd, A. K.; Mackintosh, S. G.; Wahls, W. P.; Taverna, S. D.; Raney, K. D.; Tackett, A. J. A CRISPR-based approach for proteomic analysis of a single genomic locus. *Epigenetics* **2014**, *9* (9), 1207–1211.

(57) Déjardin, J.; Kingston, R. E. Purification of proteins associated with specific genomic loci. *Cell* **2009**, *136* (1), 175–186.

(58) Antão, J. M.; Mason, J. M.; Déjardin, J.; Kingston, R. E. Protein landscape at Drosophila melanogaster telomere-associated sequence repeats. *Mol. Cell. Biol.* **2012**, *32* (12), 2170–2182.

(59) Saksouk, N.; Barth, T. K.; Ziegler-Birling, C.; Olova, N.; Nowak, A.; Rey, E.; Mateos-Langerak, J.; Urbach, S.; Reik, W.; Torres-Padilla, M.-E.; et al. Redundant mechanisms to form silent chromatin at pericentromeric regions rely on BEND3 and DNA methylation. *Mol. Cell* **2014**, *56* (4), 580–594.

(60) Ide, S.; Dejardin, J. End-targeting proteomics of isolated chromatin segments of a mammalian ribosomal RNA gene promoter. *Nat. Commun.* **2015**, *6*, 6674.

(61) Gauchier, M.; Kan, S.; Barral, A.; Sauzet, S.; Agirre, E.; Bonnell, E.; Saksouk, N.; Barth, T. K.; Ide, S.; Urbach, S.; et al. SETDB1-dependent heterochromatin stimulates alternative lengthening of telomeres. *Sci. Adv.* **2019**, *5* (5), eaav3673.

(62) Qi, L. S.; Larson, M. H.; Gilbert, L. A.; Doudna, J. A.; Weissman, J. S.; Arkin, A. P.; Lim, W. A. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **2013**, *152* (5), 1173–1183.

(63) Fujita, T.; Fujii, H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem. Biophys. Res. Commun.* **2013**, *439* (1), 132–136.

(64) Liu, X.; Zhang, Y.; Chen, Y.; Li, M.; Zhou, F.; Li, K.; Cao, H.; Ni, M.; Liu, Y.; Gu, Z.; et al. In situ capture of chromatin interactions by biotinylated dCas9. *Cell* **2017**, *170* (5), 1028–1043.

(65) Gao, X. D.; Tu, L.-C.; Mir, A.; Rodriguez, T.; Ding, Y.; Leszyk, J.; Dekker, J.; Shaffer, S. A.; Zhu, L. J.; Wolfe, S. A.; et al. C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9–APEX2. *Nat. Methods* **2018**, *15* (6), 433–436.

(66) Myers, S. A.; Wright, J.; Peckner, R.; Kalish, B. T.; Zhang, F.; Carr, S. A. Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat. Methods* **2018**, *15* (6), 437–439.

(67) West, J. A.; Davis, C. P.; Sunwoo, H.; Simon, M. D.; Sadreyev, R. I.; Wang, P. I.; Tolstorukov, M. Y.; Kingston, R. E. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* **2014**, *55* (5), 791–802.

(68) Chu, C.; Zhang, Q. C.; da Rocha, S. T.; Flynn, R. A.; Bharadwaj, M.; Calabrese, J. M.; Magnuson, T.; Heard, E.; Chang, H. Y. Systematic discovery of Xist RNA binding proteins. *Cell* **2015**, *161* (2), 404–416.

(69) McHugh, C. A.; Guttman, M. RAP-MS: a method to identify proteins that interact directly with a specific RNA molecule in cells. In *RNA Detection: Methods and Protocols*, Methods in Molecular Biology, vol. 1649; Gaspar, I., Ed.; Humana Press: New York, NY, 2018; pp 473–488.

(70) Kennedy-Darling, J.; Holden, M. T.; Shortreed, M. R.; Smith, L. M. Multiplexed programmable release of captured DNA. *ChemBioChem* **2014**, *15* (16), 2353–2356.

(71) Knoener, R.; Evans III, E.; Becker, J. T.; Scalf, M.; Benner, B.; Sherer, N. M.; Smith, L. M. Identification of host proteins differentially associated with HIV-1 RNA splice variants. *eLife* **2021**, *10*, e62470.

(72) Kennedy-Darling, J.; Guillen-Ahlers, H.; Shortreed, M. R.; Scalf, M.; Frey, B. L.; Kendziorski, C.; Olivier, M.; Gasch, A. P.; Smith, L. M. Discovery of chromatin-associated proteins via sequence-specific capture and mass spectrometric protein identification in Saccharomyces cerevisiae. *J. Proteome Res.* **2014**, *13* (8), 3810–3825.

(73) Guillen-Ahlers, H.; Rao, P. K.; Levenstein, M. E.; Kennedy-Darling, J.; Perumalla, D. S.; Jadhav, A. Y. L.; Glenn, J. P.; Ludwig-Kubinski, A.; Drigalenko, E.; Montoya, M. J.; et al. HyCCAPP as a tool to characterize promoter DNA-protein interactions in Saccharomyces cerevisiae. *Genomics* **2016**, *107* (6), 267–273.

(74) Dai, Y.; Kennedy-Darling, J.; Shortreed, M. R.; Scalf, M.; Gasch, A. P.; Smith, L. M. Multiplexed sequence-specific capture of chromatin and mass spectrometric discovery of associated proteins. *Anal. Chem.* **2017**, *89* (15), 7841–7846.

(75) Buxton, K. E.; Kennedy-Darling, J.; Shortreed, M. R.; Zaidan, N. Z.; Olivier, M.; Scalf, M.; Sridharan, R.; Smith, L. M. Elucidating protein-DNA interactions in human alphoid chromatin via hybridization capture and mass spectrometry. *J. Proteome Res.* **2017**, *16* (9), 3433–3442.

(76) Knoener, R. A.; Becker, J. T.; Scalf, M.; Sherer, N. M.; Smith, L. M. Elucidating the in vivo interactome of HIV-1 RNA by hybridization capture and mass spectrometry. *Sci. Rep.* **2017**, *7* (1), 16965.

(77) Spiniello, M.; Knoener, R. A.; Steinbrink, M. I.; Yang, B.; Cesnik, A. J.; Buxton, K. E.; Scalf, M.; Jarrard, D. F.; Smith, L. M. HyPR-MS for multiplexed discovery of MALAT1, NEAT1, and NORAD lncRNA protein interactomes. *J. Proteome Res.* **2018**, *17* (9), 3022–3038.

(78) Spiniello, M.; Steinbrink, M. I.; Cesnik, A. J.; Miller, R. M.; Scalf, M.; Shortreed, M. R.; Smith, L. M. Comprehensive in vivo identification of the c-Myc mRNA interactome using HyPR-MS. *RNA* **2019**, *25* (10), 1337–1352.

(79) Henke, K. B.; Miller, R. M.; Knoener, R. A.; Scalf, M.; Spiniello, M.; Smith, L. M. Identifying protein interactomes of target RNAs using HyPR-MS. In *Post-Transcriptional Gene Regulation*, 3rd ed., Methods in Molecular Biology, vol. 2404; Dassi, E., Ed.; Humana Press: New York, NY, 2022; pp 219–244.

(80) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; et al. How many human proteoforms are there? *Nat. Chem. Biol.* **2018**, *14* (3), 206–214.

(81) Smith, L. M.; Kelleher, N. L.; The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10* (3), 186–187.

(82) Shortreed, M. R.; Frey, B. L.; Scalf, M.; Knoener, R. A.; Cesnik, A. J.; Smith, L. M. Elucidating proteoform families from proteoform intact-mass and lysine-count measurements. *J. Proteome Res.* **2016**, *15* (4), 1213–1221.

(83) Jenuwein, T.; Allis, C. D. Translating the histone code. *Science* **2001**, *293* (5532), 1074–1080.

(84) Mylona, A.; Theillet, F.-X.; Foster, C.; Cheng, T. M.; Miralles, F.; Bates, P. A.; Selenko, P.;

Treisman, R. Opposing effects of Elk-1 multisite phosphorylation shape its response to ERK activation. *Science* **2016**, *354* (6309), 233–237.

(85) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates III, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394.

(86) Gillet, L. C.; Leitner, A.; Aebersold, R. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annu. Rev. Anal. Chem.* **2016**, *9* (1), 449–472.

(87) Chait, B. T. Mass spectrometry: bottom-up or top-down? *Science* **2006**, *314* (5796), 65–66.

(88) Schaffer, L. V; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; et al. Identification and quantification of proteoforms by mass spectrometry. *Proteomics* **2019**, *19* (10), 1800361.

(89) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693.

(90) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* **2016**, *9* (1), 499–519.

(91) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-down proteomics: ready for prime time? *Anal. Chem.* **2018**, *90* (1), 110–127.

(92) Donnelly, D. P.; Rawlins, C. M.; DeHart, C. J.; Fornelli, L.; Schachner, L. F.; Lin, Z.; Lippens, J. L.; Aluri, K. C.; Sarin, R.; Chen, B.; et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* **2019**, *16* (7), 587–594.

(93) Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L. On the scalability and requirements of whole protein mass spectrometry. *Anal. Chem.* **2011**, *83* (17), 6868–6874.

(94) Millea, K. M.; Krull, I. S.; Cohen, S. A.; Gebler, J. C.; Berger, S. J. Integration of multidimensional chromatographic protein separations with a combined "top-down" and "bottom-up" proteomic strategy. *J. Proteome Res.* **2006**, *5* (1), 135–146.

(95) Jefferys, S. R.; Giddings, M. C. Baking a mass-spectrometry data PIE with McMC and simulated annealing: predicting protein post-translational modifications from integrated top-down and bottom-up data. *Bioinformatics* **2011**, *27* (6), 844–852.

(96) Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D.; et al. Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Mol. Cell. Proteomics* **2016**, *15* (1), 45–56.

(97) Dai, Y.; Shortreed, M. R.; Scalf, M.; Frey, B. L.; Cesnik, A. J.; Solntsev, S.; Schaffer, L. V; Smith,

L. M. Elucidating Escherichia coli proteoform families using intact-mass proteomics and a global PTM discovery database. *J. Proteome Res.* **2017**, *16* (11), 4156–4165.

(98) Schaffer, L. V; Shortreed, M. R.; Cesnik, A. J.; Frey, B. L.; Solntsev, S. K.; Scalf, M.; Smith, L. M. Expanding proteoform identifications in top-down proteomic analyses by constructing proteoform families. *Anal. Chem.* **2018**, *90* (2), 1325–1333.

(99) Schaffer, L. V; Rensvold, J. W.; Shortreed, M. R.; Cesnik, A. J.; Jochem, A.; Scalf, M.; Frey, B. L.; Pagliarini, D. J.; Smith, L. M. Identification and quantification of murine mitochondrial proteoforms using an integrated top-down and intact-mass strategy. *J. Proteome Res.* **2018**, *17* (10), 3526–3536.

(100) Schaffer, L. V; Millikin, R. J.; Shortreed, M. R.; Scalf, M.; Smith, L. M. Improving proteoform identifications in complex systems through integration of bottom-up and top-down data. *J. Proteome Res.* **2020**, *19* (8), 3510–3517.

(101) Schaffer, L. V; Anderson, L. C.; Butcher, D. S.; Shortreed, M. R.; Miller, R. M.; Pavelec, C.; Smith, L. M. Construction of human proteoform families from 21 Tesla Fourier transform ion cyclotron resonance mass spectrometry top-down proteomic data. *J. Proteome Res.* **2021**, *20* (1), 317–325.

(102) Dai, Y.; Buxton, K. E.; Schaffer, L. V; Miller, R. M.; Millikin, R. J.; Scalf, M.; Frey, B. L.; Shortreed, M. R.; Smith, L. M. Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. *J. Proteome Res.* **2019**, *18* (10), 3671–3680.

CHAPTER 2

ELUCIDATING PROTEIN-DNA INTERACTIONS IN HUMAN ALPHOID CHROMATIN VIA HYBRIDIZATION CAPTURE AND MASS SPECTROMETRY

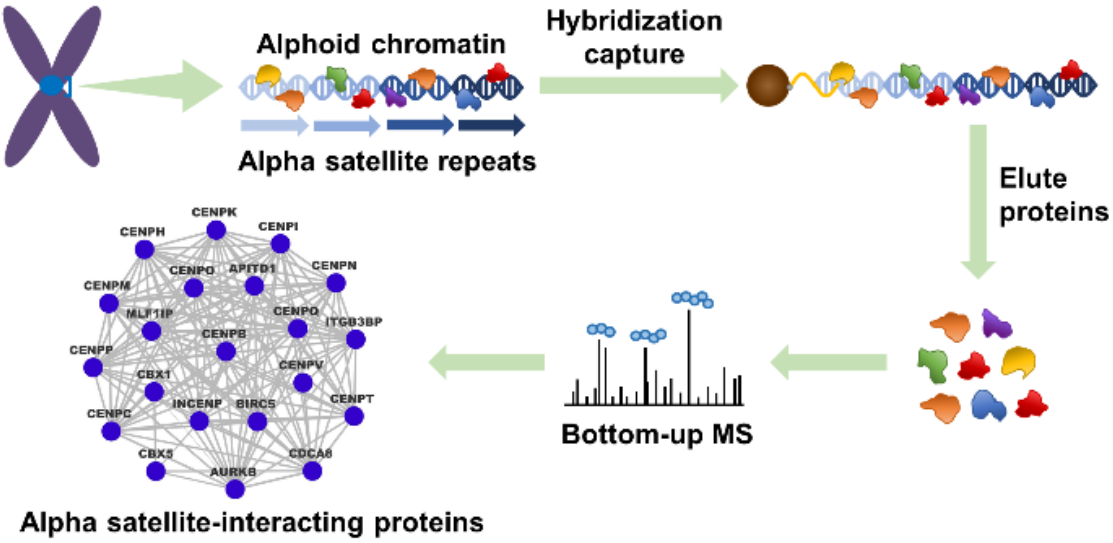
This chapter has been adapted from a publication and was reprinted with permission from:

Buxton, K. E.; Kennedy-Darling, J.; Shortreed, M. R.; Zaidan, N. Z.; Olivier, M.; Scalf, M.; Sridharan, R.; Smith, L. M. Elucidating protein-DNA interactions in human alphoid chromatin via hybridization capture and mass spectrometry. *J. Proteome Res.* **2017**, *16* (9), 3433-3442. https://doi.org/10.1021/acs.jproteome.7b00448.

2.1 ABSTRACT

The centromere is the chromosomal locus where the kinetochore forms and is critical for ensuring proper segregation of sister chromatids during cell division. A substantial amount of effort has been devoted to understanding the characteristic features and roles of the centromere, yet some fundamental aspects of the centromere, such as the complete list of elements that define it, remain obscure. It is well-known that human centromeres include a highly repetitive class of DNA known as alpha satellite, or alphoid, DNA. We present here the first DNA-centric examination of human protein-alpha satellite interactions, employing an approach known as HyCCAPP (hybridization

capture of chromatin-associated proteins for proteomics) to identify the protein components of alphoid chromatin in a human cell line. Using HyCCAPP, cross-linked alpha satellite chromatin was isolated from cell lysate, and captured proteins were analyzed via mass spectrometry. After being compared to proteins identified in control pulldown experiments, 90 proteins were identified as enriched at alphoid DNA. This list included many known centromere-binding proteins in addition to multiple novel alpha satellite-binding proteins, such as LRIF1, a heterochromatin-associated protein. The ability of HyCCAPP to reveal both known as well as novel alphoid DNA-interacting proteins highlights the validity and utility of this approach.



2.2 INTRODUCTION

The centromere is the chromosomal region that is responsible for linking replicated sister chromatids, appearing as the primary constriction in condensed, mitotic chromosomes. The centromere is also where the kinetochore, the proteinaceous structure that serves as the attachment point for spindle microtubules, forms and is therefore critical for ensuring proper chromosomal segregation during cell division. Human centromeres and their surrounding pericentromeric regions include a highly repetitive class of DNA known as alpha satellite, or alphoid, DNA.[1] The repeating unit of alpha satellite DNA is 171 base pairs in length, and these monomeric units are arranged tandemly in a head-to-tail fashion within the centromeric region.[2] An integral number of repeated alpha satellite monomers make up a structure known as a higher-order repeat (HOR), which is itself tandemly repeated so that the alpha satellite array may extend for ~0.2-5 Mb (Figure 2.1).[2] Because this large block of alpha satellite DNA is found within the centromeric region of every chromosome, it is estimated that alphoid DNA makes up as much as ~3-5% of the human genome as a whole.[1] Individual alpha satellite monomers are relatively heterogeneous in nature, varying in sequence by as much as ~30-50%, but the HORs that make up an array show much more sequence similarity, varying by less than ~2%.[2,3] Flanking the HOR array, the pericentromeric region contains disordered alpha satellite monomers, which become more heterogeneous in sequence further away from the centromeric core and gradually lead into nonalphoid genomic DNA.[2,4]

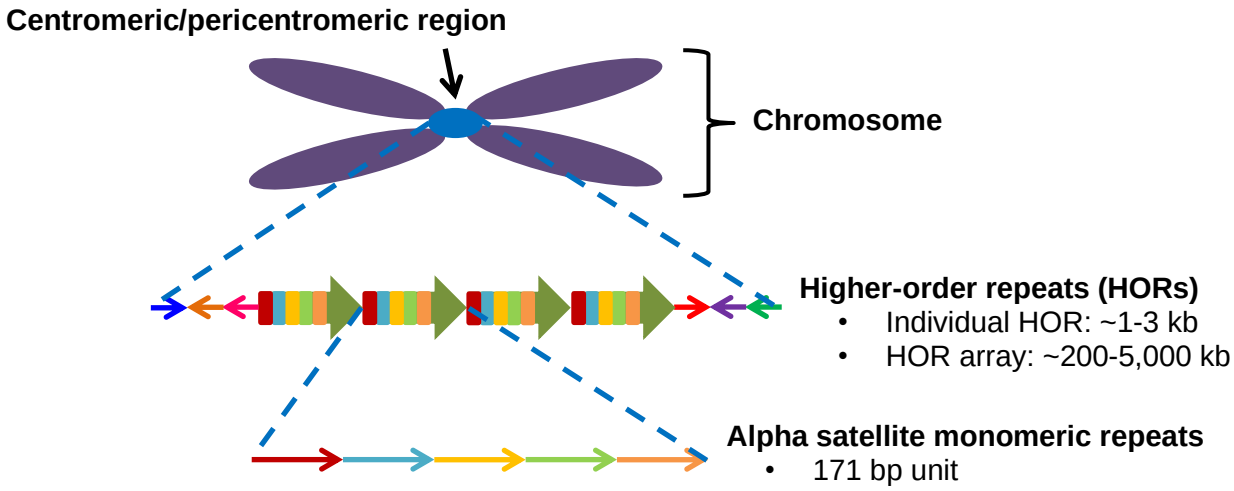


Figure 2.1 Alpha satellite repeats are found in the centromeric/pericentromeric region of every human chromosome. Individual alpha satellite monomers (depicted here as small arrows) are 171 bp in length and are relatively heterogeneous in sequence (heterogeneity is represented by different colors). These monomers are arranged in a tandem, head-to-tail fashion within the centromere, forming higher-order repeat (HOR) structures (depicted here as large arrows). HORs show much more sequence homogeneity than their individual monomeric constituents and are themselves repeated to form large alpha satellite arrays within the centromeric region. The HOR array is flanked by a pericentromeric region containing disordered alpha satellite monomers. HOR and HOR array size estimates for this figure are from Aldrup-MacDonald and Sullivan.[2]

Highly repetitive centromeric DNA is not unique to human cells. In fact, with few exceptions including the point centromeres of budding yeast and the holocentric chromosomes of *Caenorhabditis elegans*, repetitive centromeric sequences have been identified in most other animal, plant, and fungal species studied to date.[3,5] This suggests that such repetitive sequences may be critical to centromere identity or kinetochore function, a hypothesis supported by the finding that both alpha satellite DNA and the CENP-B box, a 17 bp sequence found within a subset of alpha satellite repeats to which centromere protein B binds,[6] are required for the *de novo* formation of functional human artificial

chromosomes (HACs).[7] These results suggest that alpha satellite DNA may have intrinsic properties without which a functional human centromere cannot form.

However, the conclusion that repetitive DNA is required for normal centromere function is called into question when one considers the phenomenon of neocentromeres. Neocentromeres, first identified by Voullaire et al. in 1993,[8] are novel, ectopic centromeres that may form after the disruption or inactivation of a natural centromere.[9] Unlike natural centromeres, neocentromeres form at diverse, nonalphoid genomic loci.[9,10] These neocentromeres appear to be functional, supporting chromosomal segregation during cell division,[8,10] which refutes the idea that alpha satellite DNA is required for centromere formation.

Together, the observations outlined above raise questions about what role, if any, alpha satellite DNA actually plays in centromeric identity or function. The discovery of neocentromeres at nonalphoid sequences indicates that the human centromere is not defined solely by a particular DNA sequence, suggesting that the specification of the centromeric locus may involve other factors.

Much work has already been completed to shed light on the nongenetic elements, particularly proteins and associated post-translational modifications (PTMs), that may be important for centromere specification.[4,11-13] Centromere protein A, or CENP-A, is just one example of a protein that appears to be important for centromere function. CENP-A is a histone H3 variant that is found in some centromeric nucleosomes,[14-16] and homologues of CENP-A have been shown to be important for viability in budding yeast, mice, and *Drosophila*.[17-19] Furthermore, extensive affinity purification

experiments in human HeLa cells have revealed that CENP-A nucleosomes recruit a set of six proteins (CENP-C, -H, -M, -N, -T, and -U), collectively referred to as the CENP-A nucleosome-associated complex (NAC).[20] The CENP-A NAC, in turn, recruits a set of seven additional proteins (CENP-K, -L, -O, -P, -Q, -R, and -S), which are collectively referred to as the CENP-A distal (CAD) complex, and it has been observed that the disruption of CENP-A NAC formation results in mitotic errors.[20] However, despite its clearly important role, it does not appear that the presence of CENP-A alone is sufficient for centromere specification, as CENP-A nucleosomes are also found at ectopic sites, where functional centromeres do not form.[21]

CENP-A and its associated NAC and CAD proteins are just a few examples of the many players that have been identified as important for centromere function and kinetochore recruitment. Additionally, the transcriptional status of the centromeric region appears to be critical for proper centromere and kinetochore performance, requiring a balance between open, euchromatic chromatin and closed, heterochromatic chromatin.[22] Still, with all of this information we do not yet have conclusive answers to questions about why human centromeres almost invariably form on repetitive, alpha satellite DNA, when other genomic loci are clearly also capable of performing this task.

In this study, we seek to further investigate the protein components of alpha satellite chromatin to reveal novel protein-alpha satellite interactions that may be important for centromere specification or function. We have used an approach known as HyCCAPP, or hybridization capture of chromatin-associated proteins for proteomics.[23,24] Briefly, human K562 cells are cross-linked using formaldehyde

to fix protein-DNA interactions, then cells are lysed and chromatin is sheared. The lysate is incubated with biotinylated capture oligonucleotides, whose sequences are designed to complement the human alpha satellite consensus sequence,[25] and hybridized complexes are isolated using streptavidin-coated magnetic beads. Following washing steps, the hybridized complexes are released from the beads, and associated proteins are identified via bottom-up mass spectrometry.[26] The HyCCAPP approach is distinct from other strategies commonly used to study protein-DNA interactions, such as affinity purification or immunoprecipitation, in that it is DNA- rather than protein-centric. HyCCAPP has proven to be a useful tool, revealing both known as well as novel protein-DNA interactions in yeast.[23,24] Here, we apply HyCCAPP to the human genome for the first time, providing a DNA-centric lens through which to view protein-alphoid DNA interactions. Using this approach, 90 proteins were identified as enriched at the alpha satellite repeats, including many known centromere-binding proteins in addition to a number of novel alpha satellite-binding proteins, which may provide new insights into centromere structure or function.

2.3 METHODS

A detailed account of all experimental procedures employed in this work can be found in the Supplementary Information. Brief summaries of these procedures are provided here.

HyCCAPP

The HyCCAPP procedure was performed essentially as described previously[23] with relatively minor adjustments. Human K562 cells were cross-linked for 30 min at room temperature via gentle shaking with a final concentration of 3% (w/v) formaldehyde, and excess formaldehyde was quenched with the addition of Tris buffer pH = 8. Cells were washed, pelleted, and stored at -45°C for up to one month prior to use. For each HyCCAPP experiment, ~3.3×10^9 cells were resuspended in lysis buffer containing 200 mM NaCl, 20 mM EDTA pH = 8, 50 mM Tris pH = 7, and protease inhibitors diluted 200× from a concentrated cocktail (Sigma-Aldrich #P8340). The cells were lysed using a Constant Systems TS Series cell disruptor at 18 kpsi, and then SDS was added to a final concentration of 1% and the lysate was heated in a 65°C water bath for 8 min to aid in chromatin solubilization. The chromatin was sheared via sonication to a median fragment size of ~3 kb (Misonix Ultrasonic Processor S4000) and the lysate was cleared of insoluble cellular debris via centrifugation. The supernatant was diluted five-fold with lysis buffer and RNA was digested via incubation with RNase A for 1 h at 37°C. The lysate was centrifuged again to remove any remaining particulates, and streptavidin-coated magnetic beads (Fisher Scientific #09-981-140) were added to the supernatant to clear endogenously biotinylated moieties. The lysate was incubated with the beads for 1 h at room temperature, after which the beads were removed and discarded. At this point, a 1 mL aliquot of the lysate was removed and stored at -20°C for eventual qPCR analysis.

Next, biotinylated alpha satellite capture oligonucleotides were added to the lysate. Eight capture oligonucleotides were designed to complement the human alpha satellite consensus sequence[25] and added at equimolar concentrations. As a control, a biotinylated scrambled sequence capture oligonucleotide was also added to the lysate at this point. The sequences of all nine capture oligonucleotides used in HyCCAPP experiments are listed in Supplementary Table S-2.1. The lysate was incubated with the oligonucleotides for 3 h at 37°C, then allowed to cool to room temperature. To capture the hybridized complexes, streptavidin-coated magnetic beads were added and the lysate was incubated at room temperature for 1 h. The beads were isolated from the lysate using a magnet stand.

The beads were washed four times for 5 min each with wash buffer containing 50 mM Tris pH = 8, 200 mM NaCl, and 0.2% SDS. The beads were then concentrated into a smaller volume and washed once for 5 min and once for 1 h. To release the hybridized complexes, a toehold-mediated strand displacement strategy was employed (see Supplementary Information for details). This strategy was developed and demonstrated for the selective release of individual target loci from cross-linked yeast chromatin,[27,28] and here we extended the strategy to cross-linked human chromatin. First, alpha satellite release solution containing equimolar concentrations of the eight alpha satellite release oligonucleotides was added to the beads, and the resultant bead slurry was gently mixed for 15 min at room temperature. The release solution was then removed and stored at 4°C, and the beads were washed twice for 5 min each. Next, scrambled release solution containing the scrambled release

oligonucleotide was added to the beads, and the slurry was again gently mixed for 15 min at room temperature. The scrambled release solution was removed and stored at 4°C, and the beads were discarded. The sequences of all nine release oligonucleotides used in HyCCAPP experiments are listed in Supplementary Table S-2.1.

After removing a small aliquot of each release solution for eventual qPCR analysis, the proteins in both the alpha satellite and scrambled release solutions were precipitated using cold trichloroacetic acid (TCA). An aliquot of lysate material removed after hybridization capture was also precipitated at this time. The three pellets were resuspended in buffer containing 8 M urea and 0.1% deoxycholic acid, and the proteins were prepared for mass spectrometric analysis using the eFASP method.[29] Note that an explicit cross-link reversal step was not employed during the preparation of peptides for mass spectrometric analysis as, given the known rates of spontaneous formaldehyde cross-link reversal,[30] such a step was not deemed necessary. After removal of surfactant via extraction with ethyl acetate, the peptides in the aqueous phase were dried down and resuspended in 0.1% trifluoroacetic acid. The samples were then desalted using C18 solid-phase extraction pipet tips (Agilent Technologies #A57003100K), and the desalted peptides were dried down and resuspended in 95:5 water/acetonitrile with 0.2% formic acid.

Mass spectrometric analysis of HyCCAPP samples

Three replicates of the HyCCAPP experiment were carried out as described. Samples were analyzed via HPLC-ESI-MS/MS using a system consisting of a high performance liquid chromatograph (nanoAcquity, Waters) connected to an electrospray ionization (ESI) orbitrap mass spectrometer (LTQ Velos, Thermo Fisher Scientific). HPLC separation employed a 100 × 365 µm fused silica capillary microcolumn packed with 20 cm of 1.7 µm diameter, 130 Å pore size C18 beads (Waters BEH), with an emitter tip pulled to approximately 1 µm using a laser puller (Sutter instruments). Peptides were loaded on-column at a flow-rate of 400 nL/min for 30 min, then eluted over 125 min at a flow-rate of 300 nL/min with a gradient from 2% to 30% acetonitrile in 0.1% formic acid. Full-mass profile scans were performed in the orbitrap between 300 and 1,500 *m/z* at a resolution of 60,000, followed by MS/MS HCD scans of the ten highest intensity parent ions with $z > 1$ at 42% relative collision energy and 7,500 resolution, with a mass range starting at 100 *m/z*. Dynamic exclusion was enabled with a repeat count of two over a duration of 30 s and an exclusion window of 120 s. All raw data files are available from the MassIVE repository under the name "AlphaSatelliteHyCCAPP".

The raw mass spectrometric data were analyzed using MaxQuant (version 1.5.3.30).[31] The detected features were searched against the human canonical protein database from UniProt (downloaded on April 22, 2016) and a list of potential contaminants. A false discovery rate (FDR) of 1% at both the peptide and protein levels was allowed. Relative quantification was performed for each capture sample (alpha satellite and scrambled) of each HyCCAPP experiment (1-3) using the protein

intensities reported in the "proteinGroups.txt" file produced by MaxQuant. Individual protein intensities were normalized to the median protein intensity of the corresponding capture sample after removing potential contaminants, and these normalized protein intensities were loaded into Perseus (version 1.5.3.2)[32] for downstream statistical analysis. Two-sided Welch's *t* tests were performed to compare protein abundances between groups (alpha satellite capture versus scrambled capture). The statistical tests were corrected for multiple hypothesis testing using a permutation-based FDR cutoff of 5% (S0 = 2) (see Supplementary Table S-2.2 for output of Perseus analysis).

Quantitative real-time PCR analysis of HyCCAPP samples

Quantitative real-time PCR (qPCR) was performed to assess the efficiency and specificity with which the alpha satellite repeats were captured in HyCCAPP experiments. Because of the heterogeneity of alpha satellite monomeric sequences, it was not feasible to design qPCR assays to target every individual alpha satellite monomer. Instead, a TaqMan assay was designed to target a single alpha satellite sequence, and the capture efficiency of this sequence was used as a proxy for the capture efficiency of alpha satellite DNA as a whole. Additionally, a qPCR assay for an off-target sequence (28S rDNA) was designed to assess capture specificity. These assays were ordered from IDT, and the sequences of the primers and probes are listed in Supplementary Table S-2.3. qPCR was performed essentially as described previously[23] to measure the concentrations of the target alpha satellite sequence and the off-target 28S rDNA sequence in the alpha satellite release, scrambled release, and lysate (input) samples. Capture efficiency was then calculated by dividing the amount of alpha

satellite DNA or 28S rDNA present in the release solution of interest by the amount of that sequence present in the lysate sample (prior to capture).

Chromatin immunoprecipitation with qPCR (ChIP-qPCR)

Human K562 cells were cross-linked for 10 min at room temperature via gentle shaking with a final concentration of 1% (w/v) formaldehyde, and excess formaldehyde was quenched with the addition of Tris buffer pH = 8. Cells were washed, pelleted, and stored at -80°C prior to use. To prepare chromatin for ChIP experiments, cells were resuspended in ice-cold buffer containing 50 mM HEPES pH = 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 substitute (Amresco #M158), 0.25% Triton X-100, and protease/phosphatase inhibitors diluted 100× from a concentrated cocktail (Thermo Fisher Scientific #1861281). The resuspended cells were incubated on ice for 20 min, and the nuclear pellet was collected via centrifugation and washed. The pellet was resuspended at a final concentration of ~7.4×10^6 nuclei/mL in ice-cold shearing buffer containing 0.1% SDS, 1 mM EDTA, 10 mM Tris pH = 8.1, and protease/phosphatase inhibitors, and the chromatin was sheared to ~200-500 bp using a Covaris S220 water bath sonicator. The sheared chromatin was centrifuged to remove insoluble debris, and the supernatant was aliquoted into 10-µg and 1-µg portions (based on Nanodrop measurement of DNA concentration) and stored with 10% glycerol at -80°C for up to two months prior to use.

For each ChIP experiment, a 10-µg aliquot of chromatin was thawed on ice and diluted 10-fold in ice-cold buffer containing 16.7 mM Tris pH = 8, 0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 333 mM NaCl, and protease/phosphatase inhibitors. Then, 2 µg of either CENP-B antibody (abcam

#ab25734), LRIF1 antibody (Atlas Antibodies #HPA044515), or IgG (Jackson ImmunoResearch #312-005-003) were added to the chromatin and incubated overnight at 4°C with gentle mixing. Next, 10 µL each of protein A- and protein G-coated magnetic beads (Thermo Fisher Scientific #10002D and #10004D) were washed, then added to the sample and gently mixed for 2 h at 4°C. The beads were isolated using a magnet stand and the supernatant was discarded. The beads were then washed thoroughly to remove nonspecifically-bound proteins/DNA, and the captured chromatin was eluted from the beads by heating at 65°C for 10 min in 100 µL of elution buffer containing 50 mM Tris pH = 8, 1 mM EDTA, 1% SDS, and protease/phosphatase inhibitors. Next, 80 µL of TE buffer containing 0.67% SDS and protease/phosphatase inhibitors were added to the sample, and the sample was heated overnight at 65°C. At this time, a 1-µg aliquot of chromatin ("10% input sample") was thawed on ice and diluted with 100 µL of elution buffer and 80 µL of TE buffer containing 0.67% SDS and protease/phosphatase inhibitors and heated overnight at 65°C.

The ChIP and 10% input samples were treated with RNase A and proteinase K, and the beads were removed from the ChIP sample using a magnet stand. The DNA in both samples was then purified via phenol-chloroform extraction and ethanol precipitation, and the DNA pellets were resuspended in 10 mM Tris pH = 7. The DNA was then analyzed via qPCR using two assays designed to target alpha satellite DNA and one assay designed to target 28S rDNA (negative control sequence). These assays were ordered from IDT, and the sequences of the primers and probes are listed in Supplementary Table S-2.3. For each assay, a standard curve was generated by preparing serial dilutions

of the 10% input sample, and each standard/ChIP sample was analyzed in duplicate on a 96-well plate using a Roche 480 LightCycler.

2.4 RESULTS AND DISCUSSION

Evaluation of hybridization capture efficiency and specificity

Quantitative real-time PCR was employed to evaluate the efficiency and specificity of hybridization capture in HyCCAPP. Here, capture efficiency was defined as the percentage of input alpha satellite DNA present in the final capture sample. Using a qPCR assay designed to target an alpha satellite sequence, the efficiency with which alphoid DNA was captured in the alpha satellite pulldown experiments was determined to be ~0.22%, which is comparable to capture efficiencies reported previously for HyCCAPP in yeast.[23] The efficiency with which alphoid DNA was captured in the scrambled pulldown experiments was approximately 44-times lower, or ~0.0049% (Figure 2.2). In small-scale HyCCAPP experiments, the effect of release order (alpha satellite first followed scrambled or vice versa) on alpha satellite capture efficiency was evaluated, and no dramatic differences were apparent (Supplementary Figure S-2.1).

To evaluate the specificity of capture, the capture efficiency of an off-target sequence (28S rDNA) was also calculated for each sample. Neither the alpha satellite nor the scrambled capture oligonucleotides captured the off-target sequence with appreciable efficiency (Figure 2.2).

Additionally, the enrichment of alpha satellite DNA after hybridization capture was determined by comparing the ratio of alpha satellite DNA to 28S rDNA in whole cell lysate (precapture) to the ratio in the alpha satellite capture sample. This gave an average enrichment factor of ~140-fold, which indicated that the hybridization capture step effectively enriched alphoid DNA.

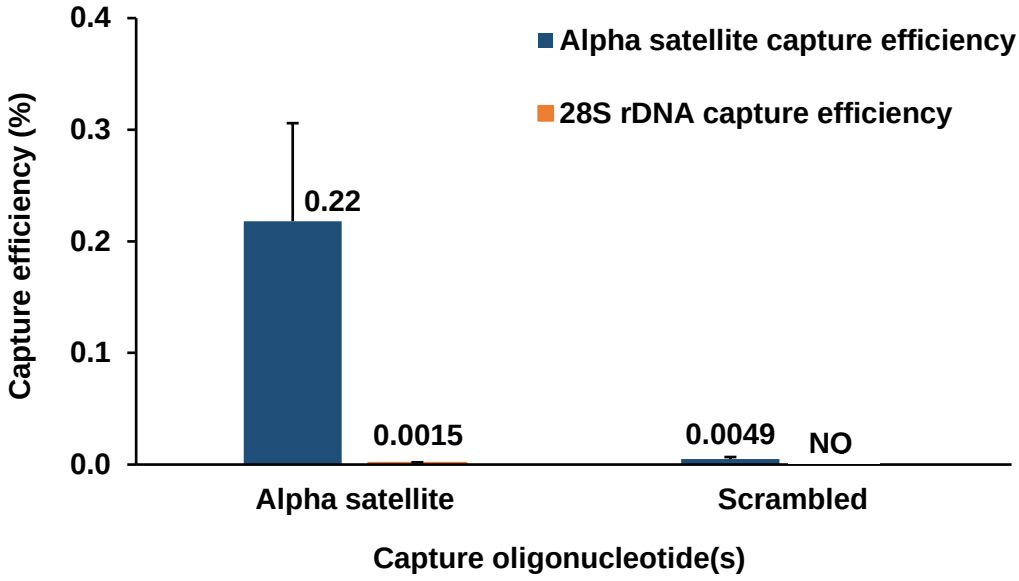


Figure 2.2 qPCR was used to measure the capture efficiencies of alphoid DNA and an off-target sequence (28S rDNA) in both the alpha satellite capture and scrambled capture samples. These measurements were made for all three HyCCAPP experiments, and the labels denote average values. Error bars represent +1 standard deviation of the three experimental replicates. Note that the average capture efficiency of 28S rDNA in the scrambled capture samples could not be calculated because the concentration of 28S rDNA in scrambled capture samples was too low to be accurately quantified (NQ = not quantified).

The success of a HyCCAPP experiment hinges on its ability to specifically isolate the target genomic locus from cross-linked chromatin in sufficient quantities that its associated proteins are

detectable via mass spectrometry. Because of the exceedingly high copy number of alpha satellite DNA, a capture efficiency of ~0.22% provided enough material to meet this criterion, allowing for the detection of locus-specific proteins by standard bottom-up proteomics. Given that we expect capture efficiency in HyCCAPP to be related to the presence of single-stranded regions within the input chromatin, it is interesting to consider the mechanism(s) by which single-stranded regions are available in the first place. An earlier generation of the HyCCAPP technology, known as GENECAPP (global exonuclease-based enrichment of chromatin-associated proteins for proteomics), employed exonuclease treatment to produce single-stranded regions amenable to capture. This approach was successful in *in vitro* studies on a model system,[33] but transitioning to the capture of chromatin fragments from whole cell lysate proved challenging. Initially, a strategy involving restriction enzyme digestion followed by exonuclease treatment was employed to generate single-stranded regions in a sequence-specific manner. However, the restriction enzyme digestion step proved inefficient in the context of whole cell lysate, and sonication was instead selected as the means of chromatin fragmentation. It was assumed that exonuclease digestion would still be necessary to generate single-stranded regions in the chromatin, but control experiments showed that hybridization occurred equally well either with or without exonuclease treatment. This observation led to the development of HyCCAPP, which does not rely on enzymatic digestions, thereby reducing the cost and complexity of the procedure. Still, the mechanism by which chromatin is available for capture in HyCCAPP is not yet completely clear. One hypothesis is that the sonication step employed early in the HyCCAPP

experiment serves not only to shear DNA into smaller pieces, but also to introduce single-stranded regions[34,35] and overhanging ends, which could be amenable to hybridization capture. Additional work will be required to fully understand the mechanism(s) by which single-stranded regions of chromatin are available in HyCCAPP and to take advantage of these mechanisms to increase capture efficiency.

Identification of proteins enriched at the alpha satellite repeats

Identification and extracted ion chromatogram-based label-free quantification of the proteins present in the alpha satellite capture and scrambled capture samples were performed using MaxQuant software. Perseus software was then used to perform *t* tests to determine which proteins were enriched in the alpha satellite capture samples as compared to the scrambled capture samples. Comparison to the scrambled oligonucleotide pulldown allows one to control for background proteins which may be carried through to the final alpha satellite capture sample, perhaps due to their high abundance or through nonspecific binding to the beads, as these proteins are presumably present in the alpha satellite and scrambled capture samples in similar amounts. The comparison between the alpha satellite capture and scrambled capture samples yielded 90 proteins which were significantly enriched at the alpha satellite repeats, with an FDR of 5% and a minimum fold-change of 5.8 (Figure 2.3). This list of 90 proteins will henceforth be referred to simply as "alpha satellite-enriched proteins" (a complete list of these proteins can be found in Supplementary Table S-2.4).

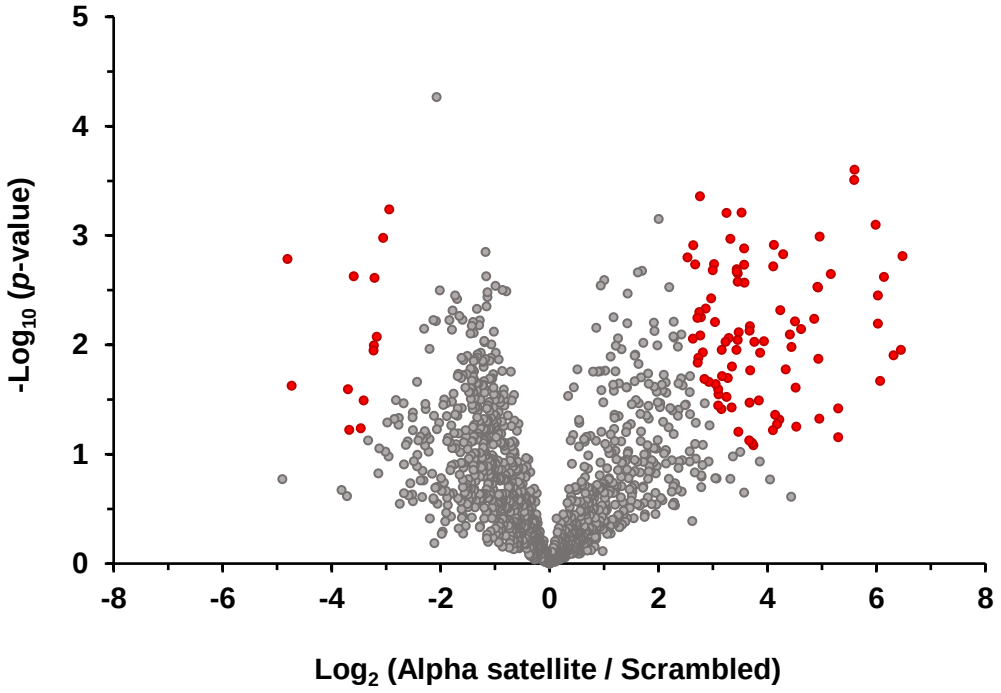


Figure 2.3 Volcano plot showing the results of *t* tests performed to compare protein abundances between the alpha satellite capture and scrambled capture samples. Data points represent individual proteins. Red points indicate those proteins that were found to be significantly enriched either in the alpha satellite or scrambled capture samples, with a permutation-based FDR ≤ 5% and a minimum fold-change of 5.8. Subject to these criteria, 90 proteins were found to be significantly enriched at the alpha satellite repeats. The data used to generate this volcano plot, including the identity of each protein and its associated *p*-value, permutation-based *q*-value, and fold-change, can be found in Supplementary Table S-2.2.

Analysis of the alpha satellite-enriched protein list

To ascertain whether the final list of 90 alpha satellite-enriched proteins was reasonable, the list was analyzed via the PANTHER over-representation test (released on July 15, 2016).[36,37] This test finds GO terms that are more abundant among input proteins than would be expected based on a random list of proteins from the reference *Homo sapiens* database. Three over-representation tests were

performed to find 65 enriched cellular component, 15 enriched molecular function, and 100 enriched biological process GO terms, all with Bonferroni-corrected *p*-values less than 0.05. The list of enriched GO terms included terms like "chromosome, centromeric region", "mitotic cell cycle", "sister chromatid cohesion", "kinetochore", "CENP-A containing chromatin organization", and "pericentric heterochromatin". These terms make sense based on what is known about the alpha satellite repeats such as their localization at the centromeric/pericentromeric region of human chromosomes and the centromere's relationship to the kinetochore and cell cycle. This GO term analysis thus served as an initial indication that the HyCCAPP experiments and subsequent data analysis steps were working as expected to identify true alpha satellite-interacting proteins. Table 2.1 lists the 30 enriched GO terms with the smallest corrected *p*-values (see Supplementary Table S-2.5 for a complete list of enriched GO terms). Because the HyCCAPP experiments were performed using asynchronous cells, we expect the set of enriched proteins to reflect activities of the centromeric region at various points of the cell cycle.

Table 2.1 Selection of GO terms over-represented in the list of 90 alpha satellite-enriched proteins identified by HyCCAPP. The PANTHER over-representation tests generated a list of 180 over-represented GO terms with Bonferroni-corrected *p*-values less than 0.05. This table shows the 30 GO terms from that list with corrected *p*-values less than 3.5×10^{-24}.

GO Term	*p*-value	GO Term	*p*-value
Chromosome	1.53×10^{-64}	Organelle organization	7.15×10^{-31}
Chromosomal region	7.25×10^{-57}	Cell cycle	1.12×10^{-30}
Chromosomal part	7.26×10^{-53}	Condensed chromosome	1.12×10^{-30}
Chromosome organization	6.39×10^{-48}	Nuclear chromosome	9.51×10^{-30}
Nuclear part	2.20×10^{-41}	Nucleus	3.21×10^{-28}
Nuclear lumen	1.26×10^{-38}	Sister chromatid segregation	4.01×10^{-28}
Intracellular non-membrane-bounded organelle	6.21×10^{-38}	Nuclear chromosome part	6.42×10^{-28}
Non-membrane-bounded organelle	6.21×10^{-38}	Protein-DNA complex assembly	2.72×10^{-26}
DNA conformation change	9.82×10^{-36}	Sister chromatid cohesion	4.60×10^{-26}
Chromosome, centromeric region	2.23×10^{-35}	Nuclear chromosome segregation	4.75×10^{-26}
Organelle lumen	8.22×10^{-34}	DNA packaging	9.19×10^{-26}
Intracellular organelle lumen	8.22×10^{-34}	Intracellular organelle part	4.28×10^{-25}
Membrane-enclosed lumen	8.22×10^{-34}	Protein-DNA complex subunit organization	8.32×10^{-25}
Cell cycle process	3.78×10^{-32}	Chromosome segregation	2.39×10^{-24}
Nucleoplasm	6.28×10^{-31}	Organelle part	3.04×10^{-24}

As another way to validate that the final list of alpha satellite-enriched proteins was nonrandom and reflected real associations existing within the cell, we used STRING (version 10.0)[38] to identify interactions between these 90 proteins (Figure 2.4). The resultant network consisted of 464 edges, with

an average of 10.3 interactions per protein. The expected number of edges for a random network of this size is 76, meaning that this collection of proteins is very likely nonrandom.

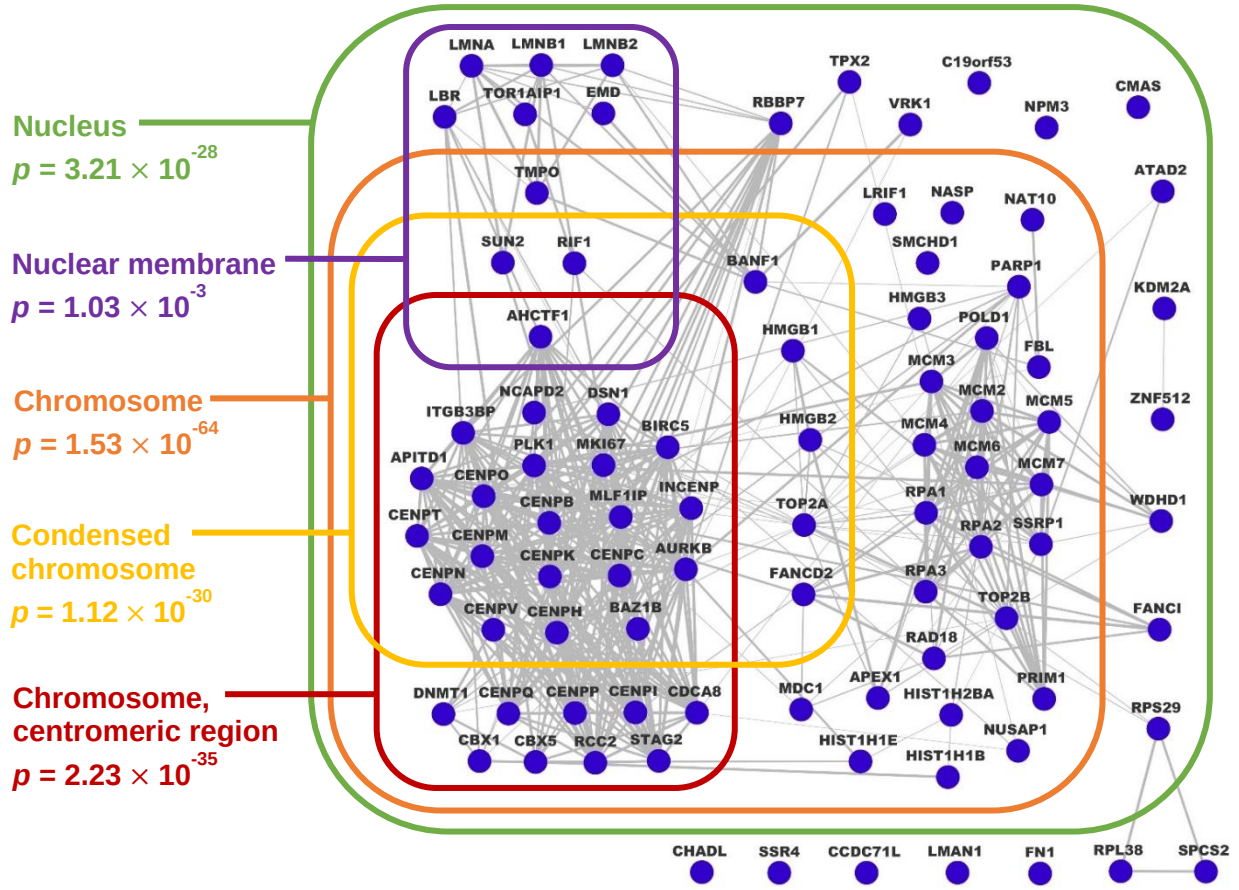


Figure 2.4 Interaction network for the 90 alpha satellite-enriched proteins identified through HyCCAPP experiments. The network was created in Cytoscape (version 3.5.1),[39] using interactions identified by STRING (version 10.0).[38] Textmining, experiments, and databases were used as sources of evidence. The network reflects interactions that STRING identified with at least medium confidence, and the thickness of the line connecting two nodes reflects the strength of the data supporting that interaction. Proteins with selected GO terms are outlined to aid in interpretation of the network, and associated Bonferroni-corrected *p*-values for GO term enrichment are indicated.

HyCCAPP identifies known centromeric proteins as enriched at the alpha satellite repeats

After the list of alpha satellite-enriched proteins was evaluated as a whole, individual proteins on the list were explored further. It was immediately obvious that HyCCAPP had identified a number of known alpha satellite-binding proteins as enriched at alphoid DNA (~30-50, depending on the criteria used for classification). For example, all six members of the CENP-A nucleosome-associated complex (CENP-C, -H, -M, -N, -T, and -U) and six out of seven members of the CENP-A distal complex (CENP-K, -O, -P, -Q, -R, and -S) were identified as enriched. Interestingly, despite the presence of all of these proteins which are, at least to some extent, recruited by CENP-A,[20] CENP-A itself was not identified by MaxQuant in any of the samples analyzed.

Although it is well-known that CENP-A is enriched at the centromere, it is not unreasonable that HyCCAPP did not identify this protein as one of the 90 alpha satellite-enriched proteins. It is important to understand that the HyCCAPP technology is fundamentally limited by the sensitivity of the mass spectrometric analysis. The detection limit for a given peptide in bottom-up proteomics is dependent on factors such as its solubility and ionization efficiency, and if none of the tryptic peptides corresponding to a protein are present at levels above their detection limits, the protein will not be detected. Additionally, even if a peptide is present at quantities above its detection limit, it still may not be detected if it coelutes from the LC column with other, more abundant peptides that are preferentially selected for fragmentation. It is also worth mentioning that although CENP-A is enriched at the centromere compared to other regions of the genome, it is estimated that CENP-A

nucleosomes represent only ~4% of all centromeric nucleosomes and that a typical human centromere contains about 400 molecules of CENP-A.[21] Thus, even within the centromere, histone H3 nucleosomes outnumber CENP-A nucleosomes by approximately 25:1, which means that HyCCAPP likely pulls down many more histone H3 than CENP-A nucleosomes (indeed, histone H3 was detected in all three alpha satellite and scrambled capture samples). On top of this, we must consider that histones are often decorated with a number of post-translational modifications, and indeed a search of CENP-A in the neXtProt database[40] reveals a number of annotated PTMs. The presence of multiple post-translationally modified forms of a base peptide sequence serves to "dilute" the signal of that peptide, thereby lowering the signal-to-noise ratio and making the peptide more difficult to detect. Taken together, these considerations present a plausible explanation for the absence of detected CENP-A peptides in HyCCAPP samples via mass spectrometry.

Besides CENP-A NAC and CAD proteins, many other known centromere-binding proteins were identified as enriched at alphoid DNA via HyCCAPP. Such proteins include CENP-B, -I, and -V. CENP-B in particular is a well-known alpha satellite-binding protein, and has been observed to bind to a specific 17 bp sequence, known as the CENP-B box, which is contained within a subset of alpha satellite repeats.[6] Additionally, the four members of the chromosomal passenger complex (CPC), INCENP, CDCA8, BIRC5, and AURKB, were all enriched at the alpha satellite repeats, which makes sense given that the CPC is known to localize to the centromere during certain phases of mitosis.[41] Multiple nuclear membrane-associated proteins, including LMNA, LMNB1, LMNB2, LBR, EMD,

TMPO, and TOR1AIP1, were also identified as enriched at the alpha satellite repeats, which could be explained by the observation that centromeres tend to localize near the nuclear periphery.[42] Furthermore, two members of the heterochromatin protein 1 (HP1) family, HP1α and HP1β, were identified as enriched at the alpha satellite repeats, which validates prior evidence of interactions between these proteins and the centromere.[43-45]

While it is clear from this discussion that the list of proteins identified by HyCCAPP as enriched at alphoid DNA reflects legitimate centromeric protein-DNA interactions, the results are further corroborated when compared to those from a related study performed in mouse embryonic stem cells.[46] In that study, a technique similar to HyCCAPP known as PICh (proteomics of isolated chromatin segments) was used to identify protein-DNA interactions occurring at mouse major satellite repeats. Similar to human cells, mouse centromeres are made up of repetitive satellite sequences. However, although alpha satellite DNA is found within both the centromeric and pericentromeric regions of human chromosomes, the mouse centromeric and pericentromeric regions are made up of two distinct types of satellite DNA. The 234 bp major satellite repeats make up the pericentromeric region, while the 120 bp minor satellite repeats make up the centromeric region.[3,47] Using PICh, the authors identified 135 proteins as enriched at the mouse major satellites and divided those proteins into five main categories including "histone/DNA regulation and chromosome organization", "DNA replication and cell cycle control", "RNA processing/nucleolus", "DNA damage", and "lamina". It is worth noting that the criteria used to classify a protein as enriched at the major satellites in that work

were different than the criteria employed in this work. For the PICh experiments, proteins identified in major satellite pulldown experiments were compared to those identified in "input chromatin" samples. A protein was denoted as "pericentromere-enriched" if it was found only in the major satellite pulldown and not in the input chromatin sample in at least three of seven independent experiments or if, on average, three times more peptide counts were observed in the major satellite pulldown experiments than in input chromatin samples. This approach contrasts with the more rigorous statistical analysis of extracted ion chromatogram-based label-free quantification data described herein to identify alpha satellite-enriched proteins from HyCCAPP experiments.

Despite the differences in these two approaches, of the 135 proteins identified as enriched at the mouse major satellites, the human homologues of 39 were found in the list of 90 proteins identified as enriched at alphoid DNA via HyCCAPP. It was particularly interesting to note that two proteins, SMCHD1 and MKI67, whose interactions with the major satellites were suggested by PICh experiments and validated via immuno-FISH experiments, were also found in the list of alpha satellite-enriched proteins. The overlap between proteins identified at mouse major satellites by PICh and those identified at human alphoid DNA using HyCCAPP provides corroborating evidence for and strengthens confidence in the existence of these protein-DNA interactions. However, it is also interesting to consider proteins that were only found in one of these two experiments. For example, it was intriguing to find that, though HyCCAPP identified many CENP-A NAC and CAD proteins as enriched at the alpha satellite repeats, these proteins were entirely missing from the list of proteins

identified as enriched at the mouse major satellites. This is likely due to the fact that the major satellites make up the pericentromeric region of mouse chromosomes, while the centromeric region is composed of minor satellites. Because NAC and CAD proteins are recruited by CENP-A, which is primarily localized to the minor satellites in mouse cells, the absence of these proteins at the major satellites is expected. In human cells, however, where alphoid DNA is found both within the core, CENP-A-containing centromere as well as within the pericentromeric region, the identification of these proteins by HyCCAPP makes sense.

HyCCAPP identifies novel alpha satellite-binding proteins

In addition to many well-known centromere-binding proteins, HyCCAPP also identified a number of proteins as enriched at the alpha satellite repeats for which there is not, to our knowledge, prior evidence suggesting an interaction with the centromere (~20-40, depending on the criteria used for classification). These novel protein hits are exciting because their interactions with alpha satellite DNA may provide insight into additional, previously unsuspected factors important for centromere structure or function. Here, we discuss one such protein—LRIF1.

LRIF1 (ligand-dependent nuclear receptor-interacting factor 1) was found to be enriched at alpha satellite DNA with a fold-change of ~63 and a permutation-based *q*-value of 0. The LRIF1 protein had previously been identified in a yeast two-hybrid screen for potential retinoic acid receptor (RAR)-interacting proteins.[48] RARs are members of the steroid/nuclear receptor (NR) family of ligand-dependent transcription factors, which regulate genes important for a variety of essential biological

processes.[48] In their screen for RAR-interacting proteins, Li et al. used RARα as a bait protein to pull down LRIF1 and subsequently showed that LRIF1 also interacts with a number of other members of the NR family of transcription factors.[48] They further showed that LRIF1 is a nuclear protein, and that it represses RARα-mediated transcriptional activation in a ligand-dependent manner, possibly through the recruitment of histone deacetylases. Additional studies showed that LRIF1 associates with HP1α, HP1β, HP1γ, and SMCHD1 in human T-REx 293 cells and is concentrated on the inactive X chromosome in female human fibroblasts.[49] The inactive X chromosome (Xi), or Barr body, is an example of facultative heterochromatin and is transcriptionally silenced in mammalian female cells. LRIF1 has been observed to play an important role in Xi compaction, by linking the H3K9me3 and XIST-associated H3K27me3 regions of Xi via interactions with HP1 and SMCHD1, respectively.[49] Furthermore, a recent study in mouse cells showed that LRIF1, in conjunction with HP1γ, can mediate the interaction between SMCHD1 and H3K9me3, potentially facilitating transcriptional repression.[50] Taken together, these studies provide evidence that LRIF1 may play a role in the formation and/or maintenance of heterochromatin.

It has been well-established that human centromeres are embedded within constitutive heterochromatin, and therefore the finding that LRIF1 is enriched at alpha satellite DNA is intriguing. On the basis of what is already known about LRIF1, it seems plausible that its enrichment on centromeric/pericentromeric alphoid DNA indicates a role for LRIF1 in the maintenance or formation of centromeric heterochromatin. This hypothesis is supported by the observation that HP1α, HP1β,

and SMCHD1, proteins which appear to play a role in transcriptional silencing by LRIF1, were also found to be enriched at alpha satellite DNA via HyCCAPP. To validate the LRIF1-alpha satellite interaction suggested by HyCCAPP, ChIP-qPCR experiments were performed to see whether alpha satellite DNA was enriched in LRIF1 immunoprecipitations compared to control IgG pulldowns. The ChIP-qPCR experiments revealed that approximately twice as much alpha satellite DNA was precipitated using an LRIF1 antibody compared to IgG (Figure 2.5B). Such enrichment was not observed for 28S rDNA, a nonalphoid sequence, suggesting that LRIF1 may indeed preferentially localize to alpha satellite DNA (Figure 2.5B). As a positive control, ChIP-qPCR experiments also showed substantial enrichment of alpha satellite DNA in CENP-B pulldown experiments as compared to the IgG control, but no enrichment of 28S rDNA, a sequence that CENP-B is not expected to bind (Figure 2.5A).

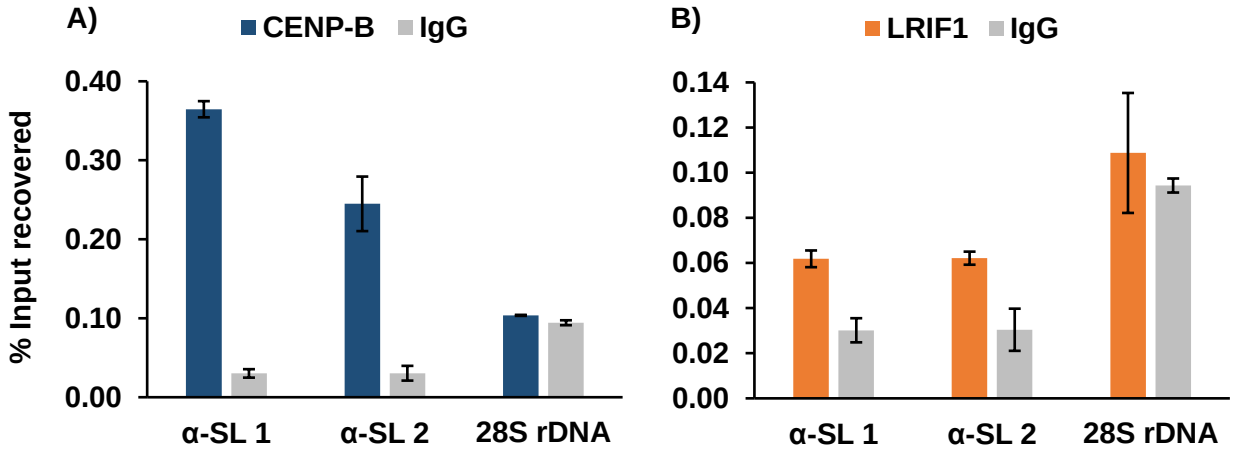


Figure 2.5 ChIP-qPCR experiments were performed using antibodies against CENP-B, a known alpha satellite-binding protein (panel A), as well as LRIF1, a novel alpha satellite interactor suggested by HyCCAPP (panel B). IgG was used as a control for the background level of the immunoprecipitation. The DNA recovered from each ChIP experiment was analyzed using two qPCR assays designed to target alpha satellite DNA (α-SL 1 and 2) and one qPCR assay designed to target a nonalphoid sequence (28S rDNA). The amount of each sequence recovered is plotted as a percentage of the input chromatin material. Average values of two experimental replicates are plotted, and error bars represent ±1 standard deviation.

It is worth noting that we attempted to validate a second novel alpha satellite-interacting protein suggested by HyCCAPP experiments, NASP, via ChIP-qPCR, but did not see enrichment of alphoid DNA in the NASP pulldown as compared to the IgG control (i.e., both NASP and IgG pulled down approximately the same amount of alpha satellite DNA; data not shown). Although we were unable to validate the NASP-alpha satellite interaction, these NASP ChIP-qPCR experiments did prove informative, as they showed that not all proteins show enrichment of alphoid DNA as compared to the IgG control. This provides support for our interpretation that the enrichment of alpha satellite

DNA seen in the CENP-B and LRIF1 ChIP experiments reflects genuine protein-alpha satellite interactions.

2.5 CONCLUSION

HyCCAPP was used in this study to identify protein-alpha satellite interactions. This constitutes the first application of the HyCCAPP technology in mammalian cells and, to our knowledge, is the first DNA-centric study of human protein-alpha satellite interactions. Many known centromere-binding proteins were identified as enriched at the alpha satellite repeats, validating the approach and providing confidence in the results obtained. HyCCAPP also identified a number of novel alpha satellite-binding proteins, including LRIF1, a heterochromatin-associated protein. Further study will be required to elucidate the roles that the novel alpha satellite-binding proteins revealed in this work may play in centromere structure and/or function.

2.6 SUPPLEMENTARY INFORMATION

Large supplementary tables are not included here, but are available online at https://doi.org/10.1021/acs.jproteome.7b00448: Supplementary Table S-2.2, Output from Perseus analysis comparing alpha satellite capture and scrambled capture samples; Supplementary Table S-2.5, Complete list of over-represented GO terms for the 90 alpha satellite-enriched proteins.

Supplementary methods

HyCCAPP

Cell culture and cross-linking: Human K562 cells were cultured at 37°C under 5% CO_2 in Iscove's Modified Dulbecco's Medium (Life Technologies #12440061) supplemented with a final concentration of 10% (v/v) fetal bovine serum (Gemini Bio-Products #900-108) and 1% (v/v) antibiotic/antimycotic solution (Gemini Bio-Products #400-101). Cells were grown to a density of ~6-7 × 10^5 cells/mL and cross-linked for 30 min at room temperature via gentle shaking with a final concentration of 3% (w/v) formaldehyde. Excess formaldehyde was quenched for 10 min at room temperature via gentle shaking with a final concentration of ~670 mM Tris pH = 8. Cells were collected via centrifugation, washed once with 1× PBS, collected again via centrifugation, snap-frozen in liquid nitrogen, and stored at -45°C for up to one month prior to use.

Cell lysis and chromatin solubilization: The HyCCAPP experimental procedure employed in this work is similar to the procedure described by Kennedy-Darling et al.[23] For each HyCCAPP experiment, frozen pellets collectively containing ~3.3 × 10^9 K562 cells were thawed briefly and resuspended in 100 mL of lysis buffer containing 200 mM NaCl, 20 mM EDTA pH = 8, 50 mM Tris pH = 7, and protease inhibitors diluted 200× from a concentrated cocktail (Sigma-Aldrich #P8340). Cells were lysed using a Constant Systems TS Series cell disruptor at 18 kpsi, and then SDS was added to a final concentration of 1% and the lysate was heated in a 65°C water bath for 8 min to aid in chromatin solubilization. The chromatin was then sheared to a median fragment size of ~3 kb by sonicating the lysate on ice in ~50-

mL aliquots for 6 min with 4 s on/off intervals at 20% amplitude using a Misonix Ultrasonic Processor S4000. After sonication, the lysate was centrifuged at 8,000 *g* for 12 min at 4°C to clear insoluble cellular debris. The supernatant was diluted five-fold with lysis buffer and RNase A (Thermo Fisher Scientific #12091021) was added to a final concentration of 60 µg/mL. The lysate was incubated at 37°C for 1 h at 150 rpm, then centrifuged at 15,000 *g* for 15 min at room temperature to remove any remaining particulates.

Hybridization capture: In order to remove endogenously biotinylated moieties from the lysate, 1.3 mL of streptavidin-coated Sera-Mag SpeedBeads (Fisher Scientific #09-981-140) were washed once with ~4 mL of wash buffer (50 mM Tris pH = 8, 200 mM NaCl, and 0.2% SDS), resuspended in 1.7 mL of wash buffer, and added to the lysate. The lysate was then incubated at room temperature for 1 h at 150 rpm. The lysate and beads were separated in 50-mL aliquots using a DynaMag-50 magnet (Thermo Fisher Scientific #12302D) and the beads were discarded. At this point, a 1-mL aliquot of the lysate was removed and stored at -20°C for eventual qPCR analysis.

Next, a total of 530 pmol of biotinylated alpha satellite capture oligonucleotides were added to the lysate. As eight capture oligonucleotides were designed to complement the alpha satellite consensus sequence,[25] ~66 pmol of each oligonucleotide were added. Additionally, 530 pmol of a biotinylated scrambled sequence capture oligonucleotide were added to the lysate at this point. The sequences of all nine capture oligonucleotides used in HyCCAPP experiments are listed in Supplementary Table S-2.1. The lysate was incubated with the oligonucleotides for 3 h at 37°C at 150

rpm, after which it was allowed to cool to room temperature. To capture the hybridized complexes, 5.3 mL of the Sera-Mag SpeedBeads were washed once with ~16 mL of wash buffer, resuspended in 6.6 mL wash buffer, and added to the lysate. The lysate was incubated at room temperature for 1 h at 150 rpm, after which the beads were isolated by continually adding lysate aliquots to five 50-mL Falcon tubes in DynaMag-50 magnet stands. The beads were allowed to separate for a period of ~4 min before the lysate was removed and fresh aliquots of the lysate-bead slurry were added to the Falcon tubes (~107 mL of slurry per Falcon tube were processed in this way).

Washing and elution: The Sera-Mag beads in each of the five Falcon tubes were washed four times with 45 mL of wash buffer, allowing 5 min of gentle mixing at room temperature per wash and discarding the used wash buffer each time. After washing the beads for the fourth time, the beads in each Falcon tube were concentrated into 1.5 mL of wash buffer and transferred to five 2-mL low-retention tubes. The beads were washed once for 5 min and once for 1 h at room temperature in this smaller volume, collecting the beads against a Magna-Sep magnet (Life Technologies #K1585-01) after each wash.

To release the hybridized complexes, two different release solutions were added to the beads in succession (the toehold-mediated strand displacement strategy used as the release method here is detailed elsewhere in the Supplementary Information). First, 1.2 mL of alpha satellite release solution containing 1.3 nmol of each of the eight alpha satellite release oligonucleotides in wash buffer were added to each tube, and the tubes were rocked gently at room temperature for 15 min. The beads were

isolated using the Magna-Sep magnet and the release solution from each of the five tubes was removed and stored temporarily at 4°C. After washing the beads twice for 5 min with 1.5 mL of wash buffer, 1.2 mL of scrambled release solution containing 10.4 nmol of the scrambled release oligonucleotide in wash buffer were added to each of the five tubes, and the tubes were again gently rocked for 15 min. The release solution from each of the five tubes was then removed and stored temporarily at 4°C. The sequences of all nine release oligonucleotides used in HyCCAPP experiments are listed in Supplementary Table S-2.1.

Sample preparation for mass spectrometry: One tube of alpha satellite release solution and one tube of scrambled release solution were removed from 4°C, and 100 µL of both solutions were aliquoted and stored at -20°C for eventual qPCR analysis. Then, 350 µL of cold trichloroacetic acid (TCA) were added to each tube of release solution. The two samples were placed in an ice bath for 10 min, centrifuged at 20,000 *g* for 20 min at 4°C, and the supernatants were discarded. During the centrifugation step, second tubes of both the alpha satellite and scrambled release solutions were removed from 4°C, and 350 µL of cold TCA were added to each tube. These tubes were placed on ice for 10 min, and the solutions were then added to the appropriate precipitate of the first set of tubes. These tubes were again centrifuged, thereby combining the precipitates from two tubes into one. This was repeated until the precipitates from all five tubes of a given release solution were combined into a single tube. During this process, the proteins in a single 1.2-mL aliquot of lysate material removed after hybridization capture were precipitated, as well. The three pellets were then washed twice with

500 µL of cold acetone, with 5 min centrifugation steps after each wash. Residual acetone was removed from the pellets by briefly heating the tubes at 95°C, and the three dried pellets were stored at -80°C overnight.

Samples were prepared for mass spectrometric analysis using the eFASP method.[29] For each sample, one collection tube and one filter (EMD Millipore #UFC505096) were passivated with 1% CHAPS detergent overnight, then rinsed thoroughly. Each TCA-precipitated pellet was resuspended in 900 µL of exchange buffer containing 8 M urea and 0.1% deoxycholic acid. For each sample, a passivated filter was placed in a non-passivated collection tube, 450 µL of the sample were transferred to the filter unit, and the units were centrifuged at 14,000 *g* for 10 min. The flow-through was discarded, and the remaining 450 µL of sample were added and centrifuged. The flow-through was again discarded, and 200 µL of exchange buffer were added to each filter and centrifuged at 14,000 *g* for 10 min. This step was repeated twice more, discarding the flow-through each time. Next, 200 µL of reducing buffer containing 8 M urea and 20 mM DL-dithiothreitol were added and allowed to incubate at room temperature for 30 min before centrifuging again at 14,000 *g* for 10 min. The flow-through was discarded, and 200 µL of alkylation buffer containing 8 M urea, 50 mM iodoacetamide, and 50 mM ammonium bicarbonate were added and allowed to incubate at room temperature for 1 h in the dark. DTT was added to a concentration of 75 mM, and the samples were incubated for 10 additional minutes before centrifuging at 14,000 *g* for 10 min. The flow-through was discarded and 200 µL of digestion buffer containing 1 M urea, 50 mM ammonium bicarbonate, and 0.1% deoxycholic

acid were added before centrifuging again at 14,000 *g* for 10 min. This step was repeated twice more, removing the flow-through after each centrifugation.

The filters were then transferred to clean, passivated collection tubes, 100 µL of digestion buffer containing trypsin (0.5 µg for capture samples and 1 µg for lysate sample) (Promega #V5111) were added to each sample, and the samples were incubated at 37°C overnight. The samples were centrifuged at 14,000 *g* for 10 min, and the flow-through was allowed to collect in the tube. Next, 50 µL of 50 mM ammonium bicarbonate were added to each filter, and the units were centrifuged at 14,000 *g* for 10 min. This step was repeated once more, again allowing the flow-through to collect in the tube. The 200 µL of flow-through were transferred to a clean low-retention tube and 200 µL of ethyl acetate were added. Trifluoroacetic acid (TFA) was added to a final concentration of 0.5%, and the tubes were vortexed for 1 min before centrifuging at 15,700 *g* for 2 min. The ethyl acetate layer was removed, and this extraction step was repeated twice more. The samples were then dried in a Savant SVC-100H SpeecVac Concentrator for ~2.5 h. The samples were resuspended in 180 µL of 0.1% TFA and stored at -20°C overnight.

Finally, each sample was desalted using a C18 solid-phase extraction pipet tip (Agilent Technologies #A57003100K). The pipet tip was activated by slowly pipetting 70% acetonitrile then 0.1% TFA up and down at least three times each. The tip was then placed in the peptide sample, and the sample was slowly pipetted up and down at least five times. The tip was washed three times with 0.1% TFA, and the peptides were eluted by slowly pipetting 70% acetonitrile/0.1% TFA up and down

at least five times. The samples were dried in the Savant SpeedVac Concentrator for ~1 h, and the peptides were resuspended in 95:5 water/acetonitrile with 0.2% formic acid (18 µL for capture samples and 1.8 mL for lysate sample). Two technical replicate injections of 9 µL were performed for each sample.

Mass spectrometry analysis: Samples were analyzed via HPLC-ESI-MS/MS using a system consisting of a high performance liquid chromatograph (nanoAcquity, Waters) connected to an electrospray ionization (ESI) orbitrap mass spectrometer (LTQ Velos, Thermo Fisher Scientific). HPLC separation employed a 100 × 365 µm fused silica capillary microcolumn packed with 20 cm of 1.7 µm diameter, 130 Å pore size C18 beads (Waters BEH), with an emitter tip pulled to approximately 1 µm using a laser puller (Sutter instruments). Peptides were loaded on-column at a flow-rate of 400 nL/min for 30 min, then eluted over 125 min at a flow-rate of 300 nL/min with a gradient from 2% to 30% acetonitrile in 0.1% formic acid. Full-mass profile scans were performed in the orbitrap between 300 and 1,500 *m/z* at a resolution of 60,000, followed by MS/MS HCD scans of the ten highest intensity parent ions with $z > 1$ at 42% relative collision energy and 7,500 resolution, with a mass range starting at 100 *m/z*. Dynamic exclusion was enabled with a repeat count of two over a duration of 30 s and an exclusion window of 120 s. All raw data files are available from the MassIVE repository under the name "AlphaSatelliteHyCCAPP".

Mass spectrometry data analysis: Three replicates of the HyCCAPP experiment were carried out as described. For each given sample (alpha satellite capture, scrambled capture, or lysate), raw files from

both technical replicates of a given experimental replicate were treated as a single "experiment", and the resultant nine experiments were analyzed together using MaxQuant (version 1.5.3.30).[31] The "match between runs" algorithm was employed to allow for the identification of peptides that were not selected for fragmentation in a given run, but were selected for fragmentation in a separate run, permitting the recalibrated retention times to deviate by a maximum of 0.7 min. The detected features were searched against the human canonical protein database from UniProt (downloaded on April 22, 2016) and a list of potential contaminants. Precursor and fragment ion mass tolerances were set to 4.5 and 20 ppm, respectively. Carbamidomethylation of cysteine (+57.0215 Da) was set as a fixed modification, and oxidation of methionine (+15.9949 Da) was set as a variable modification. Only tryptic peptides with at least seven amino acids and up to two missed cleavages were considered, and one peptide was required per protein. A false discovery rate (FDR) of 1% at both the peptide and protein levels was allowed. Relative quantification was performed for each capture sample (alpha satellite and scrambled) of each experimental replicate (1-3) using the protein intensities reported in the "proteinGroups.txt" file produced by MaxQuant. Individual protein intensities were normalized to the median protein intensity of the given capture sample after removing potential contaminants.

Perseus (version 1.5.3.2)[32] was used for downstream statistical analysis of the normalized protein intensities. Proteins from the reverse database and proteins only identified by site were removed, and the normalized protein intensities were $\log_2$-transformed. The three experimental replicates of a given capture sample (alpha satellite or scrambled) were grouped, and the proteins were

again filtered to keep only those proteins which had three observations in at least one group (i.e., those proteins which were observed in all three experimental replicates of at least one capture sample type). Missing intensity values were imputed based on the normal distributions, and two-sided Welch's *t* tests were carried out to compare protein abundances between groups. The statistical tests were corrected for multiple hypothesis testing using a permutation-based FDR cutoff of 5% (S0 = 2) (output of Perseus analysis can be found in Supplementary Table S-2.2).

Quantitative real-time PCR analysis: Quantitative real-time PCR (qPCR) was performed to assess the efficiency and specificity with which the alpha satellite repeats were captured in HyCCAPP experiments. Because of the heterogeneity of alpha satellite monomeric sequences, it was not feasible to design qPCR assays to target every individual alpha satellite monomer. Instead, a TaqMan assay was designed to target a single alpha satellite sequence, and the capture efficiency of this sequence was used as a proxy for the capture efficiency of alpha satellite DNA as a whole. Additionally, a qPCR assay for an off-target sequence (28S rDNA) was designed to assess capture specificity. These assays were ordered from IDT, and the sequences of the primers and probes are listed in Supplementary Table S-2.3.

Formaldehyde cross-links were reversed prior to qPCR analysis by heating samples at 95°C for 25 min. For the alpha satellite and scrambled release samples, 20-µL aliquots were removed from the 100-µL aliquot taken prior to TCA precipitation. For the lysate sample, a 50-µL aliquot was removed from the 1-mL aliquot taken prior to hybridization capture. A small volume (1.5 µL or less) of Tris buffer pH = 8 was added prior to heating to bring each sample to ~190 mM Tris. After heating, two

technical replicates of each sample were prepared by diluting 5-µL aliquots 10× with 10 mM Tris buffer pH = 7. Additionally, standards were prepared using human genomic DNA (Promega #G3041) serially diluted with 10 mM Tris buffer. All samples and standards were analyzed in duplicate on a 96-well plate (4titude #4ti-0951) covered with a thermal seal (Excel Scientific #TS-RT2-100). Each well contained 5 µL of sample or standard, 10 µL of LightCycler 480 probes master (Roche #04707494001), 4.5 µL of water, and 0.5 µL of 40× primer-probe mix. Each plate was centrifuged at 89 *g* for 2 min after pipetting and analyzed using a Roche 480 LightCycler.

Chromatin immunoprecipitation with qPCR (ChIP-qPCR)

Human K562 cells were cultured under the same conditions used for HyCCAPP experiments. Cells were cross-linked for 10 min at room temperature via gentle shaking with a final concentration of 1% (w/v) formaldehyde. Excess formaldehyde was quenched for 5 min at room temperature via gentle shaking with a final concentration of ~250 mM Tris pH = 8. Cells were collected via centrifugation, washed once with 1× PBS, collected again via centrifugation, snap-frozen in liquid nitrogen, and stored at -80°C prior to use.

Approximately 5×10^7 cross-linked cells were resuspended in 25 mL of cold buffer containing 50 mM HEPES pH = 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 substitute (Amresco #M158), 0.25% Triton X-100, and protease/phosphatase inhibitors diluted 100× from a concentrated cocktail (Thermo Fisher Scientific #1861281). The resuspended cells were incubated on ice for 20 min, then centrifuged at 3,200 *g* for 5 min at 4°C. The supernatant was discarded and the pellet was

resuspended in 16.8 mL of cold wash buffer containing 10 mM Tris pH = 8.1, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, and protease/phosphatase inhibitors. The suspension was centrifuged at 3,200 *g* for 5 min at 4°C, and the supernatant was discarded. The interior of the tube was gently rinsed with 8.4 mL of cold shearing buffer containing 0.1% SDS, 1 mM EDTA, 10 mM Tris pH = 8.1, and protease/phosphatase inhibitors, taking care not to disturb the pellet. The sample was again centrifuged and the supernatant was discarded. The nuclear pellet was resuspended in cold shearing buffer to a final concentration of ~7.4×10^6 nuclei/mL, and the chromatin was sonicated in 850 µL aliquots using a Covaris S220 water bath sonicator. The water bath was set to 5-8°C, and each aliquot was sonicated for 10 cycles of 60 sec of sonication followed by a 45 sec resting period with the following settings: peak power = 170.0, duty factor = 10.0, and 200 cycles/burst. The size of the chromatin fragments following sonication was determined to be ~200-500 bp by agarose gel electrophoresis, and the concentration of DNA in the sample was measured using a Nanodrop. The sheared chromatin was centrifuged at maximum speed in a tabletop microcentrifuge for 10 min at 4°C to remove insoluble debris, and the supernatant was aliquoted into 10-µg and 1-µg portions (based on Nanodrop measurement) and stored with 10% glycerol at -80°C for up to two months prior to use.

For each ChIP experiment, a 10-µg aliquot of chromatin was thawed on ice and diluted 10-fold in cold buffer containing 16.7 mM Tris pH = 8, 0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 333 mM NaCl, and protease/phosphatase inhibitors. Then, 2 µg of either CENP-B antibody (abcam #ab25734), LRIF1 antibody (Atlas Antibodies #HPA044515), or IgG (Jackson ImmunoResearch #312-

005-003) were added to the chromatin and incubated overnight at 4°C with gentle mixing. Next, 10 µL each of protein A- and protein G-coated magnetic beads (Thermo Fisher Scientific #10002D and #10004D) were washed once with cold 1× PBS pH = 7.4 with 0.2% Tween-20 and protease/phosphatase inhibitors and once with cold buffer containing 15 mM Tris pH = 8, 0.01% SDS, 1% Triton X-100, 1 mM EDTA, 300 mM NaCl, and protease/phosphatase inhibitors, then resuspended in their original volume of the latter wash buffer, added to the sample, and gently mixed for 2 h at 4°C. The beads were isolated using a magnet stand and the supernatant was discarded. The beads were then washed twice for 10 min each at room temperature with 500 µL of the following wash buffers: wash buffer A (50 mM HEPES pH = 7.9, 0.1% SDS, 1% Triton X-100, 0.1% deoxycholate, 1 mM EDTA pH = 8, 140 mM NaCl, and protease/phosphatase inhibitors), wash buffer B (50 mM HEPES pH = 7.9, 0.1% SDS, 1% Triton X-100, 0.1% deoxycholate, 1 mM EDTA pH = 8, 500 mM NaCl, and protease/phosphatase inhibitors), LiCl wash buffer (20 mM Tris pH = 8, 0.5% NP-40 substitute, 0.5% deoxycholate, 1 mM EDTA pH = 8, 250 mM LiCl, and protease/phosphatase inhibitors), and TE (10 mM Tris pH = 8, 1 mM EDTA pH = 8, and protease/phosphatase inhibitors). The captured chromatin was then eluted from the beads by heating at 65°C for 10 min in 100 µL of elution buffer containing 50 mM Tris pH = 8, 1 mM EDTA, 1% SDS, and protease/phosphatase inhibitors. Next, 80 µL of TE buffer containing 0.67% SDS and protease/phosphatase inhibitors were added to the sample, and the sample was heated overnight at 65°C. At this time, a 1-µg aliquot of chromatin ("10% input sample") was thawed on ice

and diluted with 100 µL of elution buffer and 80 µL of TE buffer containing 0.67% SDS and protease/phosphatase inhibitors and heated overnight at 65°C.

Next, RNase A (1.8 µL of a 10 mg/mL stock) was added to both the ChIP sample and the 10% input sample and incubated at 37°C for 30 min. Proteinase K (1.8 µL of a 20 mg/mL stock) was added to each sample and incubated at 55°C for 2 h, then the beads were removed from the ChIP sample using a magnet stand. The DNA in both samples was then purified via phenol-chloroform extraction and ethanol precipitation, and the DNA pellets were resuspended in 50 µL of 10 mM Tris pH = 7. The DNA was then analyzed via qPCR using two assays designed to target alpha satellite DNA and one assay designed to target 28S rDNA (negative control sequence). These assays were ordered from IDT, and the sequences of the primers and probes are listed in Supplementary Table S-2.3. For each assay, a standard curve was generated by preparing serial dilutions of the 10% input sample, and each standard/ChIP sample was analyzed in duplicate on a 96-well plate (4titude #4ti-0951) covered with a thermal seal (Excel Scientific #TS-RT2-100). Each well contained 2.5 µL of sample/standard, 5 µL of LightCycler 480 probes master (Roche #04707494001), 2.3 µL of water, and 0.25 µL of 40× primer-probe mix. Each plate was centrifuged at 89 *g* for 2 min after pipetting and analyzed using a Roche 480 LightCycler.

Implementation of toehold-mediated strand displacement as an elution strategy for HyCCAPP

In this work, we describe the first application of the HyCCAPP technology in a mammalian system. HyCCAPP has previously proven its utility in yeast,[23,24] and the extension of this tool to a

human cell line represents a significant step forward. To make this transition, relatively few alterations were made to the originally reported procedure.[23] Perhaps the most significant alteration was the release strategy employed to elute the protein-DNA complexes from the streptavidin-coated magnetic beads following hybridization capture. Previously, 30 nt-long desthiobiotinylated capture oligonucleotides were used to capture the target genomic locus and isolate the hybridized complexes on the beads. Free biotin was then introduced to elute the complexes, taking advantage of the fact that biotin binds streptavidin with much higher affinity than does desthiobiotin. While this strategy worked well, we sought to develop an approach that would allow the HyCCAPP experiment to be multiplexed, enabling multiple loci to be studied in parallel from the same collection of cells. This would reduce the cost and time required to perform a HyCCAPP experiment by a factor of the degree of multiplexing, thereby increasing the return-on-investment for this technology.

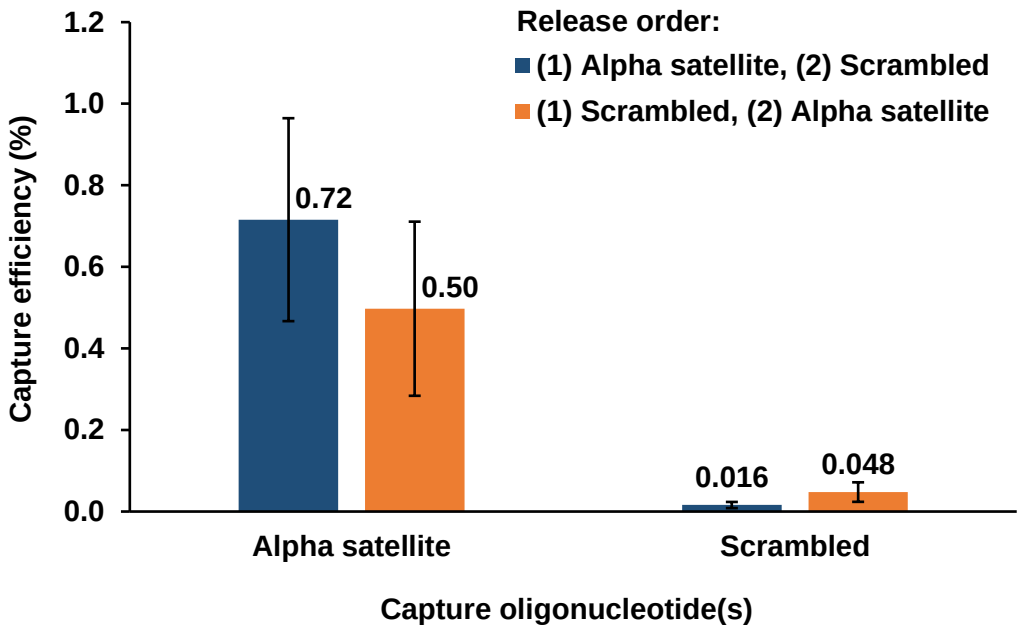
In our eyes, the key to multiplexing HyCCAPP was developing a strategy which would allow for the selective, sequential release of captured loci from the streptavidin-coated magnetic beads. Unfortunately, the biotin release method did not meet these criteria, as the introduction of free biotin released all desthiobiotinylated moieties from the beads at once in a non-selective manner. This led us to investigate other strategies for elution, and one promising such strategy that emerged is based on the principle of toehold-mediated strand displacement. For the purposes of HyCCAPP, this approach involves the use of a 38 nt biotinylated capture oligonucleotide. This oligonucleotide is composed of a 30 nt stretch which is complementary to a target DNA sequence, and an 8 nt stretch at the 5' end which

is not complementary, but serves as the toehold. This oligonucleotide is allowed to hybridize to the target, and the hybridized complex is isolated on a solid support. Following capture, the target DNA can be selectively eluted by introducing a 38 nt release oligonucleotide which is completely complementary to the capture oligonucleotide, taking advantage of the fact that the extra eight nucleotides make the capture oligonucleotide-release oligonucleotide hybrid more thermodynamically stable than the capture oligonucleotide-target hybrid. We previously developed and demonstrated this approach for simple mixtures of synthetic oligonucleotides, for oligonucleotides hybridized to DNA microarrays, and for the selective, multiplexed release of individual target regions from cross-linked yeast chromatin.[27,28]

Here, we utilized this strategy to perform target (alpha satellite capture) and control (scrambled capture) experiments from the same cellular material. A concentration of release oligonucleotides 100-fold higher than the concentration of capture oligonucleotides was used to achieve fast release kinetics, and two washing steps were employed between the release steps to remove excess alpha satellite release oligonucleotides prior to the introduction of scrambled release oligonucleotides. Using qPCR, the efficiency with which alphoid DNA was captured in the HyCCAPP experiments was determined by dividing the amount of alpha satellite DNA present in the final alpha satellite capture sample by the amount of alpha satellite DNA present in the lysate prior to capture. Using this method, the efficiency with which alphoid DNA was captured in the alpha satellite capture experiment was determined to be ~0.22%, which is comparable to capture efficiencies reported using the biotin release method.[23] By an

analogous calculation, the efficiency with which alpha satellite DNA was captured in the scrambled control experiment was determined to be ~0.0049%, approximately 44 times lower than in the target capture experiment (Figure 2.2). To evaluate the specificity of capture, the capture efficiency of an off-target sequence (28S rDNA) was also calculated for each sample. Neither the alpha satellite nor the scrambled capture oligonucleotides captured the off-target sequence with appreciable efficiency (Figure 2.2). In small-scale HyCCAPP experiments, the effect of release order (alpha satellite first followed scrambled or vice versa) on alpha satellite capture efficiency was evaluated, and no dramatic differences were apparent (Supplementary Figure S-2.1).

Supplementary figures and tables



Supplementary Figure S-2.1 The impact of toehold-mediated release order (alpha satellite first followed scrambled or vice versa) on alpha satellite capture efficiency was assessed in small-scale HyCCAPP experiments. For each release order, the efficiency with which alphoid DNA was captured in both the alpha satellite capture and scrambled capture samples was measured via qPCR. These measurements were made for three small-scale HyCCAPP experiments per release order, and the labels denote average values. Error bars represent ±1 standard deviation of the three replicates. From these data, it does not appear that release order has a dramatic impact on alpha satellite capture efficiency. In full-scale HyCCAPP experiments, the alpha satellite target was released first.

Supplementary Table S-2.1 Sequences for the nine capture oligonucleotides and nine release oligonucleotides used in HyCCAPP experiments. All oligonucleotides were ordered from Sigma-Aldrich.

Oligonucleotide	Sequence (5' - 3')	Modification
Alpha satellite capture 1	AGT GTC ACC ATT CTC AGA AAC TTC TTT GTG ATG TGT GC	3'-Biotin-TEG
Alpha satellite capture 2	AGT GTC ACG TGA TGT GTG CAT TCA ACT CAC AGA GTT GA	3'-Biotin-TEG
Alpha satellite capture 3	AGT GTC ACA CAG AGT TGA ACC TTT CTT TTG ATA GAG CA	3'-Biotin-TEG
Alpha satellite capture 4	AGT GTC ACT GAT AGA GCA GTT TTG AAA CAC TCT TTT TG	3'-Biotin-TEG
Alpha satellite capture 5	AGT GTC ACA CTC TTT TTG TAG AAT CTG CAA GTG GAT AT	3'-Biotin-TEG
Alpha satellite capture 6	AGT GTC ACA AGT GGA TAT TTG GAG CGC TTT GAG GCC TA	3'-Biotin-TEG
Alpha satellite capture 7	AGT GTC ACT TGA GGC CTA TGG TGG AAA AGG AAA TAT CT	3'-Biotin-TEG
Alpha satellite capture 8	AGT GTC ACA AAT ATC TTC ACA TAA AAA CTA GAC AGA AG	3'-Biotin-TEG
Alpha satellite release 1	GCA CAC ATC ACA AAG AAG TTT CTG AGA ATG GTG ACA CT	None
Alpha satellite release 2	TCA ACT CTG TGA GTT GAA TGC ACA CAT CAC GTG ACA CT	None
Alpha satellite release 3	TGC TCT ATC AAA AGA AAG GTT CAA CTC TGT GTG ACA CT	None
Alpha satellite release 4	CAA AAA GAG TGT TTC AAA ACT GCT CTA TCA GTG ACA CT	None
Alpha satellite release 5	ATA TCC ACT TGC AGA TTC TAC AAA AAG AGT GTG ACA CT	None
Alpha satellite release 6	TAG GCC TCA AAG CGC TCC AAA TAT CCA CTT GTG ACA CT	None
Alpha satellite release 7	AGA TAT TTC CTT TTC CAC CAT AGG CCT CAA GTG ACA CT	None
Alpha satellite release 8	CTT CTG TCT AGT TTT TAT GTG AAG ATA TTT GTG ACA CT	None
Scrambled capture	GTT TAC CCT CGC AAC TGA ACA CAT GAG CTA GTC AAA TA	3'-Biotin-TEG
Scrambled release	TAT TTG ACT AGC TCA TGT GTT CAG TTG CGA GGG TAA AC	None

Supplementary Table S-2.3 Primer and probe sequences for all qPCR assays used in this study. All assays were ordered from IDT.

Primer/probe	Sequence (5' - 3')	Modifications	Application
Alpha satellite assay 1 primer 1	CTT CGT TTC AAA ACT AGA CA	None	ChIP-qPCR
Alpha satellite assay 1 primer 2	ACT GCT CTA TGA AAA GAA AG	None	
Alpha satellite assay 1 probe	TTG AAT GAA CAC ATC ACA ACG CAG TT	5'FAM-3'TAMRA	
Alpha satellite assay 2 primer 1	TGG ATA GCT TTG AGG ATT TCG	None	ChIP-qPCR
Alpha satellite assay 2 primer 2	GCT CTA TGA AAG GGA ATG TTC A	None	
Alpha satellite assay 2 probe	CAC AAA GAA GTT TCT GAG AAT GCT TCT GTC	5'FAM-3'TAMRA	
28S rDNA primer 1	CAG CCG ACT TAG AAC TGG TG	None	HyCCAPP/ ChIP-qPCR
28S rDNA primer 2	CAC TGG GCA GAA ATC ACA TC	None	
28S rDNA probe	ACA AAG CAT CGC GAA GGC CC	5'FAM-3'TAMRA	
Alpha satellite assay 3 primer 1	GCC TCA AAC TGC TCA CAA GT	None	HyCCAPP
Alpha satellite assay 3 primer 2	TGT GTG CAT TCT TCT CAC AGA G	None	
Alpha satellite assay 3 probe	CCG ACA AAC AGA CTG TTT CCA AAC TGC	5'FAM-3'TAMRA	

Supplementary Table S-2.4 List of 90 proteins identified as enriched at the alpha satellite repeats via HyCCAPP experiments. For each protein, the UniProt ID and gene name are given.

UniProt ID	Gene	Protein name
Q8WYP5	AHCTF1	Protein ELYS
P27695	APEX1	DNA-(apurinic or apyrimidinic site) lyase
Q8N2Z9	APITD1	Centromere protein S
Q6PL18	ATAD2	ATPase family AAA domain-containing protein 2
Q96GD4	AURKB	Aurora kinase B
O75531	BANF1	Barrier-to-autointegration factor
Q9UIG0	BAZ1B	Tyrosine-protein kinase BAZ1B
O15392	BIRC5	Baculoviral IAP repeat-containing protein 5
Q9UNZ5	C19orf53	Leydig cell tumor 10 kDa protein homolog
P83916	CBX1	Chromobox protein homolog 1
P45973	CBX5	Chromobox protein homolog 5
Q8N9Z2	CCDC71L	Coiled-coil domain-containing protein 71L
Q53HL2	CDCA8	Borealin
P07199	CENPB	Major centromere autoantigen B
Q03188	CENPC	Centromere protein C
Q9H3R5	CENPH	Centromere protein H
Q92674	CENPI	Centromere protein I

UniProt ID	Gene	Protein name
Q9BS16	CENPK	Centromere protein K
Q9NSP4	CENPM	Centromere protein M
Q96H22	CENPN	Centromere protein N
Q9BU64	CENPO	Centromere protein O
Q6IPU0	CENPP	Centromere protein P
Q7L2Z9	CENPQ	Centromere protein Q
Q96BT3	CENPT	Centromere protein T
Q71F23	CENPU	Centromere protein U
Q7Z7K6	CENPV	Centromere protein V
Q6NUI6	CHADL	Chondroadherin-like protein
Q8NFW8	CMAS	N-acylneuraminate cytidylyltransferase
P26358	DNMT1	DNA (cytosine-5)-methyltransferase 1
Q9H410	DSN1	Kinetochore-associated protein DSN1 homolog
P50402	EMD	Emerin
Q9BXW9	FANCD2	Fanconi anemia group D2 protein
Q9NVI1	FANCI	Fanconi anemia group I protein
P22087	FBL	rRNA 2-O-methyltransferase fibrillarin
P02751	FN1	Fibronectin
P16401	HIST1H1B	Histone H1.5
P10412	HIST1H1E	Histone H1.4
Q96A08	HIST1H2BA	Histone H2B type 1-A
P09429	HMGB1	High mobility group protein B1
P26583	HMGB2	High mobility group protein B2
O15347	HMGB3	High mobility group protein B3
Q9NQS7	INCENP	Inner centromere protein
Q13352	ITGB3BP	Centromere protein R
Q9Y2K7	KDM2A	Lysine-specific demethylase 2A
Q14739	LBR	Lamin-B receptor
P49257	LMAN1	Protein ERGIC-53
P02545	LMNA	Prelamin-A/C
P20700	LMNB1	Lamin-B1
Q03252	LMNB2	Lamin-B2
Q5T3J3	LRIF1	Ligand-dependent nuclear receptor-interacting factor 1
P49736	MCM2	DNA replication licensing factor MCM2
P25205	MCM3	DNA replication licensing factor MCM3
P33991	MCM4	DNA replication licensing factor MCM4
P33992	MCM5	DNA replication licensing factor MCM5
Q14566	MCM6	DNA replication licensing factor MCM6

UniProt ID	Gene	Protein name
P33993	MCM7	DNA replication licensing factor MCM7
Q14676	MDC1	Mediator of DNA damage checkpoint protein 1
P46013	MKI67	Antigen KI-67
P49321	NASP	Nuclear autoantigenic sperm protein
Q9H0A0	NAT10	N-acetyltransferase 10
Q15021	NCAPD2	Condensin complex subunit 1
O75607	NPM3	Nucleoplasmin-3
Q9BXS6	NUSAP1	Nucleolar and spindle-associated protein 1
P09874	PARP1	Poly [ADP-ribose] polymerase 1
P53350	PLK1	Serine/threonine-protein kinase PLK1
P28340	POLD1	DNA polymerase delta catalytic subunit
P49642	PRIM1	DNA primase small subunit
Q9NS91	RAD18	E3 ubiquitin-protein ligase RAD18
Q16576	RBBP7	Histone-binding protein RBBP7
Q9P258	RCC2	Protein RCC2
Q5UIP0	RIF1	Telomere-associated protein RIF1
P27694	RPA1	Replication protein A 70 kDa DNA-binding subunit
P15927	RPA2	Replication protein A 32 kDa subunit
P35244	RPA3	Replication protein A 14 kDa subunit
P63173	RPL38	60S ribosomal protein L38
P62273	RPS29	40S ribosomal protein S29
A6NHR9	SMCHD1	Structural maintenance of chromosomes flexible hinge domain-containing protein 1
Q15005	SPCS2	Signal peptidase complex subunit 2
P51571	SSR4	Translocon-associated protein subunit delta
Q08945	SSRP1	FACT complex subunit SSRP1
Q8N3U4	STAG2	Cohesin subunit SA-2
Q9UH99	SUN2	SUN domain-containing protein 2
P42166	TMPO	Lamina-associated polypeptide 2, isoform alpha
P11388	TOP2A	DNA topoisomerase 2-alpha
Q02880	TOP2B	DNA topoisomerase 2-beta
Q5JTV8	TOR1AIP1	Torsin-1A-interacting protein 1
Q9ULW0	TPX2	Targeting protein for Xklp2
Q99986	VRK1	Serine/threonine-protein kinase VRK1
O75717	WDHD1	WD repeat and HMG-box DNA-binding protein 1
Q96ME7	ZNF512	Zinc finger protein 512

2.7 ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health Center of Excellence in Genomic Science grant 1P50HG004952. K.E.B. was supported by the National Human Genome Research Institute grant to the Genomic Sciences Training Program, 5T32HG002760. The authors would like to thank Professor Emery H. Bresnick and members of his group for allowing us to use their cell culture facilities and for providing a stock of human K562 cells, Professor Ronald T. Raines for use of the cell disruptor, and Professor Eric R. Strieter for use of the Misonix sonicator. We are grateful to Dr. Yunxiang Dai for helpful conversations about HyCCAPP experiments.

2.8 REFERENCES

(1) Willard, H. F. Evolution of alpha satellite. *Curr. Opin. Genet. Dev.* **1991**, *1*, 509-514.

(2) Aldrup-MacDonald, M. E.; Sullivan, B. A. The past, present, and future of human centromere genomics. *Genes* **2014**, *5*, 33-50.

(3) Choo, K. H. *The Centromere*; Oxford University Press: Oxford, 1997.

(4) McKinley, K. L.; Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 16-29.

(5) Eichler, E. E. Repetitive conundrums of centromere structure and function. *Hum. Mol. Genet.* **1999**, *8*, 151-155.

(6) Masumoto, H.; Masukata, H.; Muro, Y.; Nozaki, N.; Okazaki, T. A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *J. Cell Biol.* **1989**, *109*, 1963-1973.

(7) Ohzeki, J.; Nakano, M.; Okada, T.; Masumoto, H. CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *J. Cell Biol.* **2002**, *159*, 765-775.

(8) Voullaire, L. E.; Slater, H. R.; Petrovic, V.; Choo, K. H. A. A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? *Am. J. Hum. Genet.* **1993**, *52*, 1153-1163.

(9) Fukagawa, T.; Earnshaw, W. C. The centromere: chromatin foundation for the kinetochore machinery. *Dev. Cell* **2014**, *30*, 496-508.

(10) Burrack, L. S.; Berman, J. Neocentromeres and epigenetically inherited features of centromeres. *Chromosome Res.* **2012**, *20*, 607-619.

(11) Sullivan, B. A.; Karpen, G. H. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat. Struct. Mol. Biol.* **2004**, *11*, 1076-1083.

(12) McAinsh, A. D.; Meraldi, P. The CCAN complex: linking centromere specification to control of kinetochore-microtubule dynamics. *Semin. Cell Dev. Biol.* **2011**, *22*, 946-952.

(13) Bailey, A. O.; Panchenko, T.; Shabanowitz, J.; Lehman, S. M.; Bai, D. L.; Hunt, D. F.; Black, B. E.; Foltz, D. R. Identification of the post-translational modifications present in centromeric chromatin. *Mol. Cell. Proteomics* **2016**, *15*, 918-931.

(14) Palmer, D. K.; O'Day, K.; Wener, M. H.; Andrews, B. S.; Margolis, R. L. A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. *J. Cell Biol.* **1987**, *104*, 805-815.

(15) Palmer, D. K.; O'Day, K.; Trong, H. L.; Charbonneau, H.; Margolis, R. L. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 3734-3738.

(16) Yoda, K.; Ando, S.; Morishita, S.; Houmura, K.; Hashimoto, K.; Takeyasu, K.; Okazaki, T. Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution *in vitro*. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 7266-7271.

(17) Stoler, S.; Keith, K. C.; Curnick, K. E.; Fitzgerald-Hayes, M. A mutation in *CSE4*, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis. *Genes Dev.* **1995**, *9*, 573-586.

(18) Howman, E. V.; Fowler, K. J.; Newson, A. J.; Redward, S.; MacDonald, A. C.; Kalitsis, P.; Choo, K. H. A. Early disruption of centromeric chromatin organization in centromere protein A (*Cenpa*) null mice. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 1148-1153.

(19) Blower, M. D.; Karpen, G. H. The role of *Drosophila* CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. *Nat. Cell Biol.* **2001**, *3*, 730-739.

(20) Foltz, D. R.; Jansen, L. E. T.; Black, B. E.; Bailey, A. O.; Yates III, J. R.; Cleveland, D. W. The human CENP-A centromeric nucleosome-associated complex. *Nat. Cell Biol.* **2006**, *8*, 458-469.

(21) Bodor, D. L.; Mata, J. F.; Sergeev, M.; David, A. F.; Salimian, K. J.; Panchenko, T.; Cleveland, D. W.; Black, B. E.; Shah, J. V.; Jansen, L. E. T. The quantitative architecture of centromeric chromatin. *eLife* **2014**, *3*, e02137.

(22) Nakano, M.; Cardinale, S.; Noskov, V. N.; Gassmann, R.; Vagnarelli, P.; Kandels-Lewis, S.; Larionov, V.; Earnshaw, W. C.; Masumoto, H. Inactivation of a human kinetochore by specific targeting of chromatin modifiers. *Dev. Cell* **2008**, *14*, 507-522.

(23) Kennedy-Darling, J.; Guillen-Ahlers, H.; Shortreed, M. R.; Scalf, M.; Frey, B. L.; Kendziorski, C.; Olivier, M.; Gasch, A. P.; Smith, L. M. Discovery of chromatin-associated proteins via sequence-specific capture and mass spectrometric protein identification in *Saccharomyces cerevisiae*. *J. Proteome Res.* **2014**, *13*, 3810-3825.

(24) Guillen-Ahlers, H.; Rao, P. K.; Levenstein, M. E.; Kennedy-Darling, J.; Perumalla, D. S.; Jadhav, A. Y. L.; Glenn, J. P.; Ludwig-Kubinski, A.; Drigalenko, E.; Montoya, M. J.; Göring, H. H.; Anderson, C. D.; Scalf, M.; Gildersleeve, H. I. S.; Cole, R.; Greene, A. M.; Oduro, A. K.; Lazarova, K.; Cesnik, A. J.; Barfknecht, J.; Cirillo, L. A.; Gasch, A. P.; Shortreed, M. R.; Smith, L. M.; Olivier, M. HyCCAPP as a tool to characterize promoter DNA-protein interactions in *Saccharomyces cerevisiae*. *Genomics* **2016**, *107*, 267-273.

(25) Choo, K. H.; Vissel, B.; Nagy, A.; Earle, E.; Kalitsis, P. A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res.* **1991**, *19*, 1179-1182.

(26) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates III, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**, *113*, 2343-2394.

(27) Kennedy-Darling, J.; Holden, M. T.; Shortreed, M. R.; Smith, L. M. Multiplexed programmable release of captured DNA. *ChemBioChem* **2014**, *15*, 2353-2356.

(28) Dai, Y.; Kennedy-Darling, J.; Shortreed, M. R.; Scalf, M.; Gasch, A. P.; Smith, L. M. Multiplexed sequence-specific capture of chromatin and mass spectrometric discovery of associated proteins. *Anal. Chem.* **2017**, *89*, 7841-7846.

(29) Erde, J.; Loo, R. R. O.; Loo, J. A. Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. *J. Proteome Res.* **2014**, *13*, 1885-1895.

(30) Kennedy-Darling, J.; Smith, L. M. Measuring the formaldehyde protein-DNA cross-link reversal rate. *Anal. Chem.* **2014**, *86*, 5678-5681.

(31) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367-1372.

(32) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **2016**, *13*, 731-740.

(33) Wu, C.-H.; Chen, S.; Shortreed, M. R.; Kreitinger, G. M.; Yuan, Y.; Frey, B. L.; Zhang, Y.; Mirza, S.; Cirillo, L. A.; Olivier, M.; Smith, L. M. Sequence-specific capture of protein-DNA complexes for mass spectrometric protein identification. *PLoS One* **2011**, *6*, e26217.

(34) Pinamonti, S.; Caruso, A.; Mazzeo, V.; Zebini, E.; Rossi, A. DNA damage from pulsed sonication of human leukocytes *in vitro*. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **1986**, *33*, 179-185.

(35) Elsner, H. I.; Lindblad, E. B. Ultrasonic degradation of DNA. *DNA* **1989**, *8*, 697-701.

(36) Mi, H.; Muruganujan, A.; Casagrande, J. T.; Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **2013**, *8*, 1551-1566.

(37) Mi, H.; Huang, X.; Muruganujan, A.; Tang, H.; Mills, C.; Kang, D.; Thomas, P. D. PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **2017**, *45*, D183-D189.

(38) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447-D452.

(39) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498-2504.

(40) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J. J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D177-D182.

(41) Carmena, M.; Wheelock, M.; Funabiki, H.; Earnshaw, W. C. The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 789-803.

(42) Weierich, C.; Brero, A.; Stein, S.; von Hase, J.; Cremer, C.; Cremer, T.; Solovei, I. Three-dimensional arrangements of centromeres and telomeres in nuclei of human and murine lymphocytes. *Chromosome Res.* **2003**, *11*, 485-502.

(43) Wreggett, K. A.; Hill, F.; James, P. S.; Hutchings, A.; Butcher, G. W.; Singh, P. B. A mammalian homologue of *Drosophila* heterochromatin protein 1 (HP1) is a component of constitutive heterochromatin. *Cytogenet. Cell Genet.* **1994**, *66*, 99-103.

(44) Minc, E.; Allory, Y.; Worman, H. J.; Courvalin, J.-C.; Buendia, B. Localization and phosphorylation of HP1 proteins during the cell cycle in mammalian cells. *Chromosoma* **1999**, *108*, 220-234.

(45) Minc, E.; Courvalin, J.-C.; Buendia, B. HP1γ associates with euchromatin and heterochromatin in mammalian nuclei and chromosomes. *Cytogenet. Cell Genet.* **2000**, *90*, 279-284.

(46) Saksouk, N.; Barth, T. K.; Ziegler-Birling, C.; Olova, N.; Nowak, A.; Rey, E.; Mateos-Langerak, J.; Urbach, S.; Reik, W.; Torres-Padilla, M.-E.; Imhof, A.; Déjardin, J. Redundant mechanisms to form silent chromatin at pericentromeric regions rely on BEND3 and DNA methylation. *Mol. Cell* **2014**, *56*, 580-594.

(47) Joseph, A.; Mitchell, A. R.; Miller, O. J. The organization of the mouse satellite DNA at centromeres. *Exp. Cell Res.* **1989**, *183*, 494-500.

(48) Li, H. J.; Haque, Z. K.; Chen, A.; Mendelsohn, M. RIF-1, a novel nuclear receptor corepressor that associates with the nuclear matrix. *J. Cell. Biochem.* **2007**, *102*, 1021-1035.

(49) Nozawa, R.-S.; Nagao, K.; Igami, K.-T.; Shibata, S.; Shirai, N.; Nozaki, N.; Sado, T.; Kimura, H.; Obuse, C. Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* **2013**, *20*, 566-573.

(50) Brideau, N. J.; Coker, H.; Gendrel, A.-V.; Siebert, C. A.; Bezstarosti, K.; Demmers, J.; Poot, R. A.; Nesterova, T. B.; Brockdorff, N. Independent mechanisms target SMCHD1 to trimethylated histone H3-modified chromatin and the inactive X chromosome. *Mol. Cell. Biol.* **2015**, *35*, 4053-4068.

CHAPTER 3

CONSTRUCTING HUMAN PROTEOFORM FAMILIES USING INTACT-MASS AND TOP-DOWN PROTEOMICS WITH A MULTI-PROTEASE GLOBAL POST-TRANSLATIONAL MODIFICATION DISCOVERY DATABASE

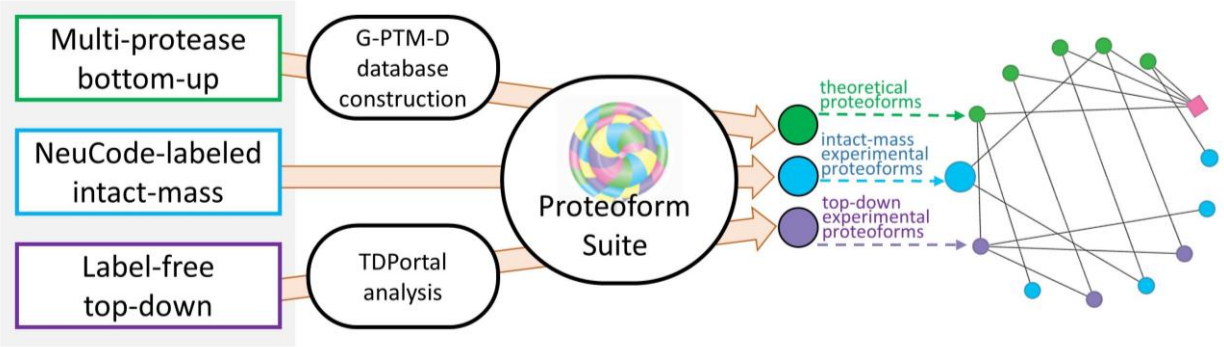
This chapter has been adapted from a publication and was reprinted with permission from:

Dai, Y.;[#] Buxton, K. E.;[#] Schaffer, L. V.; Miller, R. M.; Millikin, R. J.; Scalf, M.; Frey, B. L.; Shortreed, M. R.; Smith, L. M. Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. *J. Proteome Res.* **2019**, *18* (10), 3671-3680. https://doi.org/10.1021/acs.jproteome.9b00339. [#]Y.D. and K.E.B. contributed equally to this work.

3.1 ABSTRACT

Complex human biomolecular processes are made possible by the diversity of human proteoforms. Constructing proteoform families, groups of proteoforms derived from the same gene, is one way to represent this diversity. Comprehensive, high-confidence identification of human proteoforms remains a central challenge in mass spectrometry-based proteomics. We have previously reported a strategy for proteoform identification using intact-mass measurements, and we have since improved that strategy by mass calibration based on search results, the use of a global post-translational

modification discovery database, and the integration of top-down proteomics results with intact-mass analysis. In the present study, we combine these strategies for enhanced proteoform identification in total cell lysate from the Jurkat human T lymphocyte cell line. We collected, processed, and integrated three types of proteomics data (NeuCode-labeled intact-mass, label-free top-down, and multi-protease bottom-up) to maximize the number of confident proteoform identifications. The integrated analysis revealed 5,950 unique experimentally observed proteoforms, which were assembled into 848 proteoform families. Twenty percent of the observed proteoforms were confidently identified at a 3.9% false discovery rate, representing 1,207 unique proteoforms derived from 484 genes.



3.2 INTRODUCTION

The complex biological processes essential for cell survival, development, and homeostasis require a wide variety of proteins. The human proteome originates from roughly 20,000 protein-coding genes, but the complexity of the proteome is then expanded through genetic variations,

alternative splicing, and post-translational modifications (PTMs).[1] Capturing and organizing this molecular complexity is aided by the concepts of the "proteoform", referring to a defined amino acid sequence with a specific set of PTMs, and the "proteoform family", the set of proteoforms derived from the same gene.[2,3] Characterization of proteoforms and elucidation of proteoform families are important emerging areas of proteomic research.

Mass spectrometry (MS)-based analysis of intact protein molecules has developed into a robust and efficient approach to proteoform identification in complex samples.[4] Many studies of intact proteoforms have utilized the top-down strategy, where whole proteins are fragmented in the gas phase and analyzed by tandem MS.[5] Top-down proteomics is advantageous for proteoform analysis because the molecular context of the PTMs is preserved and fragmentation data can provide sequence evidence for identification.[4,6] However, challenges still exist for top-down data acquisition and analysis, as a large fraction of precursor ions are not selected for fragmentation[7,8] and limitations in fragmentation can lead to ambiguous proteoform identifications.[9,10] Proteoform identification without fragmentation is also possible, by inferring identity from accurate proteoform intact-mass measurements. In such a strategy, identification is achieved by relating the masses of experimentally observed proteoforms to those of theoretical proteoforms in databases.[3,7,11] We have previously explored this intact-mass approach to identify proteoforms and proteoform families in both prokaryotic and eukaryotic proteomes[3,12-14] and have streamlined the procedure in the Proteoform Suite software (available at https://smith-chem-wisc.github.io/ProteoformSuite).[15] Although intact-mass proteomics compensates

for some of the limitations of traditional top-down proteomic strategies by making more efficient use of precursor ion data, it is nevertheless challenging to confidently identify proteoforms based on intact-mass measurements alone given the complexity of the proteoform-ome.

We have recently implemented several innovations to improve the number and confidence of proteoform identifications using Proteoform Suite. One of these was NeuCode SILAC (Stable Isotope Labeling by Amino acids in Cell culture),[16-18] which was used to count the number of lysine residues in a proteoform as a second piece of information to leverage during identification.[3,12,15] Post-acquisition mass calibration based on the software lock-mass concept[19] was also introduced to increase the accuracy of intact-mass data, contributing to more proteoform identifications with lower false discovery rates (FDRs).[13,14] We also employed bottom-up proteomics to build global PTM discovery (G-PTM-D)[20,21] databases for enhanced intact-mass proteoform identification.[12] Although bottom-up proteomics by itself does not generally provide sufficient information to identify proteoforms, as intact sequence and PTM context are lost after protease digestion,[22] it is an extremely powerful strategy to produce detailed peptide-level data. Tools such as G-PTM-D allow novel PTM sites to be discovered from bottom-up data, information which may then be used to construct richer and more accurate databases of theoretical proteoforms. We have previously shown that the use of a G-PTM-D database generated from tryptic bottom-up data increased the number of *Escherichia coli* proteoforms that could be confidently identified from intact-mass data[12] and have integrated intact-mass and conventional top-down proteomic analyses to increase proteoform identifications in yeast and murine mitochondria.[13,14]

In the present study, we combine these strategies (NeuCode SILAC, post-acquisition intact-mass calibration, G-PTM-D, and incorporation of top-down data) to identify intact proteoforms in human samples using the Jurkat T lymphocyte cell line as a model system.[23] We extended the G-PTM-D strategy to use multiple proteases instead of only trypsin digestion to increase proteome coverage.[24] Proteoform Suite's functionality was also expanded to accommodate processing of NeuCode-labeled and label-free data together. Multiple recent studies have explored global human proteoform investigation using conventional top-down proteomics,[25-32] and top-down and bottom-up data have been integrated for the purpose of proteoform analysis for more than a decade.[29,33,34] Here, we further explored how different types of proteomics schemes (intact-mass, top-down, and bottom-up) can be integrated to yield the most proteoform-level information from the data collected (Figure 3.1).

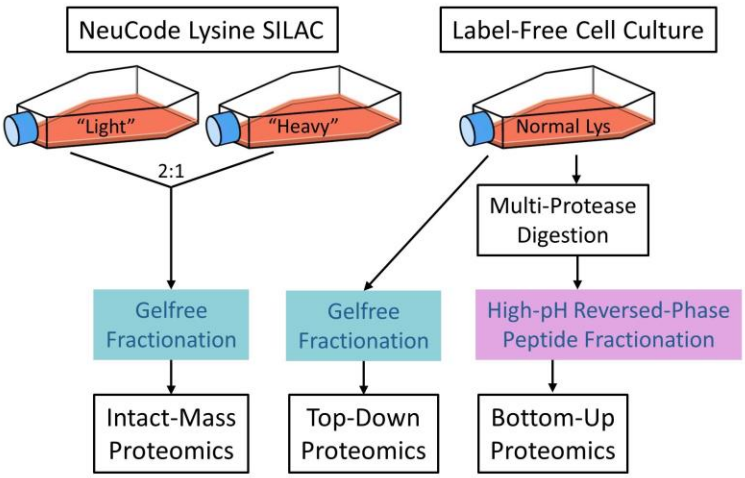


Figure 3.1 Schematic of sample preparation for intact-mass, top-down, and bottom-up proteomics. In this study, "intact-mass proteomics" refers to MS1-only analysis with no precursor fragmentation, while "top-down proteomics" refers to tandem MS analysis with precursor fragmentation and MS2 analysis.

3.3 METHODS

A detailed account of all materials, including their sources, and experimental procedures employed in this work can be found in the Supplementary Information. Brief summaries of these procedures are provided here.

Intact-mass proteomics of NeuCode-labeled Jurkat cells

NeuCode SILAC cell culture: Jurkat cells were cultured at 37 °C under 5% CO_2 in SILAC RPMI-1640 medium supplemented with 10% fetal bovine serum, 1× antibiotic-antimycotic solution, 10 mM HEPES buffer, 1 mM sodium pyruvate, 2 mM GlutaMAX, 1.2 mM L-arginine, and 0.5 mM of either one of two NeuCode lysine isotopologues: "light" ($^{15}N_2{}^{13}C_6$) or "heavy" ($^{2}H_8$).[18] Cells were grown to a density of ~10^6 cells/mL at which time they were washed, pelleted, snap-frozen in liquid nitrogen, and stored at -80 °C until use. Cellular incorporation of NeuCode lysine reached ~99% in cells after approximately five doublings, as determined by bottom-up mass spectrometry.

Protein purification and fractionation: Sample preparation was similar to that described in our previous NeuCode proteoform studies of yeast and *E. coli*.[3,12,15] Briefly, light and heavy NeuCode-labeled Jurkat cells were lysed separately and proteins were reduced and alkylated. Proteins were then precipitated with acetone, resuspended, and mixed in a 2:1 light/heavy ratio (Figure 3.1). The proteins were separated based on molecular weight (MW) using a Gelfree system (Expedeon),[35] and 11 fractions were collected. Prior to mass spectrometric analysis, sodium dodecyl sulfate was removed from the

fractions via methanol–chloroform precipitation[36] and proteins were reconstituted with 5% acetonitrile (ACN) and 0.2% formic acid in water. Three biological replicates of this experiment were performed.

Liquid chromatography/mass spectrometry (LC/MS): All fractions were analyzed by HPLC-ESI-MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Fisher Scientific). Two technical replicate injections of each fraction were performed, yielding a total of 66 raw data files (3 biological replicates × 11 fractions × 2 injections).

Bottom-up proteomics of label-free Jurkat cells

Cell culture and protein digestion using multiple proteases in parallel: Bottom-up proteomic data had been collected previously for five aliquots of Jurkat cell lysate, each of which had been digested with a different protease (chymotrypsin, GluC, ArgC, AspN, or LysC).[37] Briefly, cells were cultured in medium containing normal (i.e., not isotopically labeled) lysine and lysed. Aliquots of lysate were transferred to separate filter units for filter-aided sample preparation (FASP)[38] using different proteases. The resultant peptide samples were each separated into 11 fractions via high-pH reversed-phase liquid chromatography. Each fraction was then dried down and reconstituted in 2% ACN and 0.2% formic acid in water. Additionally, as part of a separate work, this process was repeated to collect 10 fractions of peptides from label-free Jurkat cell lysate digested with trypsin.[39]

LC/MS: Bottom-up analysis was performed via HPLC-ESI-MS/MS (nanoAcquity, Waters and LTQ Velos Orbitrap, Thermo Fisher Scientific) as described previously.[37] The top 10 most intense precursor ions were selected for higher-energy collisional dissociation (HCD) fragmentation via data-dependent acquisition. Dynamic exclusion was enabled. A total of 65 raw data files were collected (10 fractions for trypsin, and 11 fractions for each of the other five proteases).

Top-down proteomics of label-free Jurkat cells

Cell culture and sample preparation: Label-free Jurkat cells were cultured as described for the NeuCode-labeled cells, except that normal lysine was substituted for the heavy lysine isotopologues. Cells were lysed and proteins were extracted as described for the NeuCode-labeled samples. After acetone precipitation, proteins were separated via Gelfree and 11 fractions were prepared for mass spectrometry as described for the NeuCode-labeled samples.

LC/MS: Top-down analysis was performed using HPLC-ESI-MS/MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Fisher Scientific). The top three most intense precursor ions were selected for HCD fragmentation via data-dependent acquisition. Dynamic exclusion was enabled. One biological and two technical replicates were performed, generating 22 raw data files.

Processing and integration of intact-mass, bottom-up, and top-down data sets

The overall workflow for data processing and proteoform analysis is shown in Figure 3.2.

Intact-mass raw data deconvolution: Intact-mass data files (.raw) were deconvoluted into monoisotopic mass components using Thermo Protein Deconvolution 4.0. The outputs of deconvolution (monoisotopic masses) are referred to herein as "raw mass components".

Bottom-up data analysis in MetaMorpheus: The global PTM discovery search workflow[20] was performed on the bottom-up raw data files in MetaMorpheus (v0.0.297, available at https://smith-chem-wisc.github.io/MetaMorpheus).[21] This strategy, which enables discovery of PTMs that are not annotated in UniProt proteome databases, was applied in previous studies to bottom-up data from a single protease (trypsin).[12,20] Here, we adapted the MetaMorpheus software to allow raw data files from samples digested with different proteases to be calibrated and searched at the same time. Protease type was specified for each file in the file-specific search parameters. Data were searched against a UniProt human proteome XML database (73,928 entries, downloaded February 2019) and calibrated based on peptide mapping results. The calibrated data were searched again with selected mass errors allowed. These mass errors reflected common biological PTMs, such as phosphorylation, acetylation, and methylation, as well as common artifacts, such as deamidation, sodium adduction, and ammonia loss (Supplementary Information, Table S-3.1). MetaMorpheus added the modification sites revealed by this G-PTM-D search into the database, thereby generating the "multi-protease G-PTM-D database" in XML format. Finally, all of the calibrated files were searched against this new database. Integrating data from multiple proteolytic digestions increases proteome coverage[24] and improves protein inference in this final search, thereby decreasing the number of ambiguous protein identifications.[39]

A "pruned" version of the multi-protease G-PTM-D database was created, limiting the entries to only those proteins that had confidently identified peptides in deep bottom-up data (1% FDR), along with any UniProt-documented modifications and confident G-PTM-D modifications (1% FDR) for those proteins. This pruned version of the multi-protease G-PTM-D database was utilized by Proteoform Suite during the analysis of intact-mass and top-down data, if not otherwise specified.

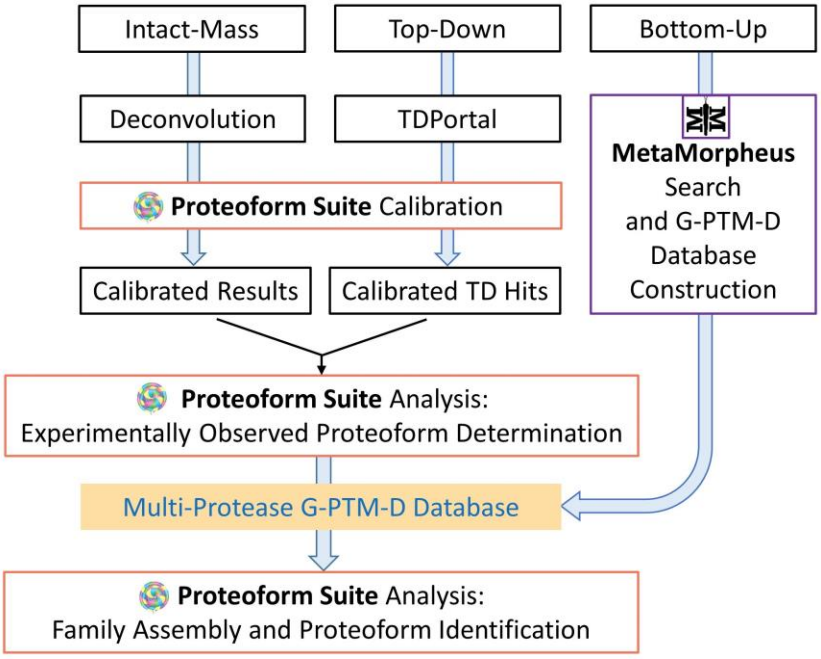


Figure 3.2 Schematic of data processing and analysis for proteoform identification and family construction using intact-mass, top-down, and bottom-up proteomics data.

Top-down raw data processing: Top-down raw data files were analyzed using TDPortal (National Resource for Translational and Developmental Proteomics, NRTDP, Northwestern University, Evanston, IL) as previously described.[13] Files were searched against the human proteome, and

carbamidomethylation of cysteine was set as a fixed modification. A search result table containing all top-down hits (proteoform spectral matches) observed at 1% FDR was generated and used for subsequent data calibration in Proteoform Suite.

Data calibration with Proteoform Suite: Deconvoluted NeuCode intact-mass data and top-down hits obtained from TDPortal were calibrated using Proteoform Suite (v0.3.4) to improve mass accuracy for subsequent proteoform family construction (Figure 3.2). This post-acquisition calibration process utilized a search result-dependent strategy that we initially developed for bottom-up proteomics[21] and have since implemented in Proteoform Suite for intact-mass and top-down proteomics. Its previous application for label-free intact-mass data calibration provided an improvement in mass accuracy, which resulted in an increased number of proteoform identifications and a decreased FDR for identifications.[13,14] Here, we extend this strategy to calibrate intact-mass data collected from NeuCode-labeled proteoforms. The theoretical light NeuCode-labeled mass was determined for each high-confidence (C-score > 40)[40] top-down hit based on the identified sequence's lysine count. Raw mass components from intact-mass measurements were then selected as calibration points if within 10 ppm and 5 min retention time (RT) of a top-down hit from the same Gelfree fraction number. A random forest machine learning algorithm determined the mass error as a function of *m/z*, RT, scan total ion current (TIC), and scan injection time to perform a global calibration for each raw file.[13] New result tables were generated containing calibrated deconvoluted NeuCode-labeled intact-mass data as well as

calibrated label-free top-down hits (Figure 3.2). The resultant 2,021,232 calibrated raw mass components and 39,382 calibrated top-down hits were used for the subsequent analysis.

Proteoform family construction with Proteoform Suite: Proteoform Suite (v0.3.4) was used to further filter intact-mass and top-down data to generate a list of intact-mass and top-down experimental proteoforms. Additionally, Proteoform Suite was used to make a catalog of theoretical proteoforms from the pruned multi-protease G-PTM-D database generated using bottom-up data.

Deconvoluted and calibrated NeuCode intact-mass data files, calibrated top-down hits, and the pruned multi-protease G-PTM-D database were loaded into Proteoform Suite (Figure 3.2). Raw mass components from the intact-mass files were first filtered and merged to eliminate errors from missed monoisotopic masses and charge-state harmonics.[13,15] Proteoform isotopologue pairs (light and heavy NeuCode pairs) were then identified from the processed mass components. Only those NeuCode pairs with light/heavy intensity ratios between 1.8:1 and 2.5:1 were retained based on the most abundant intensity ratio observed at 2.15:1 (Supplementary Information, Figure S-3.1). The number of lysine residues for each NeuCode pair was calculated using the 36 mDa per lysine residue mass difference. NeuCode pairs were then aggregated to eliminate redundant observations of the same proteoform, allowing mass deviations of up to 10 ppm and RT deviations of up to 5 min. In this way, a list of 5,615 intact-mass experimental proteoforms was created, each with a monoisotopic mass, lysine count, and RT.[12-14]

Imported top-down hits were filtered by C-score. Those larger than 40 were retained, as they were judged to be confidently identified and extensively characterized.[40] The filtered hits were then aggregated using two criteria: (i) the same proteoform record (PFR) number assigned by TDPortal and (ii) an RT tolerance of 5 min,[13] generating a list of top-down experimental proteoforms. Each experimental mass was converted to the corresponding light NeuCode-labeled mass based on the number of lysine residues in the identified sequence. This list was combined with the list of intact-mass experimental proteoforms to make a final list of experimental proteoforms.

A catalog of theoretical proteoforms was generated from the pruned multi-protease G-PTM-D protein database, allowing combinations of up to four PTMs on each protein. Note that these theoretical proteoform sequences do not include N-terminal methionine. The strategy of constructing proteoform families has been described previously.[3,12-15] Briefly, all experimental proteoforms were compared with the theoretical proteoforms containing the same number of lysines, forming experimental-theoretical (ET) pairs (Supplementary Information, Figure S-3.2). Experimental proteoforms with the same number of lysines and RT differences of less than 2.5 min were also compared to each other, generating experimental-experimental (EE) pairs (Supplementary Information, Figure S-3.3). ET and EE pairs with FDRs no larger than 25% and mass differences corresponding to ~0 Da (exact matches, ET only), known PTMs, PTM combinations, and amino acid residues were accepted. The average FDR of the accepted pairs was determined to be 5% for ET and 8% for EE as previously described.[3,12] Proteoforms in accepted pairs were grouped into proteoform

families, which were visualized in Cytoscape[41,42] (v3.6.0) as networks with nodes representing proteoforms and edges representing mass differences between proteoforms.

This strategy of proteoform family construction is flexible and can accommodate different combinations of input data sets and various protein databases. In this study, we have performed analyses of the NeuCode intact-mass data using a UniProt database, a pruned trypsin-only G-PTM-D database, and a pruned multi-protease G-PTM-D database. The analysis with the trypsin-only G-PTM-D database yielded more proteoform identifications at a fixed FDR than the analysis with the UniProt database, and the number of identifications was further increased when a multi-protease G-PTM-D database was employed. We also integrated NeuCode intact-mass data with label-free top-down data as described to further improve the analysis. Several other types of analyses were performed (i.e., using an unpruned multi-protease G-PTM-D database, using uncalibrated data, and using top-down data only with MS1 spectra as "label-free intact-mass" data) and are presented in the Supplementary Information for the interested reader. Proteoform Suite analysis of the data described in this study typically takes ~100 min (Supplementary Information, Table S-3.2)

3.4 RESULTS AND DISCUSSION

NeuCode-labeled intact-mass experimental proteoforms

Mass spectra from the 66 raw data files obtained from analysis of NeuCode-labeled intact protein samples were deconvoluted and calibrated to provide 2,021,232 mass components. After Proteoform Suite removed the missed monoisotopic and charge-state harmonic errors,[13,15] a total of 283,634 NeuCode pairs were revealed. Proteoform Suite accepted 113,762 of these pairs falling within the selected intensity ratio range of 1.8:1 to 2.5:1. This range was a parameter decision seeking to retain the highest number of true NeuCode pairs possible while eliminating likely false NeuCode pairs. The accepted NeuCode pairs were aggregated by mass and RT, yielding 5,615 intact-mass experimental proteoforms (Supplementary Information, Table S-3.3). In each section below, we examine the impact of various analysis strategies on the number of these experimental proteoforms that can be confidently identified.

Multi-protease G-PTM-D database improves proteoform identification

The G-PTM-D strategy was developed and implemented in MetaMorpheus to identify PTMs in bottom-up data and subsequently add newly discovered PTMs to a sample-specific protein database.[20,21] We have previously reported that using a G-PTM-D database improves identification of *E. coli* proteoforms from intact-mass data.[12] Here, we demonstrate a further improvement of this strategy by utilizing a pruned multi-protease G-PTM-D database. It is important that the pruned

version of the database was used here as the full, unpruned G-PTM-D database contained many proteins from the original UniProt database that were not confidently observed in bottom-up data and therefore were less likely to be observed in intact-mass data. Pruning the database helps to limit the size of the theoretical proteoform catalog, preventing large FDRs in ET comparisons (see the Supplementary Information for results from an analysis using an unpruned multi-protease G-PTM-D database). The pruned multi-protease G-PTM-D database contained ~83% fewer proteins than the original UniProt database (12,767 vs 73,928 sequences). The sequences in the pruned database contained 14,559 modified residues that were not documented in the UniProt database (Supplementary Information, Table S-3.4), as these modifications were identified during global PTM discovery. Using the sequences and PTMs from this multi-protease G-PTM-D database (allowing combinations of up to four PTMs on each sequence), a catalog of theoretical proteoforms containing 121,602 entries was built by Proteoform Suite.

Employing the strategy described in the Methods section, Proteoform Suite constructed 614 proteoform families from accepted ET pairs (6% FDR) and EE pairs (8% FDR) (Supplementary Information, Table S-3.5). A total of 157 families were unambiguously identified, meaning that they were associated with a single gene. These families contained 532 experimental proteoforms. There were also five ambiguous families (associated with multiple genes) assembled, containing 150 experimental proteoforms. Proteoform Suite determined a subset of the proteoforms in these unambiguous and ambiguous families to be "identified experimental proteoforms", as the program

automatically searches for potentially false EE connections (e.g., delta-mass indicating loss of a PTM that is not found in that family) and excludes such proteoforms from the identified list. Proteoform Suite also removes duplicated proteoforms with the same sequence and PTMs but with RT differences larger than 5 min (this RT tolerance is applied in the upstream aggregation step described above), further consolidating this list to 442 "unique proteoform identifications" (Supplementary Information, Table S-3.6). These identifications are depicted in Figure 3.3 and compared to those obtained from analyses using other databases (discussed below).

The same analysis was repeated using the original UniProt database and a pruned trypsin-only G-PTM-D database (detailed results of these additional Proteoform Suite analyses can be found in the Supplementary Information, Table S-3.7). We found that using the multi-protease database increased the number of unique proteoform identifications by 23% as compared to the original, unmodified UniProt database (442 vs 360), and by 13% as compared to the trypsin-only G-PTM-D database (442 vs 392). In general, the number of PTMs on identified proteoforms also increased (Figure 3.3), as did the average number of experimental proteoforms in identified families (2.8 in UniProt, 3.0 in trypsin-only G-PTM-D, and 3.4 in multi-protease G-PTM-D). The better performance of the G-PTM-D databases is due to decreased database size, which reduces FDR, as well as to the incorporation of additional PTMs discovered in bottom-up data. The multi-protease G-PTM-D analysis provided more identifications than the trypsin-only G-PTM-D analysis, as the use of multiple proteases provides better proteome coverage, leading to a more comprehensive protein database for proteoform analysis. The superiority

of the multi-protease G-PTM-D analysis is also reflected by the highest number of ET proteoform exact matches (~0 Da mass difference) (225 for UniProt, 252 for trypsin-only G-PTM-D, and 317 for multi-protease G-PTM-D).

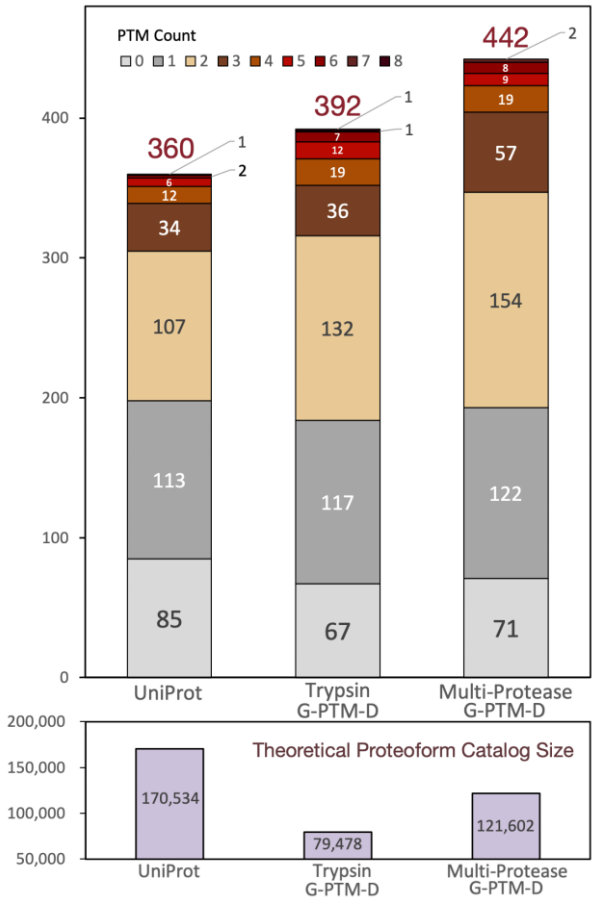


Figure 3.3 Number of identified NeuCode intact-mass experimental proteoforms from Proteoform Suite analyses using three different protein databases. Identified proteoforms were grouped by PTM count (upper panel). The theoretical proteoform catalog size for each analysis is indicated (lower panel). The overall identification FDR for these three analyses was maintained at ~5%.

We also examined the results of proteoform family construction from these analyses. In selecting ET pairs, pairs are grouped into "ET peaks". Each of these peaks has an associated FDR, which reflects the proportion of ET pairs within that peak that are likely false relationships. The ET pairs formed in any given analysis depend on the database used to generate the catalog of theoretical proteoforms. This has a direct impact on the FDR of ET peaks, affecting how many and which of these peaks can be accepted while maintaining the same FDR threshold. This, in turn, determines which theoretical proteoforms are included in proteoform families, and therefore how many experimental proteoforms can be identified. Figure 3.4 demonstrates how two example families evolved when changing the database utilized in the analysis. The non-histone chromosomal protein HMG-14 family gained three new members when the trypsin-only G-PTM-D database was used (and no further growth was observed in a multi-protease-assisted analysis). The G-PTM-D-assisted analyses updated the identity of the 10,752.8 Da proteoform, as it was not directly connected to a theoretical proteoform in the UniProt analysis but could be connected to a theoretical proteoform in the G-PTM-D analyses (identifications via direct ET connections are used for PTM annotation instead of identifications resulting from daisy-chaining EE connections). The 60S ribosomal protein L28 family gradually increased in size by adding first one and then two identified experimental proteoforms when utilizing the trypsin and multi-protease G-PTM-D databases, respectively. The multi-protease-assisted analysis updated the identity of the 15,792.8 Da proteoform, as it became an exact match to a new theoretical

proteoform in the database. These results illustrate how the G-PTM-D strategy improves human proteoform analysis, and how the use of data from multiple proteases further enhances this strategy.

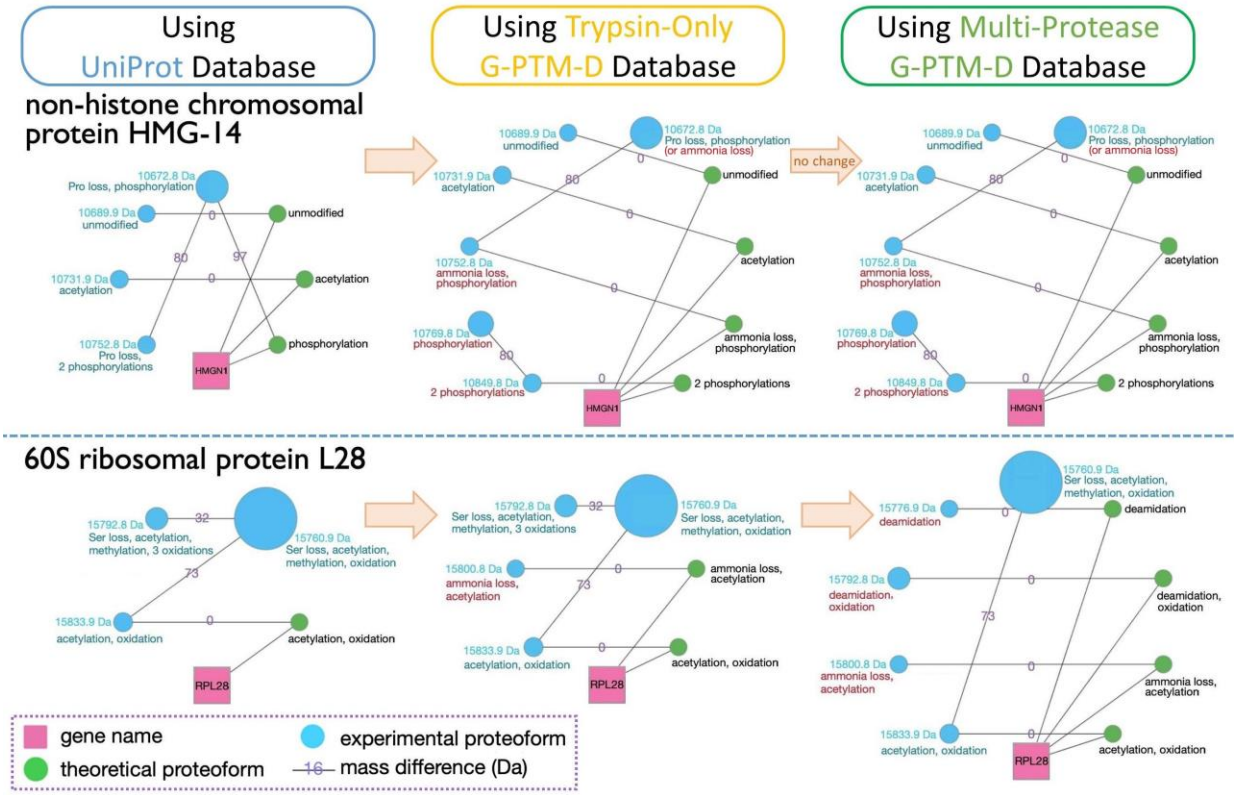


Figure 3.4 Two examples of proteoform families constructed using NeuCode intact-mass data. Three separate Proteoform Suite analyses were performed using UniProt, trypsin-only G-PTM-D, and multi-protease G-PTM-D databases. Gene names (pink squares) connect to all theoretical proteoforms (green nodes) in the family. Theoretical proteoforms are labeled "unmodified" or with PTM information and any terminal amino acid losses. Intact-mass experimental proteoforms (blue nodes) are labeled with their masses and PTMs, as deduced by Proteoform Suite. Experimental proteoforms are arranged counterclockwise in ascending order of mass. The size of each node corresponds to the integrated intensity of that proteoform's spectral peaks. The edges are labeled with the mass difference of the two connected proteoforms (Da). The accepted mass differences are the result of selecting low-FDR ET and EE pairs during the Proteoform Suite analyses. Turquoise annotations are from the UniProt analysis, while red annotations are new findings or PTM corrections gleaned from analyses using G-PTM-D databases.

Top-down experimental proteoforms

The 39,382 calibrated top-down hits obtained from TDPortal analysis of label-free top-down data contained 2,602 unique proteoform record (PFR) numbers. However, this decreased to 1,194 proteoforms (defined by unique PFRs) after filtering for C-scores above 3—the score cutoff that indicates at least partially characterized identifications.[40] A total of 711 proteoforms were identified with a C-score above 40, indicating confident characterizations.[40] These results were similar to those obtained in a previous top-down study of human proteoforms from a single Gelfree separation.[27] In this study, we opted to apply a stringent C-score cutoff of 40 to retain only high-confidence top-down experimental proteoforms for subsequent family construction.

About 100 top-down experimental proteoforms had accession numbers not found in Proteoform Suite's catalog of theoretical proteoforms. There are a few explanations for this. First, the pruned multi-protease G-PTM-D database used to generate the catalog of theoretical proteoforms only contains proteins that were confidently observed in bottom-up data (i.e., proteins that had a peptide observed at 1% FDR). Thus, proteins that may be present in the sample but did not have a confidently identified peptide would not be included in the pruned database. This explanation accounts for the majority of the 100 accessions that were observed in top-down data but not included in the catalog of theoretical proteoforms. Furthermore, the process of protein inference has a significant influence on which protein sequences are included in the pruned database. When multiple proteins have shared subsequences, those shared peptides are mapped to several possible accessions during the protein

inference process. However, based on the principle of Occam's razor, some of these protein sequences may be excluded from the pruned database if there is stronger support for an alternative protein according to peptide-level evidence (e.g., if a unique peptide was also observed for one of the proteins under consideration). To address this discrepancy between the accessions observed in top-down data and the accessions included in the theoretical proteoform catalog, we made a separate database that contained the sequences and PTMs of the proteoforms that were identified by TDPortal but were not found in the pruned multi-protease G-PTM-D database (see the Supplementary Information for further discussion of the entries in this additional database). This "patch database" was imported into Proteoform Suite together with the multi-protease G-PTM-D database. Additionally, top-down identifications whose corresponding theoretical proteoforms were not already present in the catalog were added. The resultant comprehensive catalog contained 123,110 theoretical proteoforms, a modest 1.2% increase in size from the previous catalog, and this catalog was used for the integrated intact-mass/top-down analysis in the next section.

Proteoform family construction using NeuCode intact-mass and top-down experimental proteoforms

Both NeuCode intact-mass and label-free top-down proteomics are useful strategies to identify proteoforms.[3,12,15,26-29] However, each of these approaches has its own advantages. NeuCode intact-mass proteomics generates MS1 spectra only, which means that it provides more proteoform observations than top-down proteomics, where instrument time is spent fragmenting precursors and acquiring MS2

spectra. Top-down proteomics, on the other hand, provides better characterized proteoforms because sequence tags can be identified from fragmentation data. Integrating these two types of data for analysis combines the advantages of each strategy. Previously, we were able to expand label-free top-down proteoform identifications by leveraging additional information contained in the MS1 spectra of the top-down data set.[13,14] This is in contrast to a typical top-down analysis workflow where a substantial number of peaks in MS1 spectra are ignored because they are never selected for fragmentation. In the current work, we enabled Proteoform Suite to construct families by integrating label-free top-down identifications and NeuCode-labeled intact-mass proteoforms that were obtained from separate MS runs. In this integrated intact-mass/top-down analysis, each identified label-free top-down mass was converted to the corresponding light NeuCode mass using the lysine count of the identified sequence. Proteoform Suite merged the 814 top-down experimental proteoforms with the 5,615 intact-mass experimental proteoforms using a mass tolerance of 10 ppm and an RT tolerance of 5 min. As part of this process, 306 intact-mass experimental proteoforms were replaced by top-down experimental proteoforms of the same mass and RT since the identities of these proteoforms had already been deduced from top-down data. After merging, a final list of 6,123 accepted experimental proteoforms was generated (Supplementary Information, Tables S-3.8 and S-3.9).

Using this list and the catalog of 123,110 theoretical proteoforms, 848 families were constructed. These included 438 unambiguously identified families, 10 ambiguous families, and 400 unidentified families (Figure 3.5 and Supplementary Information, Table S-3.10). Overall, we found 526

unique proteoform identifications from the NeuCode intact-mass portion of this integrated analysis, which was an increase from the 442 unique proteoforms identified in the intact-mass-only analysis (Figure 3.3). The reason for this increase is that when the top-down experimental proteoforms were included in the ET and EE comparisons, they generally increased the number of pairs grouped in delta-mass histogram peaks while providing an overall decrease in the FDR of those peaks. This allowed more peaks to have a low enough FDR to be accepted (Supplementary Information, Table S-3.11), which increased the number of proteoforms that were identified. Numerous proteoforms with multiple PTMs were still present in this analysis (Figure 3.5, right). The number of identified proteoforms with 2 PTMs or more increased from the intact-mass-only analysis (Figure 3.3), including one identification with 9 PTMs.

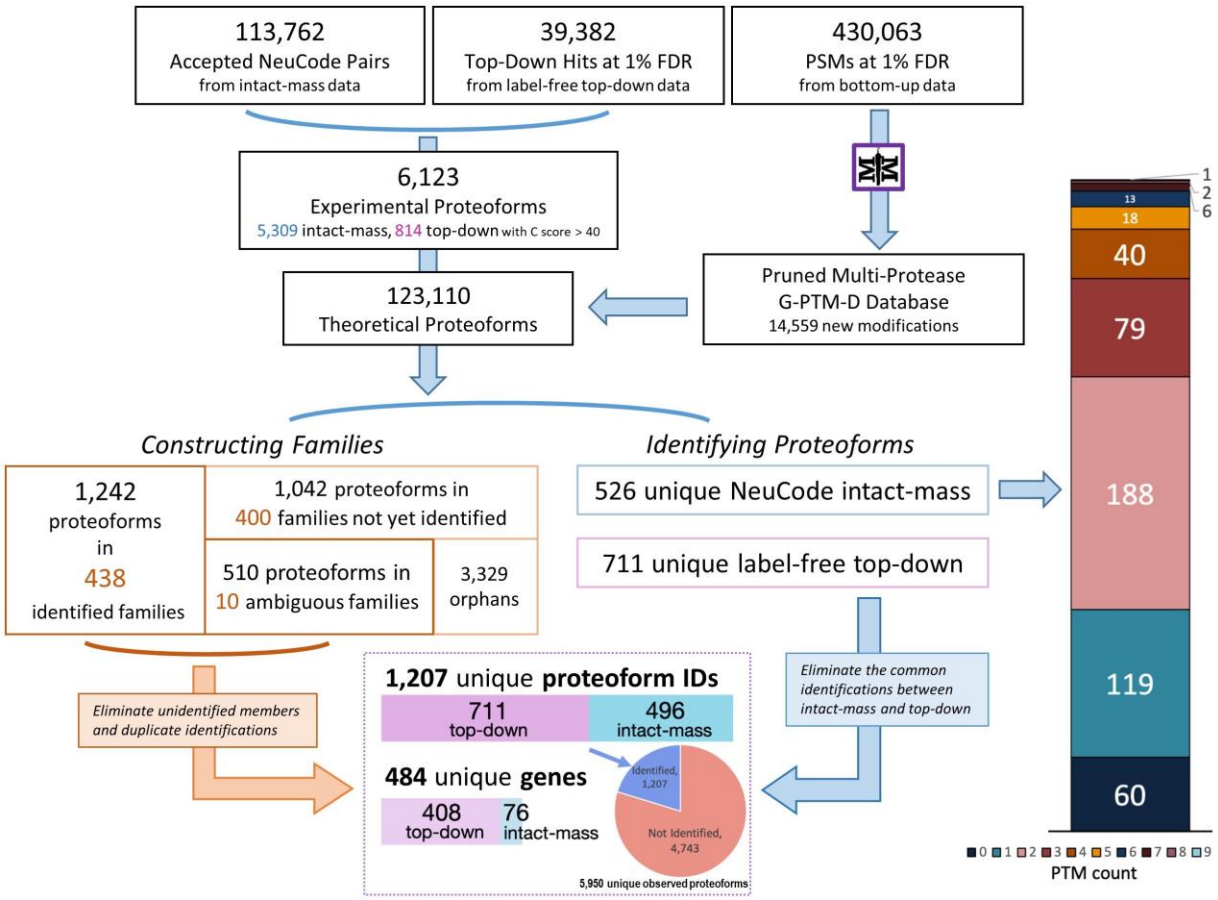


Figure 3.5 Stepwise results of the Proteoform Suite integration of intact-mass, top-down, and bottom-up data. Overall, 1,207 unique proteoforms were identified, representing 484 genes. In the bottom box, only the 496 unique intact-mass proteoforms are depicted, so as to eliminate the common identifications between intact-mass and top-down. See the Supplementary Information, Table S-3.12 for more detailed results of this analysis.

Within the 526 unique intact-mass proteoform identifications, 496 were new additions to the list of 711 unique top-down identifications. This represents a 70% increase in identifications as compared to the TDPortal analysis of top-down data alone (C-score cutoff at 40). Thus, a total of 1,207 unique proteoforms representing 484 genes were identified by integrating intact-mass and top-down data (Supplementary Information, Table S-3.13). The overall FDR of the identified proteoforms was

3.9%. After manual removal of the redundant identifications from the original 6,123 experimentally observed proteoforms, 5,950 unique experimental proteoforms remain, 1,207 (20%) of which were confidently identified (Figure 3.5, bottom box and Supplementary Information, Table S-3.12). These were in a MW range between 1.9 and 30.5 kDa (Supplementary Information, Figure S-3.4). Higher MW proteoforms are not well represented in this study due to limitations of the Orbitrap mass analyzer,[43] generally decreased signal-to-noise ratio for high mass proteoforms,[44] and the elimination of higher MW species in the Gelfree separation employed.[35] The identified proteoforms contained numerous biologically relevant PTMs, including but not limited to methylation, acetylation, and phosphorylation (Supplementary Information, Figure S-3.5). In addition, PTMs that could have either biological or artificial (i.e., sample handling) origin, such as oxidation and deamidation, were also present.

Various functional classes of protein were represented by the proteoforms identified in this integrated intact-mass/top-down analysis, including histones (see the Supplementary Information for a discussion of histone proteoforms and Supplementary Information, Figure S-3.6 for a histone H3 family), ribosomal proteins, RNA-/DNA-binding proteins, transcription and translation factors, transmembrane transporters, and ubiquitin-associated proteins (Supplementary Information, Tables S-3.14 and S-3.15). Among the 438 unambiguously identified families from this integrated analysis (Figure 3.6), 368 families (84%) contained top-down proteoforms. These included the previously introduced non-histone chromosomal protein HMG-14 family (Figure 3.4, upper panel) to which three

top-down experimental proteoforms were added as part of this analysis (Figure 3.6A). The acetylated and the singly phosphorylated top-down proteoforms replaced the intact-mass proteoforms with the same mass and RT. The unmodified top-down proteoform did not merge with the unmodified intact-mass proteoform because the two had RTs larger than 5 min apart; nonetheless, these two proteoforms only count as one unique identified proteoform among the 1,207 reported. In addition, the HMG-14 family gained a new intact-mass proteoform: 10,787.9 Da, containing one methylation and two acetylations. This proteoform was identified through the 98 Da EE pair connection with the unmodified top-down proteoform. Figure 3.6B shows another previously identified family, 60S ribosomal protein L28 (Figure 3.4, lower panel), which also increased in size as compared to the intact-mass-only analysis. Although no top-down proteoforms were added to the family, the incorporation of top-down data and the changes associated with those data led to an additional intact-mass ET pair that fell within an ET peak with sufficiently low FDR for acceptance. Thus, the 15,817.8 Da intact-mass proteoform was identified via an ET match to the theoretical proteoform with one acetylation.

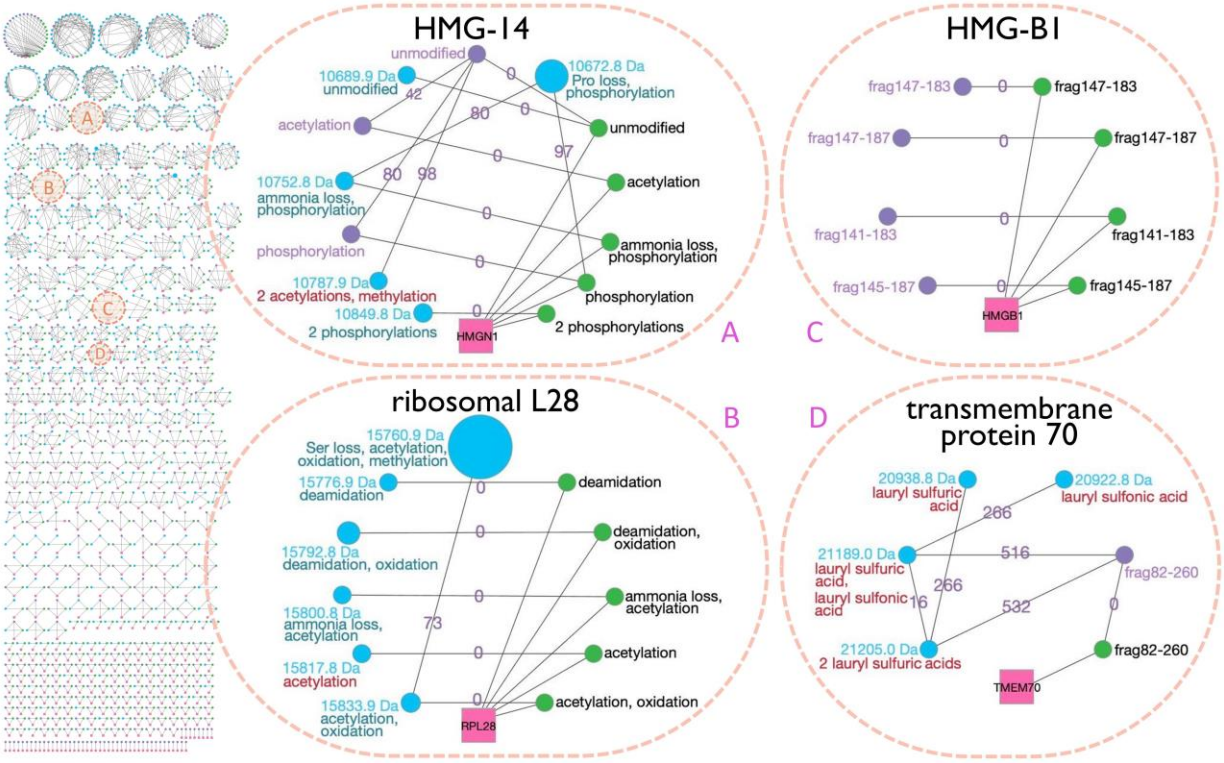


Figure 3.6 Array of the 438 unambiguously identified (i.e., assigned to a single gene) proteoform families that were constructed by the integrated intact-mass/top-down analysis (left) and four example families (right). In addition to the symbols utilized in Figure 3.4, here we add purple nodes to represent top-down experimental proteoforms, and the blue nodes with red annotations denote new intact-mass identifications arising from the inclusion of top-down data in the Proteoform Suite analysis. Previous versions of families A and B were presented in Figure 3.4. The versions presented here show new developments in the families upon integrating top-down data. Families C and D are newly identified proteoform families. Note: the families in this figure were modified slightly from the automated output of Proteoform Suite (i.e., some nodes and edges were removed), as described in the Supplementary Experimental Methods.

The integrated intact-mass/top-down analysis also revealed proteoform families not seen in the intact-mass-only analysis, such as high mobility group protein B1 (Figure 3.6C) and mitochondrial transmembrane protein 70 (Figure 3.6D). The former contained only experimental proteoforms identified by top-down, which were four protein fragments. The latter family contained one top-down

proteoform, which enabled the identification of four intact-mass proteoforms with lauryl sulfuric and lauryl sulfonic acid adducts.

3.5 CONCLUSION

Analysis of all intact-mass and top-down proteomic data files revealed the presence of 5,950 unique experimental proteoforms. Twelve percent (711) of these were identified and extensively characterized by traditional top-down proteomic analysis. The strategy of constructing intact-mass proteoform families with Proteoform Suite and G-PTM-D further increased the fraction of identified proteoforms to 20% (1,207). Development of new strategies for identification of the remaining 80% of observed experimental proteoforms presents an important challenge to the field of proteomics. These proteoforms, which are manifested in high quality mass spectrometric data, yet remain unidentified, represent the front line in top-down/intact-mass analysis as they have already been observed. Overall, in this study, we identified 1,207 human proteoforms at 3.9% FDR. This work demonstrates that the integration of different types of proteomic data at a high confidence level is an effective strategy to substantially increase the quantity and quality of proteoform identifications.

3.6 SUPPLEMENTARY INFORMATION

Large supplementary tables are not included here, but are available online at https://doi.org/10.1021/acs.jproteome.9b00339: Table S-3.3, List of 5,615 aggregated intact-mass experimental proteoforms; Table S-3.4, New modifications in the multi-protease G-PTM-D database (not in UniProt); Table S-3.5, Selected ET and EE mass differences in intact-mass analyses with three different databases; Table S-3.6, List of 442 unique proteoform identifications from intact-mass-only analysis; Table S-3.7, Summary of results and settings for intact-mass-only analyses using three databases; Table S-3.8, List of 6,123 aggregated experimental proteoforms (5,309 intact-mass) from integrated analysis; Table S-3.9, List of 6,123 aggregated experimental proteoforms (814 top-down) from integrated analysis; Table S-3.10, Proteoform families and orphans from integrated intact-mass/top-down analysis; Table S-3.11, Selected ET and EE mass differences in integrated intact-mass/top-down analysis; Table S-3.12, Summary of results and settings for integrated intact-mass/top-down analysis; Table S-3.13, List of 1,207 unique proteoform identifications from integrated intact-mass/top-down analysis; Table S-3.14, List of genes represented by proteoform identifications; Table S-3.15, Gene ontology term analysis of genes represented by proteoform identifications.

Supplementary methods

Materials:

- Jurkat cells (TIB-152) were purchased from the American Type Culture Collection (ATCC) (Manassas, VA).
- SILAC RPMI-1640 medium (A2494401), fetal bovine serum (26400036), antibiotic-antimycotic solution (15240062), HEPES buffer (15630080), sodium pyruvate solution (11360070), GlutaMAX (35050061), 100× HALT protease/phosphatase inhibitor cocktail (78441), and methanol (A452) were purchased from Thermo Fisher Scientific (Waltham, MA).
- L-arginine (A5006), DL-dithiothreitol (DTT) (D5545), sodium butyrate (B5887), iodoacetamide (I1149), acetone (270725), unlabeled L-lysine (62840), 10% sodium dodecyl sulfate (SDS) (71736), and chloroform (319988) were purchased from Sigma-Aldrich Corp. (St. Louis, MO).
- L-lysine:2HCl, $^{13}C_6{}^{15}N_2$ (CNLM-291-H) and L-lysine:2HCl, 3,3,4,4,5,5,6,6 D_8 (DLM-2641) were purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA).
- 4 M Tris-HCl pH=7.5 (T5575), 10× phosphate buffered saline (PBS) (P0195), and 500 mM EDTA pH=8.0 (E0306) were purchased from Teknova (Hollister, CA).
- Gelfree 8100 12% Tris-acetate cartridge (42402), Tris-acetate 5× sample buffer (42302), and HEPES running buffer (42202) were purchased from Expedeon (San Diego, CA).
- Acetonitrile (ACN) (AH015-4) was purchased from Honeywell (Morris Plains, NJ).
- Formic acid (11670) was purchased from EMD Millipore (Burlington, MA).

Cell lysis for NeuCode intact-mass proteomics: For each replicate NeuCode intact-mass experiment, approximately 10^7 "light" and 10^7 "heavy" labeled Jurkat cells were thawed and lysed separately in 1 mL buffer containing 4% (w/v) SDS, 100 mM Tris-HCl pH = 7.5, 10 mM DTT, 10 mM sodium butyrate, 20 mM EDTA, and 1 × HALT protease/phosphatase inhibitors. The cells were incubated at room temperature for 10 min with frequent vortexing, then sonicated in a water bath sonicator (FS20, Fisher Scientific) for 5 min with 20 s on/off intervals. The lysate was incubated for an additional 30 min at

room temperature, then proteins were alkylated with 20 mM iodoacetamide for 30 min. Residual iodoacetamide was quenched via a 15 min incubation with a final concentration of 20 mM DTT. Proteins were precipitated with acetone at -20°C and resuspended in 200 µL of 1% SDS. Proteins from the two NeuCode-labeled samples were then mixed in a 2:1 "light":"heavy" ratio by volume.

Gelfree fractionation: Two ~65 µL aliquots of this mixed protein sample were added to new 1.7 mL tubes. Sample buffer (30 µL), 1 M DTT (8 µL), and water were added to bring the total volume in each tube up to 150 µL. The tubes were incubated at 50 °C for 10 min, then cooled to room temperature. The contents of each tube were fractionated in separate 12% Tris-acetate Gelfree cartridge channels using the manufacturer's recommended procedure. To prepare the channels, storage buffer was removed and replaced with running buffer. Each of the 150 µL samples was loaded, and a standard running method was used to separate the samples into fractions based on molecular weight. Between each step in the method, fractions in the collection chambers of the two channels were combined into a new 2 mL low-retention tube. The collection chambers were rinsed and replenished with new running buffer. These steps were repeated 11 times for each fraction collection. Throughout the run, the running buffer was changed twice according to the standard procedure. The collected fractions were stored at –20°C or 4°C for subsequent sample preparation. Prior to mass spectrometric analysis, SDS was removed from each of the fractions via methanol-chloroform precipitation[36] and proteins were reconstituted with 16 µL of 5% ACN and 0.2% formic acid in water. Intact protein solutions were gently vortexed and centrifuged on a bench-top centrifuge for 1 min. Solutions were carefully

transferred into HPLC sample vials, leaving behind undissolved substances. Three biological replicates of this experiment were performed.

LC/MS methods for NeuCode intact-mass proteomics: All fractions were analyzed by HPLC-ESI-MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Fisher Scientific). HPLC separation employed a 100 × 365 μm fused silica capillary microcolumn packed with 20 cm of 5 μm diameter, 1000 Å pore size PLRP-S resin (Agilent) with an emitter tip pulled to approximately 1 μm using a laser puller (Sutter Instruments). Proteins were loaded on-column at a flow rate of 500 nL/min for 30 min, then eluted at 500 nL/min over 67 min with a gradient of 5% to 85% ACN in 0.2% formic acid. Full-mass profile scans were performed between 500 and 1,600 m/z at a resolution of 240,000. Seven microscans were averaged, using an AGC target of 3×10^6 with a maximum injection time of 200 ms. Source-induced dissociation was set to 15.0 eV. Two technical replicate injections of each fraction were performed, yielding a total of 66 raw data files (3 biological replicates × 11 fractions × 2 injections).

Bottom-up proteomics: Bottom-up proteomics data were collected previously for five aliquots of Jurkat cell lysate, each of which was digested with a different protease (chymotrypsin, GluC, ArgC, AspN, and LysC).[37] Methods for cell culture, lysis, FASP (including different digestion conditions for different proteases), peptide fractionation via high-pH reversed-phase liquid chromatography, and LC/MS are described in detail elsewhere.[37,45] As part of a separate work, this process was repeated to collect MS data for tryptic peptides from unlabeled Jurkat cell lysate.[39] The only alterations to the aforementioned procedure were as follows: proteins were digested in 50 mM ammonium bicarbonate buffer (pH = 7.8)

using a 1:50 trypsin:protein ratio, 10 fractions of peptides were collected (instead of 11), and peptides were reconstituted with 5% ACN/1% formic acid in water prior to LC/MS.

Gelfree fractionation and LC/MS methods for top-down proteomics: Proteins from label-free Jurkat cell lysate were fractionated by Gelfree as described for the NeuCode-labeled samples, except ~110 µL of resuspended protein were fractionated in a single Gelfree channel. Top-down analysis of each of the 11 Gelfree fractions was performed via HPLC–ESI–MS/MS (nanoAcquity, Waters and QE-HF Orbitrap, Thermo Fisher Scientific). The LC method was the same as the intact-mass experiments. MS1 scans were performed between 500 and 1,600 *m/z* at a resolution of 240,000. Seven microscans were averaged, using an AGC target of 1×10^6 with a maximum injection time of 100 ms. The top three most intense peaks in the MS1 with $z > 2$ were selected for HCD fragmentation with a normalized collision energy setting of 25. The MS2 resolution was 120,000, the isolation window was 4 *m/z* units, and three microscans were averaged. Dynamic exclusion was enabled with a duration of 30 s. Source-induced dissociation was set to 15.0 eV. One biological replicate and two technical replicates were performed for this analysis, generating 22 raw data files.

Intact-mass data deconvolution: Intact-mass raw files were deconvoluted into monoisotopic components using Protein Deconvolution 4.0 software (Thermo Fisher Scientific) (minimum S/N = 2, minimum number of detected charge states = 2, fit factor = 70%, remainder threshold = 10%, target average spectrum width = 0.18 min, target average spectrum offset = 34%). Different charge state ranges were selected for deconvoluting different fractions: +5 to +30 for fractions 1–9 and +5 to +50 for

fractions 10–11. Each raw file was split into three to nine retention time (RT) ranges for deconvolution so that the output tables did not exceed allowed spreadsheet size. In the subsequent mass calibration process, calibrated deconvolution files for fractions 1 and 2 were not obtained due to insufficient data points. Therefore, fractions 1 and 2 were not analyzed further.

Proteoform family visualization with description of alterations: The proteoform family figures shown in the main manuscript (Figures 3.4 and 3.6) were modified slightly from the default Cytoscape output to eliminate false connections and improve clarity. In the Figure 3.4 bottom-right family (L28 family built with multi-protease G-PTM-D database), an edge (32 Da) between the 15,760.9 Da and 15,792.8 Da proteoforms was removed, as the new annotation of the 15,792.8 Da proteoform with the multi-protease G-PTM-D database is no longer a doubly-oxidized version of the 15,760.9 Da proteoform. In Figure 3.6 family A, a 10,720.8 Da proteoform was removed, as it was connected to the 10,752.8 Da proteoform with a 32 Da mass difference, but the latter did not contain two oxidations. In families A, C, and D, the edges between gene names and top-down proteoforms were removed for clarity. In both Figures 3.4 and 3.6, the slope of the linear function of node size vs. proteoform intensity has been adjusted for example families to enhance contrast.

Previous publications by the Smith group also contain helpful descriptions of experimental methods for proteoform identification.[3,12-15]

Raw MS data files and G-PTM-D databases

All raw data files and G-PTM-D databases are available on the MassIVE platform (MSV000083768, ftp://massive.ucsd.edu/MSV000083768, and MSV000083304, ftp://massive.ucsd.edu/MSV000083304). There are 66 MS files (in .raw format) from NeuCode-labeled intact-mass proteomics (11 Gelfree fractions of 3 biological replicates with 2 technical replicate injections each). There are 22 MS files from label-free top-down proteomics (11 fractions of one biological replicate with 2 technical replicate injections each). There are also 65 MS files from multi-protease bottom-up proteomics (10 fractions for trypsin and 11 fractions for each of the other five proteases). The pruned trypsin-only G-PTM-D database and the pruned multi-protease G-PTM-D database built from these bottom-up data are also included. The intact-mass data, top-down data, and G-PTM-D databases can be found in data set MSV000083768. The bottom-up data can be found in data set MSV000083304.

Supplementary results

NeuCode intact-mass-only analysis using unpruned multi-protease G-PTM-D database: We performed a Proteoform Suite analysis with calibrated NeuCode intact-mass data but using the full multi-protease G-PTM-D database instead of the pruned database. The full G-PTM-D database contained a large number of protein sequences (~61,000) and PTMs that were not detected in our bottom-up data, which expanded the search space significantly. The full G-PTM-D database was used to construct a theoretical proteoform catalog that contained 426,170 entries, which is 3.5 times larger than pruned database-

derived catalog. Using the same data filtering parameters and selecting the same ET and EE peaks, 561 unique proteoforms were identified in this analysis at 30% FDR (selecting peaks with the same low FDRs as those in the pruned database-assisted analysis was not possible). This high FDR is a result of the large search space of the theoretical proteoform catalog, which contributed to high FDRs in ET pairs (the highest being 134%). This analysis illustrates how using a pruned G-PTM-D database can help to prevent high FDRs during proteoform identification.

Integrated intact-mass/top-down analysis using uncalibrated data: We performed a Proteoform Suite analysis using uncalibrated intact-mass and top-down data with a catalog generated from the pruned multi-protease G-PTM-D database. Raw mass components from these uncalibrated data were filtered with the same parameters reported in the main manuscript. ET comparisons revealed that only 103 pairs grouped to the 0.0185 Da peak (10% FDR), the only peak indicating exact matches. This is only 12% of the 863 exact-matching pairs in the analysis with calibrated data, leading to many fewer identifications in this analysis. Therefore, mass calibration is crucial to effectively identify proteoforms.

Top-down-only analysis utilizing top-down MS1 spectra as "label-free intact-mass" data: We attempted to maximize the utility of the proteomics data collected, so we delved deeper into our label-free top-down data to look for unidentified proteoforms observed in the MS1 spectra. This data analysis strategy was previously employed in yeast and murine mitochondrial proteoform analyses.[13,14]

Precursor ion spectra (MS1) of the top-down data files were extracted. These new top-down MS1-only files were referred to as "label-free intact-mass data files". They were deconvoluted like

NeuCode intact-mass data in Thermo Protein Deconvolution 4.0 to provide label-free raw mass components. The resultant mass components were calibrated as described in the main manuscript. Processed intact-mass data were imported into Proteoform Suite together with calibrated top-down hits. A theoretical proteoform catalog was constructed using the same multi-protease G-PTM-D database and the "patch" databases as described in the main manuscript, except only one PTM was allowed instead of four to prevent high FDRs in the subsequent ET comparison stage. NeuCode pair determination and lysine count calculation were skipped, so ET and EE comparisons proceeded with aggregated proteoform masses. ET and EE pairs were grouped with smaller intervals (0.005 and 0.01 Da, respectively), so that mass difference peaks with relatively low FDRs would be revealed and accepted for family construction.

The label-free intact-mass data further increased the number of unique proteoform identifications beyond those revealed by the integrated NeuCode intact-mass/label-free top-down analysis discussed in the main manuscript. In this label-free intact-mass/top-down analysis, Proteoform Suite identified 880 proteoforms representing 444 genes. The overall FDR for proteoform identification was 5.4%, which is low considering the high noise level of ET and EE comparisons in label-free proteoform analysis (where we do not have the knowledge of lysine count to help limit the number of comparisons). We identified 169 unique intact-mass experimental proteoforms not found in the top-down hits. Among those, 120 proteoforms were new identifications not found in the NeuCode intact-mass proteoform identifications. These proteoforms represented 28 new genes not

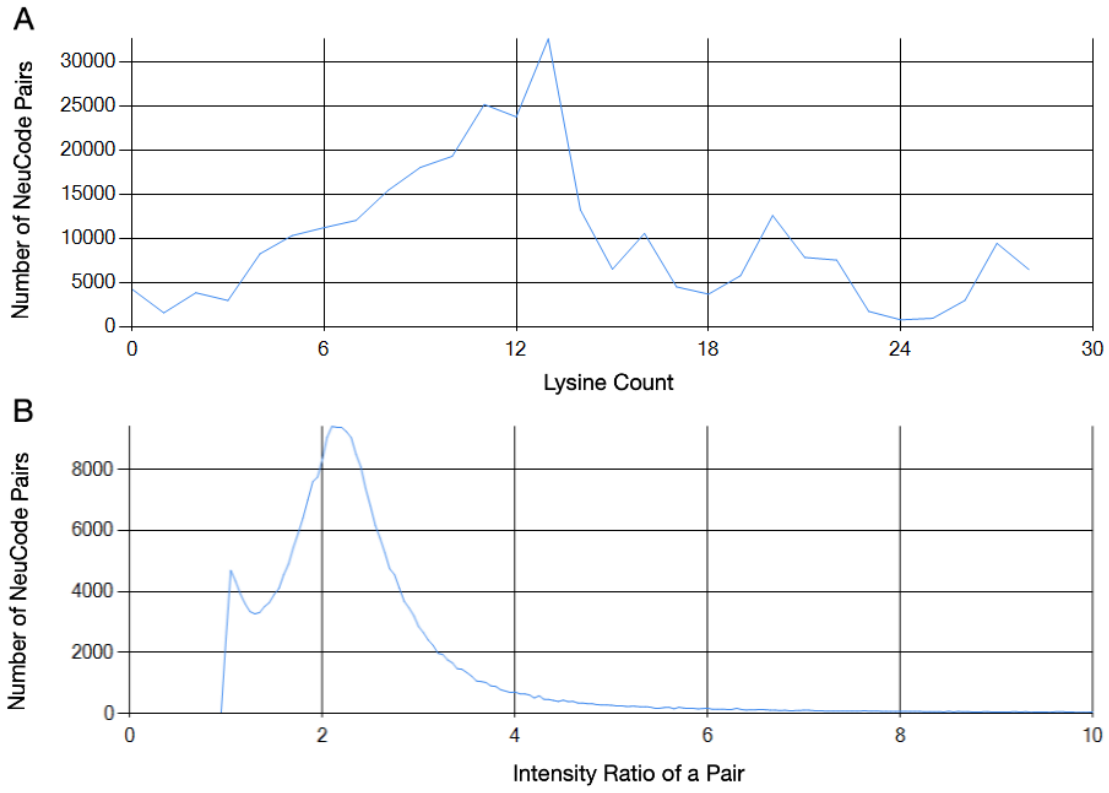
found in the previous analysis. These new identifications contributed to a 10% increase in total unique proteoform identifications, and a 5.8% increase in total genes represented, yielding a final result of 1,327 (1,207 + 120) proteoforms representing 512 (484 + 28) genes. Over 45% of the proteoform identifications came from intact-mass measurements (NeuCode-labeled and/or label-free).

Protein entries in the patch database: The patch database contained 101 protein accession numbers. Fifty-six of these accessions had peptide-level evidence in the multi-protease bottom-up data, but the peptide of interest was observed above a 1% FDR, and therefore the accession was not included in the pruned multi-protease G-PTM-D database. Furthermore, 29 of the 101 accessions in the patch database were included as isoforms. The original UniProt XML database used to search the bottom-up data did not include isoform sequences, which explains why these 29 accessions were not included in the pruned multi-protease G-PTM-D database. That being said, peptide-level evidence for the canonical form of 26 of these proteins was observed in the bottom-up data (FDR > 2%). We posit that many of the remaining 16 accessions in the patch database were eliminated from the pruned multi-protease G-PTM-D database as a result of the protein inference process.

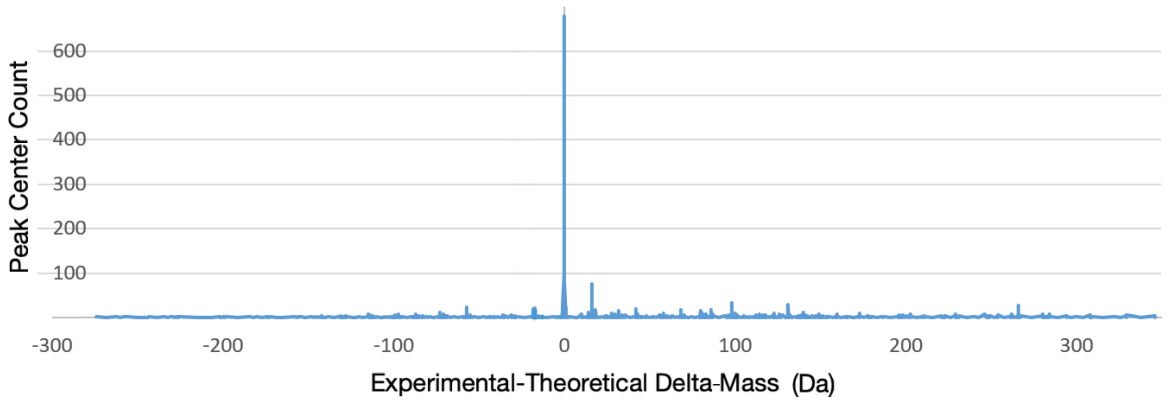
Histone proteoforms and PTMs: Histone proteoforms are known to play important roles in gene expression. In this study, we identified 289 histone proteoforms among the 1,207 reported. These histone proteoforms contained PTMs including methylation, acetylation, phosphorylation, malonylation, succinylation, oxidation, deamidation, and carbamylation. Many of these PTM types play key roles in histone-mediated modulation of gene expression. There were 6 histone proteoform

families assembled, including 5 ambiguously identified families and 1 unambiguously identified family (Supplementary Information, Table S-3.10). Ambiguous histone families account for half of the ambiguous families reported (5 out of 10), and this ambiguity is largely the result of the similar molecular weights of numerous types of histones H2A, H2B, and H3 (e.g., H2B type 1-A, B, C, etc.). It is important to note that, although most histone proteoforms were in the ambiguous families, many of them were still unambiguously identified through direct connections to known theoretical histone proteoforms. The one unambiguous histone family is the Histone H3.1t proteoform family, containing 3 top-down and 3 intact-mass experimental proteoforms (Supplementary Information, Figure S-3.6).

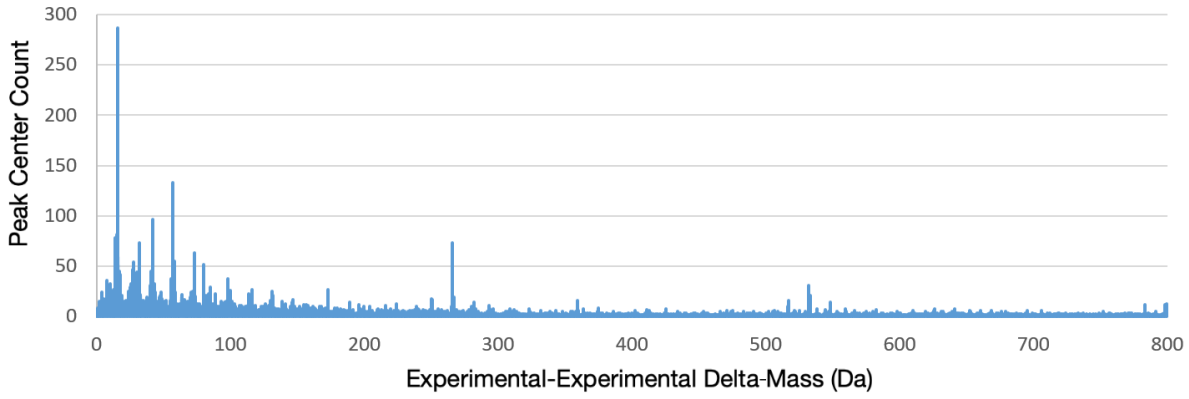
Supplementary figures and tables



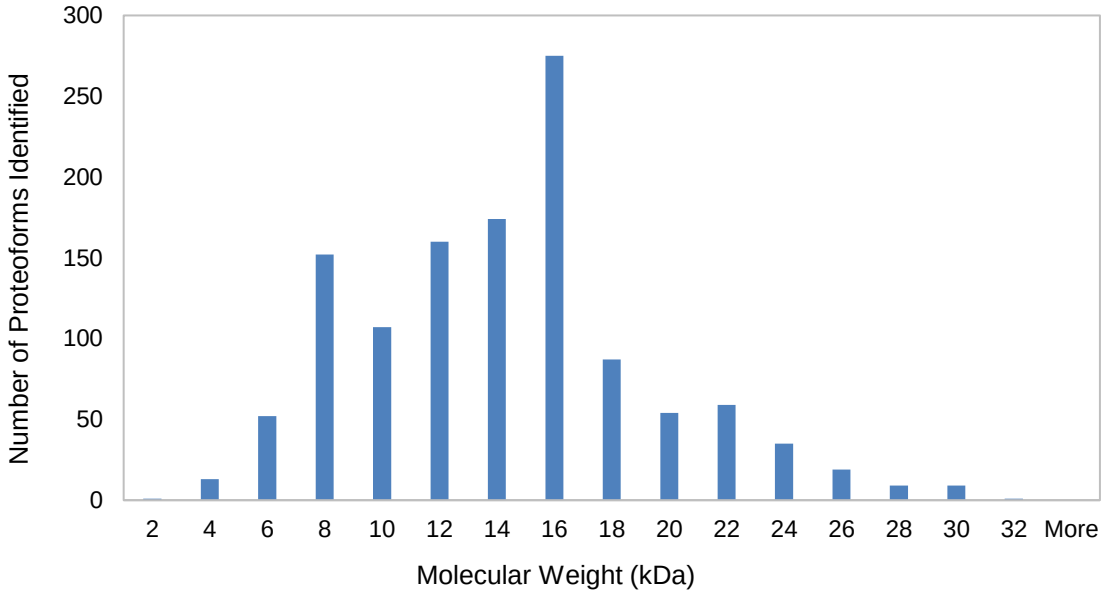
Supplementary Figure S-3.1 Lysine count (A) and intensity ratio (B) distributions of the 283,634 NeuCode pairs that Proteoform Suite identified in this study. The peak intensity ratio was 2.15:1, which was close to the mixing ratio of "light" and "heavy" protein samples at 2:1. NeuCode pairs whose intensity ratios were between 1.8:1 to 2.5:1 were accepted for aggregation into experimental proteoforms.



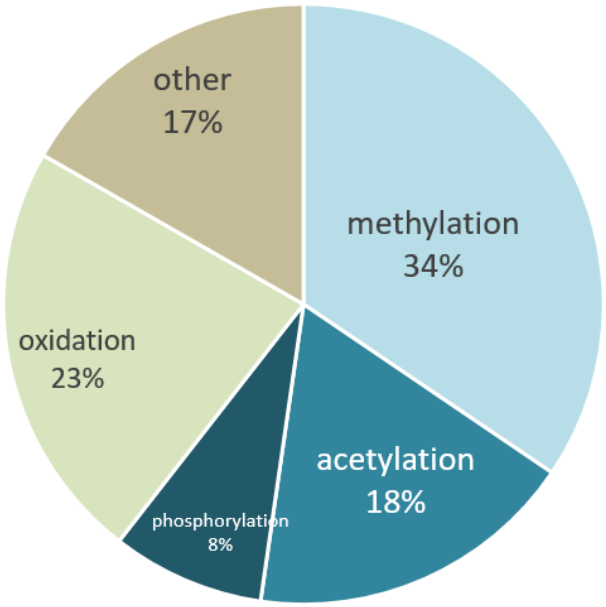
Supplementary Figure S-3.2 Mass difference histogram from experimental-theoretical proteoform comparisons. Analysis was performed using intact-mass and top-down data with a catalog of theoretical proteoforms derived from the pruned multi-protease G-PTM-D database.



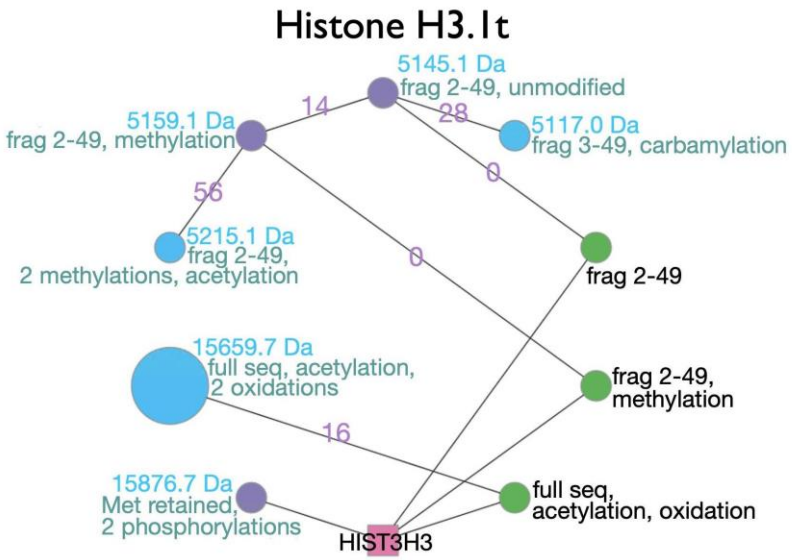
Supplementary Figure S-3.3 Mass difference histogram from experimental-experimental proteoform comparisons. Analysis was performed using intact-mass and top-down data.



Supplementary Figure S-3.4 Molecular weight distribution of the 1,207 identified proteoforms. This result is from the integrated intact-mass/top-down analysis.



Supplementary Figure S-3.5 PTM types found in the 1,207 identified proteoforms. This result is from the integrated intact-mass/top-down analysis.



Supplementary Figure S-3.6 Histone H3.1t proteoform family. Proteoforms in this family contain PTMs such as methylation, acetylation, and phosphorylation that are known to contribute to histone function.

Supplementary Table S-3.1 PTMs and artifacts searched in building the G-PTM-D databases.

PTM/artifact/metal	Residues allowed
Phosphorylation	S, T, or Y
Acetylation	K or protein N-terminal
Methylation	K or R
Demethylation	K or R
Trimethylation	K
Pyroglutamate	Q at protein N-terminal or E at peptide N-terminal
Deamidation	N or Q
Ammonia loss	N (anywhere) or C (peptide N-terminal)
Sodium	D or E
Carbamylation	K, R, C, M, or peptide N-terminal
Potassium	D or E
Calcium	D or E
Iron(II)	D or E
Iron(III)	D or E

Supplementary Table S-3.2 Proteoform Suite data analysis time. Computation time was recorded on an eight-core computer with 32 GB RAM.

Step	Time
Importing files	Manual, ~1 min
Constructing theoretical proteoform catalog	2 min
Aggregating top-down experimental proteoforms	3 min
Extracting raw mass components	30 min
Aggregating intact-mass experimental proteoforms	6 min
Experimental-Theoretical (ET) comparison	5 min
ET peak selection	Manual, ~10 min
Experimental-Experimental (EE) comparison	1 min
EE peak selection	Manual, ~20 min
Proteoform family construction	15 min
Saving data	Manual, ~5 min
Total time	~100 min

3.7 ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences, NIH grants R01GM114292 and R35GM126914. K.E.B. and R.J.M. were supported in part by the National Human Genome Research Institute grant to the Genomic Science Training Program, 5T32HG002760. L.V.S. was supported by the Biotechnology Training Program, T32GM008349. R.M.M. was supported in part by the NIH Chemistry-Biology Interface Training Grant, T32GM008505.

3.8 REFERENCES

(1) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How many human proteoforms are there? *Nat. Chem. Biol.* **2018**, *14*, 206-214.

(2) Smith, L. M.; Kelleher, N. L.; The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186-187.

(3) Shortreed, M. R.; Frey, B. L.; Scalf, M.; Knoener, R. A.; Cesnik, A. J.; Smith, L. M. Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. *J. Proteome Res.* **2016**, *15*, 1213-1221.

(4) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **2016**, *9*, 499-519.

(5) Siuti, N.; Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **2007**, *4*, 817-821.

(6) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018**, *90*, 110-127.

(7) Durbin, K. R.; Tran, J. C.; Zamdborg, L.; Sweet, S. M. M.; Catherman, A. D.; Lee, J. E.; Li, M.; Kellie, J. F.; Kelleher, N. L. Intact mass detection, interpretation, and visualization to automate Top-Down proteomics on a large scale. *Proteomics* **2010**, *10*, 3589-3597.

(8) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. Coupling Capillary Zone Electrophoresis to a Q Exactive HF Mass Spectrometer for Top-down Proteomics: 580 Proteoform Identifications from Yeast. *J. Proteome Res.* **2016**, *15*, 3679-3685.

(9) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolić, N.; Paša-Tolić, L.; Smith, R. D.; Payne, S. H.; Kim, S. Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **2017**, *14*, 909-914.

(10) Dang, X.; Scotcher, J.; Wu, S.; Chu, R. K.; Tolić, N.; Ntai, I.; Thomas, P. M.; Fellers, R. T.; Early, B. P.; Zheng, Y.; Durbin, K. R.; Leduc, R. D.; Wolff, J. J.; Thompson, C. J.; Pan, J.; Han, J.; Shaw, J. B.; Salisbury, J. P.; Easterling, M.; Borchers, C. H.; Brodbelt, J. S.; Agar, J. N.; Paša-Tolić, L.; Kelleher, N. L.; Young, N. L. The first pilot project of the consortium for top-down proteomics: a status report. *Proteomics* **2014**, *14*, 1130-1140.

(11) Karabacak, N. M.; Li, L.; Tiwari, A.; Hayward, L. J.; Hong, P.; Easterling, M. L.; Agar, J. N. Sensitive and specific identification of wild type and variant proteins from 8 to *669* kDa using top-down mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8*, 846-856.

(12) Dai, Y.; Shortreed, M. R.; Scalf, M.; Frey, B. L.; Cesnik, A. J.; Solntsev, S.; Schaffer, L. V.; Smith, L. M. Elucidating Escherichia coli Proteoform Families Using Intact-Mass Proteomics and a Global PTM Discovery Database. *J. Proteome Res.* **2017**, *16*, 4156-4165.

(13) Schaffer, L. V.; Shortreed, M. R.; Cesnik, A. J.; Frey, B. L.; Solntsev, S. K.; Scalf, M.; Smith, L. M. Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families. *Anal. Chem.* **2018**, *90*, 1325-1333.

(14) Schaffer, L. V.; Rensvold, J. W.; Shortreed, M. R.; Cesnik, A. J.; Jochem, A.; Scalf, M.; Frey, B. L.; Pagliarini, D. J.; Smith, L. M. Identification and Quantification of Murine Mitochondrial Proteoforms Using an Integrated Top-Down and Intact-Mass Strategy. *J. Proteome Res.* **2018**, *17*, 3526-3536.

(15) Cesnik, A. J.; Shortreed, M. R.; Schaffer, L. V.; Knoener, R. A.; Frey, B. L.; Scalf, M.; Solntsev, S. K.; Dai, Y.; Gasch, A. P.; Smith, L. M. Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. *J. Proteome Res.* **2018**, *17*, 568-578.

(16) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376-386.

(17) Rhoads, T. W.; Prasad, A.; Kwiecien, N. W.; Merrill, A. E.; Zawack, K.; Westphall, M. S.; Schroeder, F. C.; Kimble, J.; Coon, J. J. NeuCode Labeling in Nematodes: Proteomic and Phosphoproteomic Impact of Ascaroside Treatment in Caenorhabditis elegans. *Mol. Cell. Proteomics* **2015**, *14*, 2922-2935.

(18) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **2013**, *10*, 332-334.

(19) Cox, J.; Michalski, A.; Mann, M. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1373-1380.

(20) Li, Q.; Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Schaffer, L. V.; Scalf, M.; Smith, L. M. Global Post-Translational Modification Discovery. *J. Proteome Res.* **2017**, *16*, 1383-1390.

(21) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17*, 1844-1851.

(22) Chait, B. T. Chemistry. Mass spectrometry: bottom-up or top-down? *Science* **2006**, *314*, 65-66.

(23) Schneider, U.; Schwenk, H. U.; Bornkamm, G. Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int. J. Cancer* **1977**, *19*, 621-626.

(24) Swaney, D. L.; Wenger, C. D.; Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **2010**, *9*, 1323-1329.

(25) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480*, 254-258.

(26) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics* **2013**, *12*, 3465-3473.

(27) Anderson, L. C.; DeHart, C. J.; Kaiser, N. K.; Fellers, R. T.; Smith, D. F.; Greer, J. B.; LeDuc, R. D.; Blakney, G. T.; Thomas, P. M.; Kelleher, N. L.; Hendrickson, C. L. Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *J. Proteome Res.* **2017**, *16*, 1087-1096.

(28) Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L. Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa. *J. Proteome Res.* **2016**, *15*, 976-982.

(29) Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D.; Ellis, M. J. C.; Ruggles, K. V.; Fenyö, D.; Boja, E. S.; Rodriguez, H.; Townsend, R. R.; Kelleher, N. L. Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol. Cell. Proteomics* **2016**, *15*, 45-56.

(30) Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y. Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Anal. Chem.* **2017**, *89*, 5467-5475.

(31) Toby, T. K.; Fornelli, L.; Srzentić, K.; DeHart, C. J.; Levitsky, J.; Friedewald, J.; Kelleher, N. L. A comprehensive pipeline for translational top-down proteomics from a single blood draw. *Nat. Protoc.* **2019**, *14*, 119-152.

(32) Li, Z.; He, B.; Kou, Q.; Wang, Z.; Wu, S.; Liu, Y.; Feng, W.; Liu, X. Evaluation of top-down mass spectral identification with homologous protein sequences. *BMC Bioinformatics* **2018**, *19*, 494.

(33) Millea, K. M.; Krull, I. S.; Cohen, S. A.; Gebler, J. C.; Berger, S. J. Integration of multidimensional chromatographic protein separations with a combined "top-down" and "bottom-up" proteomic strategy. *J. Proteome Res.* **2006**, *5*, 135-146.

(34) Jefferys, S. R.; Giddings, M. C. Baking a mass-spectrometry data PIE with McMC and simulated annealing: predicting protein post-translational modifications from integrated top-down and bottom-up data. *Bioinformatics* **2011**, *27*, 844-852.

(35) Tran, J. C.; Doucette, A. A. Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. *Anal. Chem.* **2008**, *80*, 1568-1573.

(36) Wessel, D.; Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **1984**, *138*, 141-143.

(37) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **2014**, *13*, 228-240.

(38) Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6*, 359-362.

(39) Miller, R. M.; Millikin, R. J.; Hoffmann, C. V.; Solntsev, S. K.; Sheynkman, G. M.; Shortreed, M. R.; Smith, L. M. Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data. *J. Proteome Res.* **2019**, *18*, 3429-3438.

(40) LeDuc, R. D.; Fellers, R. T.; Early, B. P.; Greer, J. B.; Thomas, P. M.; Kelleher, N. L. The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. *J. Proteome Res.* **2014**, *13*, 3231-3240.

(41) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011**, *27*, 431-432.

(42) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498-2504.

(43) Makarov, A.; Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1486-1495.

(44) Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L. On the scalability and requirements of whole protein mass spectrometry. *Anal. Chem.* **2011**, *83*, 6868-6874.

(45) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-seq. *Mol. Cell. Proteomics* **2013,** *12*, 2341-2353.

CHAPTER 4

IDENTIFYING PROTEIN INTERACTOMES OF TARGET RNAs USING HyPR-MS

This chapter has been adapted from a publication and was reproduced with permission from Springer Nature.

Henke, K. B.; Miller, R. M.; Knoener, R. A.; Scalf, M.; Spiniello, M.; Smith, L. M. Identifying protein interactomes of target RNAs using HyPR-MS. In *Post-Transcriptional Gene Regulation*, 3rd ed., Methods in Molecular Biology, vol. 2404; Dassi, E., Ed.; Humana Press: New York, NY, 2022; pp 219-244. https://doi.org/10.1007/978-1-0716-1851-6_12.

4.1 ABSTRACT

RNA—protein interactions are integral to maintaining proper cellular function and homeostasis, and the disruption of key RNA—protein interactions is central to many disease states. HyPR-MS (hybridization purification of RNA—protein complexes followed by mass spectrometry) is a highly versatile and efficient technology which enables multiplexed discovery of specific RNA—protein interactomes. This chapter provides extensive guidance for successful application of HyPR-MS to the system and target RNA(s) of interest, as well as a detailed description of the fundamental HyPR-MS procedure, including: (1) experimental design of controls, capture oligonucleotides, and qPCR

assays; (2) formaldehyde cross-linking of cell culture; (3) cell lysis and RNA solubilization; (4) isolation of target RNA(s); (5) RNA purification and RT-qPCR analysis; (6) protein preparation and mass spectrometric analysis; and (7) mass spectrometric data analysis.

4.2 INTRODUCTION

RNA—protein interactions are crucial to multiple aspects of cellular function and homeostasis. Proteins bind to both coding and noncoding RNA sequences to mediate processes including RNA transcription, splicing, localization, translation, and degradation *(**1–7**)*, and disruptions in RNA—protein interactions are central to many different disease states *(**8–10**)*. Characterizing the protein interactome of a specific RNA is therefore essential to understanding its biology both under normal conditions and in pathological states, and may aid in the discovery of therapeutic targets.

Strategies for the interrogation of RNA—protein interactions can be broadly classified as either protein-centric or RNA-centric *(**11**)*. Protein-centric approaches isolate a protein of interest using immunoprecipitation then identify its associated RNAs, generally through high-throughput RNA sequencing. Such approaches include CLIP (cross-linking immunoprecipitation) *(**12**)* and variants like HITS-CLIP *(**13**)*, PAR-CLIP *(**14**)*, iCLIP *(**15**)*, eCLIP *(**16**)*, and fCLIP *(**17**)*, among others *(**11**)*. Conversely, RNA-centric approaches utilize sequence-specific hybridization probes to isolate a specific RNA, then identify the associated proteins by mass spectrometry. These techniques include CHART-MS *(**18**)*, ChIRP-MS *(**19**)*, RAP-MS *(**20**)*, and HyPR-MS *(**21–24**)*, described here.

HyPR-MS (hybridization purification of RNA—protein complexes followed by mass spectrometry) is a versatile strategy for probing the *in vivo* protein interactomes of one or more target RNAs (Figure 4.1). Briefly, cell culture is subjected to *in vivo* formaldehyde cross-linking to covalently stabilize RNA—protein and protein—protein interactions. After cell lysis, biotinylated capture oligonucleotides, designed to be complementary to the target RNA(s), facilitate the capture of the target RNA—protein complexes. The oligonucleotide—target hybrids are isolated using streptavidin-coated magnetic beads and then released using a toehold-mediated release strategy. The proteins associated with the RNA target(s) are then purified, digested with a protease, and analyzed via mass spectrometry. Strengths of HyPR-MS include its high efficiency and specificity, its versatility, and its capacity for multiplexed discovery of specific RNA—protein interactomes.

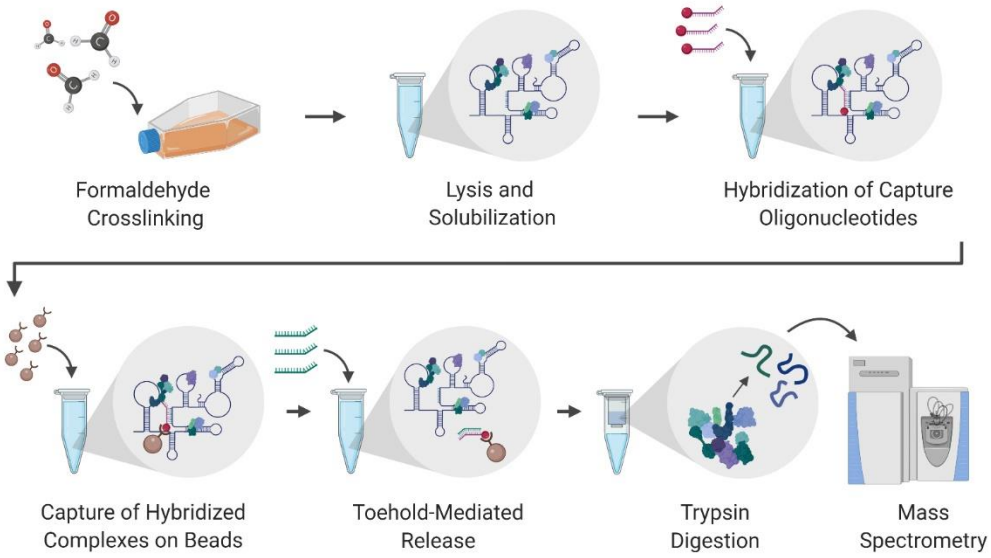


Figure 4.1 Overview of the HyPR-MS technology. HyPR-MS begins with formaldehyde cross-linking of cell culture to covalently fix protein—RNA and protein—protein interactions. Cells are then lysed and RNA—protein complexes are solubilized. Biotinylated capture oligonucleotides (magenta), which include a sequence that is specifically complementary to the target RNA, are then added to the lysate and hybridize to the target RNA. Hybridized complexes are captured on streptavidin-coated magnetic beads, and the beads are washed to remove nonspecific interactors. Release oligonucleotides (green), which are complementary to the entire sequence of the capture oligonucleotide, are then added, and the target RNA—protein complexes are released from the beads via toehold-mediated strand displacement. The proteins are then digested with trypsin prior to mass spectrometric analysis to identify the protein interactome of the target RNA. (This figure was adapted from Knoener et al. ***(24)*** under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

The protocol presented here describes the general design and execution of a HyPR-MS experiment, and offers guidance for successful application to the system and target RNA(s) of interest. The Methods section (Subheading 4.4) describes the HyPR-MS workflow, including (1) experimental design of controls, capture oligonucleotides, and qPCR assays; (2) formaldehyde cross-linking of cells in culture; (3) cell lysis and RNA solubilization; (4) isolation of target RNA(s); (5) RNA purification and RT-qPCR analysis; (6) protein preparation and mass spectrometric analysis; and (7) mass

spectrometric data analysis (Figure 4.2). Importantly, some HyPR-MS parameters are target RNA-specific and may require optimization through empirical testing. Before performing full-scale HyPR-MS experiments for protein identification, we recommend first performing small-scale experiments on relatively few cells to evaluate capture oligonucleotide performance and to establish appropriate capture parameters. These small-scale experiments will provide enough RNA to monitor capture efficiency and specificity via RT-qPCR, but will not provide enough protein for mass spectrometric analysis. Once appropriate capture parameters have been established, one can scale-up to perform the entire HyPR-MS experiment, including mass spectrometric analysis, on a larger number of cells.

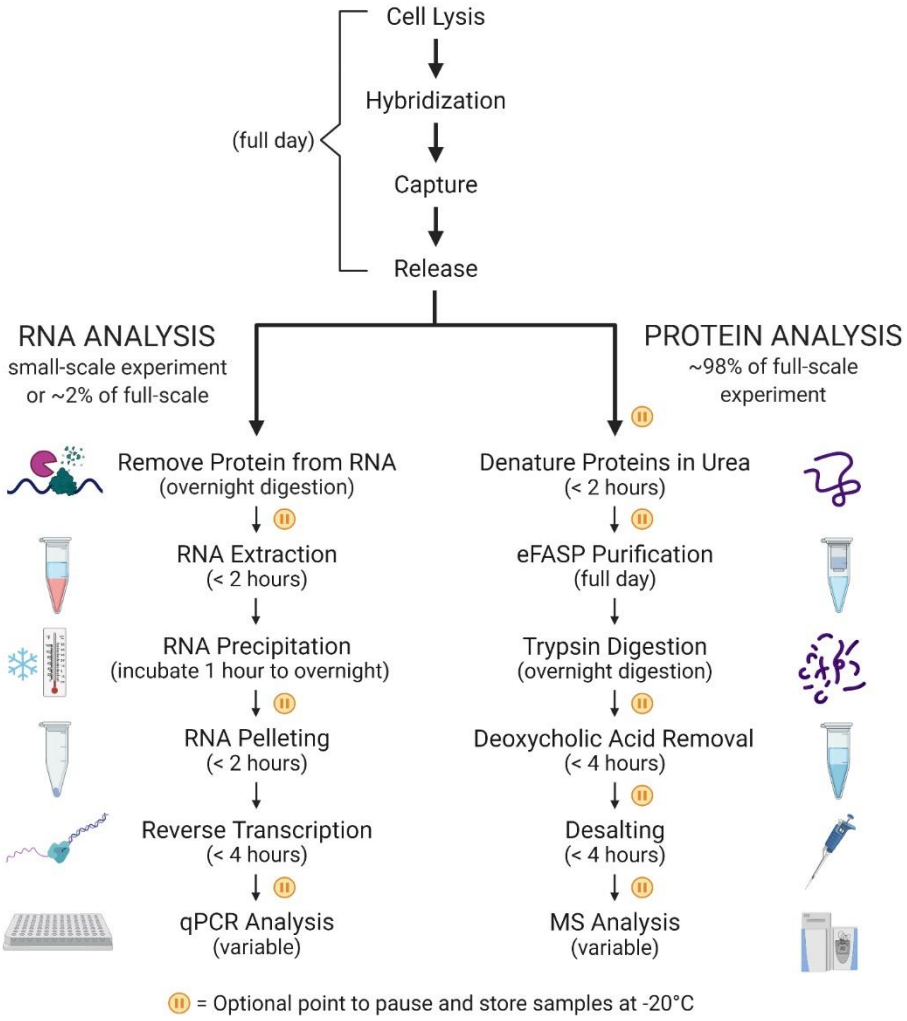


Figure 4.2 RNA and protein purification/analysis steps in HyPR-MS. After isolating the target RNA—protein complexes from cell lysate, the sample should be divided into two aliquots for RNA and protein analysis. For a full-scale experiment, the majority of the sample (~98%) should be used for protein analysis, while ~2% should be used for RNA analysis. For a small-scale experiment, the entire sample can be used for RNA analysis. The major steps for both RNA and protein purification/analysis are indicated, with the approximate duration of each step noted. Optional places to pause the experiment and store the samples at –20 °C are indicated. In general, we recommend completing the entire experiment (from cell lysis through qPCR and mass spectrometric analysis) in the span of approximately one week and minimizing the number of freeze/thaw cycles.

4.3 MATERIALS

4.3.1 Formaldehyde cross-linking

1. Cell line of interest growing in culture.
2. Formaldehyde solution.
3. Tris—HCl solution pH 8.0.
4. Phosphate-buffered saline (PBS).
5. Cell scraper or trypsin (for adherent cell lines).
6. Liquid nitrogen.
7. Orbital shaker.
8. Centrifuge with a swinging-bucket rotor.

4.3.2 Cell lysis, target RNA isolation, and RT-qPCR

All solutions used in this portion of the experiment should be prepared using certified RNase-free components. Similarly, all pipette tips and tubes should be certified RNase-free. The use of RNase decontamination wipes to wipe down pipettes, lab benches, and other lab surfaces prior to beginning this portion of the experiment is highly recommended.

1. Lysis buffer (prepare fresh): 469 mM LiCl, 62.5 mM Tris—HCl pH = 7.5, 1.25% lithium dodecyl sulfate (LiDS), 1.25% Triton X-100, 12.5 mM ribonucleoside vanadyl complex, 12.5 mM dithiothreitol (DTT), 125 U/mL RNasin Plus (Promega), 1.25× protease/phosphatase inhibitor cocktail.
2. Nuclease-free water.
3. Target RNA and control capture oligonucleotides (see Subheading 4.4.1.2 for discussion of capture oligonucleotide design).
4. Streptavidin-coated magnetic Sera-Mag SpeedBeads (Thermo Fisher Scientific).
5. Wash buffer: 375 mM LiCl, 50 mM Tris—HCl pH = 7.5, 0.2% LiDS, 0.2% Triton X-100.
6. Release buffer: 375 mM LiCl, 50 mM Tris—HCl pH = 7.5, 0.1% LiDS, 0.1% Triton X-100.
7. Target RNA and control release oligonucleotides (complementary to respective capture oligonucleotides).
8. Proteinase K.
9. $CaCl_2$ solution.
10. TRI Reagent.
11. Chloroform.
12. Ethanol.

13. Reverse transcription kit.
14. qPCR master mix.
15. Target-specific hydrolysis probe qPCR assays (see Subheading 4.4.1.3 for discussion of qPCR assay design).
16. Vortex mixer.
17. Probe sonicator.
18. Benchtop centrifuge.
19. Low-protein-binding tubes.
20. Nutating mixer.
21. Laboratory incubator.
22. Magnetic tube rack.
23. Low-retention PCR tubes.
24. Thermal cycler.
25. qPCR plates and sealing film.
26. Real-time PCR instrument.

4.3.3 Protein preparation and mass spectrometric analysis

1. 1% 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) solution.
2. LC-MS grade water.

3. Urea.
4. Deoxycholic acid.
5. DTT solution.
6. eFASP exchange buffer (prepare fresh): 8 M urea, 50 mM ammonium bicarbonate, 0.1% deoxycholic acid.
7. eFASP reducing buffer (prepare fresh): 8 M urea, 50 mM ammonium bicarbonate, 20 mM DTT.
8. eFASP alkylation buffer (prepare fresh and protect from light): 8 M urea, 50 mM ammonium bicarbonate, 50 mM iodoacetamide.
9. eFASP digestion buffer (prepare fresh): 1 M urea, 50 mM ammonium bicarbonate, 0.1% deoxycholic acid.
10. Sequencing- or mass spectrometry-grade trypsin.
11. 50 mM ammonium bicarbonate.
12. Trifluoroacetic acid (TFA).
13. Ethyl acetate.
14. LC-MS grade acetonitrile.
15. Formic acid (FA).
16. 50 kDa molecular weight cutoff filters (0.5 mL) and collection tubes.
17. Benchtop centrifuge.
18. Laboratory incubator.

19. Low-protein-binding tubes.
20. Vortex mixer.
21. Vacuum centrifuge concentrator.
22. C18 solid-phase extraction pipette tips (100 µL).
23. nanoAcquity high performance liquid chromatography system (Waters) coupled on-line to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific) (or similar LC-MS setup).
24. LC-MS column: 100 µm id × 365 µm od fused silica capillary microcolumn packed with 20 cm of 1.7 µm diameter, 130 Å pore size C18 beads with an emitter tip pulled to ~1 µm using a laser puller (or similar column).
25. Column oven.

4.4 METHODS

4.4.1 Experimental design

4.4.1.1 Design of control experiment(s): A successful HyPR-MS experimental design must include a control(s) to help determine which proteins identified in a target RNA pulldown sample are target-specific interactors (*see* **Note 1**). Three potential controls are listed below, with descriptions provided in the

corresponding Notes. Alternative controls may be appropriate depending on the goal(s) of the particular HyPR-MS experiment.

1. Scrambled oligonucleotide pulldown control (*see* **Note 2**).
2. Poly(dT) oligonucleotide pulldown control (*see* **Note 3**).
3. Lysate control (*see* **Note 4**).

4.4.1.2 Design of capture oligonucleotides: In HyPR-MS, biotinylated DNA oligonucleotides are used to target and capture RNA molecules of interest in a sequence-specific manner. Designing quality capture oligonucleotides is therefore critical to the success of a HyPR-MS experiment. Several factors to consider when designing capture oligonucleotides are outlined below, with in-depth discussion provided in the associated Notes.

1. Number of capture oligonucleotides required for a target RNA (*see* **Note 5**).
2. Secondary structure of the target RNA (*see* **Note 6**).
3. Capture oligonucleotide specificity and potential for off-target hybridization (*see* **Note 7**).
4. Capture oligonucleotide—target RNA melting temperature (T_m) under HyPR-MS experimental conditions (*see* **Note 8**).
5. Secondary structure of the capture oligonucleotide (*see* **Note 9**).

6. Potential for oligonucleotide—oligonucleotide hybridization (*see* **Note 10**).
7. Toehold-mediated release of the target RNA—protein complexes from the beads (*see* **Note 11**).

4.4.1.3 Design of qPCR assay(s): It is important to monitor RNA capture efficiency and specificity in any HyPR-MS experiment, and RT-qPCR is a useful technique for making these measurements. "Capture efficiency" is the percentage of target RNA present in the lysate at the beginning of the experiment that is captured on and subsequently released from the streptavidin-coated beads. "Capture specificity" is the ratio of target RNA captured using the target RNA capture oligonucleotide(s) relative to that captured by a scrambled oligonucleotide or other negative control (*see* **Note 12**). Typically, one qPCR assay should be designed within ~500 nt of each capture oligonucleotide to monitor the capture efficiency of that region of the target RNA (*see* **Notes 13 and 14**). Designing at least one qPCR assay to monitor the presence of a nontarget, housekeeping RNA, such as GAPDH, is also recommended.

4.4.2 Formaldehyde cross-linking

All mammalian cell lines investigated thus far have been amenable to HyPR-MS, including both suspension and adherent cell lines ***(21–24)***. Cells have been cultured using

standard medium (e.g., DMEM with 10% fetal bovine serum and 1% penicillin—streptomycin). Appropriate cell culture conditions should be determined for the specific cell line of interest. In general, we have found that small-scale HyPR-MS experiments require on the order of 10^5-10^6 cells, while full-scale experiments require on the order of 10^7-10^8 cells (*see* **Note 15**).

1. Add formaldehyde to the cell culture medium to a final concentration of 1% and gently shake for 10 min at room temperature (*see* **Note 16**).
2. Add Tris—HCl pH = 8.0 to 250 mM and gently shake for 10 min at room temperature to quench excess formaldehyde.
3. Remove the culture medium and wash the cells twice with cold 1× PBS. For suspension cell culture, collect the cells via centrifugation (~125 × *g* for 10 min at 4 °C) after each wash and pipet off and discard the supernatant. For adherent cell culture, washing can be done in the culture plate. After the second wash, use trypsin or a cell scraper to detach the cells from the plate and transfer the cell suspension to a centrifuge tube. Collect the cells via centrifugation (~125 × *g* for 10 min at 4 °C) and pipet off and discard the supernatant.
4. Flash-freeze the cell pellet with liquid nitrogen and store at –80 °C (*see* **Note 17**).

4.4.3 Cell lysis and RNA solubilization

1. Thaw the cell pellet on ice and resuspend the cells in freshly prepared, cold lysis buffer to a concentration of 5×10^6 cells/mL (*see* **Notes 15 and 18**).
2. Lyse the cells on ice for 10 min, vortexing periodically to help break up the cell pellet.
3. Sonicate the lysate with a probe sonicator to break up chromatin and solubilize RNA—protein complexes (*see* **Note 19**).
4. Centrifuge the lysate at 1,000 × *g* for 2 min at 4 °C to pellet any insoluble material. Transfer the supernatant to a new low-protein-binding tube.
5. Reserve two aliquots of the clarified cell lysate for downstream RT-qPCR (~2% of the total lysate volume) and mass spectrometric (~20 µL aliquot) analyses. Store these aliquots in low-protein-binding tubes at 4 °C until ready for use.

4.4.4 Hybridization capture and elution

1. Add the appropriate amount of capture oligonucleotide(s) to the lysate (*see* **Note 20**), then add RNase-free water to increase the lysate volume by 25% (new buffer component concentrations for hybridization: 375 mM LiCl, 50 mM Tris—HCl, 1% LiDS, 1% Triton X-100, 10 mM ribonucleoside vanadyl complex, 10 mM DTT, 100 U/mL RNasin Plus, 1× protease/phosphatase inhibitors).
2. Gently rock the lysate at 37 °C for 3 h to allow for hybridization (*see* **Note 21**).

3. Toward the end of the hybridization period (~20 min remaining), transfer an appropriate volume of streptavidin-coated magnetic beads to a low-protein-binding tube (henceforth, this volume will be referred to as the "bead volume") (*see* **Note 22**). Place the tube on a magnet stand and wait 3-4 min for the beads to be drawn to the magnet. Remove the supernatant with a pipette and discard.
4. Remove the tube of beads from the magnet stand and resuspend the beads in one bead volume of wash buffer. Place the tube on a magnet stand and wait 3-4 min for the beads to be drawn to the magnet. Remove the supernatant with a pipette and discard.
5. Repeat step 4 twice more, for a total of three washes.
6. Remove the tube of beads from the magnet stand and resuspend the beads in one bead volume of 37 °C wash buffer.
7. Once the 3 h hybridization period is complete, add the bead slurry to the lysate and gently rock the lysate—bead mixture at 37 °C for 1 h to capture the hybridized RNA—protein complexes (*see* **Note 23**). Ensure that the rocking is sufficient to prevent the beads from aggregating at the bottom of the tube during capture.
8. After the 1 h incubation period, place the tube containing the lysate—bead mixture on a magnet stand and wait 3-4 min for the beads to be drawn to the

magnet. Remove the supernatant with a pipette and store at 4 °C in a low-protein-binding tube for eventual RT-qPCR analysis.

9. Resuspend the beads in one bead volume of 37 °C wash buffer and gently rock at 37 °C for 15 min.

10. Place the tube on a magnet stand and wait 3-4 min for the beads to be drawn to the magnet. Remove the supernatant with a pipette and discard.

11. Repeat steps 9-10 for a second wash.

12. Resuspend the beads in one bead volume of release buffer and gently rock at room temperature for 5 min.

13. Place the tube on a magnet stand and wait 3-4 min for the beads to be drawn to the magnet. Remove the supernatant with a pipette and discard.

14. Resuspend the beads in one bead volume of release buffer and add release oligonucleotide(s) (*see* **Note 24**).

15. Gently rock the bead slurry for 30 min at room temperature to release the target RNA—protein complexes from the beads.

16. Place the tube on a magnet stand and wait 3-4 min for the beads to be drawn to the magnet. Remove the supernatant with a pipette and transfer to a new low-protein-binding tube. Store the supernatant at 4 °C for eventual RT-qPCR and mass spectrometric analyses (*see* **Note 25**).

17. Resuspend the beads in one bead volume of release buffer and store the bead slurry at 4 °C for eventual RT-qPCR analysis (*see* **Note 26**).

4.4.5 RNA purification and RT-qPCR analysis

Small aliquots of the precapture lysate sample (from Subheading 4.4.3, step 5), the postcapture lysate sample (from Subheading 4.4.4, step 8), the RNA capture sample(s) (from Subheading 4.4.4, step 16), and the bead sample (from Subheading 4.4.4, step 17) should be prepared for RT-qPCR analysis. It is important to note what proportion (percentage) of each sample is used for RT-qPCR, as this information is critical for downstream calculations. In general, aliquoting ~2% of the volume of each sample for RT-qPCR analysis is sufficient (*see* **Note 27**) (Figure 4.2).

4.4.5.1 RNA purification

1. Bring all RT-qPCR aliquots to the same final volume (300 µL) and same buffer component concentrations as the postcapture lysate sample (from Subheading 4.4.4, step 8). Add $CaCl_2$ to 4 mM and proteinase K to 1 mg/mL.
2. Gently rock the samples at 37 °C overnight to digest proteins.
3. After allowing the samples to cool to room temperature, add 500 µL of TRI Reagent to each sample and vortex.

4. Allow the samples to sit at room temperature for 5 min, vortexing periodically.
5. Add 100 µL of chloroform to each sample, shake the samples vigorously for 15 s, and allow to sit at room temperature for 10 min.
6. Centrifuge the samples at 12,000 × *g* for 15 min at 4 °C and quantitatively transfer the top, aqueous layer to a clean tube (*see* **Note 28**).
7. Add ethanol to each sample to a final concentration of 75% (*see* **Note 29**) and incubate at –20 °C for at least 1 h or overnight.
8. Centrifuge the samples at 20,800 × *g* for 15 min at 4 °C to pellet the RNA.
9. Carefully remove the supernatant with a pipette and discard (*see* **Note 30**).
10. Wash each RNA pellet with 750 µL of room temperature 75% ethanol.
11. Centrifuge the samples at 20,800 × *g* for 15 min at room temperature to pellet the RNA.
12. Carefully remove the supernatant with a pipette and discard (*see* **Note 30**).

13. Allow the RNA pellets to air dry for ~5 min, then resuspend each pellet in 15 µL of RNase-free water. Store samples at 4 °C prior to performing reverse transcription.

4.4.5.2 Reverse transcription

1. Reverse transcription should be performed according to the kit manufacturer's protocol (*see* **Note 31**). We advise including several control reactions (*see* **Note 32**).
2. After reverse transcription, store the cDNA samples at 4 °C if qPCR will be performed on the same day, or at –20 °C if performed on a different day.

4.4.5.3 qPCR

1. Prepare qPCR plate(s) according to the qPCR master mix manufacturer's protocol using the cDNA samples and controls from Subheading 4.4.5.2 (*see* **Note 33**).
2. Perform qPCR using cycling parameters appropriate for each qPCR assay of interest (*see* **Note 34**).
3. Calculate capture efficiency (*see* **Note 35**), release efficiency (*see* **Note 36**), capture specificity (*see* **Note 37**), and target RNA enrichment (*see*

Note 38) based on the qPCR results. Details about these calculations are discussed in the corresponding Notes.

4.4.6 Protein preparation and mass spectrometric analysis

Samples to prepare for mass spectrometric analysis include the precapture lysate sample (from Subheading 4.4.3, step 5) and all target RNA/control capture samples (from Subheading 4.4.4, step 16). The eFASP procedure described below has been adapted from the method described by Erde et al. *(25)* (Figure 4.2).

4.4.6.1 eFASP

1. Prepare one CHAPS-passivated molecular weight cutoff filter and collection tube per sample one day prior to performing eFASP. Passivation is achieved by filling the collection tube with ~1 mL of 1% CHAPS, inserting the filter, filling the filter with ~0.5 mL of 1% CHAPS, and letting the tube/filter sit overnight at room temperature. Prior to use, the CHAPS solution should be discarded, and the tube and filter rinsed at least five times to remove excess CHAPS. For each rinse, the tube/filter should be placed in a clean beaker containing a large volume of LC-MS grade water and gently stirred for 30 min.
2. Add solid urea and deoxycholic acid to bring each protein sample to 8 M urea, 0.1% deoxycholic acid.

3. For each sample, place a passivated filter inside a non-passivated collection tube and add 450 µL of sample to the filter. Centrifuge the sample at 14,000 × *g* for 10 min at room temperature and discard the flow-through. Continue passing the sample through the filter in this manner until the entire sample volume has passed through.
4. Add 400 µL of eFASP exchange buffer to the filter and centrifuge at 14,000 × *g* for 10 min at room temperature. Discard the flow-through. Repeat twice more, for a total of three washes.
5. Add 200 µL of eFASP reducing buffer to the filter and incubate at room temperature for 30 min.
6. Centrifuge the sample at 14,000 × *g* for 10 min at room temperature and discard the flow-through.
7. Add 200 µL of eFASP alkylation buffer to the filter and incubate at room temperature in the dark for 1 h.
8. After the 1 h incubation period, add DTT to 75 mM and incubate at room temperature for an additional 10 min.
9. Centrifuge the sample at 14,000 × *g* for 10 min at room temperature and discard the flow-through.

10. Add 400 µL of eFASP digestion buffer to the filter and centrifuge at 14,000 × *g* for 10 min at room temperature. Discard the flow-through. Repeat twice more, for a total of three washes.
11. Transfer the filter to a clean, passivated collection tube and add 100 µL of eFASP digestion buffer containing an appropriate amount of trypsin to the filter (*see* **Note 39**).
12. Close the tube cap and seal with parafilm. Incubate the tube/filter apparatus overnight at 37 °C without rocking.
13. Remove the parafilm and centrifuge the sample at 14,000 × *g* for 10 min at room temperature.
14. Leaving the flow-through in the bottom of the tube, add 50 µL of 50 mM ammonium bicarbonate to the filter and centrifuge at 14,000 × *g* for 10 min at room temperature. Repeat once more for a total of two washes, each time allowing the wash volume to accumulate in the bottom of the tube.
15. Transfer the complete flow-through volume (200 µL) to a clean, low-protein-binding tube and add 200 µL of ethyl acetate.
16. Add trifluoroacetic acid (TFA) to 0.5% and vortex the sample for 1 min.
17. Centrifuge the sample at 15,800 × *g* for 2 min at room temperature.
18. Remove the top (ethyl acetate) layer with a pipette and discard.

19. Repeat the ethyl acetate extraction twice more, for a total of three extractions. Each time, add 200 µL of ethyl acetate and vortex the sample for 1 min prior to repeating steps 17-18.
20. Dry the peptide sample in a vacuum centrifuge concentrator.

4.4.6.2 C18 solid-phase extraction

1. Reconstitute the dried peptide sample in 150 µL of 0.1% TFA.
2. Condition a C18 solid-phase extraction pipette tip by washing it at least three times with 150 µL aliquots of 70% acetonitrile (ACN). All pipetting with the C18 tip should be performed slowly.
3. Equilibrate the tip by washing it at least three times with 150 µL aliquots of 0.1% TFA.
4. Load the peptides onto the tip by pipetting the complete peptide sample up and down at least five times.
5. Wash the tip at least ten times with 150 µL aliquots of 0.1% TFA.
6. Elute the peptides from the tip by pipetting a 150 µL aliquot of 70% ACN/0.1% TFA up and down at least five times.
7. Dry the desalted peptides in a vacuum centrifuge concentrator and reconstitute in 95:5 H_2O:ACN with 0.2% formic acid (FA). Store the sample at –20 °C prior to mass spectrometric analysis.

4.4.6.3 Mass spectrometry

We describe here the analysis of HyPR-MS samples using a high-performance liquid chromatography system (nanoAcquity, Waters) coupled to an electrospray ionization (ESI) orbitrap mass spectrometer (Q Exactive HF, Thermo Fisher Scientific). The column (described in Subheading 4.3.3) should be operated at 60 °C using a column oven. Comparable LC-MS setups could also be used. Similarly, the LC-MS/MS method described below has worked well in our hands for the analysis of HyPR-MS samples, but a variety of routine bottom-up LC-MS/MS methods may be acceptable. Provided that the samples contain enough peptides, we recommend performing two technical replicate LC-MS/MS injections of each sample. We also recommend running ACN and water blank injections between samples to minimize carryover from previous injections.

1. LC method: Load peptides on-column with 2% ACN in 0.2% FA at a flow rate of 400 nL/min for 30 min. Elute peptides over 120 min at a flow rate of 300 nL/min with the following gradient (all in 0.2% FA): 8% ACN at time 1 min; 34% ACN at time 81 min; 44% ACN at time 91 min; 64% ACN from 92-99 min; equilibrate to 2% ACN from 103-120 min.

2. MS/MS method: Perform full-mass profile scans (375-1,500 *m/z*) in the orbitrap at a resolution of 120,000, automatic gain control (AGC) target of 1×10^6, and maximum injection time of 100 ms. Follow each full-mass profile scan by MS/MS HCD scans of the 10 highest intensity parent ions with $z > 1$ at 30% relative collision energy and 15,000 resolution with a mass range starting at 100 *m/z*, AGC target of 1×10^5, and a maximum injection time of 50 ms. Enable dynamic exclusion with a repeat count of one over a duration of 15 s.

4.4.7 Mass spectrometric data analysis

In general, the goal of mass spectrometric data analysis in HyPR-MS is to identify which proteins are significantly enriched in the target RNA capture sample as compared to control samples, indicating their interaction with the target RNA. Because the determination of enriched proteins is a statistical process, consideration should be given to how many biological replicates will be necessary for drawing meaningful conclusions from the experiment. Typically, at least three replicates are necessary, though more replicates are helpful for obtaining greater confidence in the results.

1. Obtain a protein database for the relevant organism (*see* **Note 40**).
2. Load the protein database and all spectral files from the target and control samples into a suitable proteomic search software program and perform the

search (*see* **Note 41**). The default search parameters for many search software programs will be sufficient. Ensure that oxidation of methionine is set as a variable modification and that carbamidomethylation of cysteine is set as a fixed modification. Also ensure that trypsin is selected as the protease to be used for *in silico* digestion and set the maximum number of missed cleavages to two and the minimum peptide length to seven amino acids.

3. Filter the search results to remove low-confidence identifications (e.g., apply a 1% false discovery rate (FDR) for both peptides and proteins) (*see* **Note 42**).

4. Perform label-free peptide and protein quantification using the results obtained from the search software (*see* **Note 43**).

5. Using the protein abundances calculated by the quantification software, perform statistical analyses to determine which proteins are significantly enriched in the target RNA capture sample as compared to the control. Large-scale statistical analyses can be performed using a software platform such as Perseus ***(26)***. The best approach for normalizing the quantification data (*see* **Note 44**) and determining proper statistical thresholds will be specific to each application of HyPR-MS and experimental design. There are multiple acceptable ways to analyze the data, and there are no universal parameters that guarantee meaningful results. The most appropriate method, and the confidence one can have in the results, will depend on the data. One standard approach is to begin

by log_2-transforming the protein intensity values for each sample to obtain normal distributions. The protein intensity values can then be grouped by condition (i.e., target RNA capture samples and control samples), and the proteins can be filtered to remove those with an insufficient number of observations (e.g., proteins observed in fewer than two-thirds of the biological replicates for a particular sample type). Any remaining missing values can be imputed (*see* **Note 45**). Next, two-sample T-tests can be used to compare protein abundances between the target and control samples to determine which proteins are differentially abundant at a specified *p*-value (e.g., $p = 0.05$). A permutation-based FDR cutoff (e.g., 1-10%) can be used to correct for multiple hypothesis testing. Proteins within this FDR cutoff are statistically differentially abundant between the target RNA capture and control samples, subject to the criteria applied (*see* **Note 46**). Additional approaches for analyzing quantitative proteomics data from a HyPR-MS experiment can be found in our previous publications ***(21–24)***.

4.5 NOTES

1. Carefully designed, target-specific capture oligonucleotides and stringent wash steps in the HyPR-MS protocol reduce the presence of nonspecific protein binders in the target RNA pulldown samples; however, nonspecific binders cannot be completely eliminated.
2. A scrambled pulldown control utilizes a scrambled sequence oligonucleotide which is not significantly complementary to any region of the genome or transcriptome (determined using BLAST *(27)* (https://blast.ncbi.nlm.nih.gov/Blast.cgi)), and therefore does not target any specific RNA. Generally, this oligonucleotide is designed to have approximately the same G/C content and T_m as the target capture oligonucleotide(s). The proteins identified in this pulldown are likely nonspecific binders inherent to the HyPR-MS procedure. Therefore, proteins that overlap between this control and the target RNA pulldown are probably not specific binders of the target RNA.
3. A poly(dT) pulldown control captures RNA molecules with poly(A) tails, such as mRNAs. We have found that a poly(dT) oligonucleotide containing ~20 Ts, in addition to the 8 nt toehold sequence, is sufficient for capture. This control experiment enables one to determine which of the proteins identified in the target RNA pulldown experiment are distinct to that particular RNA, and which are general RNA-binding proteins.

4. A lysate control involves the proteomic analysis of whole cell lysate to determine the most abundant proteins present in the sample. Proteins which are highly abundant in cell lysate have the potential to be carried through to the final RNA pulldown sample simply due to their abundance and not due to specific interaction with the target RNA. If many of the proteins identified in the target RNA pulldown sample are the same as the most abundant proteins in whole cell lysate, this may be a sign that the pulldown parameters require further optimization and/or that more stringent washing steps should be included.

5. It is advisable to design capture oligonucleotides complementary to multiple regions of the target RNA. The sonication step of HyPR-MS may cause RNA fragmentation, thus the use of multiple capture oligonucleotides spanning the length of the target RNA provides a more complete characterization of the proteins bound along the entire length of the RNA than would the use of a single capture oligonucleotide. The appropriate number of capture oligonucleotides depends greatly on the target RNA length and sonication intensity. Begin by designing several (~3-8) capture oligonucleotides and qPCR assays to span the length of the target RNA. Then, monitor the capture efficiency of different regions of the target transcript when different combinations of capture oligonucleotides are used. Ideally, one should use the minimum number of capture oligonucleotides that are necessary to ensure sufficient capture along the entire length of

the transcript. Additional capture oligonucleotides may marginally increase capture efficiency, but they also increase the potential for off-target hybridization and undesirable oligonucleotide—oligonucleotide interactions. If it proves difficult to obtain acceptable levels of capture along the entire length of the target transcript, it may be helpful to investigate the integrity of the RNA in the sample by purifying the RNA and analyzing it on an agarose gel or Bioanalyzer. If the RNA looks very degraded (determined by analyzing the ratio of the 28S:18S rRNA band intensities or the RNA integrity number *(**28**)*), it may be necessary to use more RNase inhibitors in the experiment and/or to decrease sonication intensity.

6. It is important to consider the secondary structure of the target RNA when designing capture oligonucleotides. Capture oligonucleotides provide better capture efficiency when they are designed to complement regions of the target RNA that are single-stranded. If the secondary structure of the target RNA is uncharacterized, software tools like Mfold *(**29**)* (www.unafold.org) can be useful for making predictions.

7. Capture oligonucleotides should be designed so that they have minimal potential for hybridization to off-target RNAs. Capture of off-target RNAs and their associated protein interactors cause false positives in the HyPR-MS protein data analysis. BLAST should be

used to assess the specificity of all proposed capture oligonucleotides in the context of the relevant complete genome/transcriptome.

8. Hybridization parameters (temperature, salt concentration, incubation time, etc.) for HyPR-MS will need to be optimized on a case-by-case basis to allow for stable oligonucleotide—target RNA hybridization whilst minimizing the formation of off-target hybrids. In general, capture oligonucleotides ~30 nt in length, a hybridization buffer containing 375 mM LiCl, a hybridization temperature of 37 °C, and a hybridization time of 3 h have provided satisfactory stability, specificity, and efficiency results. The T_m of a given capture oligonucleotide—target RNA hybrid under various hybridization conditions can be assessed using freely available webtools such as the IDT OligoAnalyzer (https://www.idtdna.com/calc/analyzer) or OligoCalc ***(30)*** (http://biotools.nubic.northwestern.edu/OligoCalc.html).

9. It is important to design capture oligonucleotides that do not form stable secondary structures, such as hairpins, under HyPR-MS hybridization conditions. It is more favorable for an oligonucleotide which does not have stable intramolecular interactions to hybridize to the target RNA than it is for an oligonucleotide which adopts a stable secondary structure. There are multiple freely available webtools that can be used to determine the propensity of any given capture oligonucleotide to form such structures,

such as the IDT OligoAnalyzer (https://www.idtdna.com/calc/analyzer) or OligoCalc ***(30)*** (http://biotools.nubic.northwestern.edu/OligoCalc.html).

10. Capture oligonucleotides should be designed so that stable homodimers and heterodimers do not form under HyPR-MS experimental conditions. These dimers compete with the desired capture oligonucleotide—target RNA hybrid. There are multiple freely available webtools that can be used to determine the propensity of the capture oligonucleotides used in any given experiment to form such hybrids, such as the IDT OligoAnalyzer (https://www.idtdna.com/calc/analyzer) or the Thermo Fisher Scientific Multiple Primer Analyzer (https://www.thermofisher.com/us/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html).

11. HyPR-MS utilizes a "toehold-mediated" strategy to release the purified target RNA—protein complexes from the streptavidin-coated magnetic beads ***(21–24)***. This "toehold" must be incorporated into the design of the capture oligonucleotide. Typically, a 30 nt sequence complementary to the target RNA is designed first. Then, on the 5' or 3' end of that sequence, an 8 nt sequence which is not complementary to the target RNA is added, making the entire capture oligonucleotide 38 nt in length. With this design, a 30 nt

stretch of the capture oligonucleotide forms a hybrid with the target RNA, while the 8 nt toehold remains single-stranded (Figure 4.1). During the release step of HyPR-MS, a release oligonucleotide which is completely complementary to the capture oligonucleotide (38 nt in length, in this example) is added to the bead slurry. The 8 nt, single-stranded region of the capture oligonucleotide serves as a toehold for hybridization of the entire 38 nt capture oligonucleotide with the 38 nt release oligonucleotide. Because the capture oligonucleotide—release oligonucleotide hybrid is more thermodynamically stable than the capture oligonucleotide—target RNA hybrid (a 38 nt hybrid is more stable than a 30 nt hybrid), the release oligonucleotide displaces the target RNA from the capture oligonucleotide, thereby releasing the target RNA—protein complexes into solution (Figure 4.1). A notable feature of toehold-mediated release is that it enables multiplexing of HyPR-MS experiments, allowing for the analysis of multiple RNA targets from the same cell lysate preparation ***(21–24)***. When designing capture oligonucleotides, the entire 38 nt sequence (complementary region plus toehold region) should be assessed for off-target hybridization via BLAST.

12. An alternative measure of capture specificity is the ratio of target to nontarget RNA in the target RNA capture sample. RNA-seq is better suited to making these measurements than is RT-qPCR.

13. There are benefits to designing qPCR assays to amplify regions both adjacent to and further away from the locations targeted by capture oligonucleotides. A qPCR assay designed to amplify a region directly adjacent to a capture oligonucleotide gives the best measurement of the capture of that region of the transcript. However, that qPCR assay gives very little information about the length of the captured RNA molecules. For this information, a qPCR assay which amplifies a region further away from the capture oligonucleotide should be designed. If the capture efficiency measured by this qPCR assay is lower than desired, adjusting sonication parameters to decrease RNA fragmentation may be helpful.

14. The use of hydrolysis probe qPCR assays (rather than SYBR Green) is recommended to maximize the specificity and sensitivity of qPCR measurements. Information on how to design qPCR assays is available elsewhere *(**31**)*.

15. The number of cells required per HyPR-MS experiment will vary and should be determined empirically for each distinct RNA target and cell line. The goal of a HyPR-MS experiment is to identify the ensemble of proteins that are associated with a target RNA, therefore a successful HyPR-MS experiment must yield enough protein in the target RNA capture sample for analysis via bottom-up mass spectrometry. The amount of any given protein in the capture sample is a function of the stoichiometry with which the

protein binds the target RNA, the abundance of the target RNA in the cell line of interest, and target RNA capture efficiency. Higher protein:RNA stoichiometry, target RNA abundance, and/or capture efficiency will decrease the number of cells required per HyPR-MS experiment. To determine the number of cells required for any given experiment, the abundance of the target RNA in the cell line of interest should first be estimated via RNA-seq and/or RT-qPCR. Then, "small-scale" RNA capture experiments should be performed using relatively few cells (on the order of 10^5-10^6) to optimize capture efficiency and specificity. Note that these small-scale experiments will provide enough RNA for RT-qPCR analyses, but will not provide enough protein for mass spectrometric analysis. Once capture parameters have been optimized in small-scale experiments, one can estimate the number of cells required for full-scale HyPR-MS experiments by taking into account the measured target RNA abundance and capture efficiency, and by making a few approximations regarding protein:target RNA stoichiometry and mass spectrometric detection limit (e.g., a 1:1 protein:target RNA stoichiometry and a mass spectrometric detection limit of ~1 fmol of peptide). Previous studies have successfully used on the order of 10^7-10^8 cells for full-scale HyPR-MS experiments ***(21–24)***.

16. Formaldehyde concentration and cross-linking time may need to be optimized based on the cell line and RNA target of interest.

17. Cells should be pelleted in quantities appropriate for individual HyPR-MS experiments because cross-linked cells are difficult to resuspend without lysing. To preserve the integrity of the RNA, cross-linked cells should not be thawed and then refrozen.

18. We have found that it is important to lyse cells at a consistent concentration of 5×10^6 cells/mL. In our hands, this concentration has provided a sufficient protease/RNase inhibitor:cell ratio to maintain protein and RNA integrity.

19. Appropriate sonication parameters will need to be determined for each system and sonicator. Sonication aids in breaking up chromatin and solubilizing RNA—protein complexes, but it also causes RNA fragmentation. Ideal sonication parameters will effectively solubilize the target RNA whilst minimizing RNA fragmentation. Optimal parameters should be determined empirically by monitoring a) RNA solubilization via RT-qPCR and/or absorbance at 260 nm and b) the general degree of RNA fragmentation via analysis of sonicated RNA on an agarose gel or Bioanalyzer (to investigate RNA integrity via the ratio of the 28S:18S rRNA band intensities or the RNA integrity number *(**28**)*). In general, ~4-12 s of light sonication is appropriate for a ~1 mL aliquot of lysate. Because sonication heats the lysate and heat can reverse formaldehyde cross-links *(**32, 33**)*, sonicating the lysate on ice and performing the sonication in ~4 s bursts with ~4 s of rest between each burst is recommended.

20. Appropriate capture oligonucleotide concentrations will depend on target RNA abundance and will need to be determined empirically. To start, try small-scale experiments to test oligonucleotide concentrations ranging from ~1-15 nM. Following RT-qPCR analysis, determine which concentration provides the optimal capture efficiency and specificity. Additional experiments to titrate up or down may be necessary.
21. Hybridization time and temperature may be optimized for the specific melting temperature(s) of the capture oligonucleotide—target RNA hybrids and the abundance of the target RNA.
22. The appropriate volume of streptavidin-coated magnetic beads should be determined empirically. Typically, 3 µL of beads for every picomole of capture oligonucleotide is sufficient.
23. If the bead volume exceeds ~20% of the hybridization volume, it is best to first transfer the bead slurry to clean tube(s), remove the supernatant, then transfer the lysate to the tube(s) containing the beads. This way, the RNase and protease inhibitors present in the lysate do not get diluted by a large bead volume.
24. Appropriate release oligonucleotide concentrations will depend on capture oligonucleotide concentrations and will need to be determined empirically. To start, try using 100-1,000× more release oligonucleotide than the corresponding capture

oligonucleotide and adjust based on measurements of capture and release efficiency (*see* **Notes 35 and 36**).

25. If multiplexing the HyPR-MS experiment to analyze multiple RNA targets from the same cell lysate preparation, or if including a scrambled or poly(dT) oligonucleotide control, one should perform any additional RNA release steps prior to proceeding to step 17. For each sequential release step, the beads should first be washed by resuspending in one bead volume of release buffer and gently rocking at room temperature for 5 min prior to repeating steps 13-16.

26. The RNA from an aliquot of these resuspended beads will be analyzed via RT-qPCR to measure the amount of target RNA that remains on the beads after all release steps. If a significant amount of target RNA remains on the beads, release conditions may require further optimization (e.g., increasing release oligonucleotide concentration, increasing release time, or decreasing release temperature).

27. For small-scale experiments, the entire sample volume can be used.

28. It is important to avoid transferring any of the interphase (containing DNA) or organic phase (containing protein). Additionally, transferring a consistent volume for each sample (e.g., 500 µL for the parameters described here) will ensure accurate downstream calculations of capture efficiency, release efficiency, etc.

29. The addition of a coprecipitant, such as glycogen, is recommended to facilitate RNA precipitation and to make the RNA pellet easier to observe in subsequent steps.

30. It is important not to disturb the RNA pellet while removing the supernatant. To minimize disruption of the pellet, first use a large (~1,000 µL) pipette tip to remove the majority of the supernatant, then centrifuge the sample briefly and remove the remainder of the supernatant using a smaller (~10 µL) pipette tip.

31. It is important that each reverse transcription reaction contain an appropriate amount of RNA, which will be dictated by the kit manufacturer. We recommend using a spectrophotometer to measure the concentration of RNA in each sample following ethanol precipitation.

32. The following controls should be considered for reverse transcription experiments: (1) A reaction without any RNA to verify that the reverse transcription reactions are not contaminated with exogenous RNA/DNA. For this control, omit RNA from the reaction and use water to bring the reaction to the appropriate final volume; (2) A no-reverse transcriptase control to monitor for genomic DNA contamination in the RNA samples. Most qPCR polymerases will not amplify RNA, therefore any qPCR signal from a reaction without reverse transcriptase may be attributed to genomic DNA contamination. For this control, omit the reverse transcriptase enzyme from the reaction and use water to bring

the reaction to the appropriate final volume; (3) Serial dilutions of the RNA samples of interest. Reverse transcription efficiency is sensitive to the complexity of the RNA sample being reverse transcribed. If the qPCR signal from prereverse transcription serial dilutions of an RNA sample is not dropping linearly, it can indicate that the initial RNA sample may be too complex and qPCR results from that sample may be unreliable. For these controls, prepare serial dilutions (e.g., three 10-fold dilutions) of the RNA sample in water and perform reverse transcription on each of the dilutions.

33. Technical duplicate or triplicate qPCR reactions should be performed for each cDNA sample/qPCR assay combination. Additionally, a control qPCR reaction without any cDNA should be included to verify that the qPCR reactions are not contaminated with exogenous DNA. For this control, omit cDNA from the reaction and use water to bring the reaction to the appropriate final volume.

34. Appropriate qPCR cycling parameters should be determined empirically for each qPCR assay. qPCR master mixes often come with a recommended protocol, which can be a helpful place to start. Online resources can also be useful for approximating appropriate annealing temperatures for each qPCR assay.

35. Target RNA capture efficiency can be determined using a target-specific qPCR assay and a calibration curve made up of serial dilutions of the precapture lysate sample (from

Subheading 4.4.3, step 5). These dilutions can be made prior to performing reverse transcription. By comparing the quantification cycle (C_q) of the target RNA capture sample (from Subheading 4.4.4, step 16) to the calibration curve, one can quantify the relative amount of target RNA present in that sample and calculate capture efficiency. The same approach can be used to quantify the percentage of the target RNA present in the postcapture lysate sample (from Subheading 4.4.4, step 8), any other RNA capture samples (from Subheading 4.4.4, step 16), and the bead sample (from Subheading 4.4.4, step 17). Capture efficiencies can vary substantially depending on the RNA target and capture oligonucleotide(s) used, and capture efficiencies ranging from ~10% to 60% are commonly observed. If the observed capture efficiency is lower than desired, increasing capture oligonucleotide concentration(s), adjusting hybridization time and temperature, and/or designing new capture oligonucleotide(s) to target more open regions of the target RNA may be helpful.

36. Release efficiency can be calculated by dividing the percentage of the target RNA in the target RNA capture sample by the summed percentage in the target RNA capture sample, any other capture sample(s), and the bead sample. Release efficiencies ranging from ~50% to 90% are commonly observed. If the observed release efficiency is lower than desired,

adjusting release time or temperature or increasing the concentration of release oligonucleotides used may be helpful.

37. Capture specificity can be calculated by dividing the capture efficiency of the target RNA in the target RNA capture sample by the capture efficiency of the target RNA in a nontarget RNA capture sample. The nontarget RNA capture sample could be a scrambled oligonucleotide control capture sample or a capture sample for a different RNA target.

38. Target RNA enrichment can be calculated by comparing the ratio of target RNA:GAPDH (or some other housekeeping RNA) in the lysate prior to hybridization capture to the same ratio in the final RNA capture sample. Enrichment can be calculated either by using the $2^{-\Delta\Delta Ct}$ method ***(34)*** or by comparing to a genomic DNA standard curve (for absolute transcript quantification).

39. Trypsin should be added so that the final trypsin:protein ratio is ~1:20 to 1:100.

40. Reference protein databases are commonly used for mass spectrometric data analysis. Reference databases can be obtained from UniProt ***(35)*** (https://www.uniprot.org/proteomes/), GENCODE ***(36)*** (https://www.gencodegenes.org/), and RefSeq ***(37)*** (https://www.ncbi.nlm.nih.gov/refseq/). In the proteomics community, UniProt databases are often used. We also recommend including a database of common

contaminants in the search, so that exogenous proteins present in the samples can be identified (e.g., trypsin, streptavidin, keratin, etc.).

41. Search software programs such as MetaMorpheus ***(38)*** (https://github.com/smith-chem-wisc/MetaMorpheus/), MSFragger ***(39)*** (https://msfragger.nesvilab.org/), and Andromeda ***(40)*** (http://coxdocs.org/doku.php?id=maxquant:andromeda:start), among others, can be used to obtain peptide and protein identifications from the acquired mass spectra.

42. Depending on the search software used, there may be an option during search setup to prevent low-confidence identifications from being written to the output file. If selected, this would remove the need to perform this step manually after the search is complete.

43. Software programs capable of label-free peptide and protein quantification include FlashLFQ ***(41)*** (https://github.com/smith-chem-wisc/FlashLFQ) and MaxQuant ***(42)*** (http://coxdocs.org/doku.php?id=maxquant:start), among others.

44. If using a normalization algorithm that is built into the quantification software, ensure that the assumptions made by the algorithm are appropriate for the HyPR-MS experimental design.

45. Imputed values should maintain the normal distribution of the data and not create a bimodal distribution. If this occurs, adjust the parameters of how imputed values are determined.

46. We recommend applying a fold-change cutoff at this point, which enables one to specify a minimum difference in abundance that is required for a protein to be considered enriched in the target RNA capture sample. There is no single fold-change cutoff that is suitable for all experiments. In general, a smaller sample variance and more sample replicates enable one to apply a lower fold-change cutoff and still discover biologically significant changes.

4.6 ACKNOWLEDGEMENTS

This work was supported by NIH-NCI grant R01CA193481. K.B.H. was supported in part by the National Human Genome Research Institute grant to the Genomic Science Training Program, 5T32HG002760. R.M.M. was supported in part by the NIH Chemistry-Biology Interface Training Grant, T32GM008505. The authors would like to thank members of the Smith lab for helpful discussions and guidance in the development of HyPR-MS. The figures in this chapter were created with BioRender.com.

4.7 REFERENCES

1. Moore MJ (2005) From birth to death: the complex lives of eukaryotic mRNAs. Science 309:1514–1518
2. Glisovic T, Bachorik JL, Yong J et al (2008) RNA-binding proteins and post-transcriptional gene

regulation. FEBS Lett 582:1977–1986

3. Mitchell SF, Parker R (2014) Principles and properties of eukaryotic mRNPs. Mol Cell 54:547–558

4. Re A, Joshi T, Kulberkyte E et al (2014) RNA-protein interactions: an overview. In: Gorodkin J, Ruzzo WL (eds) RNA sequence, structure, and function: computational and bioinformatic methods, Methods in molecular biology, vol 1097. Humana Press, Totowa, NJ, pp. 491–521

5. Matera AG, Terns RM, Terns MP (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol 8:209–220

6. Mayr C (2017) Regulation by 3'-untranslated regions. Annu Rev Genet 51:171–194

7. Marchese FP, Raimondi I, Huarte M (2017) The multidimensional mechanisms of long noncoding RNA function. Genome Biol 18:206

8. Allerson CR, Cazzola M, Rouault TA (1999) Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. J Biol Chem 274:26439–26447

9. Lukong KE, Chang KW, Khandjian EW et al (2008) RNA-binding proteins in human genetic disease. Trends Genet 24:416–425

10. Corbett AH (2018) Post-transcriptional regulation of gene expression and human disease. Curr Opin Cell Biol 52:96–104

11. Ramanathan M, Porter DF, Khavari PA (2019) Methods to study RNA–protein interactions. Nat Methods 16:225–234

12. Ule J, Jensen KB, Ruggiu M et al (2003) CLIP identifies Nova-regulated RNA networks in the brain. Science 302:1212–1215

13. Licatalosi DD, Mele A, Fak JJ et al (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456:464–469

14. Hafner M, Landthaler M, Burger L et al (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141:129–141

15. König J, Zarnack K, Rot G et al (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 17:909–915

16. Van Nostrand EL, Pratt GA, Shishkin AA et al (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods 13:508–514

17. Kim B, Kim VN (2019) fCLIP-seq for transcriptomic footprinting of dsRNA-binding proteins:

lessons from DROSHA. Methods 152:3–11

18. West JA, Davis CP, Sunwoo H et al (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol Cell 55:791–802

19. Chu C, Zhang QC, da Rocha ST et al (2015) Systematic discovery of Xist RNA binding proteins. Cell 161:404–416

20. McHugh CA, Chen CK, Chow A et al (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature 521:232–236

21. Knoener RA, Becker JT, Scalf M et al (2017) Elucidating the in vivo interactome of HIV-1 RNA by hybridization capture and mass spectrometry. Sci Rep 7:16965

22. Spiniello M, Knoener RA, Steinbrink MI et al (2018) HyPR-MS for multiplexed discovery of MALAT1, NEAT1, and NORAD lncRNA protein interactomes. J Proteome Res 17:3022–3038

23. Spiniello M, Steinbrink MI, Cesnik AJ et al (2019) Comprehensive in vivo identification of the c-Myc mRNA interactome using HyPR-MS. RNA 25:1337–1352

24. Knoener R, Evans E III, Becker JT et al (2021) Identification of host proteins differentially associated with HIV-1 RNA splice variants. eLife 10:e62470

25. Erde J, Loo RRO, Loo JA (2014) Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. J Proteome Res 13:1885–1895

26. Tyanova S, Temu T, Sinitcyn P et al (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13:731–740

27. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

28. Schroeder A, Mueller O, Stocker S et al (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3

29. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415

30. Kibbe WA (2007) OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Res 35:W43–W46

31. Shipley GL (2013) Assay design for real-time qPCR. In: Nolan T, Bustin SA (eds) PCR technology: current innovations, 3rd edn. CRC Press, Taylor & Francis Group, Boca Raton, FL, pp 177-197

32. Jackson V (1978) Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. Cell 15:945–954

33. Kennedy-Darling J, Smith LM (2014) Measuring the formaldehyde protein-DNA cross-link reversal rate. Anal Chem 86:5678–5681

34. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCt method. Methods 25:402–408

35. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515

36. Frankish A, Diekhans M, Ferreira AM et al (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47:D766–D773

37. O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745

38. Solntsev SK, Shortreed MR, Frey BL et al (2018) Enhanced global post-translational modification discovery with MetaMorpheus. J Proteome Res 17:1844–1851

39. Kong AT, Leprevost FV, Avtonomov DM et al (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 14:513–520

40. Cox J, Neuhauser N, Michalski A et al (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10:1794–1805

41. Millikin RJ, Solntsev SK, Shortreed MR et al (2018) Ultrafast peptide label-free quantification with FlashLFQ. J Proteome Res 17:386–391

42. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 11:2301–2319

CHAPTER 5

IDENTIFYING PROTEOFORMS BOUND TO TARGET RNAs WITH HyPR-MS: CURRENT STATUS AND FUTURE DIRECTIONS

5.1 ABSTRACT

RNA-protein interactions are central to many fundamental cellular processes, and disruptions of these interactions are implicated in numerous diseases. As such, technologies for the study of RNA-protein interactomes are critical to furthering our understanding of cellular biology. Multiple approaches exist both for the global identification of protein-RNA interactions and for the identification of proteins interacting with specific RNA species. However, no current technology is capable of identifying the *proteoforms* interacting with a target RNA transcript. In this chapter, we describe our work to adapt the HyPR-MS (Hybridization Purification of RNA-protein complexes followed by Mass Spectrometry) technology to enable the identification of proteoforms bound to

target RNA(s) via top-down mass spectrometry. We first describe the establishment of appropriate parameters for the capture of the MALAT1 long noncoding RNA in human Jurkat cells. We then discuss our work to develop a top-down proteomics sample preparation method suitable for the small quantities of protein present in HyPR-MS release samples. We detail our work to establish parameters for reversing formaldehyde cross-links prior to top-down proteomics analysis, and present suspension trapping (S-Trap) as a novel method of top-down proteomics sample preparation. Finally, we describe the results of full-scale top-down HyPR-MS experiments to study MALAT1-binding proteoforms. The proteoforms identified in MALAT1 HyPR-MS release samples were very similar to those identified in control samples, and therefore we were unable to ascertain a list of MALAT1-binding proteoforms from these initial top-down HyPR-MS experiments. Nevertheless, this work represents an important first step in enabling the identification of target RNA-binding proteoforms, and at the end of this chapter we discuss a number of future directions that could improve the technology such that its vision can be realized.

5.2 INTRODUCTION

Many fundamental cellular processes, including RNA transcription, translation, splicing, localization, and degradation, are facilitated by protein-RNA interactions.[1–7] Additionally, disrupted or anomalous protein-RNA interactions are responsible for numerous pathological states,[8–10] making technologies that enable the identification and study of these interactions critical. There are multiple

such technologies presently available, either for the global identification of protein-RNA interactions[11–14] or for the elucidation of the proteins interacting with a specific target RNA.[15–19]

Although numerous technologies exist for the study of protein-RNA interactomes, one limitation that remains is that none of these technologies enable the identification of the *proteoforms* that interact with a target RNA transcript. A proteoform is a specific molecular form of a protein, with a particular amino acid sequence and localized set of post-translational modifications (PTMs).[20–22] Proteoforms are the molecular actors which drive biological function, and different proteoforms arising from the same gene can perform very different roles within the cell. A classic example is histone proteoforms, where the specific combination of PTMs on a histone protein can recruit other proteins and influence the open/closed nature of chromatin, thereby helping to regulate processes such as gene expression.[23] In another example of the importance of proteoforms, a recent study[24] revealed a link between the expression of a particular proteoform of the human OAS1 protein and clinical severity of COVID-19, the disease caused by infection with the SARS-CoV-2 virus. That study found that a prenylated proteoform of OAS1 was able to interact with secondary structures within the SARS-CoV-2 RNA genome, triggering a series of events that degraded the viral RNA, thereby blocking replication.[24] A shorter OAS1 proteoform, which lacked the C-terminal prenylation site, did not exhibit this antiviral activity, and COVID-19 patients who expressed this proteoform were statistically more likely to be admitted to intensive care units than patients who expressed the antiviral proteoform.[24] In this example, both the amino acid sequence and post-translational modification

(prenylation) of OAS1 enabled the interaction of the proteoform with SARS-CoV-2 RNA and subsequent antiviral activity, exemplifying the importance of understanding not only protein-RNA interactions, but *proteoform*-RNA interactions.

The aforementioned technologies for studying protein-RNA interactomes rely on bottom-up proteomics[25,26] for protein identification. In a standard bottom-up proteomics workflow, proteins in a sample are digested by a protease and the resultant peptides are analyzed via tandem mass spectrometry (MS/MS) (Figure 5.1, top). The generated spectra are then searched against a protein database for identification, and identified peptides are used to infer the presence of proteins in the sample. While bottom-up proteomics is a robust and widely-used approach for proteome analysis,[25,26] this technique does not allow for the identification of proteoforms. This is because the proteolytic digestion inherent to bottom-up proteomics destroys the molecular context of the intact proteoform, making it impossible to decipher the parent proteoform of an identified peptide (Figure 5.1, top).[22] In contrast, top-down proteomics[22,27–30] analyzes intact proteoforms directly, without the intermediary step of protease digestion (Figure 5.1, bottom). Because the relationship between amino acid sequence and PTMs is preserved, top-down proteomics is capable of identifying proteoforms. However, top-down mass spectrometry experiments face many analytical challenges, including the low abundance of many proteoforms, low signal-to-noise ratios for large proteoforms,[31] insufficient proteoform fragmentation for PTM localization, complex data analysis, and coelution of proteoforms in standard liquid chromatography (LC)-MS/MS experiments.[22]

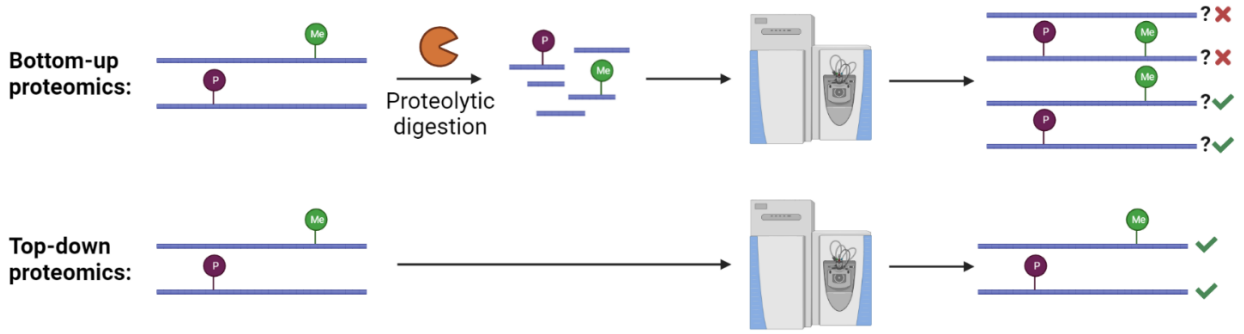


Figure 5.1 Comparison of bottom-up and top-down proteomics approaches. In bottom-up proteomics (top), proteoforms are digested into peptides which are then analyzed via mass spectrometry. Attempting to infer the presence of proteoforms from peptide identifications can lead to uncertainty and erroneous results. In top-down proteomics (bottom), intact proteoforms are analyzed directly, making correct proteoform identifications possible. This figure was created with BioRender.com.

This chapter describes our work to date in adapting HyPR-MS (Hybridization Purification of RNA-protein complexes followed by Mass Spectrometry), a technology developed in our research group for the identification of proteins interacting with specific RNA species,[18,19,32–34] to enable the identification of proteoforms interacting with target RNA transcripts. Briefly, the HyPR-MS procedure (Figure 5.2)[19] involves in vivo formaldehyde cross-linking of cells in culture to covalently fix protein-RNA interactions. Cells are then lysed and RNA-protein complexes are solubilized via light sonication. Biotinylated capture oligonucleotides, designed to specifically complement the target RNA sequence, are then introduced to the lysate and allowed to hybridize. Hybridized complexes are then captured on streptavidin-coated magnetic beads and the beads are washed to remove nonspecific binders. Hybridized complexes are then eluted from the beads via toehold-mediated strand displacement and target RNA-associated proteins are analyzed via mass spectrometry. Prior to this

work, HyPR-MS experiments exclusively used the bottom-up proteomics approach (Figure 5.1, top) for protein identification.[18,32–34]

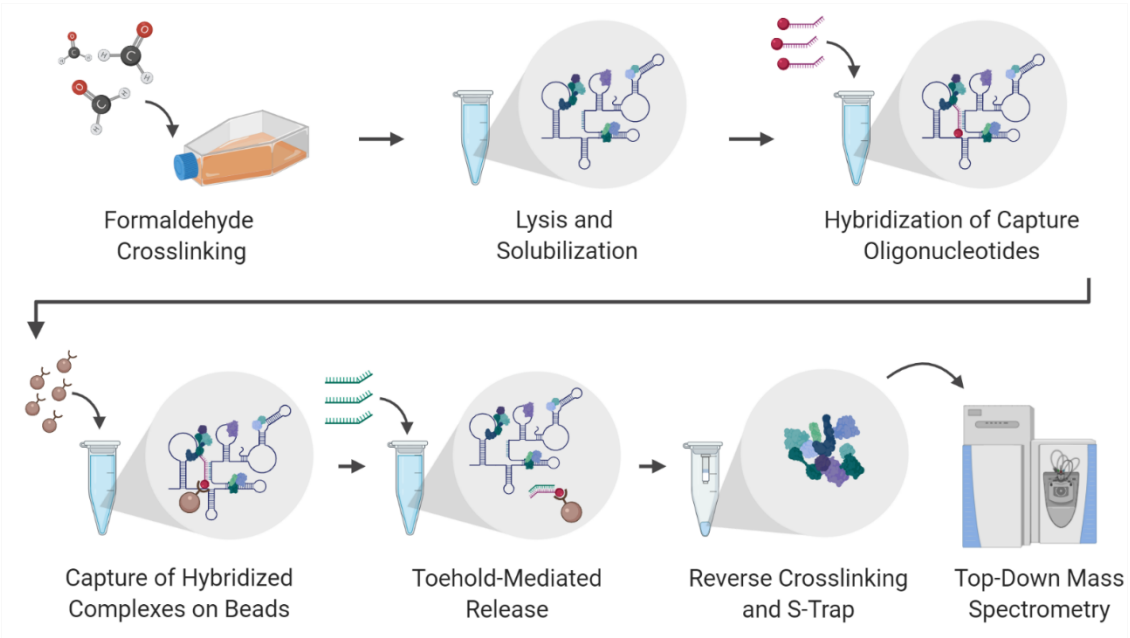


Figure 5.2 Overview of the HyPR-MS technology. Cells are cross-linked with formaldehyde in culture to covalently fix protein-RNA interactions. Cells are then lysed and RNA-protein complexes are solubilized via light sonication. Sequence-specific, biotinylated capture oligonucleotides (magenta) are then introduced to the lysate and hybridize to the target RNA. Hybridized complexes are captured on streptavidin-coated magnetic beads and the beads are washed to remove nonspecific interactors. Release oligonucleotides (green) are then added and target RNA-protein complexes are released from the beads via toehold-mediated strand displacement. In the top-down HyPR-MS protocol introduced here, formaldehyde cross-links are then reversed by heating and proteoforms in the HyPR-MS release sample are concentrated and purified via suspension trapping (S-Trap). The purified proteoforms are then analyzed directly via top-down LC-MS/MS. (This figure was created with BioRender.com and was adapted from Knoener et al.[34] under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/))

Here, we kept the target RNA purification steps of the HyPR-MS procedure, but developed a novel protein sample preparation technique to enable the use of top-down proteomics with HyPR-MS

release samples ("top-down HyPR-MS"). We began by optimizing capture conditions for MALAT1 RNA in the human Jurkat cell line.[35] MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) is an ~8.7 kb long noncoding RNA (lncRNA) that is dysregulated in many types of cancer.[36–38] MALAT1 is an abundant RNA localized to nuclear speckles and is known to play a role in alternative splicing.[36–39] Here, we selected MALAT1 as the target for top-down HyPR-MS development due to its abundant expression and apparent importance in various human pathologies.[36–38] Additionally, we have studied the MALAT1-binding proteome before via bottom-up HyPR-MS,[32] providing a reference to which our top-down HyPR-MS results could be compared. After establishing appropriate conditions for MALAT1 capture in Jurkat cells, we proceeded to investigate methods for formaldehyde cross-link reversal, protein sample concentration and clean-up, and top-down LC-MS/MS data acquisition and analysis. From this work, we developed a top-down HyPR-MS procedure and applied it to study MALAT1-binding proteoforms. While were able to identify proteoforms in MALAT1 HyPR-MS release samples, these proteoforms were very similar to those identified in a control sample and as such we were unable to specify a list of MALAT1-binding proteoforms after a thorough analysis of the data collected. However, a great deal of progress was made in addressing and overcoming the challenges associated with top-down HyPR-MS, and we discuss those challenges and the progress made herein. We also discuss future directions that could improve the current protocol such that our vision of a technology capable of identifying target RNA-binding proteoforms may become a reality.

5.3 METHODS

Cell culture and formaldehyde cross-linking

Human Jurkat cells were cultured at 37 °C under 5% CO_2 in RPMI-1640 medium (ATCC #30-2001) supplemented with 10% fetal bovine serum (GeminiBio #900-108) and 1× antibiotic-antimycotic solution (Gibco #15240-062). Cells were grown to a density of ~10^6 cells/mL, at which time they were cross-linked by adding formaldehyde to the culture medium to a final concentration of 1% (w/v) and gently shaking at room temperature for 10 min. Tris-HCl pH = 7.5 was then added to a final concentration of 250 mM and the cells were gently shaken for 10 min at room temperature to quench the formaldehyde. Cross-linked cells were pelleted by centrifuging at ~125 × *g* for 10 min at 4 °C. The cells were washed twice by gently resuspending in cold 1× phosphate-buffered saline, centrifuging at ~125 × *g* for 10 min at 4 °C and discarding the supernatant after each wash. After the last wash, the cell pellet was snap frozen in liquid nitrogen and immediately stored at -80 °C.

HyPR-MS

All solutions and plasticware used in HyPR-MS experiments (from cell lysis through RT-qPCR) were certified RNase-free. The procedures for bottom-up and top-down HyPR-MS are largely the same up until the point of protein sample preparation for mass spectrometric analysis, but subtle differences will be noted where applicable.

Cell lysis and RNA solubilization

Frozen formaldehyde cross-linked Jurkat cells were resuspended to a concentration of 5×10^6 cells/mL in freshly-prepared, cold lysis buffer containing 469 mM LiCl, 62.5 mM Tris-HCl pH = 7.5, 1.25% (w/v) lithium dodecyl sulfate (LiDS), 1.25% (w/v) Triton X-100, 12.5 mM ribonucleoside vanadyl complex (New England BioLabs #S1402S), 12.5 mM dithiothreitol (DTT), 125 U/mL RNasin Plus (Promega #N2615), and 1.25× protease/phosphatase inhibitors (Thermo Scientific #78440). For top-down HyPR-MS, 3×10^8 cells were used for each of three experimental replicates. For the single bottom-up HyPR-MS experiment, 1×10^8 cells were used. Cells were lysed on ice for 10 min with periodic vortexing to help break up the cell pellet, then the lysate was aliquoted into 1.5-mL portions. Each portion was sonicated for 12 s with 4 s on/off intervals at setting 2.5 using a Misonix Ultrasonic Processor XL 2015 equipped with a microtip, then centrifuged at $1{,}000 \times g$ for 2 min at 4 °C to clear insoluble cellular debris. The supernatants from each 1.5-mL portion of lysate were combined into a new tube.

Hybridization and capture

RNase-free water was added to increase the lysate volume by 25% (new buffer component concentrations for hybridization: 375 mM LiCl, 50 mM Tris-HCl pH = 7.5, 1% (w/v) LiDS, 1% (w/v) Triton X-100, 10 mM ribonucleoside vanadyl complex, 10 mM DTT, 100 U/mL RNasin Plus, and 1× protease/phosphatase inhibitors), and a small volume of diluted lysate was stored at 4 °C for eventual RT-qPCR analysis. Capture oligonucleotides were added to the lysate and the lysate was incubated for

3 h at 37 °C with gentle rocking to allow for hybridization. For top-down HyPR-MS experiments, MALAT1, scrambled, and poly(A) capture oligonucleotides were used, while only MALAT1 and scrambled capture oligonucleotides were used for the bottom-up HyPR-MS experiment (see Supplementary Table S-5.1). During the hybridization period, streptavidin-coated magnetic beads (Fisher Scientific #09-981-140) were washed twice with five volumes of wash buffer (375 mM LiCl, 50 mM Tris-HCl pH = 7.5, 0.2% (w/v) LiDS, and 0.2% (w/v) Triton X-100). For top-down HyPR-MS experiments, 6.525 mL of beads were used, while 1.455 mL of beads were used for the bottom-up HyPR-MS experiment. For each wash step, the tube of beads was gently rocked for 5 min at room temperature before placing on a magnetic rack and discarding the supernatant. The beads were then resuspended in one volume of 37 °C wash buffer. After the hybridization period, the beads were added to the lysate and the lysate was gently rocked at 37 °C for 1 h to capture the hybridized complexes.

Washing and release

Following the capture period, the tube of beads was placed on a magnetic rack and the supernatant was removed and stored at 4 °C for eventual RT-qPCR analysis. The beads were washed (twice for top-down HyPR-MS experiments and once for the bottom-up HyPR-MS experiment) by resuspending in five volumes of 37 °C wash buffer and gently rocking at 37 °C for 15 min then once by resuspending in three volumes of release buffer (375 mM LiCl, 50 mM Tris-HCl pH = 7.5, 0.1% (w/v) LiDS, and 0.1% (w/v) Triton X-100) and gently rocking at room temperature for 5 min, discarding the supernatant after each wash. The beads were then resuspended in release buffer and MALAT1

release oligonucleotides were added (Supplementary Table S-5.1). One bead volume of release buffer was used for top-down HyPR-MS release steps, while three bead volumes of release buffer were used for bottom-up HyPR-MS release steps. The beads were gently rocked at 37 °C for 1 h to release MALAT1 RNA-protein complexes, then the supernatant was isolated and stored at 4 °C for eventual RT-qPCR and mass spectrometric analyses. The beads were washed by resuspending in three volumes of release buffer and gently rocking at room temperature for 5 min. The supernatant was discarded and the beads were resuspended in release buffer with poly(A) release oligonucleotides (Supplementary Table S-5.1). The beads were gently rocked at 37 °C for 1 h to release polyadenylated RNA-protein complexes, then the supernatant was isolated and stored at 4 °C for eventual RT-qPCR and mass spectrometric analyses. The beads were again washed by resuspending in three volumes of release buffer and gently rocking at room temperature for 5 min. The supernatant was discarded and the beads were resuspended in release buffer with scrambled release oligonucleotides (Supplementary Table S-5.1). The beads were gently rocked at 37 °C for 1 h to release the scrambled capture oligonucleotides, then the supernatant was isolated and stored at 4 °C for eventual RT-qPCR and mass spectrometric analyses. Finally, the beads were resuspended in release buffer (one bead volume for top-down HyPR-MS experiments and three bead volumes for the bottom-up HyPR-MS experiment) and the bead slurry was stored at 4 °C for eventual RT-qPCR analysis. Note that the poly(A) release step was skipped for the bottom-up HyPR-MS experiment, as no poly(A) release oligonucleotides were added.

RNA purification and RT-qPCR

Aliquots of the release samples (MALAT1, scrambled, and poly(A) (for top-down HyPR-MS)), the resuspended beads sample, the post-capture lysate sample, and the pre-hybridization lysate sample were prepared for RT-qPCR. Each sample was brought to the same final volume (300 µL) and the same buffer component concentrations as the post-capture lysate sample. $CaCl_2$ was added to 4 mM and proteinase K was added to 1 mg/mL and the samples were gently rocked at 37 °C overnight, then heated at 65 °C for 1 h. The samples were cooled to room temperature and 500 µL of TRI Reagent (Millipore Sigma #T9424) were added. The samples were vortexed and allowed to sit at room temperature for 5 min, vortexing periodically. Chloroform (100 µL) was added to each sample and the samples were vortexed and allowed to sit at room temperature for ~15 min. The samples were centrifuged at 12,000 × *g* for 15 min at 4 °C and the top, aqueous layers were carefully pipetted into new tubes. Three volumes of ethanol and 1.5 µL of GlycoBlue coprecipitant (Thermo Fisher Scientific #AM9516) were added to each sample and the samples were stored at -20 °C overnight to precipitate the RNA. The samples were centrifuged at 20,800 × *g* for 20 min at 4 °C and the supernatants were carefully removed so as not to disrupt the RNA pellets. The pellets were washed twice with room temperature 75% ethanol, centrifuging at 20,800 × *g* for 20 min at room temperature and carefully discarding the supernatants after each wash. The pellets were allowed to air dry for ~20 min then resuspended in 15 µL of nuclease-free water and heated at 37 °C for 15 min. The concentration of RNA in each sample was determined via absorbance at 260 nm using a Nanodrop spectrophotometer.

The purified RNA samples were diluted two-fold with nuclease-free water and aliquots of each sample were reverse-transcribed using the High Capacity cDNA Reverse Transcription Kit from Thermo Fisher Scientific (#4368814) following the manufacturer's protocol. Following reverse transcription, each sample was again diluted two-fold with nuclease-free water and qPCR was performed on a CFX96 Touch real-time PCR detection system (Bio-Rad) using LightCycler 480 Probes Master (Roche #04707494001) and sequence-specific primers and hydrolysis probes (Supplementary Table S-5.2). Each qPCR run included an 8 min preincubation step at 95 °C followed by 45 amplification cycles, each consisting of a 10 s incubation at 95 °C, a 20 s incubation at 57.6 °C, and a 2 s incubation at 72 °C during which the FAM fluorophore was detected. Each sample was analyzed in duplicate using each of the qPCR assays noted in Supplementary Table S-5.2, and standard curves were generated using serial dilutions of the pre-hybridization lysate sample. Capture efficiencies were calculated by comparing the average quantification cycle (C_q) of each sample to the relevant standard curve.

Proteomics for bottom-up HyPR-MS experiment

Protein preparation and mass spectrometry: The MALAT1 and scrambled release samples were prepared for mass spectrometric analysis via eFASP.[40] To begin, each sample was brought to 8 M urea and 0.1% deoxycholic acid, then heated at 37 °C for 15 min to ensure complete dissolution of the urea. Each sample was then passed through a 1% CHAPS-passivated 50 kDa molecular weight cutoff filter (EMD Millipore #UFC505096) in 480 µL increments, centrifuging the filter at 14,000 × *g* for 5 min and

discarding the flow-through after each spin. The remainder of the eFASP procedure was performed as described in Henke et al.[19] (also Chapter 4 of this dissertation), using 0.5 µg of trypsin (Promega #V5111) per sample. The samples were then desalted via C18 solid-phase extraction as described in Henke et al.[19] and analyzed via HPLC-ESI-MS/MS using a high-performance liquid chromatography system (nanoAcquity, Waters) coupled to an electrospray ionization orbitrap mass spectrometer (Q Exactive HF, Thermo Fisher Scientific). Samples were analyzed in technical duplicate (half the volume of the sample per injection). The column employed was a 100 µm id × 365 µm od fused silica capillary microcolumn packed with 20 cm of 1.7 µm diameter, 130 Å pore size C18 beads (Waters BEH) and an emitter tip pulled to ~1 µm using a laser puller (Sutter Instruments). The column was operated at 60 °C using a column oven and peptides were loaded on-column with 2% acetonitrile (ACN) in 0.2% formic acid (FA) at a flow rate of 0.4 µL/min for 30 min. Peptides were eluted over 99 min at a flow rate of 0.3 µL/min with the following gradient (all in 0.2% FA): 8% ACN at time 1 min, 34% ACN at time 81 min, 44% ACN at time 91 min, 64% ACN from 92-99 min. The gradient was then ramped to 2% ACN in 0.2% FA at a flow rate of 0.4 µL/min over four minutes and held for an additional 17 min. Full-mass profile scans were performed between 375 and 1,500 *m/z* at a resolution of 120,000. One microscan was used, with an AGC target of 1×10^6 and a maximum injection time of 100 ms. The top ten most intense peaks in the MS1 with $1 < z < 6$ were selected for HCD fragmentation with a normalized collision energy setting of 30. The MS2 resolution was set to 15,000, the isolation window was 2.5 *m/z* units, and the AGC target was 1×10^5 with a maximum injection time of 500 ms. Dynamic exclusion was enabled with a duration of 15 s.

Mass spectrometric data analysis: Mass spectrometric data were searched using MetaMorpheus[41] version 0.0.319 (https://github.com/smith-chem-wisc/MetaMorpheus). Calibration, global post-translational modification discovery (G-PTM-D),[42] and search tasks were performed using default parameters for samples digested with trypsin. Data were searched against the pruned multi-protease G-PTM-D database generated in Dai and Buxton et al.[43] (also Chapter 3 of this dissertation) and the built-in MetaMorpheus database of common contaminants. Please see the top-down HyPR-MS mass spectrometric data analysis part of the Methods section for a description of the multi-protease G-PTM-D database and other databases used in this work. Label-free quantification was performed with FlashLFQ[44] within MetaMorpheus, allowing match-between-runs with a 5 ppm peakfinding tolerance.

Proteomics for top-down HyPR-MS experiments

Protein preparation and mass spectrometry: The three release samples (MALAT1, poly(A), and scrambled) were prepared for and analyzed by mass spectrometry on consecutive days due to the extensive amount of pipetting and centrifuging required to concentrate and purify the protein samples via S-Trap. Each sample was removed from storage at 4 °C and placed on a magnet stand for ~20 min to remove any residual streptavidin-coated magnetic beads. The supernatant was then transferred to a new tube and Tris-HCl pH = 7.5 was added to a final concentration of 350 mM. The sample was heated at 70 °C in a water bath for 3 h to reverse formaldehyde cross-links, then solid sodium dodecyl sulfate (SDS) was added to a final concentration of 5% (w/v). DTT was added to a final concentration of 20 mM and the sample was heated at 55 °C for 15 min. The sample was cooled to room temperature,

then phosphoric acid was added to a final concentration of 1.2% (w/v) and the sample was divided into 32 250 µL aliquots. Six volumes (1.5 mL) of binding buffer (90% methanol/100 mM Tris-HCl pH = 7.5) were added to one aliquot and the aliquot was vortexed well. The solution was pipetted onto an S-Trap micro device (Protifi) in 250 µL increments, briefly centrifuging the S-Trap at 4,000 × *g* and discarding the flow-through after each addition. After one aliquot of release sample had completely passed through the S-Trap, six volumes (1.5 mL) of binding buffer were added to the next 250 µL aliquot and the process was repeated until all 32 aliquots of release sample had been passed through the same S-Trap device, thereby concentrating the proteins from that release sample within the bed of the S-Trap. The S-Trap was then washed four times with 250 µL of HPLC-grade water, five times with 250 µL of binding buffer, two more times with 250 µL of HPLC-grade water, then once more with 250 µL of binding buffer. The proteins were then eluted by pipetting 12 µL of 80% formic acid (FA) onto the S-Trap, centrifuging briefly, then collecting the flow-through and passing it through the device four more times (so the same 12 µL of 80% FA were passed through the device a total of five times).

The eluted protein sample was analyzed immediately to minimize formylation artifacts. For the MALAT1 and scrambled samples, all of sample (~10 µL) was analyzed in a single injection. For the poly(A) samples, varying amounts between one-eighth to one-half of the sample were injected in an attempt to match the amount of protein loaded for the MALAT1 and scrambled injections. Samples were analyzed via HPLC-ESI-MS/MS using the same system described for bottom-up HyPR-MS experiments. The column employed was a 100 µm id × 365 µm od fused silica capillary microcolumn

packed with 20 cm of 3 μm diameter, 300 Å pore size C2 beads (Separation Methods Technologies, Inc.) and an emitter tip pulled to ~1 μm using a laser puller (Sutter Instruments). The column was operated at 60 °C using a column oven and proteins were loaded on-column with 5% ACN in 0.2% FA at a flow rate of 0.6 μL/min for 20 min. Proteins were eluted over 160 min at a flow rate of 0.5 μL/min with the following gradient (all in 0.2% FA): 20% ACN at time 5 min, 40% ACN at time 105 min, 60% ACN at time 145 min, 75% ACN from 150-160 min. The gradient was then ramped to 5% ACN in 0.2% FA at a flow rate of 0.6 μL/min over five minutes and held for an additional 15 min. Full-mass profile scans were performed between 600 and 2,000 *m/z* at a resolution of 240,000. Three microscans were averaged, using an AGC target of 1×10^6 and a maximum injection time of 50 ms. The top five most intense peaks in the MS1 with $z > 4$ were selected for HCD fragmentation with a normalized collision energy setting of 25. The MS2 resolution was set to 120,000, the isolation window was 4 *m/z* units, and the AGC target was 1×10^6 with a maximum injection time of 200 ms. Dynamic exclusion was enabled with a duration of 30 s. Source-induced dissociation was set to 15.0 eV.

Mass spectrometric data analysis: Mass spectrometric data were searched using a modification of MetaMorpheus version 0.0.318 (modification is that it reports the number of proteoform spectral matches (PrSMs) corresponding to a particular proteoform in each individual .raw file). The calibration, G-PTM-D, and search tasks were performed using default parameters for top-down data, except carbamidomethylation on cysteine was not selected as a fixed modification and methylation/dimethylation of lysine and arginine, trimethylation of lysine, and formylation of lysine

were selected as modifications to look for in G-PTM-D, in addition to the default modifications. Data were searched against the built-in MetaMorpheus database of common contaminants as well as either the pruned or protein-pruned multi-protease G-PTM-D database generated in Dai and Buxton et al.[43] (also Chapter 3 of this dissertation), the UniProt reviewed human database (20,371 protein entries, downloaded on June 23, 2021), or the UniProt reviewed plus unreviewed human database (78,120 entries, downloaded on June 23, 2021). The multi-protease G-PTM-D databases were generated by searching bottom-up data from six different proteolytic digests of Jurkat cell lysate (trypsin, chymotrypsin, GluC, ArgC, AspN, and LysC) against a UniProt reviewed plus unreviewed human protein database.[43,45,46] G-PTM-D was performed to identify unannotated PTMs in the data, and these PTMs were added to the database. The "pruned" version of the multi-protease G-PTM-D database was then created by removing any G-PTM-D modifications that were not confidently observed (1% FDR) in the bottom-up data. The "protein-pruned" version of the multi-protease G-PTM-D database was created by removing both G-PTM-D modifications and protein sequences that were not confidently observed (1% FDR) in the bottom-up data.

FlashLFQ[44] (version 1.2.0; https://github.com/smith-chem-wisc/FlashLFQ) was used to perform label-free quantification of proteoforms identified in MALAT1 and scrambled capture samples. Default parameters were used, except that four isotopic peaks were required for quantification and match-between-runs was enabled. The FlashLFQ results were filtered to remove proteoforms that were observed in fewer than three charge states, and the remaining proteoform

intensity values were loaded into Perseus[47] (version 1.6.15.0) for statistical analysis. The intensity values were log_2-transformed and proteoforms were filtered to remove those that were not observed in three biological replicates of either capture condition. Missing values were imputed based on the distribution of intensity values for all remaining proteoforms, with a width setting of 0.4 and a down shift setting of 1.6. Two-sided Student's T-tests were performed to compare proteoform abundances between the two capture conditions. The tests were corrected for multiple hypothesis testing using a permutation-based FDR cutoff of 5% (S0 = 0.1).

Formaldehyde cross-link reversal study

Non-cross-linked and 1% (w/v) formaldehyde cross-linked Jurkat cells were resuspended in freshly-prepared, cold HyPR-MS lysis buffer with an elevated Tris-HCl concentration (469 mM LiCl, 200 mM Tris-HCl pH = 7.5, 1.25% (w/v) LiDS, 1.25% (w/v) Triton X-100, 12.5 mM ribonucleoside vanadyl complex, 12.5 mM DTT, 125 U/mL RNasin Plus, and 1.25× protease/phosphatase inhibitors) to a concentration of 1×10^7 cells/mL. Cells were lysed on ice for 10 min with periodic vortexing to help break up the cell pellets and then aliquoted. Aliquots were either immediately stored at -20 °C or heated at 95 °C for 1 h, 70 °C for 2 h, or 70 °C for 3 h prior to storing at -20 °C. All samples were cleaned up via methanol/chloroform precipitation[48] and resuspended in 95:5 H_2O:ACN with 0.2% FA. Equivalent volumes of each sample were analyzed via HPLC-ESI-MS/MS with the aforementioned system. The column employed was a 100 µm id × 365 µm od fused silica capillary microcolumn packed with 20 cm of 5 µm diameter, 1000 Å pore size PLRP-S resin (Agilent) with an emitter tip pulled to ~1

μm using a laser puller (Sutter Instruments). The column was operated at 60 °C using a column oven and proteins were loaded on-column with 5% ACN in 0.2% FA at a flow rate of 0.5 μL/min for 30 min. Proteins were eluted over 74 min at a flow rate of 0.5 μL/min with the following gradient (all in 0.2% FA): 20% ACN at time 7 min, 65% ACN at time 57 min, 85% ACN from 67-74 min. The gradient was then ramped to 5% ACN in 0.2% FA at a flow rate of 0.5 μL/min over 16 min and held for an additional 15 min. Full-mass profile scans were performed between 500 and 1,600 *m/z* at a resolution of 120,000. Four microscans were averaged, using an AGC target of 1×10^6 and a maximum injection time of 100 ms. The top two most intense peaks in the MS1 with $z > 2$ were selected for HCD fragmentation with a normalized collision energy setting of 25. The MS2 resolution was set to 60,000, the isolation window was 4 *m/z* units, and the AGC target was 1×10^6 with a maximum injection time of 1,000 ms. Dynamic exclusion was enabled with a duration of 30 s. Source-induced dissociation was set to 15.0 eV.

Mass spectrometric data were searched using MetaMorpheus version 0.0.319. Calibration, G-PTM-D, and search tasks were performed using default parameters for top-down data, except carbamidomethylation on cysteine was not selected as a fixed modification and methylation/dimethylation of lysine and arginine, trimethylation of lysine, and formylation of lysine were selected as modifications to look for in G-PTM-D, in addition to the default modifications. Data were searched against the pruned multi-protease G-PTM-D database generated in Dai and Buxton et al.[43] (also Chapter 3 of this dissertation) and the built-in MetaMorpheus database of common contaminants.

Evaluation of suspension trapping (S-Trap) for top-down proteomics sample preparation

Fractions four and five from GELFrEE-fractionated[49] Jurkat cell lysate were used as the protein sample to evaluate S-Trap for top-down proteomics. To obtain these GELFrEE fractions, non-cross-linked Jurkat cells (2×10^7) were resuspended in 2 mL of lysis buffer containing 4% (w/v) SDS, 100 mM Tris-HCl pH = 7.5, 10 mM DTT, and 1× protease/phosphatase inhibitors. The cells were lysed on ice for 10 min with frequent vortexing and sonicated for 20 s with 10 s on/off intervals at setting 2 using a Misonix Ultrasonic Processor XL 2015 equipped with a microtip. The lysate was centrifuged at 6,000 × *g* for 5 min at room temperature to clear insoluble cellular debris and the supernatant was incubated for 30 min at room temperature to allow for reduction of protein disulfide bonds. Three volumes of ice-cold acetone were added to the sample and the sample was incubated at -20 °C for 1 h to precipitate proteins. The sample was centrifuged at 20,800 × *g* for 15 min at 4 °C and the supernatant was discarded. The pellet was washed with acetone and centrifuged again at 20,800 × *g* for 15 min at 4 °C. The supernatant was discarded and the pellet was allowed to air dry for 5 min. The pellet was resuspended in 400 µL of 1% SDS, heating at 60 °C for 5 min to promote solubilization. Protein concentration was measured using the BCA assay (Thermo Scientific #23225). Two 100 µg aliquots of protein were then reduced and prepared for GELFrEE fractionation as described in Dai and Buxton et al.[43] (also Chapter 3 of this dissertation). The two aliquots were fractionated based on molecular weight in separate lanes of a 12% Tris-acetate GELFrEE cartridge (Expedeon) as described[43] and the fractions were stored in low-protein-binding tubes at -80 °C. Fractions four and five from the two GELFrEE

lanes were then combined and cleaned up via two rounds of methanol/chloroform precipitation. Protein concentration was measured via BCA assay and the sample was aliquoted and stored at -20 °C in 95:5 H_2O:ACN with 0.2% FA for use in S-Trap experiments.

To evaluate S-Trap for top-down proteomics sample preparation, 300 ng of protein from GELFrEE fractions four and five were diluted in 1.575 mL of buffer containing 346 mM LiCl, 351 mM Tris-HCl pH = 7.5, 0.092% LiDS, 0.092% Triton X-100, and 5.1 µM release oligonucleotide. This buffer composition was chosen to match the buffer composition of the HyPR-MS release buffer after extra Tris-HCl is added to facilitate formaldehyde cross-link reversal, and this volume is ~22% the volume of a top-down HyPR-MS release sample after adding excess Tris-HCl. The sample was split into two tubes and 20% (w/v) SDS was added to each tube to a final concentration of 5% (w/v). Phosphoric acid (12%) was added to each tube to a final concentration of 1.2%, then each tube was aliquoted into 233 µL portions (10 total). The remainder of the S-Trap procedure was performed as described in the top-down HyPR-MS protein sample preparation part of the Methods section

The eluted protein sample (~10 µL) was analyzed immediately to minimize formylation artifacts. Additionally, a 150 ng sample of protein from GELFrEE fractions four and five was analyzed directly (no S-Trap) to serve as a control for the S-Trap sample. Samples were analyzed via HPLC-ESI-MS/MS using the aforementioned setup. The column employed was a 100 µm id × 365 µm od fused silica capillary microcolumn packed with 20 cm of 3 µm diameter, 300 Å pore size C2 beads (Separation Methods Technologies, Inc.) and an emitter tip pulled to ~1 µm using a laser puller (Sutter

Instruments). The column was operated at 60 °C using a column oven and proteins were loaded on-column with 5% ACN in 0.2% FA at a flow rate of 0.6 µL/min for 20 min. Proteins were eluted over 57 min at a flow rate of 0.5 µL/min with the following gradient (all in 0.2% FA): 20% ACN at time 2 min, 65% ACN at time 37 min, 85% ACN from 47-57 min. The gradient was then ramped to 5% ACN in 0.2% FA at a flow rate of 0.6 µL/min over three minutes and held for an additional 10 min. Full-mass profile scans were performed between 600 and 2,000 *m/z* at a resolution of 240,000. Three microscans were averaged, using an AGC target of 1×10^6 and a maximum injection time of 50 ms. The top eight most intense peaks in the MS1 with $z > 4$ were selected for HCD fragmentation with a normalized collision energy setting of 25. The MS2 resolution was set to 120,000, the isolation window was 4 *m/z* units, and the AGC target was 1×10^6 with a maximum injection time of 200 ms. Dynamic exclusion was enabled with a duration of 30 s. Source-induced dissociation was set to 15.0 eV.

Mass spectrometric data were searched using MetaMorpheus. Calibration and G-PTM-D were performed using MetaMorpheus version 0.0.319, while searching was performed using a modification of MetaMorpheus version 0.0.318 (modification is that it reports the number of PrSMs corresponding to a particular proteoform in each individual .raw file). The calibration, G-PTM-D, and search tasks were performed using default parameters for top-down data, except carbamidomethylation on cysteine was not selected as a fixed modification and methylation/dimethylation of lysine and arginine, trimethylation of lysine, and formylation of lysine were selected as modifications to look for in G-PTM-D, in addition to the default modifications. Data were searched against the pruned multi-protease G-

PTM-D database generated in Dai and Buxton et al.[43] (also Chapter 3 of this dissertation) and the built-in MetaMorpheus database of common contaminants.

5.4 RESULTS AND DISCUSSION

HyPR-MS efficiently and selectively purifies MALAT1 from Jurkat cell lysate and identifies MALAT1-binding proteins in bottom-up experiment

The HyPR-MS protocol described herein was adapted from the protocol described in Spiniello et al.[32] There, bottom-up HyPR-MS was employed to study the protein interactomes of three lncRNAs (MALAT1, NEAT1, and NORAD) in the human PC-3 prostate cancer cell line. Here, we sought to expand the utility of HyPR-MS by employing top-down mass spectrometry to identify the ensemble of proteoforms bound to a particular RNA transcript (MALAT1) for the first time. We chose to perform this work using the human Jurkat acute lymphoblastic leukemia (ALL) cell line[35] rather than PC-3 cells, as Jurkat is a suspension cell line and therefore is better suited than PC-3 (adherent) to growing the large number of cells required to develop and perform top-down HyPR-MS experiments. As in prostate cancer, upregulation of MALAT1 has been observed in a variety of leukemias,[50–52] including relapsed ALL.[53]

Five capture oligonucleotides were designed to complement the MALAT1 transcript. The sequences of these oligonucleotides were selected to meet and balance several criteria: 1) high

specificity (assessed via BLAST[54]); 2) ability to form stable hybrids with MALAT1 under the conditions employed in HyPR-MS; 3) low propensity to form homodimers/heterodimers; 4) complementarity to regions of the MALAT1 transcript predicted by Mfold[55] to be at least partially single-stranded; and 5) complementarity to different regions along the MALAT1 transcript (from 5' end to 3' end). Additionally, a scrambled sequence capture oligonucleotide with approximately the same melting temperature and G/C content as the MALAT1 capture oligonucleotides was designed. This scrambled capture oligonucleotide was assessed via BLAST to ensure that it was not significantly complementary to any part of the human genome/transcriptome, thereby allowing it to serve as a control to identify nonspecific binders present in the MALAT1 HyPR-MS pulldowns. Finally, a poly(dT) capture oligonucleotide was designed to enable capture of polyadenylated RNAs, such as mRNAs. A poly(A) HyPR-MS pulldown can serve as a control to help determine which proteoforms identified in the MALAT1 pulldowns are MALAT1-specific, and which are general RNA-binding proteoforms. Each of the seven capture oligonucleotides used in this study contained a biotin-TEG tag on the 3' end to enable capture on streptavidin-coated magnetic beads and an 8 nt toehold sequence on the 5' end to enable toehold-mediated release and multiplexing of the HyPR-MS capture experiments.[32,56] Finally, seven release oligonucleotides were designed to be complementary to their respective capture oligonucleotides. The sequences of all capture and release oligonucleotides used in this study can be found in Supplementary Table S-5.1.

Before attempting to identify proteoforms bound to MALAT1 using top-down HyPR-MS, we sought to verify that the experimental parameters employed were appropriately isolating the target transcript and its associated proteins. We established RNA capture/release conditions by performing a series of "small-scale" HyPR-MS experiments, wherein a relatively small number of cells (1×10^6) were used and capture/release conditions (e.g., buffer concentrations, capture/release oligonucleotide concentrations, hybridization/release temperature, washing conditions, sonication time/intensity, etc.) were varied (see the Supplementary Information and Supplementary Figure S-5.1 for a discussion of sonication parameters and RNA integrity in HyPR-MS experiments). RT-qPCR was used to assess capture efficiency and specificity, but no mass spectrometry was attempted due to the very small amount of protein in the release sample.

After establishing appropriate RNA capture/release conditions through these small-scale experiments, we performed a full-scale bottom-up HyPR-MS experiment to verify that the capture parameters worked on a larger scale and that the proteins identified in the MALAT1 release sample were truly MALAT1-binding proteins. RT-qPCR was performed to assess capture efficiency, specificity, and other aspects of the MALAT1 pulldown experiment. Four different qPCR assays targeting different regions of the MALAT1 transcript were used (Supplementary Table S-5.2), and the average MALAT1 capture efficiency measured in the MALAT1 release sample was ~49% (Figure 5.3A). Here, capture efficiency is defined as the percentage of MALAT1 RNA present in the lysate at the beginning of the experiment that is captured on and subsequently released from the streptavidin-coated

beads. There was no appreciable capture of MALAT1 in the scrambled sample, and no capture of a control, housekeeping RNA (GAPDH) in either release sample (Figure 5.3A). Some MALAT1 did remain on the beads following release (Figure 5.3A), but the average MALAT1 release efficiency was high (~77%) (Figure 5.3D). Here, release efficiency was calculated by dividing the percentage of the MALAT1 transcript in the MALAT1 release sample by the summed percentage in the MALAT1 release sample, the scrambled release sample, and the beads sample. Furthermore, to ensure that the nucleic acid isolated in this experiment was, in fact, RNA and not DNA, we measured MALAT1 DNA capture efficiency in the various samples. For this experiment, DNA was isolated from each sample via phenol/chloroform extraction (rather than TRI Reagent extraction), and the reverse transcriptase enzyme was omitted from the reverse transcription reactions. Figure 5.3B shows that very little (less than ~1%) MALAT1 or GAPDH DNA was captured in any sample, as expected. We also measured MALAT1 capture specificity, defined as the amount of MALAT1 RNA in the MALAT1 release sample relative to the amount of MALAT1 RNA in the scrambled release sample. Figure 5.3C shows that, on average, there was ~73 times more MALAT1 RNA in the MALAT1 release sample than in the scrambled release sample. The results of this RT-qPCR analysis indicate that the MALAT1 transcript was selectively and efficiently isolated in the MALAT1 release sample, validating that the capture/release conditions optimized in small-scale experiments are appropriate for full-scale HyPR-MS.

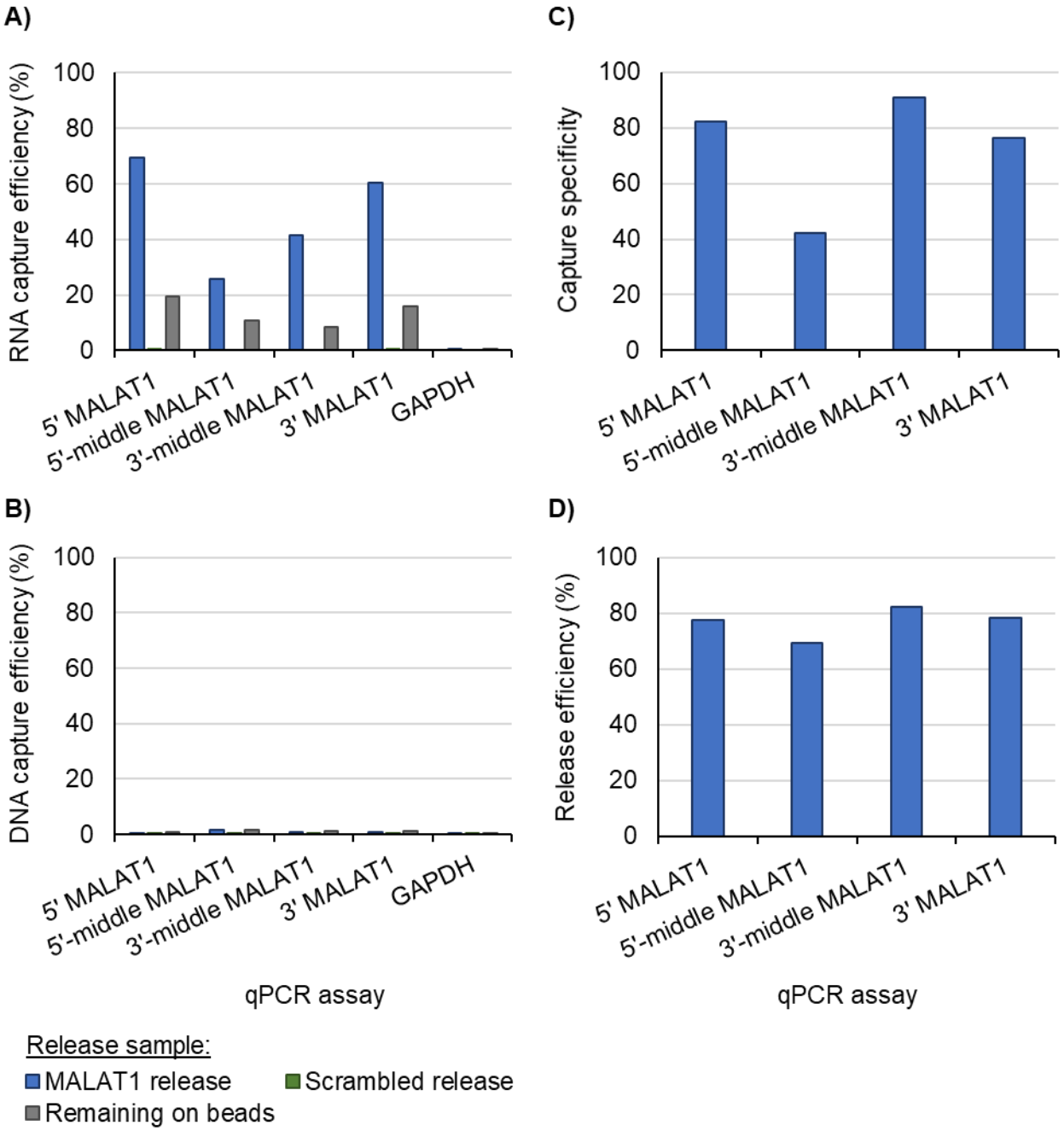


Figure 5.3 Bottom-up HyPR-MS capture efficiency and specificity. A) RNA and B) DNA capture efficiencies were measured using qPCR assays which target various regions of MALAT1 and GAPDH. Capture efficiencies measured in the MALAT1 and scrambled release samples are plotted, as well as the amount of each region that remained on the beads following all release steps. C) Capture specificity was calculated by taking the ratio of MALAT1 RNA capture efficiency in the MALAT1 release sample

(**Figure 5.3 continued**) to MALAT1 RNA capture efficiency in the scrambled release sample. Capture specificity was measured using each MALAT1 qPCR assay. D) RNA release efficiency was calculated by comparing the amount of MALAT1 RNA in the MALAT1 release sample to the sum of all MALAT1 RNA in the MALAT1 release sample, the scrambled release sample, and the beads sample. RNA release efficiency was measured using each MALAT1 qPCR assay.

Next, the proteomics data collected from this bottom-up HyPR-MS experiment were evaluated to see whether the proteins identified in the MALAT1 release sample matched those identified in our previous MALAT1 bottom-up HyPR-MS work.[32] Overall, 2,042 protein groups were observed at 1% FDR across all four samples (two MALAT1 and two scrambled technical replicates), with 1,926 non-contaminant protein groups quantified in at least one MALAT1 technical replicate. To obtain a rough estimate of which of these protein groups were more abundant in the MALAT1 release sample than the scrambled release sample, individual protein group intensities were normalized to the sum of all protein group intensities for each respective release sample. The normalized intensity for each protein group was then averaged across technical replicates and used to calculate a MALAT1/scrambled fold-change for each protein group. From this analysis, 321 protein groups had MALAT1/scrambled fold-changes greater than 3. We will henceforth refer to these 321 protein groups as the "MALAT1-enriched proteins" identified in this analysis. The MALAT1-enriched proteins from this study contained 52 (41%) of the 127 MALAT1-enriched proteins identified in our previous work,[32] with an additional 19 of those proteins (56% total) identified in this experiment with a fold-change greater than 2 or not computed because the protein was not quantified in the scrambled release sample.

A gene ontology (GO) term analysis of the MALAT1-enriched proteins from this study was performed using PANTHER[57,58] (overrepresentation test released on February 24, 2021). GO terms which reflect the known function and localization of MALAT1,[37,39,59] including "RNA-binding", "RNA splicing", "spliceosomal complex", "ribonucleoprotein complex", "nucleus", "nuclear speck", "paraspeckles", and "extracellular vesicle", were overrepresented at 1% FDR, providing confidence that the proteins identified in the MALAT1 release sample are truly MALAT1-binding proteins (Figure 5.4A).

We next assessed the masses of the MALAT1-enriched proteins from this study. Identifying high-molecular weight proteoforms is a well-established challenge in the field of top-down proteomics,[31] so it was important to assess whether suspected MALAT1-binding proteoforms could feasibly be identified from top-down data collected using an orbitrap mass analyzer (the mass analyzer employed in this study). The 321 MALAT1-enriched proteins had unmodified masses ranging from 5.0 to 308.6 kDa (Figure 5.4B). While many of these proteins had high molecular weights (> 40 kDa) and are therefore unlikely to be identified via standard top-down proteomics, about 25% had masses below 30 kDa (Figure 5.4B), a more feasible molecular weight range for analysis via top-down mass spectrometry.[31] We performed a GO term analysis on this subset of MALAT1-enriched proteins with masses less than 30 kDa and found many of the same overrepresented GO terms that we found in the full analysis including all 321 proteins (Figure 5.4A). Overall, the results from this bottom-up HyPR-MS experiment provided confidence that the parameters employed effectively isolated the target

MALAT1 transcript and its associated proteins, and that at least a portion of these MALAT1-associated proteins have masses that are amenable to analysis via top-down mass spectrometry.

A)

GO term	All proteins		Proteins ≤ 30 kDa	
	q-value	Number of proteins	*q*-value	Number of proteins
Nucleus	1.55×10^{-53}	260	6.9×10^{-5}	39
RNA-binding	1.01×10^{-118}	197	1.02×10^{-10}	26
RNA splicing	4.86×10^{-84}	100	2.97×10^{-6}	12
Ribonucleoprotein complex	3.31×10^{-58}	97	6.65×10^{-6}	13
Nuclear speck	3.76×10^{-44}	69	4.95×10^{-6}	11
Spliceosomal complex	6.19×10^{-54}	62	2.67×10^{-7}	10
Extracellular vesicle	6.38×10^{-4}	59	-	-
Paraspeckles	7.78×10^{-6}	5	-	-

B)

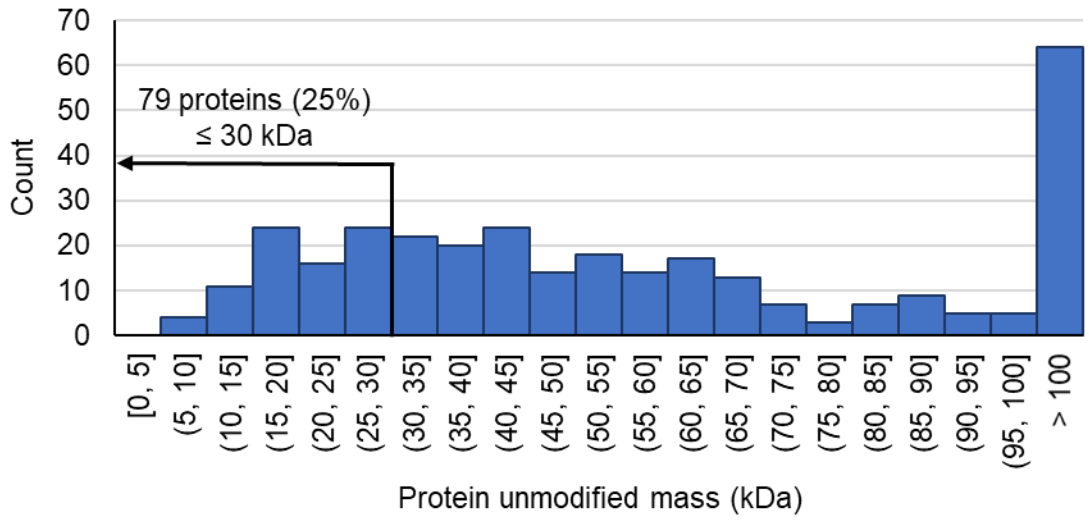


Figure 5.4 Analysis of proteomics data from bottom-up HyPR-MS experiment. A) GO term analysis of proteins identified as enriched in the MALAT1 release sample. Select overrepresented GO terms are listed, along with the number of enriched proteins with that particular GO term and the *q*-value (Benjamini-Hochberg correction). This GO term analysis was performed first with all 321 MALAT1-enriched proteins, then with just the 79 MALAT1-enriched proteins with masses ≤ 30 kDa. B) Histogram of the masses (unmodified) of the 321 MALAT1-enriched proteins. Seventy-nine proteins had masses ≤ 30 kDa, a rough cutoff for proteins amenable to standard top-down proteomics.

Heat reverses formaldehyde cross-links to enable proteoform identification via top-down mass spectrometry

Cells used for HyPR-MS experiments are treated with formaldehyde in culture to covalently fix in vivo protein-RNA interactions that might otherwise be disrupted by the use of denaturing detergents and high-salt buffers. It is important that these cross-links stay in place throughout cell lysis, hybridization, capture, and elution, but they must be completely reversed prior to performing top-down proteomics. This is because in traditional top-down proteomics, proteoforms are identified by matching the observed precursor and fragment ion masses to a theoretical proteoform in a database. Any cross-links remaining on a proteoform will shift its mass such that it does not match that of a theoretical proteoform, thereby rendering the proteoform unidentifiable. The same general principle applies in bottom-up proteomics, though there the requirement for complete cross-link reversal is more lenient. In bottom-up proteomics (Figure 5.1, top), confident identification of a single unique peptide is sufficient to infer the presence of the parent protein. Therefore, as long as one peptide from a parent protein is free of unexpected mass shifts, that protein can be identified. For this reason, bottom-up proteomics is more permissive to incomplete cross-link reversal than is top-down proteomics, where a single unexpected mass shift anywhere on the proteoform can make identification impossible.

It is well-known that formaldehyde cross-links can be reversed with heat.[60,61] However, we were unable to find any guidance in the literature for performing top-down proteomics on

formaldehyde cross-linked samples. Thus, we assessed several cross-link reversal conditions by heating non-cross-linked and cross-linked Jurkat cell lysate at either 70 °C or 95 °C for varying amounts of time. After the reversal step, samples were cleaned up via methanol/chloroform precipitation and analyzed via mass spectrometry. We noted the number of confident PrSMs, proteoform identifications, and protein identifications in each sample, and the results are presented in Figure 5.5. It is clear that foregoing cross-link reversal (i.e., not heating) is not a viable option, as far fewer PrSMs (Figure 5.5A) and proteoform identifications (Figure 5.5B) are observed in the cross-linked, no heat sample as compared to the not cross-linked, no heat control sample. Of the reversal conditions assessed here, heating at 70 °C for 3 h appeared to be the best option, as the PrSM and proteoform identification counts for this condition were most comparable to those of the not cross-linked, no heat control sample. Furthermore, the majority of the proteins identified in these two samples were the same (Figure 5.5C). Based on these results, we decided to heat HyPR-MS release samples at 70 °C for 3 h to reverse cross-links prior to performing top-down proteomics.

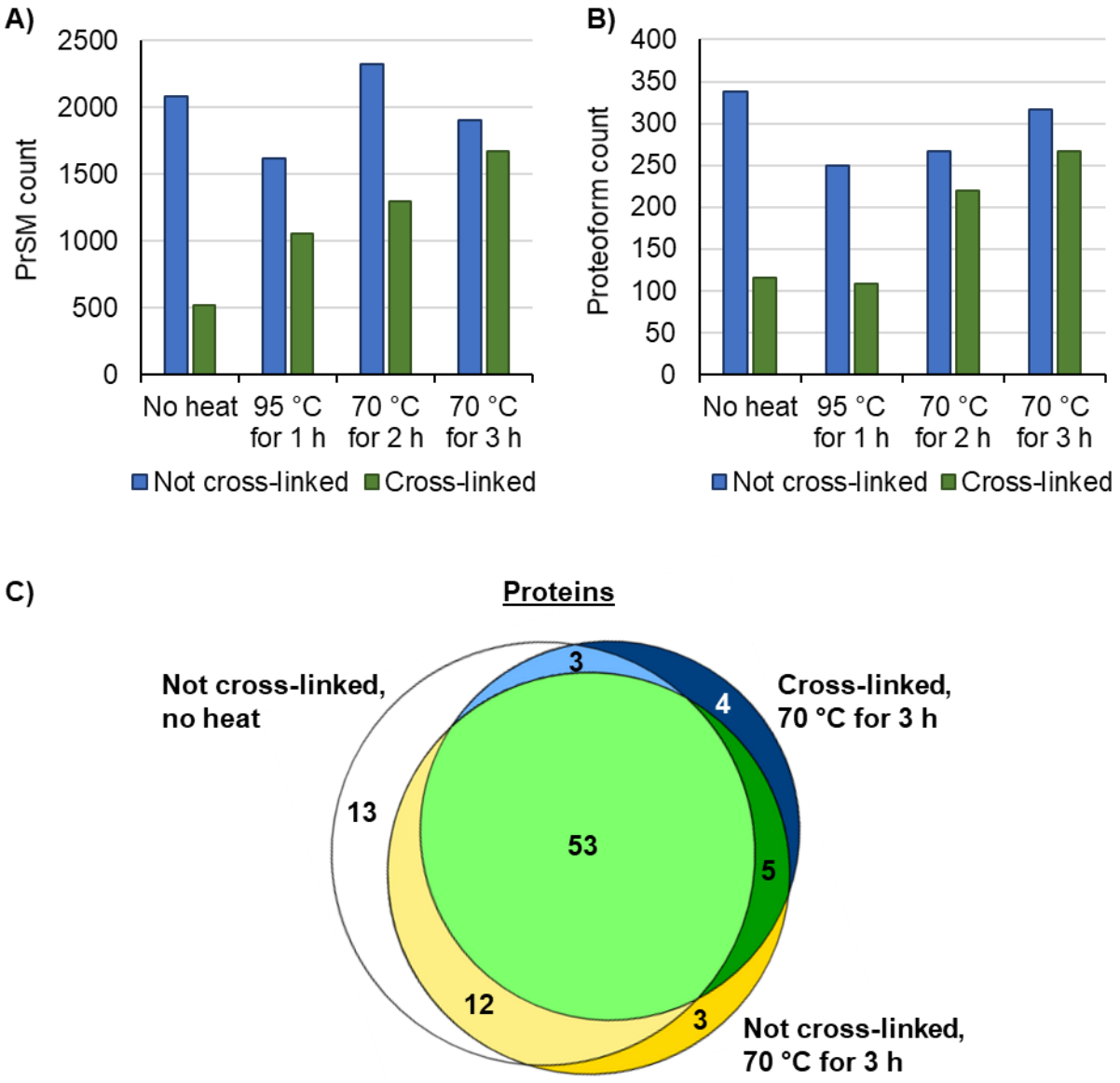


Figure 5.5 Non-cross-linked Jurkat cells and Jurkat cells cross-linked with 1% (w/v) formaldehyde were lysed and heated for varying amounts of time prior to sample cleanup and top-down mass spectrometric analysis. The data were searched and the number of proteoform spectral matches (PrSMs) (A) and proteoform identifications (B) at 1% FDR are plotted for each sample. Additionally, the proteins identified in the non-cross-linked, no heat; non-cross-linked, heated at 70 °C for 3 h; and cross-linked, heated at 70 °C for 3 h samples were compared and the resulting Venn diagram is shown in (C). The Venn diagram was generated using a tool developed at the Pacific Northwest National Laboratory (https://github.com/PNNL-Comp-Mass-Spec/Venn-Diagram-Plotter).

Suspension trapping (S-Trap) is a viable technique for the preparation of small amounts of protein for top-down proteomics

In HyPR-MS, purified RNA-protein complexes are eluted from the streptavidin-coated magnetic beads and released into a relatively large volume (6.525 mL) of buffer containing salts and detergents. Before analyzing the proteoforms in a release sample via top-down mass spectrometry, the sample must be concentrated significantly and all salts and detergents, which interfere with mass spectrometric analysis,[30] must be removed (henceforth, this process will be referred to as "sample preparation"). Further complicating the matter is that HyPR-MS release samples contain a very small amount of protein (~0.3 µg and ~0.1 µg of protein for the MALAT1 and scrambled release samples, respectively, based on analysis of bottom-up data (see the Supplementary Information and Supplementary Figure S-5.2)), and this small amount of protein must be retained throughout the sample preparation process in order to produce meaningful mass spectrometric data. Note that top-down proteomics is often performed on samples containing tens to hundreds of micrograms or even milligrams of protein, as top-down proteomics is much less sensitive than bottom-up proteomics.[22]

Given these myriad challenges, it was critical to carefully evaluate protein sample preparation methods for top-down HyPR-MS. We began by assessing established sample preparation techniques including acetone precipitation,[62,63] trichloroacetic acid precipitation,[64] methanol/chloroform precipitation,[48] and centrifugation using a molecular weight cutoff (MWCO) filter.[65,66] Despite multiple iterations and method adjustments, the precipitation techniques either provided insufficient

protein recovery, inadequate detergent removal, or both. Additionally, when testing the MWCO filters, we faced issues with incomplete detergent removal and clogging filters. Given recent studies showing the efficacy of suspension trapping (S-Trap) for bottom-up proteomics sample preparation,[67–69] we sought to evaluate the potential of S-Trap as a sample preparation method for top-down proteomics. The S-Trap procedure involves treating an SDS-containing protein sample with phosphoric acid to completely denature proteins. The acidified sample is then treated with a 90% methanolic solution at near-neutral pH, which creates a fine protein particulate suspension.[67] This suspension is then loaded onto the S-Trap device, which contains a trapping bed made of derivatized silica.[70] The unit is centrifuged and the protein particulates are trapped within the S-Trap bed. The proteins trapped within the S-Trap can then be washed with the methanolic solution, thereby removing salts, detergents, and other mass spectrometry-incompatible sample components. In a bottom-up sample preparation workflow, trypsin is then added to the S-Trap to digest the trapped proteins, and the resultant peptides can then be eluted in aqueous, mass spectrometry-compatible buffer. Importantly, a previous study using a similar protocol showed that protein amounts spanning two orders of magnitude—75 ng to 7.5 µg—could be prepared for bottom-up mass spectrometry using S-Trap, yielding an average of 763 to 2,096 protein identifications (for the 75 ng and 7.5 µg protein samples, respectively).[67]

We hypothesized that S-Trap could work for top-down proteomics sample preparation if the protease digestion step was excluded. In our eyes, the challenge to overcome was finding a way to elute

intact proteins off the S-Trap bed. In bottom-up S-Trap, intact proteins are held within the S-Trap bed until they are digested, at which time the resultant peptides are easily eluted with aqueous buffer because the silica bed does not possess the same affinity for peptides as it does for intact proteins.[70] In our trials, intact proteins did not elute from the S-Trap bed with pure water or 0.2% formic acid (FA). It has been shown that concentrated FA (80%) solubilizes precipitated proteins as well as 1% SDS,[71] so we hypothesized that it might also effectively elute proteins from the S-Trap bed. To test this, we prepared 300 ng of protein from lanes four and five of a 12% GELFrEE fractionation of Jurkat cell lysate in 1.575 mL of buffer containing 346 mM LiCl, 351 mM Tris-HCl pH = 7.5, 0.092% LiDS, 0.092% Triton X-100, and 5.1 µM release oligonucleotide. We chose this buffer composition to match the composition of the HyPR-MS release buffer after formaldehyde cross-link reversal, and this volume is ~22% the volume of a top-down HyPR-MS release sample after cross-link reversal. We chose to use protein from two lanes of a GELFrEE fractionation because we estimated that the complexity of the sample would mimic the complexity of a HyPR-MS release sample better than a mixture of standard proteins or proteins from whole cell lysate. We performed S-Trap on this sample, eluting the proteins from the S-Trap bed by passing the same 10 µL of 80% FA over the S-Trap bed five times. Note that we found that these five passes were important to elute the proteins from the S-Trap, as lower recovery was observed with just one pass and no obvious increase in recovery was observed with ten passes.

The eluted proteoforms were analyzed via top-down mass spectrometry and the data searched using MetaMorpheus. The results were compared to results from a 150 ng "control" sample of the

GELFrEE-fractionated protein that was analyzed directly, without going through the S-Trap procedure. The control sample yielded 79 protein identifications at 5% FDR, while the S-Trap sample yielded 59 (Figure 5.6A). Forty-three proteins were identified in both samples, while 36/16 proteins were identified only in the control/S-Trap sample, respectively (Figure 5.6B). It is worth noting that of the 52 protein groups identified in only one of the two samples, 49 were identified as two or fewer unique proteoforms. In other words, the majority of the proteoforms identified in these two samples reflected the same protein groups.

Approximately the same number of proteoforms were identified in these two samples, with 403 proteoforms identified in the control sample and 455 identified in the S-Trap sample (5% FDR) (Figure 5.6A). We compared the proteoforms identified in the two samples and found that 140 were found in both samples, while 263/315 were found only in the control/S-Trap sample, respectively (Figure 5.6C). Because the overlap in proteoform identifications seemed relatively small, we investigated the differences further. We first thought that perhaps the 80% FA used to elute proteins from the S-Trap might be introducing formylation artifacts which would not be observed in the control sample.[72] We found that 13 (~3%) of the proteoforms identified in the control sample and 38 (~8%) of the proteoforms identified in the S-Trap sample contained formylations (Figure 5.6A). So, it is possible that the concentrated formic acid used in the S-Trap procedure is introducing formylations, but these formylations do not appear to be ubiquitous and cannot completely account for the discrepancy in observed proteoforms. We next hypothesized that the proteoforms identified in only

one of the two samples might be relatively low abundance proteoforms, perhaps only identified by one or two PrSMs. Indeed, we found that 68% of the proteoforms only identified in the control sample and 71% of the proteoforms only identified in the S-Trap sample were identified with only one PrSM. In contrast, of the 140 proteoforms that were identified in both samples, 44% were identified with only one PrSM in the control sample and 46% were identified with only one PrSM in the S-Trap sample. Therefore, there is some evidence that proteoform abundance and the stochastic nature of proteoform sampling by the mass spectrometer may explain at least part of the discrepancy in proteoform identifications between the samples, and perhaps this discrepancy would become less pronounced if more replicates were performed. Interestingly, when comparing the proteoforms identified in two technical replicate injections of the 150 ng control sample, only ~32% of the proteoforms identified were found in both samples (Figure 5.6D). This overlap is not dramatically larger than the ~19% overlap that we observed when comparing proteoform identifications between the control and S-Trap samples (Figure 5.6C).

Based on the results of this analysis, we decided to move forward with using S-Trap to prepare HyPR-MS release samples for top-down proteomics. In our hands, other sample preparation methods were not suitable for HyPR-MS release samples, given the characteristics of these samples (i.e., small amount of protein in a dilute solution containing salts and detergents). S-Trap was able to sufficiently concentrate the sample and remove salts and detergents (including the difficult-to-remove Triton X-100) while still providing reasonable protein recovery. Note that in this experiment, we prepared 300

ng of protein with the S-Trap device, which is the amount of protein we estimate to be present in a HyPR-MS release sample in an experiment performed on 1×10^8 cells (Supplementary Figure S-5.2). Top-down HyPR-MS experiments are performed on 3×10^8 cells, and we expect a corresponding increase in the amount of protein in the release sample (900 ng). We do note that S-Trap does have at least one key drawback, which is the volume capacity of the S-Trap micro device. This device can only accommodate 250 µL of sample at a time, meaning that it takes hours to spin through a complete HyPR-MS release sample (which is ~56 mL after adding extra Tris-HCl for reverse cross-linking and SDS, DTT, phosphoric acid, and six volumes of methanolic binding buffer for the S-Trap procedure). We experimented with using a larger S-Trap device, but found that the increased size of the silica bed resulted in significant sample loss. Going forward, it would be worth further experimenting with decreasing HyPR-MS release volume, decreasing the amount of S-Trap binding buffer used, and/or increasing the volume capacity of the S-Trap micro device so that more sample can be spun through at a time. It would also be informative to perform more replicates of the experiment described here to better estimate protein recovery and to further investigate any biases that might be introduced by the S-Trap.

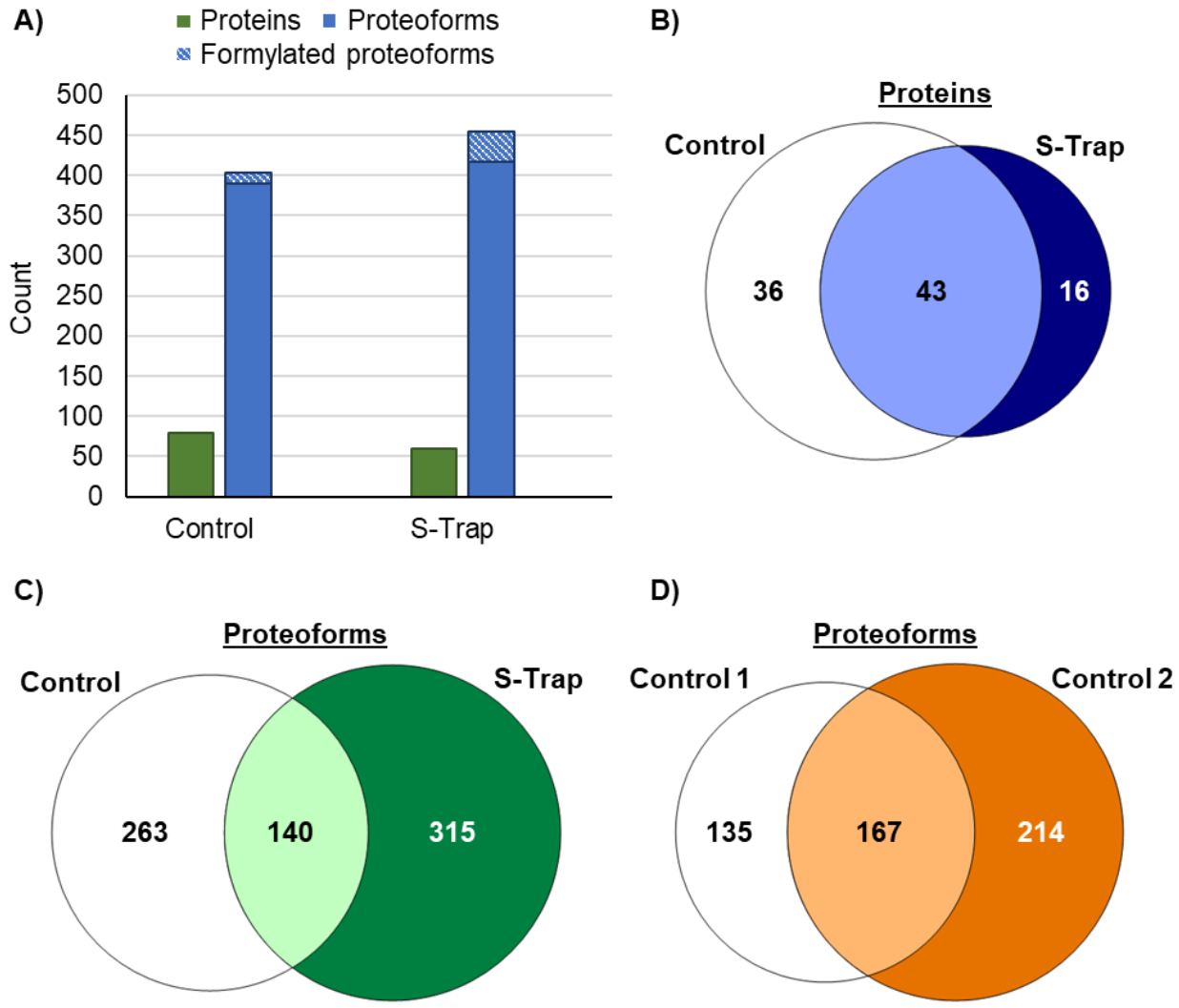


Figure 5.6 Comparison of proteins/proteoforms identified in a sample prepared for top-down mass spectrometry using S-Trap to those identified in a control sample (not subject to S-Trap). A) The number of proteins and proteoforms identified in each sample at 5% FDR is plotted. The proportion of formylated proteoforms identified in each sample is depicted with diagonal stripes. B) Venn diagram showing the overlap in protein identifications between the S-Trap and control samples. C) Venn diagram showing the overlap in proteoform identifications between the S-Trap and control samples. D) Venn diagram showing the overlap in proteoform identifications between replicate injections of the control sample. Venn diagrams were generated using a tool developed at the Pacific Northwest National Laboratory (https://github.com/PNNL-Comp-Mass-Spec/Venn-Diagram-Plotter).

Results of top-down HyPR-MS experiments for the identification of MALAT1-binding proteoforms

MALAT1 was effectively purified from Jurkat cell lysate in top-down HyPR-MS experiments

After establishing appropriate capture conditions for MALAT1 and developing S-Trap as a sample preparation technique for the analysis of proteoforms from HyPR-MS release samples, we performed top-down HyPR-MS experiments for the first time. Three biological replicates were performed to identify the proteoforms bound to MALAT1 RNA, and a summary of the capture results is shown in Figure 5.7. The average MALAT1 capture efficiency was ~15%, a reasonable number but lower than what was observed in our bottom-up study (Figure 5.3A). A small amount of MALAT1 RNA was detected in the scrambled release sample, about ~35 times less than what was observed in the MALAT1 release sample. Again, this capture specificity is reasonable, but lower than what was observed in our bottom-up study (Figure 5.3C). This lower capture specificity can be attributed to lower MALAT1 capture efficiency in the MALAT1 release sample, as no more MALAT1 RNA was observed in the scrambled release sample in this study than in the bottom-up study (Figure 5.3A). Some MALAT1 RNA was also detected in the poly(A) release sample, which perhaps can be explained by the fact that while MALAT1 is not a traditionally polyadenylated RNA, it does possess an A-rich tract near its 3' end[73] which could potentially hybridize to a poly(dT) capture oligonucleotide. As expected, GAPDH was captured by the poly(dT) capture oligonucleotide (~19% capture efficiency) and was observed in lower abundance in the MALAT1 and scrambled release samples (Figure 5.7). In all,

the capture results for these top-down HyPR-MS experiments were acceptable, but were not as good as those observed in the bottom-up experiment (Figure 5.3). We hypothesize that the lower MALAT1 capture efficiency may be due to RNA degradation, as the cells used in these top-down HyPR-MS experiments had been stored at -80 °C for approximately one year prior to use. While -80 °C is a recommended storage temperature for RNA samples,[74] it is possible that some RNA degradation still occurs at this temperature, especially when the RNA is stored in an unpurified state (i.e., in cells with cellular RNases still present). To test whether RNA integrity could be responsible for the lower capture efficiencies observed here, one could perform this experiment on freshly-grown cells and cells that had been frozen long-term and compare both capture efficiency and RNA integrity for those experiments.

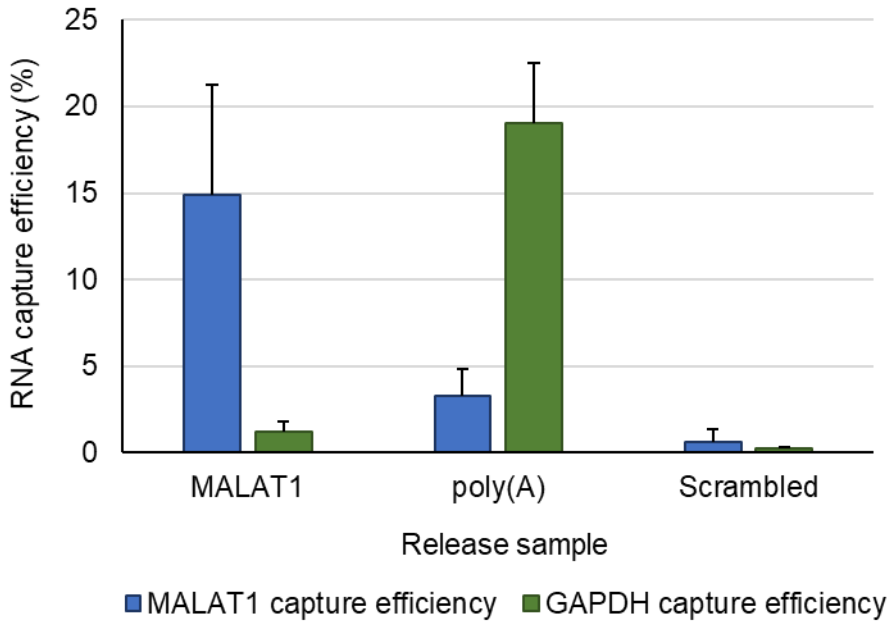


Figure 5.7 RNA capture efficiency in top-down HyPR-MS experiments. Capture efficiencies of MALAT1 and GAPDH RNA were measured in the MALAT1, poly(A), and scrambled release samples. The MALAT1 capture efficiency plotted is the average of the capture efficiencies measured using all four MALAT1 qPCR assays (Supplementary Table S-5.2). Error bars represent +1 standard deviation of the three replicate experiments.

Use of a sample-specific search database yielded the most proteoform identifications from top-down HyPR-MS data

Moving on to mass spectrometric data analysis, the first topic we addressed was choosing an appropriate search database. We evaluated several databases, including the UniProt reviewed human database (20,371 protein entries, downloaded on June 23, 2021), the UniProt reviewed plus unreviewed human database (78,120 entries, downloaded on June 23, 2021), a "pruned" multi-protease G-PTM-D database (73,928 entries), and a "protein-pruned" multi-protease G-PTM-D database (12,767 entries).

The latter two databases were generated in Dai and Buxton et al.[43] (also Chapter 3 of this dissertation), and descriptions of these databases can be found in the Methods section of this chapter. An immediate observation of the search results was that the poly(A) release samples gave fewer proteoform identifications than either the MALAT1 or scrambled release samples. We hypothesize that this is because the poly(A) release sample is more complicated than the other two samples, as it contains proteoforms bound to all polyadenylated RNAs. More complicated proteoform samples have an increased likelihood of proteoform coelution in LC-MS/MS experiments, and coelution makes proteoform identification more difficult. Additionally, during mass spectrometric analysis, we found that it was difficult to estimate how much of the poly(A) proteoform sample should be injected in order to provide good protein signal, as we observed inconsistencies in protein signal between the replicate HyPR-MS experiments. These inconsistencies were not observed with either the MALAT1 or scrambled release samples. For these reasons, we chose to exclude the poly(A) samples from the remainder of our top-down HyPR-MS proteomics analysis.

In terms of protein identifications, the protein-pruned multi-protease G-PTM-D database consistently gave the fewest (Figure 5.8A). This is most likely because that database contained the fewest protein entries and therefore may have been missing the sequences of some of the proteoforms observed in the data. The other three databases provided similar numbers of protein identifications (Figure 5.8A), with no database emerging as the clear choice. When comparing proteoform identifications, the pruned multi-protease G-PTM-D database consistently provided the most

identifications (Figure 5.8B). This was expected, as the pruned multi-protease G-PTM-D database included not only those PTMs annotated by UniProt, but also those discovered in deep bottom-up proteomics data from the Jurkat cell line.[43] This database was generated to better reflect the proteoforms that are actually present in Jurkat cells, and the search results from this work are evidence of the benefit of using such a sample-specific database.

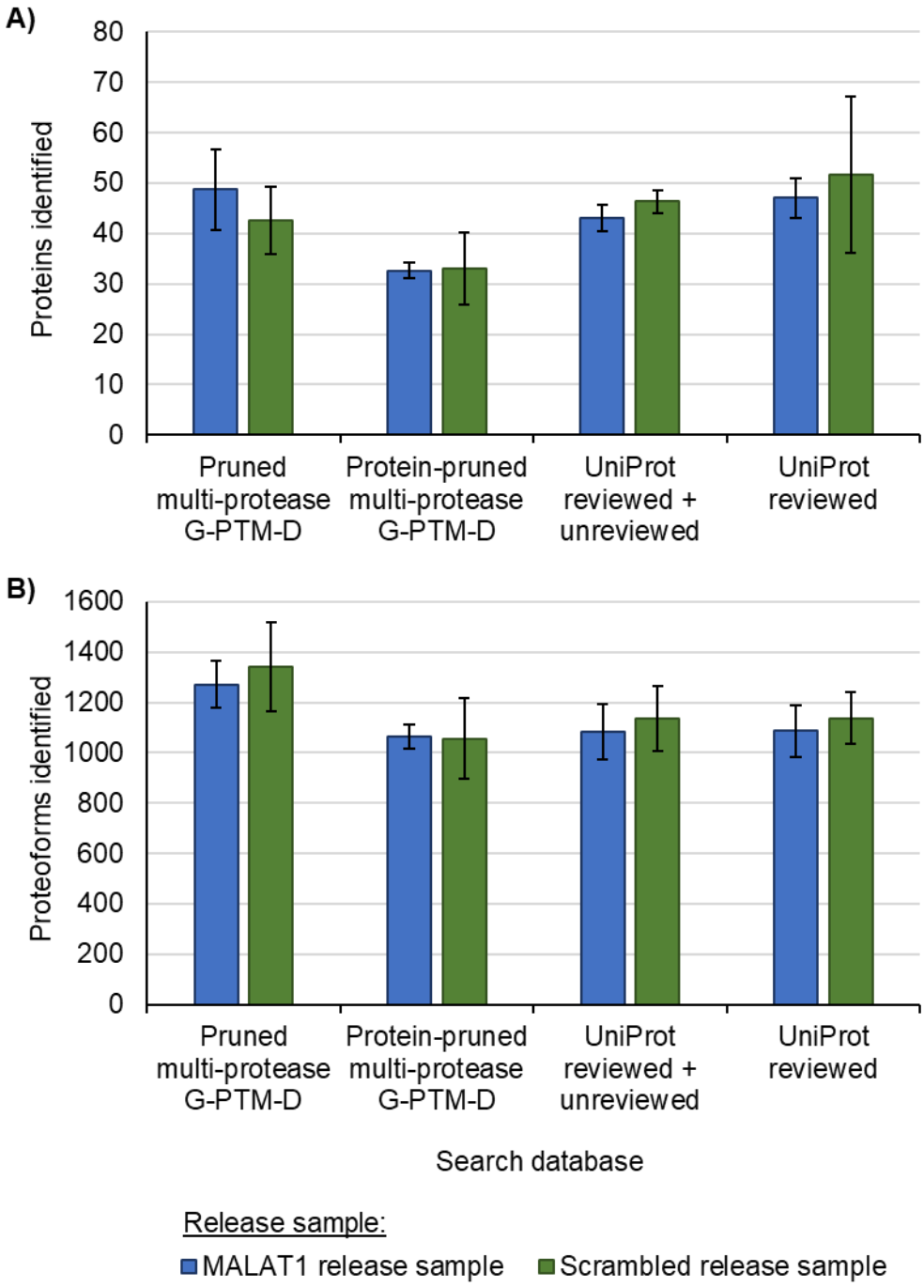


Figure 5.8 Database comparison for top-down HyPR-MS searches. Top-down data from three replicates of the MALAT1 and scrambled release samples were searched against the pruned and protein-pruned multi-protease G-PTM-D databases generated in Dai and Buxton et al.,[43] as well as the UniProt reviewed and UniProt reviewed plus unreviewed human databases. The average number of A) proteins

(**Figure 5.8 continued**) and B) proteoforms identified in the various searches at 5% FDR is plotted. Error bars show ±1 standard deviation of the three replicate samples.

Given our observation that using a sample-specific database can increase proteoform identifications, we hypothesized that a database informed by the results of our bottom-up HyPR-MS experiments might provide even more identifications from our top-down data. To test this hypothesis, we searched data from our previous bottom-up HyPR-MS study of MALAT1-binding proteins[32] and the bottom-up HyPR-MS data collected in this study against the pruned multi-protease G-PTM-D database. From this search, we created both pruned and protein-pruned databases which contained PTMs that were observed in the bottom-up HyPR-MS data but were not annotated in the original database (the new protein-pruned database also removed entries for any proteins that were not observed at 1% FDR in bottom-up HyPR-MS data). We searched our top-down HyPR-MS data against these two new databases, but found that the original pruned multi-protease G-PTM-D database performed as well as or better than either of these new databases in terms of both protein and proteoform identifications (Supplementary Figure S-5.3). Therefore, we decided to continue our analysis with the search results obtained using the pruned multi-protease G-PTM-D database from Dai and Buxton et al.[43]

Proteins identified in MALAT1 and scrambled top-down HyPR-MS release samples were similar

In HyPR-MS, RNA-binding proteins are identified by comparing the proteins identified in the RNA release sample of interest to those identified in a control sample. For this experiment, we sought

to identify MALAT1-binding proteoforms by comparing the proteoforms identified in the MALAT1 release sample to those identified in the scrambled control release sample. We began by doing qualitative comparisons of the proteins identified in these two sample types. We found good overlap in the proteins identified in different biological replicates of the same release sample type (Figure 5.9A and B). The majority of the proteins identified in all three replicates of a given sample type were histones, with 33/35 and 32/34 proteins identified in all three replicates of MALAT1 and scrambled being histones, respectively. In fact, of the 64 proteins identified in any replicate of the MALAT1 release sample, 35 (55%) were histones. Similarly, of the 54 proteins identified in any replicate of the scrambled release sample, 36 (67%) were histones. We next compared the proteins identified in the MALAT1 and scrambled samples to each other. There were 45 and 40 proteins identified in at least two biological replicates of the MALAT1 and scrambled samples, respectively. Thirty-eight of those proteins were identified in at least two biological replicates of both sample types (Figure 5.9C), and 34 of those shared proteins were histones. The complete list of proteins identified in each MALAT1 and scrambled sample can be found in Supplementary Table S-5.3.

There are numerous possible explanations for the abundance of histones observed in the HyPR-MS release samples. First, while histones are generally thought of as DNA-binding proteins, there are cases where histones do bind to RNA,[75,76] so their presence in the MALAT1 release samples may, at least to some extent, reflect meaningful biological interactions. Moreover, it has been shown that MALAT1 localizes to active chromatin sites,[15] and therefore histones within that chromatin may also

be pulled down with MALAT1 in HyPR-MS. Another explanation for the presence of histones in HyPR-MS release samples is nonspecific binding. Histones are basic proteins carrying a net positive charge and thus can be expected to interact with negatively-charged nucleic acids, such as DNA and RNA. Therefore, their presence in the release samples might be at least partially attributed to nonspecific, non-biologically-relevant interactions with captured RNA molecules and/or capture/release oligonucleotides. Additionally, the streptavidin-coated magnetic beads used in HyPR-MS may be partially responsible for the presence of histones in the release samples, as streptavidin has a slightly acidic pI of 6.1. While we do not know the exact amino acid sequence of the streptavidin coating the beads used in our HyPR-MS experiments (i.e., we do not know if/how it has been modified), if it did have a pI of 6.1, we would expect it to carry a net negative charge in HyPR-MS buffers, creating another potential source for nonspecific histone interactions.

We note that the presence of histones in HyPR-MS release samples is not a new phenomenon, and that histones have been observed in bottom-up HyPR-MS samples, as well. For example, of the 2,660 proteins identified in any replicate MALAT1 capture sample in our previous study of MALAT1-binding proteins,[32] 22 (~1%) were histones. Moreover, of the top 10% (266) most abundant proteins in those MALAT1 capture samples (as determined by intensity-based label-free quantification), six were histones. We also know that top-down mass spectrometry is best-suited to analyzing smaller proteins, such as histones (~14 kDa), because of the inverse relationship between signal-to-noise ratio and proteoform mass.[31] Therefore, the nature of histones, being both basic and relatively small, coupled

with their abundance in HyPR-MS samples makes it perhaps unsurprising that so many histones were observed in top-down HyPR-MS samples.

Finally, we compared the proteins identified in top-down HyPR-MS experiments to those identified in our previous bottom-up HyPR-MS experiments[32] to assess the overlap. Obviously, many more protein groups were identified in bottom-up experiments than in top-down experiments (2,901 at 1% FDR in bottom-up experiments and 72 at 5% FDR in top-down experiments), reflecting the better sensitivity and lack of protein mass restrictions in bottom-up proteomics. We cross-referenced the list of 47 proteins observed in at least two biological replicates of top-down MALAT1 or scrambled capture samples with the 2,901 proteins identified by bottom-up HyPR-MS and found that ~62% of the proteins identified in top-down experiments were also identified in bottom-up experiments (Figure 5.9D). Interestingly, ~72% of the proteins observed only in top-down data were histones, perhaps reflecting proteins that were erroneously removed by protein parsimony in the bottom-up search and therefore highlighting a strength of top-down proteomics, where the entire protein molecule is analyzed rather than just a short peptide (Figure 5.1, bottom). On a related note, we found a few cases where top-down proteomics was able to clarify results from the bottom-up search. The bottom-up search results reported the presence of three protein groups containing a total of eight distinct histone proteins. It was not possible to definitively state which of these histones were actually present in the samples from the bottom-up data, but the top-down data were able to confirm the presence of all eight proteins.

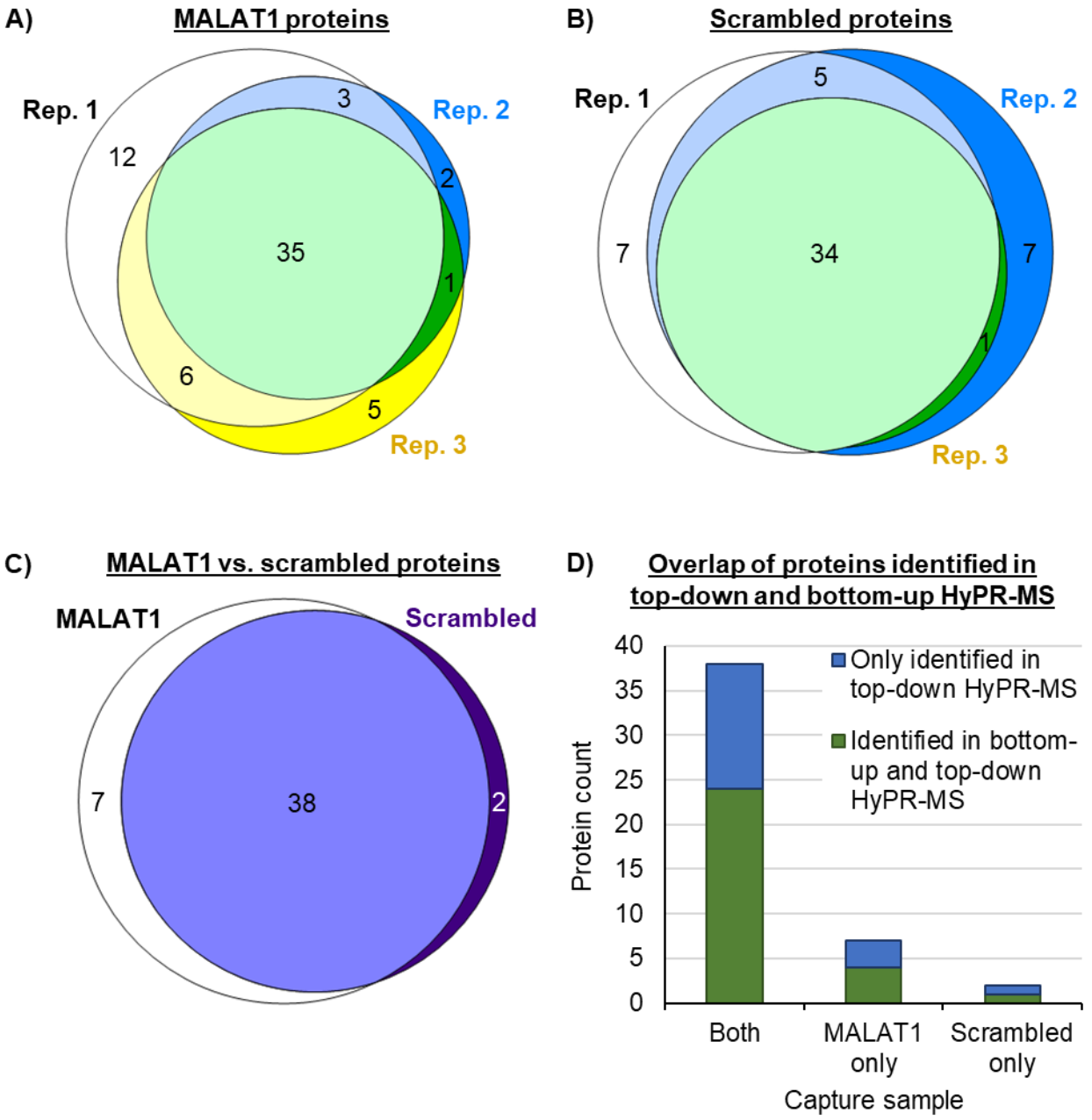


Figure 5.9 Qualitative evaluation of proteins identified in top-down HyPR-MS experiments. Venn diagrams comparing the proteins identified in different replicates of the MALAT1 and scrambled capture samples are shown in A) and B), respectively. C) Venn diagram comparing proteins identified in at least two replicates of the MALAT1 or scrambled capture samples. D) Comparison of the 47 proteins identified in at least two replicates of top-down HyPR-MS MALAT1 or scrambled capture samples to proteins identified in bottom-up HyPR-MS experiments.[32] Proteins are categorized according to whether they were identified only in MALAT1 top-down HyPR-MS samples, only in

(**Figure 5.9 continued**) scrambled top-down HyPR-MS samples, or in both. Venn diagrams were generated using a tool developed at the Pacific Northwest National Laboratory (https://github.com/PNNL-Comp-Mass-Spec/Venn-Diagram-Plotter).

Proteoforms identified in MALAT1 and scrambled top-down HyPR-MS release samples were similar

While it is informative to examine the proteins identified in MALAT1 and scrambled top-down HyPR-MS samples, the goal of the experiment was to identify the proteoforms interacting with MALAT1 RNA. To this end, we compared the proteoforms identified in the MALAT1 and scrambled capture samples. First, the proteoforms identified in different replicates of each capture sample type were compared to one another (Figure 5.10A and B). A total of 2,213 proteoforms were identified in the MALAT1 samples at 5% FDR, with 570 proteoforms (~47%) observed in at least two replicates (Figure 5.10A). A total of 2,529 proteoforms were identified in the scrambled samples at 5% FDR, with 512 proteoforms (~39%) observed in at least two replicates (Figure 5.10B). Overall, the overlap between any two replicates of a given sample type was ~30-40%, which is in good agreement with the overlap observed for replicate injections of the control protein sample used in our S-Trap studies (~32%, Figure 5.6D). Discrepancies in the proteoforms identified in replicates of the same capture sample type may be due to the introduction of artifactual PTMs during sample preparation (e.g., oxidation, deamidation, ammonia loss, formylation, etc.) or low proteoform abundance and the stochastic nature of proteoform sampling by the mass spectrometer. Overall, ~90% of the proteoforms observed in either sample type were histone proteoforms.

We next compared the proteoforms identified in the MALAT1 capture samples to those identified in the scrambled capture samples. Overall, 1,304 proteoforms were observed in at least two replicates of either the MALAT1 or scrambled capture samples, with 713 proteoforms being observed in at least two replicates of both sample types (Figure 5.10C). Notably, of the 318 proteoforms observed in at least two replicates of the MALAT1 capture sample but not in at least two replicates of the scrambled capture sample, we did find that 215 (68%) were found in one replicate of the scrambled sample, suggesting that these proteoforms were not unique to MALAT1 capture samples. A total of 103 proteoforms were observed in at least two replicates of the MALAT1 sample and no replicates of the scrambled sample, with 93 of these (90%) being histone proteoforms.

After qualitatively assessing the proteoforms present in the top-down HyPR-MS samples, we performed a quantitative analysis to determine whether any proteoforms were significantly enriched in the MALAT1 capture samples as compared to the scrambled capture samples. Proteoforms were quantified via label-free quantification with FlashLFQ[44] and Perseus[47] was used to perform Student's T-tests. Only three proteoforms showed differential abundance between the two sample types at 5% FDR (Figure 5.10D). One of these proteoforms, an acetylated and trimethylated proteoform from the RBIS (ribosome biogenesis factor) gene, was more abundant in the MALAT1 samples. The other two proteoforms were histone H2A type 1-D proteoforms, one of which was deamidated and the other of which contained a phosphoserine and a methylglutamine. We further investigated the RBIS proteoform that was reported as enriched in the MALAT1 samples. It was only identified by MS/MS

in one MALAT1 sample (it was quantified by match-between-runs in the other two MALAT1 samples), and it was not identified or quantified in any of the scrambled samples. We found only one PrSM for the proteoform in the MALAT1 sample in which it was identified via MS/MS, and this PrSM had a low MetaMorpheus score of 7.001, indicating that only seven peaks in the observed fragmentation spectrum matched the theoretical spectrum. Given this information, we do not have high confidence that this is a correct proteoform identification and did not pursue this finding any further. Overall, this quantitative comparison of the proteoforms present in the MALAT1 and scrambled capture samples did not reveal substantial differences between the target RNA capture sample and the control, and therefore we were unable to generate a list of MALAT1-binding proteoforms from the data collected in this study.

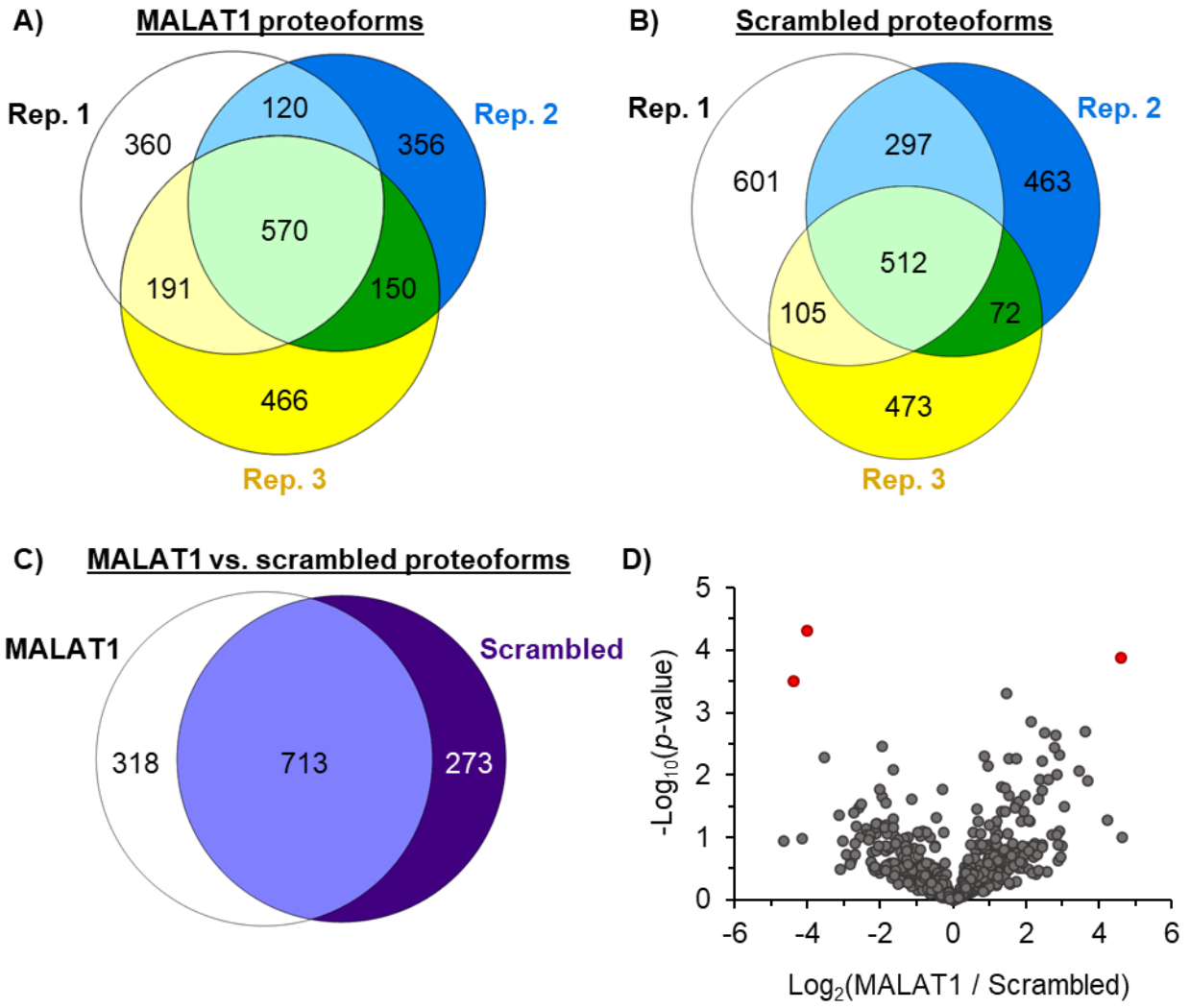


Figure 5.10 Analysis of proteoforms identified in top-down HyPR-MS experiments. Venn diagrams comparing the proteoforms identified in different replicates of the MALAT1 and scrambled capture samples are shown in A) and B), respectively. C) Venn diagram comparing proteoforms identified in at least two replicates of the MALAT1 or scrambled capture samples. D) Volcano plot showing the results of Student's T-tests performed to compare proteoform abundances between the MALAT1 and scrambled capture samples. Red points indicate proteoforms which were found to be significantly enriched in the MALAT1 or scrambled capture samples, with a permutation-based FDR of 5%. Venn diagrams were generated using a tool developed at the Pacific Northwest National Laboratory (https://github.com/PNNL-Comp-Mass-Spec/Venn-Diagram-Plotter).

Future directions for improving top-down HyPR-MS

The purpose of this study was to combine the RNA capture and isolation abilities of HyPR-MS with the proteoform identification abilities of top-down proteomics to identify the proteoforms interacting with a particular RNA species for the first time. While we were unable to accomplish this ambitious goal and generate a confident list of MALAT1-interacting proteoforms in this initial study, critical strides were made in the development of this technology and this study will serve as the foundation upon which further advancements are made. This study also suggested several potential future directions that may improve various aspects of this technology, which we will discuss below.

One of the most obvious issues encountered in this study was that the top-down proteomics data were dominated by histone proteoforms. As discussed, histones are relatively small, basic proteoforms, making them excellent analytes for top-down mass spectrometry. However, we hypothesize that the presence of at least a portion of these histones in the MALAT1 capture samples was due to nonspecific binding, rather than genuine in vivo interactions with MALAT1 RNA. This is because histones dominated the mass spectrometric data for both the MALAT1 and scrambled control capture samples, with ~90% of the proteoforms identified in either sample being histones. The first step to addressing this issue of nonspecific binding is to identify the source(s). As discussed, some potential sources include the capture/release oligonucleotides and streptavidin-coated magnetic beads used in HyPR-MS. We recommend performing a series of control experiments to pinpoint the exact source(s).

First, a "beads-only" experiment should be performed. In this experiment, the procedure for HyPR-MS would be followed except that no capture or release oligonucleotides would be added. We note that bottom-up HyPR-MS would likely be suitable for this experiment, as the goal is to get a general idea of how much histone protein is present in the sample and that can be accomplished without performing the more difficult top-down experiment. If a significant amount of histone protein is observed in the mass spectrometric data from this experiment, this would imply that histones from the cell lysate are binding to the streptavidin-coated beads during the capture phase of HyPR-MS and then later dissociating from the beads into the HyPR-MS release solution. If this proves to be the case, one could increase the stringency and/or number of washing steps between RNA capture and release to try to remove the majority of the histones from the beads prior to performing the release step. Additionally, one could attempt to perform the HyPR-MS procedure, at least from cell lysis through post-RNA capture bead washing, at a pH closer to the pI of streptavidin (6.1), rather than pH 7.5. This slightly acidic pH would serve to bring the streptavidin-coated beads to a more neutral net charge, thereby minimizing electrostatic interactions between positively-charged histones and negatively-charged beads. If nonspecific binding to the streptavidin-coated beads continues to be a problem after trying these modifications to the experiment, alternatives to using the biotin-streptavidin interaction for RNA capture could be explored. For example, alkyne-functionalized capture oligonucleotides could be used with azide-functionalized beads so that click chemistry could be used to purify target RNA-protein complexes from cell lysate.[77]

If the streptavidin-coated beads do not appear to be the source of the nonspecific histone binding, we recommend performing two additional control experiments. First, a HyPR-MS experiment with scrambled capture oligonucleotides but without release oligonucleotides should be performed. If a large amount of histone protein is observed in this experiment, but not in the "beads-only" experiment, that would suggest that the negatively-charged DNA capture oligonucleotides are the source of the nonspecific binding. If this proves to be the case, increasing the stringency and/or number of washing steps between RNA capture and release could be helpful, as could carefully titrating the amount of capture oligonucleotide required in order to minimize histone interactions with excess capture oligonucleotides. We also recommend performing a HyPR-MS experiment without capture oligonucleotides, but with release oligonucleotides. The reasoning is that if histones are not observed in the "beads-only" or "capture oligonucleotide-only" experiments, that may indicate that histones are not interacting with the beads or capture oligonucleotides, but it could also mean that histones interacting with those beads/oligonucleotides are simply not dissociating into the release solution in large numbers in the absence of release oligonucleotides. In HyPR-MS, a large excess of release oligonucleotide is used (~100× the amount of capture oligonucleotide used), thereby introducing a large number of negatively-charged molecules to the release solution. It is possible that histone-release oligonucleotide interactions could shift the histone binding equilibria enough that histones which would not dissociate from the beads/capture oligonucleotides in the absence of release oligonucleotides would dissociate in the presence of release oligonucleotides. Such histones would then become part of the release sample that is analyzed via mass spectrometry. If histones are observed

in this "release oligonucleotide-only" experiment, this would indicate that the beads are the source of the nonspecific binding, but the release oligonucleotides are required for bead-bound histones to be released into solution. This situation might be addressed by decreasing release oligonucleotide concentration and/or trying the aforementioned methods of reducing nonspecific binding of histones to the beads.

An alternate approach to reducing the presence of histones and other non-RNA-binding proteins in HyPR-MS release samples is to introduce an RNA-protein complex purification step prior to hybridization capture. In the present version of HyPR-MS, cells are lysed and target RNA-protein complexes are purified from this complex cell lysate. It could be beneficial to add an intermediate step of bulk RNA-protein complex isolation between cell lysis and RNA capture. For example, OOPS[13] (orthogonal organic phase separation) and pTex[14] (phenol toluol extraction) are methods developed to isolate cross-linked protein-RNA complexes from other components of cell lysate (e.g., free RNA/DNA, free protein, and protein-DNA complexes) based on their physiochemical properties. We therefore expect that implementing an OOPS or pTex step prior to hybridization capture in HyPR-MS would reduce the presence of many nonspecific protein binders in release samples.

In addition to making changes in the RNA capture/purification aspects of HyPR-MS, one can also imagine several ways to improve the proteoform sample preparation and mass spectrometric analysis aspects of top-down HyPR-MS. First, additional work could be done to optimize formaldehyde cross-linking/reverse cross-linking conditions. As discussed, it is critical to completely

reverse formaldehyde cross-links prior to top-down analysis, as any residual cross-links will shift the mass of the proteoform and make it unidentifiable in a database search. It may be beneficial to decrease the formaldehyde concentration and/or time used for cross-linking in order to minimize cross-links that must eventually be reversed. Furthermore, a more thorough analysis of the extent of cross-link reversal may be helpful. In addition to using top-down mass spectrometry as a readout, bottom-up mass spectrometry could be used with a cross-link-friendly search algorithm to identify the number of cross-links that remain after samples are subject to different reversal conditions.[78] Moreover, the aforementioned OOPs or pTex methods could be used to quantify the amount of RNA/protein that remains cross-linked after samples are subject to different reversal conditions. To our knowledge, there are no published accounts of top-down proteomics performed on proteoforms that were previously cross-linked with formaldehyde, making this an area ripe for exploration and advancement.

One could also imagine that pre-fractionation of proteoforms prior to performing LC-MS/MS could reduce the complexity of the sample, thereby enabling deeper proteoform sampling. In this study, online LC was the only means of proteoform separation prior to mass spectrometric analysis. The data were dominated by histone identifications, but it is possible that other proteoforms in the sample coeluted with these histones but were not selected for fragmentation because they were not as abundant. Fractionating (by ion exchange chromatography,[27,79] high-pH liquid chromatography,[80] serial size exclusion chromatography,[81,82] etc.) could help solve this problem of coelution and enable the identification of more proteoforms in HyPR-MS samples. Of course, a major tradeoff of

performing prefractionation is that introducing additional sample handling steps increases the chance of sample loss. HyPR-MS release samples contain only a small amount of protein as it is, making this a critical consideration. Should sample prefractionation be required, we recommend using a blocking agent to minimize nonspecific adsorption of proteoforms to surfaces and/or scaling up the number of cells used per HyPR-MS experiment to provide more protein in the release sample.

In addition to improving separations prior to mass spectrometric analysis, the method of mass spectrometric data collection could also be optimized. For example, targeted top-down proteomics,[83] informed by bottom-up HyPR-MS experiments, could be useful for detecting proteoforms of interest with higher sensitivity than would be possible with a traditional data-dependent acquisition strategy. Alternatively, an exclusion list[84] containing high-abundance proteoforms observed in the scrambled/"beads-only" control samples might be useful, so that instrument time could be spent analyzing the proteoforms that are more likely to be genuine MALAT1 interactors. Finally, the data described herein were collected using an orbitrap mass spectrometer (Q Exactive HF, Thermo Fisher Scientific), but others have had success in identifying higher molecular weight proteoforms and proteoform complexes when using time-of-flight (TOF) analyzers[81] or high-mass-range orbitrap instruments.[85,86]

5.5 CONCLUSION

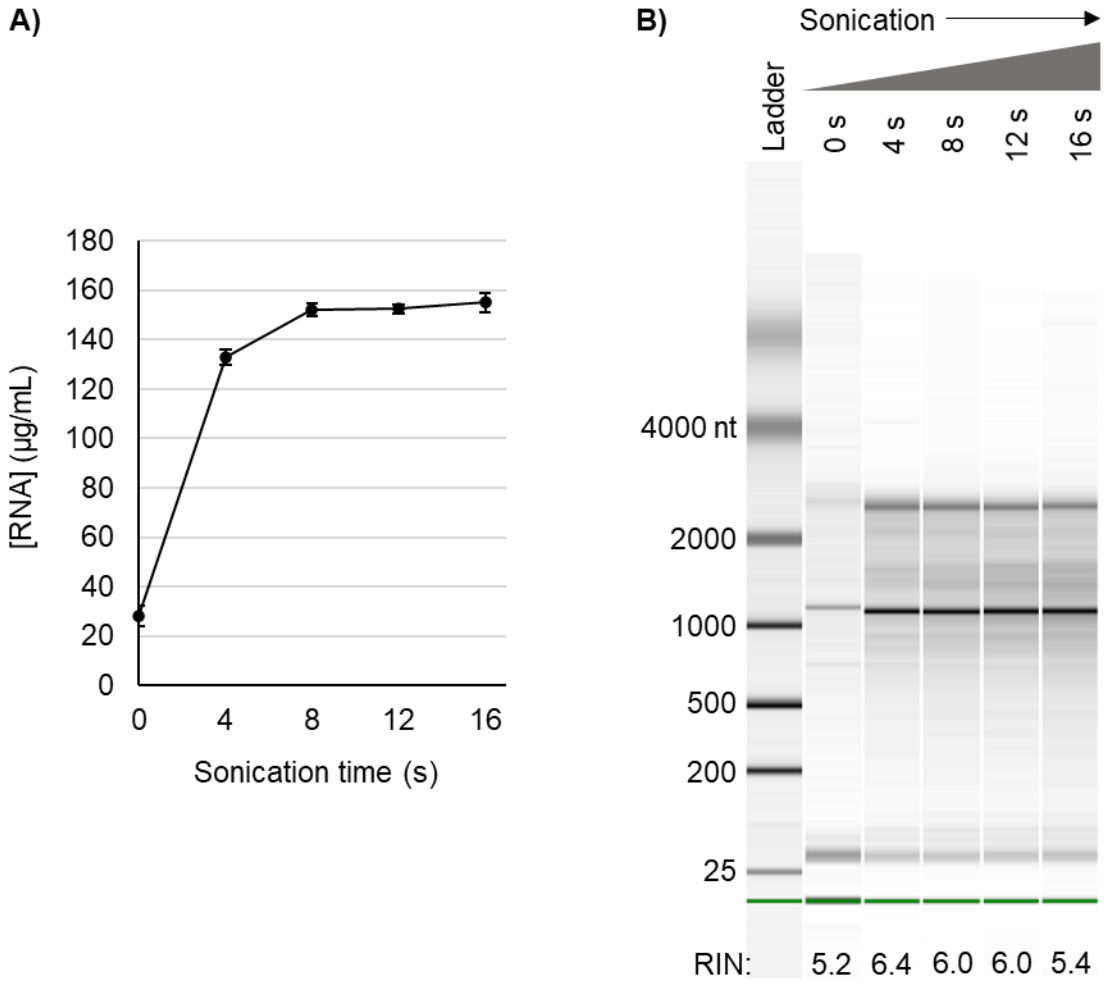
In this work, we have made important strides in the development of "top-down HyPR-MS", a technology for the identification of proteoforms interacting with target RNA species. We focused on establishing appropriate parameters for many aspects of the experiment including RNA solubilization, target RNA capture/release, formaldehyde cross-link reversal, protein sample concentration and clean-up, and mass spectrometric data analysis. We performed the first top-down HyPR-MS experiments to identify the proteoforms interacting with the MALAT1 lncRNA, and while we were unable to generate a confident list of interactors from these initial experiments, we have established a foundation upon which the technology can develop and have proposed numerous avenues to improve the technology moving forward. Given the importance of protein-RNA interactions and the fact proteoforms are crucial effectors of biological function, it is critical that we continue to work to develop tools to study these interactions on a proteoform level, such that we may better understand the complexities and intricacies of cellular biology.

5.6 SUPPLEMENTARY INFORMATION

Sonication titration and assessment of RNA integrity

In HyPR-MS, sonication is used to break up chromatin and help solubilize RNA-protein complexes. However, sonication can also cause RNA fragmentation, and therefore it is important not

to over-sonicate if one's goal is to identify the proteoforms bound along the entire length of the target transcript. In this work, sonication parameters were determined using absorbance at 260 nm to assess the extent of RNA solubilization and a Bioanalyzer electropherogram to assess RNA integrity (Supplementary Figure S-5.1). RNA solubility was very low without any sonication and plateaued after ~8 s of sonication (Supplementary Figure S-5.1A). Some degree of RNA degradation was observed in all samples (Supplementary Figure S-5.1B) (RIN values ranging from 5.2-6.4), but 28S and 18S rRNA bands are clearly visible in all samples, indicating that large molecular weight RNA species remain after sonication. To ameliorate the issue of RNA degradation/fragmentation in HyPR-MS samples, five capture oligonucleotides complementary to regions spanning the length of the MALAT1 transcript were used in HyPR-MS experiments, and four qPCR assays targeting different regions of the MALAT1 transcript were used to monitor capture efficiency. From this work, twelve seconds of sonication was selected as the condition to use in HyPR-MS experiments.

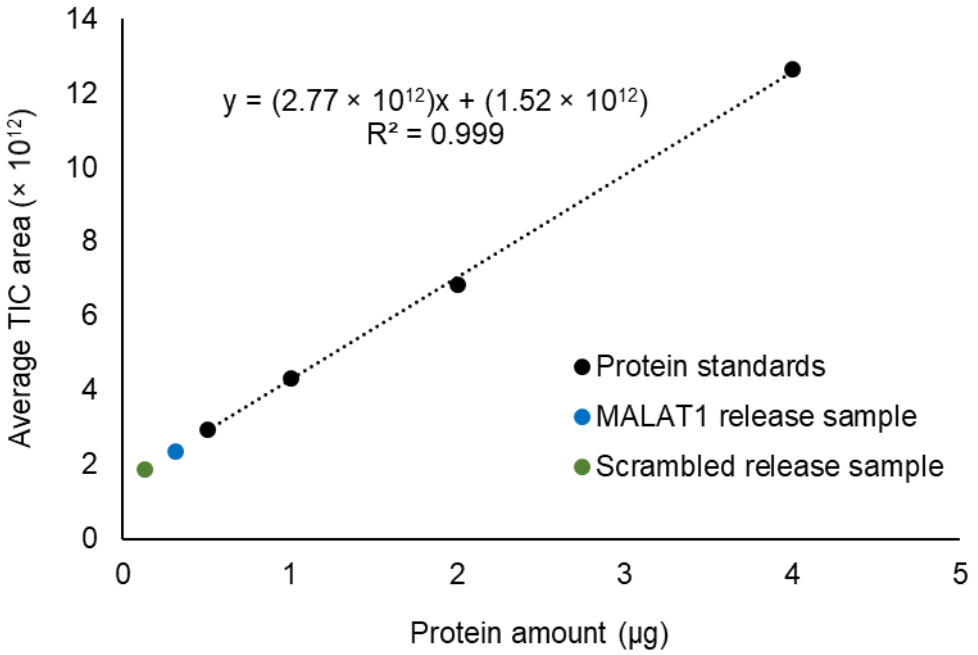


Supplementary Figure S-5.1 Jurkat cells cross-linked with 1% (w/v) formaldehyde were lysed and sonicated for varying amounts of time at intensity setting 2.5 using a Misonix Ultrasonic Processor XL 2015 equipped with a microtip. Sonication was performed in 4 s bursts with 4 s rest between each burst. After sonication, samples were centrifuged at 1,000 × *g* for 2 min at 4 °C to clear insoluble cellular debris, then treated with proteinase K and incubated at 37 °C overnight to digest proteins and reverse formaldehyde cross-links. The samples were then incubated for an additional 1 h at 65 °C and allowed to cool to room temperature. RNA was extracted and purified using TRI Reagent followed by ethanol precipitation. A) The concentration of RNA in each sample was measured via absorbance at 260 nm. Error bars represent ±1 standard deviation of duplicate measurements for each sample. B) The integrity of the RNA in each sample was assessed using an Agilent 2100 Bioanalyzer. A gel-like image of the electropherogram for each sample is depicted and the RNA integrity number (RIN) calculated for each sample is indicated. RIN values range from 1-10, with 1 indicating totally degraded RNA and 10 indicating completely intact RNA.[87]

Quantification of protein amount in HyPR-MS release samples

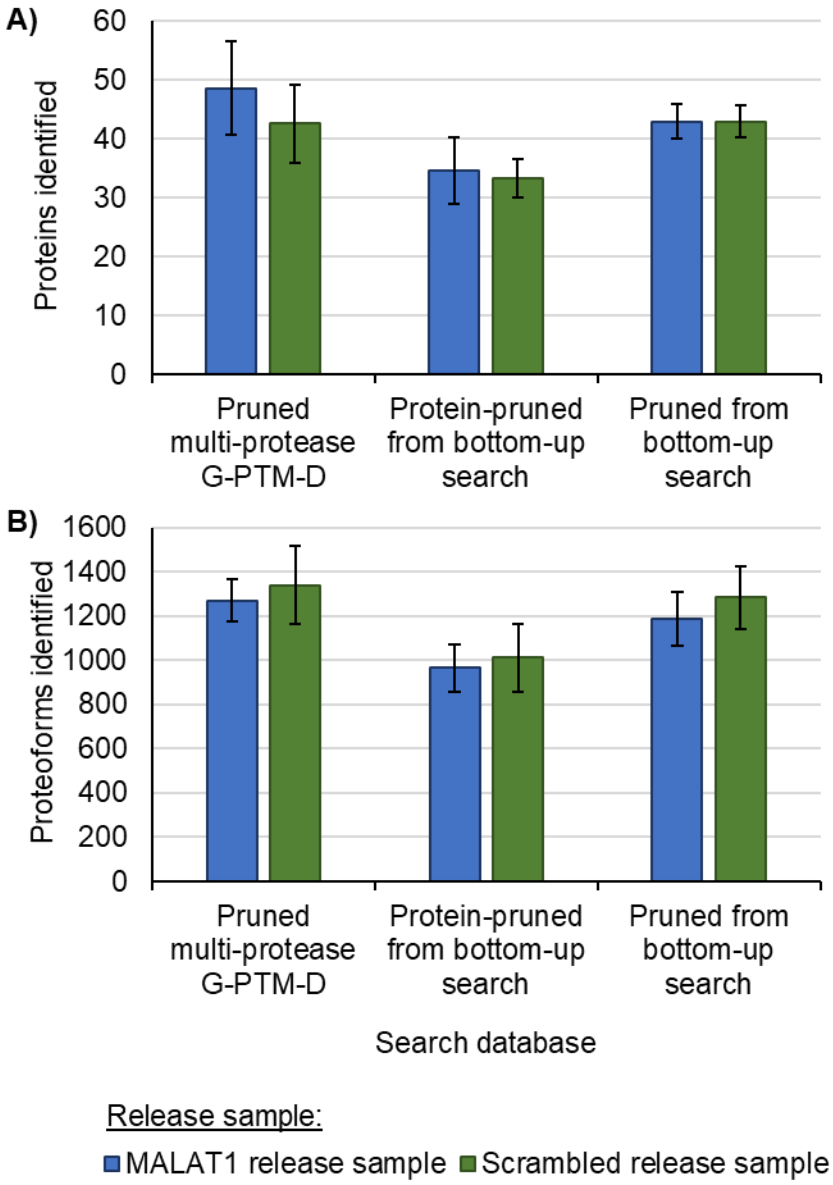
The amount of protein in a typical HyPR-MS release sample was estimated by comparing the integrated total ion chromatogram area ("TIC area") of the bottom-up HyPR-MS release samples described herein to the TIC areas of several bottom-up samples of known protein amount ("protein standards"). To prepare the protein standards, 1×10^7 1% formaldehyde cross-linked Jurkat cells were resuspended in 900 µL of buffer containing 375 mM LiCl, 50 mM Tris-HCl pH = 7.5, 1.25% (w/v) LiDS, and 1.25% (w/v) Triton X-100. The cells were lysed on ice for 10 min. The lysate was sonicated for 12 s with 4 s on/off intervals at setting 2 using a Misonix Ultrasonic Processor XL 2015 equipped with a microtip, then centrifuged at 16,000 × *g* for 6 min to clear insoluble cellular debris. The protein concentration in the supernatant was measured via the Pierce 660 nm assay (Thermo Scientific #22660) following the manufacturer's protocol. The ionic detergent compatibility reagent (Thermo Scientific #22663) was used due to the presence of LiDS in the lysis buffer. After measuring the protein concentration in the clarified cell lysate, samples containing 0.5, 1, 2, and 4 µg of protein were prepared for bottom-up mass spectrometry via eFASP and C18 solid-phase extraction, as described for bottom-up HyPR-MS. The samples were analyzed via HPLC-ESI-MS/MS in technical duplicate (half the volume of the sample per injection) using the same method described for bottom-up HyPR-MS. Qual Browser (within the Thermo Xcalibur software suite, version 3.1.66.10) was used to integrate the TIC for each .raw file, and the TIC areas for the technical duplicate injections of each protein standard were averaged and plotted to generate a standard curve (Supplementary Figure S-5.2). The average TIC areas

for the MALAT1 and scrambled bottom-up HyPR-MS samples were compared to the standard curve to estimate the amount of protein in each sample (Supplementary Figure S-5.2). As expected, the TIC area increased with increasing protein amount, and the standard curve was linear over the range assessed. From this analysis, we estimated that the MALAT1 release sample contained ~0.3 µg of protein, while the scrambled release sample contained ~0.1 µg of protein. Note that some extrapolation was required to quantify the amount of protein in the release samples as the TIC areas for those samples were slightly below that of the lowest standard.



Supplementary Figure S-5.2 Standard curve of average TIC area for various quantities of whole cell lysate protein analyzed via bottom-up mass spectrometry. The equation of the linear regression and R^2 value are shown. TIC area was measured by integrating the area under the TIC curve from 47-111 min, the time period over which peptides eluted. The average TIC area across two technical replicate injections of the bottom-up HyPR-MS MALAT1 and scrambled release samples was also calculated and the corresponding protein amount was estimated based on the standard curve.

Other supplementary figures and tables



Supplementary Figure S-5.3 Evaluation of databases informed by bottom-up HyPR-MS data for top-down HyPR-MS searches. Top-down data from three replicates of the MALAT1 and scrambled release samples were searched against the pruned multi-protease G-PTM-D database generated in Dai and Buxton et al.,[43] as well as pruned and protein-pruned versions of this database generated by performing G-PTM-D on and searching bottom-up MALAT1 and scrambled HyPR-MS data from Spiniello et al.[32]

(**Supplementary Figure S-5.3 continued**) and from this work. The average number of A) proteins and B) proteoforms identified in the various searches at 5% FDR is plotted. Error bars show ±1 standard deviation of the three replicate samples.

Supplementary Table S-5.1 Sequences of capture and release oligonucleotides used in HyPR-MS experiments. Red text denotes toehold sequence. All oligonucleotides were ordered from Integrated DNA Technologies.

Oligonucleotide	Sequence (5' – 3')	Modification	Nanomoles per top-down HyPR-MS experiment	Nanomoles per bottom-up HyPR-MS experiment
MALAT1 capture 1	TCG ATT AGA GGG GTT TTT GTT TTG CAG ATT CTG TGT TA	3'-Biotin-TEG	0.145	0.0485
MALAT1 capture 2	TCG ATC TTC AAT ATT TTC ATT TTC TAT CTT GTT TCT AT	3'-Biotin-TEG	0.145	0.0485
MALAT1 capture 3	TCG TAT CTT TGT TTT CTG TTA CAC CTT GAG TCA TTT GC	3'-Biotin-TEG	0.145	0.0485
MALAT1 capture 4	TCG TAT CTA ATT TGT CTT TCC TGC CTT AAA GTT ACA TT	3'-Biotin-TEG	0.145	0.0485
MALAT1 capture 5	TCG ATT GAT ACA TGT TCC CAC CCA GCA TTA CAG TTC TT	3'-Biotin-TEG	0.145	0.0485
MALAT1 release 1	TAA CAC AGA ATC TGC AAA ACA AAA ACC CCT CTA ATC GA	None	14.5	4.85

Oligonucleotide	Sequence (5' – 3')	Modification	Nanomoles per top-down HyPR-MS experiment	Nanomoles per bottom-up HyPR-MS experiment
MALAT1 release 2	ATA GAA ACA AGA TAG AAA ATG AAA ATA TTG AAG ATC GA	None	14.5	4.85
MALAT1 release 3	GCA AAT GAC TCA AGG TGT AAC AGA AAA CAA AGA TAC GA	None	14.5	4.85
MALAT1 release 4	AAT GTA ACT TTA AGG CAG GAA AGA CAA ATT AGA TAC GA	None	14.5	4.85
MALAT1 release 5	AAG AAC TGT AAT GCT GGG TGG GAA CAT GTA TCA ATC GA	None	14.5	4.85
Poly(A) capture	GCT TTA TGT TTT TTT TTT TTT TTT TTT TTT TTT TTT TT	3'-Biotin-TEG	0.725	0
Poly(A) release	AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA CAT AAA GC	None	72.5	0
Scrambled capture	GCT TTA TGT CTT AAG TGA TGA TAA CTG CTA GTC TGT AA	3'-Biotin-TEG	0.725	0.2425
Scrambled release	TTA CAG ACT AGC AGT TAT CAT CAC TTA AGA CAT AAA GC	None	72.5	24.25

Supplementary Table S-5.2 Primer and probe sequences for qPCR assays used in HyPR-MS experiments. All qPCR assays were ordered from Integrated DNA Technologies.

Assay	Primer/probe	Sequence (5' – 3')
5' MALAT1	Primer 1	AGA CCC AGA GCA GTG TAA A
	Primer 2	GTA GAC CAA CTA AGC GAA TGG
	Probe	/56-FAM/CTG CCC AAG/Zen/GTC TCT GTG TCT TCG/3IABkFQ/
5'-middle MALAT1	Primer 1	CAG GAT TCC AGG AAC CAG TG
	Primer 2	TTC CTA TCT TCA CCA CGA ACT G
	Probe	/56-FAM/CTA GGA CTG/Zen/AGG AGC AAG CGA GC/3IABkFQ/
3'-middle MALAT1	Primer 1	GAA CGA ATG TAA CTT TAA GGC AGG
	Primer 2	GAT CAT AAT CTC CCA CCT GTC TAA G
	Probe	/56-FAM/CCT CTA TTG/Zen/CCA TGT GCC TGG AA/3IABkFQ/
3' MALAT1	Primer 1	GGT GGG TTG AAC TAT GTT AGA AA
	Primer 2	CCA CTT ACT GGT TTA AGT TGG T
	Probe	/56-FAM/TGC CTG CAA/Zen/ATT GTT AAC AGA AGG GT/3IABkFQ/
GAPDH	Primer 1	TGC CAT CAA TGA CCC CTT C
	Primer 2	ATG ACA AGC TTC CCG TTC TC
	Probe	/56-FAM/TTG ACG GTG/Zen/CCA TGG AAT TTG CC/3IABkFQ/

Supplementary Table S-5.3 Proteins identified at 5% FDR in three biological replicates of top-down HyPR-MS MALAT1 and scrambled capture samples.

Sample	Gene	Protein name
MALAT1 replicate 1	RPS15A	40S ribosomal protein S15a
	AP3S2	AP-3 complex subunit sigma-2
	ARTN	Artemin
	CTNNA2	Catenin alpha-2
	CKLF-CMTM1	CKLF-CMTM1 readthrough
	TM7SF2	Delta(14)-sterol reductase
	DYRK4	Dual-specificity tyrosine-phosphorylation-regulated kinase 4
	EIF2D	Eukaryotic translation initiation factor 2D
	GNAO1	Guanine nucleotide-binding protein G(o) subunit alpha

Sample	Gene	Protein name
MALAT1 replicate 1 (cont.)	HIST1H2AG	Histone H2A type 1
	HIST1H2AB	Histone H2A type 1-B/E
	HIST1H2AC	Histone H2A type 1-C
	HIST1H2AD	Histone H2A type 1-D
	HIST1H2AH	Histone H2A type 1-H
	HIST1H2AJ	Histone H2A type 1-J
	HIST2H2AA3	Histone H2A type 2-A
	HIST2H2AB	Histone H2A type 2-B
	HIST2H2AC	Histone H2A type 2-C
	HIST3H2A	Histone H2A type 3
	H2AFJ	Histone H2A.J
	H2AFV	Histone H2A.V
	H2AFZ	Histone H2A.Z
	H2AFX	Histone H2AX
	HIST1H2BA	Histone H2B type 1-A
	HIST1H2BB	Histone H2B type 1-B
	HIST1H2BC	Histone H2B type 1-C/E/F/G/I
	HIST1H2BD	Histone H2B type 1-D
	HIST1H2BH	Histone H2B type 1-H
	HIST1H2BJ	Histone H2B type 1-J
	HIST1H2BK	Histone H2B type 1-K
	HIST1H2BL	Histone H2B type 1-L
	HIST1H2BM	Histone H2B type 1-M
	HIST1H2BN	Histone H2B type 1-N
	HIST1H2BO	Histone H2B type 1-O
	HIST2H2BE	Histone H2B type 2-E
	HIST2H2BF	Histone H2B type 2-F
	HIST3H2BB	Histone H2B type 3-B
	H2BFS	Histone H2B type F-S
	HIST2H3PS2	Histone H3
	HIST1H3A	Histone H3.1
	HIST3H3	Histone H3.1t
	HIST2H3A	Histone H3.2
	H3F3A	Histone H3.3
	HIST1H4A	Histone H4
	ITPKC	Kinase
	LDHA	L-lactate dehydrogenase A chain
	NEDD8-MDP1	NEDD8-MDP1 readthrough
	HMGN1	Non-histone chromosomal protein HMG-14

Sample	Gene	Protein name
MALAT1 replicate 1 (cont.)	HMGN2	Non-histone chromosomal protein HMG-17
	GALNTL6	Polypeptide N-acetylgalactosaminyltransferase-like 6
	PARP3	Protein mono-ADP-ribosyltransferase PARP3
	RHBDD1	Rhomboid-related protein 4
	RBIS	Ribosome biogenesis factor
	SREBF1	Sterol regulatory element-binding protein 1
	TMSB4X	Thymosin beta-4
	DIO1	Type I iodothyronine deiodinase
MALAT1 replicate 2	HSPE1	10 kDa heat shock protein, mitochondrial
	RPS15A	40S ribosomal protein S15a
	GGA1	ADP-ribosylation factor-binding protein GGA1
	HIST1H2AG	Histone H2A type 1
	HIST1H2AB	Histone H2A type 1-B/E
	HIST1H2AC	Histone H2A type 1-C
	HIST1H2AD	Histone H2A type 1-D
	HIST1H2AH	Histone H2A type 1-H
	HIST1H2AJ	Histone H2A type 1-J
	HIST2H2AA3	Histone H2A type 2-A
	HIST2H2AB	Histone H2A type 2-B
	HIST2H2AC	Histone H2A type 2-C
	HIST3H2A	Histone H2A type 3
	H2AFJ	Histone H2A.J
	H2AFV	Histone H2A.V
	H2AFZ	Histone H2A.Z
	H2AFX	Histone H2AX
	HIST1H2BB	Histone H2B type 1-B
	HIST1H2BC	Histone H2B type 1-C/E/F/G/I
	HIST1H2BD	Histone H2B type 1-D
	HIST1H2BH	Histone H2B type 1-H
	HIST1H2BJ	Histone H2B type 1-J
	HIST1H2BK	Histone H2B type 1-K
	HIST1H2BL	Histone H2B type 1-L
	HIST1H2BM	Histone H2B type 1-M
	HIST1H2BN	Histone H2B type 1-N
	HIST1H2BO	Histone H2B type 1-O
	HIST2H2BE	Histone H2B type 2-E
	HIST2H2BF	Histone H2B type 2-F
	HIST3H2BB	Histone H2B type 3-B
	H2BFS	Histone H2B type F-S

Sample	Gene	Protein name
MALAT1 replicate 2 (cont.)	HIST2H3PS2	Histone H3
	HIST1H3A	Histone H3.1
	HIST3H3	Histone H3.1t
	HIST2H3A	Histone H3.2
	H3F3A	Histone H3.3
	HIST1H4A	Histone H4
	HMGN1	Non-histone chromosomal protein HMG-14
	HMGN2	Non-histone chromosomal protein HMG-17
	GALNTL6	Polypeptide N-acetylgalactosaminyltransferase-like 6
	TMA7	Translation machinery-associated protein 7
MALAT1 replicate 3	NDUFAB1	Acyl carrier protein, mitochondrial
	ARTN	Artemin
	KCTD9	BTB/POZ domain-containing protein KCTD9
	CTNNA2	Catenin alpha-2
	EDF1	Endothelial differentiation-related factor 1
	EIF2D	Eukaryotic translation initiation factor 2D
	HIST1H2AG	Histone H2A type 1
	HIST1H2AB	Histone H2A type 1-B/E
	HIST1H2AC	Histone H2A type 1-C
	HIST1H2AD	Histone H2A type 1-D
	HIST1H2AH	Histone H2A type 1-H
	HIST1H2AJ	Histone H2A type 1-J
	HIST2H2AA3	Histone H2A type 2-A
	HIST2H2AB	Histone H2A type 2-B
	HIST2H2AC	Histone H2A type 2-C
	HIST3H2A	Histone H2A type 3
	H2AFJ	Histone H2A.J
	H2AFV	Histone H2A.V
	H2AFZ	Histone H2A.Z
	H2AFX	Histone H2AX
	HIST1H2BA	Histone H2B type 1-A
	HIST1H2BB	Histone H2B type 1-B
	HIST1H2BC	Histone H2B type 1-C/E/F/G/I
	HIST1H2BD	Histone H2B type 1-D
	HIST1H2BH	Histone H2B type 1-H
	HIST1H2BJ	Histone H2B type 1-J
	HIST1H2BK	Histone H2B type 1-K
	HIST1H2BL	Histone H2B type 1-L
	HIST1H2BM	Histone H2B type 1-M

Sample	Gene	Protein name
MALAT1 replicate 3 (cont.)	HIST1H2BN	Histone H2B type 1-N
	HIST1H2BO	Histone H2B type 1-O
	HIST2H2BE	Histone H2B type 2-E
	HIST2H2BF	Histone H2B type 2-F
	HIST3H2BB	Histone H2B type 3-B
	H2BFS	Histone H2B type F-S
	HIST1H3A	Histone H3.1
	HIST3H3	Histone H3.1t
	HIST2H3A	Histone H3.2
	H3F3A	Histone H3.3
	HIST1H4A	Histone H4
	NEDD8-MDP1	NEDD8-MDP1 readthrough
	HMGN1	Non-histone chromosomal protein HMG-14
	HMGN2	Non-histone chromosomal protein HMG-17
	NPC1	NPC intracellular cholesterol transporter 1
	SREBF1	Sterol regulatory element-binding protein 1
	TMA7	Translation machinery-associated protein 7
	SNRPA1	U2 small nuclear ribonucleoprotein A'
Scrambled replicate 1	RPS15A	40S ribosomal protein S15a
	CCL16	C-C motif chemokine 16
	CKLF-CMTM1	CKLF-CMTM1 readthrough
	EDF1	Endothelial differentiation-related factor 1
	EIF2D	Eukaryotic translation initiation factor 2D
	HIST1H2AG	Histone H2A type 1
	HIST1H2AB	Histone H2A type 1-B/E
	HIST1H2AC	Histone H2A type 1-C
	HIST1H2AD	Histone H2A type 1-D
	HIST1H2AH	Histone H2A type 1-H
	HIST1H2AJ	Histone H2A type 1-J
	HIST2H2AA3	Histone H2A type 2-A
	HIST2H2AB	Histone H2A type 2-B
	HIST2H2AC	Histone H2A type 2-C
	HIST3H2A	Histone H2A type 3
	H2AFJ	Histone H2A.J
	H2AFV	Histone H2A.V
	H2AFZ	Histone H2A.Z
	H2AFX	Histone H2AX
	HIST1H2BB	Histone H2B type 1-B
	HIST1H2BC	Histone H2B type 1-C/E/F/G/I

Sample	Gene	Protein name
Scrambled replicate 1 (cont.)	HIST1H2BD	Histone H2B type 1-D
	HIST1H2BH	Histone H2B type 1-H
	HIST1H2BJ	Histone H2B type 1-J
	HIST1H2BK	Histone H2B type 1-K
	HIST1H2BL	Histone H2B type 1-L
	HIST1H2BM	Histone H2B type 1-M
	HIST1H2BN	Histone H2B type 1-N
	HIST1H2BO	Histone H2B type 1-O
	HIST2H2BE	Histone H2B type 2-E
	HIST2H2BF	Histone H2B type 2-F
	HIST3H2BB	Histone H2B type 3-B
	H2BFS	Histone H2B type F-S
	HIST1H3A	Histone H3.1
	HIST3H3	Histone H3.1t
	HIST2H3A	Histone H3.2
	H3F3A	Histone H3.3
	H3F3C	Histone H3.3C
	HIST1H4A	Histone H4
	ITGB1	Integrin beta-1
	NEDD8-MDP1	NEDD8-MDP1 readthrough
	HMGN1	Non-histone chromosomal protein HMG-14
	HMGN2	Non-histone chromosomal protein HMG-17
	GALNTL6	Polypeptide N-acetylgalactosaminyltransferase-like 6
	ZNF593OS	Putative transmembrane protein ZNF593OS
	RBIS	Ribosome biogenesis factor
Scrambled replicate 2	DYRK4	Dual-specificity tyrosine-phosphorylation-regulated kinase 4
	EDF1	Endothelial differentiation-related factor 1
	EIF2D	Eukaryotic translation initiation factor 2D
	GGT5	Glutathione hydrolase 5 proenzyme
	HIST1H2AG	Histone H2A type 1
	HIST1H2AB	Histone H2A type 1-B/E
	HIST1H2AC	Histone H2A type 1-C
	HIST1H2AD	Histone H2A type 1-D
	HIST1H2AH	Histone H2A type 1-H
	HIST1H2AJ	Histone H2A type 1-J
	HIST2H2AA3	Histone H2A type 2-A
	HIST2H2AB	Histone H2A type 2-B
	HIST2H2AC	Histone H2A type 2-C
	HIST3H2A	Histone H2A type 3

Sample	Gene	Protein name
Scrambled replicate 2 (cont.)	H2AFJ	Histone H2A.J
	H2AFV	Histone H2A.V
	H2AFZ	Histone H2A.Z
	H2AFX	Histone H2AX
	HIST1H2BA	Histone H2B type 1-A
	HIST1H2BB	Histone H2B type 1-B
	HIST1H2BC	Histone H2B type 1-C/E/F/G/I
	HIST1H2BD	Histone H2B type 1-D
	HIST1H2BH	Histone H2B type 1-H
	HIST1H2BJ	Histone H2B type 1-J
	HIST1H2BK	Histone H2B type 1-K
	HIST1H2BL	Histone H2B type 1-L
	HIST1H2BM	Histone H2B type 1-M
	HIST1H2BN	Histone H2B type 1-N
	HIST1H2BO	Histone H2B type 1-O
	HIST2H2BE	Histone H2B type 2-E
	HIST2H2BF	Histone H2B type 2-F
	HIST3H2BB	Histone H2B type 3-B
	H2BFS	Histone H2B type F-S
	HIST2H3PS2	Histone H3
	HIST1H3A	Histone H3.1
	HIST3H3	Histone H3.1t
	HIST2H3A	Histone H3.2
	H3F3A	Histone H3.3
	H3F3C	Histone H3.3C
	HIST1H4A	Histone H4
	HMGN1	Non-histone chromosomal protein HMG-14
	HMGN2	Non-histone chromosomal protein HMG-17
	GALNTL6	Polypeptide N-acetylgalactosaminyltransferase-like 6
	UBB|UBC	Polyubiquitin-B|Polyubiquitin-C
	RAB17	Ras-related protein Rab-17
	TMA7	Translation machinery-associated protein 7
	RPS27A|UBA52	Ubiquitin-40S ribosomal protein S27a|Ubiquitin-60S ribosomal protein L40
Scrambled replicate 3	HIST1H2AG	Histone H2A type 1
	HIST1H2AB	Histone H2A type 1-B/E
	HIST1H2AC	Histone H2A type 1-C
	HIST1H2AD	Histone H2A type 1-D
	HIST1H2AH	Histone H2A type 1-H

Sample	Gene	Protein name
Scrambled replicate 3 (cont.)	HIST1H2AJ	Histone H2A type 1-J
	HIST2H2AA3	Histone H2A type 2-A
	HIST2H2AB	Histone H2A type 2-B
	HIST2H2AC	Histone H2A type 2-C
	HIST3H2A	Histone H2A type 3
	H2AFJ	Histone H2A.J
	H2AFV	Histone H2A.V
	H2AFZ	Histone H2A.Z
	HIST1H2BA	Histone H2B type 1-A
	HIST1H2BB	Histone H2B type 1-B
	HIST1H2BC	Histone H2B type 1-C/E/F/G/I
	HIST1H2BD	Histone H2B type 1-D
	HIST1H2BH	Histone H2B type 1-H
	HIST1H2BJ	Histone H2B type 1-J
	HIST1H2BK	Histone H2B type 1-K
	HIST1H2BL	Histone H2B type 1-L
	HIST1H2BM	Histone H2B type 1-M
	HIST1H2BN	Histone H2B type 1-N
	HIST1H2BO	Histone H2B type 1-O
	HIST2H2BE	Histone H2B type 2-E
	HIST2H2BF	Histone H2B type 2-F
	HIST3H2BB	Histone H2B type 3-B
	H2BFS	Histone H2B type F-S
	HIST1H3A	Histone H3.1
	HIST3H3	Histone H3.1t
	HIST2H3A	Histone H3.2
	H3F3A	Histone H3.3
	HIST1H4A	Histone H4
	HMGN1	Non-histone chromosomal protein HMG-14
	HMGN2	Non-histone chromosomal protein HMG-17

5.7 ACKNOWLEDGEMENTS

This work was supported by NIH-NIGMS grant #R35GM126914, NIH-NCI grant #R01CA193481, and NIH-NHLBI grant #R01HL149966. K.B.H. was supported in part by the NIH-

NHGRI grant to the Genomic Science Training Program, #5T32HG002760. R.M.M. was supported in part by the NIH Chemistry-Biology Interface Training Grant, #T32GM008505.

5.8 REFERENCES

(1) Moore, M. J. From birth to death: the complex lives of eukaryotic mRNAs. *Science* **2005**, *309* (5740), 1514–1518.

(2) Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **2008**, *582* (14), 1977–1986.

(3) Mitchell, S. F.; Parker, R. Principles and properties of eukaryotic mRNPs. *Mol. Cell* **2014**, *54* (4), 547–558.

(4) Re, A.; Joshi, T.; Kulberkyte, E.; Morris, Q.; Workman, C. T. RNA-protein interactions: an overview. In *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, Methods in Molecular Biology, vol. 1097; Gorodkin, J., Ruzzo, W. L., Eds.; Humana Press: Totowa, NJ, 2014; pp 491–521.

(5) Matera, A. G.; Terns, R. M.; Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (3), 209–220.

(6) Mayr, C. Regulation by 3'-untranslated regions. *Annu. Rev. Genet.* **2017**, *51*, 171–194.

(7) Marchese, F. P.; Raimondi, I.; Huarte, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* **2017**, *18* (1), 206.

(8) Allerson, C. R.; Cazzola, M.; Rouault, T. A. Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. *J. Biol. Chem.* **1999**, *274* (37), 26439–26447.

(9) Lukong, K. E.; Chang, K.; Khandjian, E. W.; Richard, S. RNA-binding proteins in human genetic disease. *Trends Genet.* **2008**, *24* (8), 416–425.

(10) Corbett, A. H. Post-transcriptional regulation of gene expression and human disease. *Curr. Opin. Cell Biol.* **2018**, *52*, 96–104.

(11) Baltz, A. G.; Munschauer, M.; Schwanhäusser, B.; Vasile, A.; Murakawa, Y.; Schueler, M.; Youngs, N.; Penfold-Brown, D.; Drew, K.; Milek, M.; et al. The mRNA-bound proteome and its

global occupancy profile on protein-coding transcripts. *Mol. Cell* **2012**, *46* (5), 674–690.

(12) Castello, A.; Fischer, B.; Eichelbaum, K.; Horos, R.; Beckmann, B. M.; Strein, C.; Davey, N. E.; Humphreys, D. T.; Preiss, T.; Steinmetz, L. M.; et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **2012**, *149* (6), 1393–1406.

(13) Queiroz, R. M. L.; Smith, T.; Villanueva, E.; Marti-Solano, M.; Monti, M.; Pizzinga, M.; Mirea, D.-M.; Ramakrishna, M.; Harvey, R. F.; Dezi, V.; et al. Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **2019**, *37* (2), 169–178.

(14) Urdaneta, E. C.; Vieira-Vieira, C. H.; Hick, T.; Wessels, H.-H.; Figini, D.; Moschall, R.; Medenbach, J.; Ohler, U.; Granneman, S.; Selbach, M.; et al. Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nat. Commun.* **2019**, *10* (1), 990.

(15) West, J. A.; Davis, C. P.; Sunwoo, H.; Simon, M. D.; Sadreyev, R. I.; Wang, P. I.; Tolstorukov, M. Y.; Kingston, R. E. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* **2014**, *55* (5), 791–802.

(16) Chu, C.; Zhang, Q. C.; da Rocha, S. T.; Flynn, R. A.; Bharadwaj, M.; Calabrese, J. M.; Magnuson, T.; Heard, E.; Chang, H. Y. Systematic discovery of Xist RNA binding proteins. *Cell* **2015**, *161* (2), 404–416.

(17) McHugh, C. A.; Chen, C.-K.; Chow, A.; Surka, C. F.; Tran, C.; McDonel, P.; Pandya-Jones, A.; Blanco, M.; Burghard, C.; Moradian, A.; et al. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **2015**, *521* (7551), 232–236.

(18) Knoener, R. A.; Becker, J. T.; Scalf, M.; Sherer, N. M.; Smith, L. M. Elucidating the in vivo interactome of HIV-1 RNA by hybridization capture and mass spectrometry. *Sci. Rep.* **2017**, *7* (1), 16965.

(19) Henke, K. B.; Miller, R. M.; Knoener, R. A.; Scalf, M.; Spiniello, M.; Smith, L. M. Identifying protein interactomes of target RNAs using HyPR-MS. In *Post-Transcriptional Gene Regulation*, 3rd ed., Methods in Molecular Biology, vol. 2404; Dassi, E., Ed.; Humana Press: New York, NY, 2022; pp 219–244.

(20) Smith, L. M.; Kelleher, N. L.; The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10* (3), 186–187.

(21) Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Paša-Tolić, L.; Tsybin, Y. O.; Kelleher, N. L.; The Consortium for Top-Down Proteomics. The human proteoform project: defining the human proteome. *Sci. Adv.* **2021**, *7* (46), eabk0734.

(22) Schaffer, L. V; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N.

L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; et al. Identification and quantification of proteoforms by mass spectrometry. *Proteomics* **2019**, *19* (10), 1800361.

(23) Jenuwein, T.; Allis, C. D. Translating the histone code. *Science* **2001**, *293* (5532), 1074–1080.

(24) Wickenhagen, A.; Sugrue, E.; Lytras, S.; Kuchi, S.; Noerenberg, M.; Turnbull, M. L.; Loney, C.; Herder, V.; Allan, J.; Jarmson, I.; et al. A prenylated dsRNA sensor protects against severe COVID-19. *Science* **2021**, *374* (6567), eabj3624.

(25) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates III, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **2013**, *113* (4), 2343–2394.

(26) Gillet, L. C.; Leitner, A.; Aebersold, R. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annu. Rev. Anal. Chem.* **2016**, *9* (1), 449–472.

(27) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* **2014**, *445* (4), 683–693.

(28) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem.* **2016**, *9* (1), 499–519.

(29) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-down proteomics: ready for prime time? *Anal. Chem.* **2018**, *90* (1), 110–127.

(30) Donnelly, D. P.; Rawlins, C. M.; DeHart, C. J.; Fornelli, L.; Schachner, L. F.; Lin, Z.; Lippens, J. L.; Aluri, K. C.; Sarin, R.; Chen, B.; et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* **2019**, *16* (7), 587–594.

(31) Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L. On the scalability and requirements of whole protein mass spectrometry. *Anal. Chem.* **2011**, *83* (17), 6868–6874.

(32) Spiniello, M.; Knoener, R. A.; Steinbrink, M. I.; Yang, B.; Cesnik, A. J.; Buxton, K. E.; Scalf, M.; Jarrard, D. F.; Smith, L. M. HyPR-MS for multiplexed discovery of MALAT1, NEAT1, and NORAD lncRNA protein interactomes. *J. Proteome Res.* **2018**, *17* (9), 3022–3038.

(33) Spiniello, M.; Steinbrink, M. I.; Cesnik, A. J.; Miller, R. M.; Scalf, M.; Shortreed, M. R.; Smith, L. M. Comprehensive in vivo identification of the c-Myc mRNA interactome using HyPR-MS. *RNA* **2019**, *25* (10), 1337–1352.

(34) Knoener, R.; Evans III, E.; Becker, J. T.; Scalf, M.; Benner, B.; Sherer, N. M.; Smith, L. M. Identification of host proteins differentially associated with HIV-1 RNA splice variants. *eLife* **2021**, *10*, e62470.

(35) Schneider, U.; Schwenk, H.-U.; Bornkamm, G. Characterization of EBV-genome negative

"null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int. J. Cancer* **1977**, *19* (5), 621–626.

(36) Amodio, N.; Raimondi, L.; Juli, G.; Stamato, M. A.; Caracciolo, D.; Tagliaferri, P.; Tassone, P. MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. *J. Hematol. Oncol.* **2018**, *11* (1), 63.

(37) Gutschner, T.; Hämmerle, M.; Diederichs, S. MALAT1 – a paradigm for long noncoding RNA function in cancer. *J. Mol. Med.* **2013**, *91* (7), 791–801.

(38) Wilusz, J. E. Long noncoding RNAs: re-writing dogmas of RNA processing and stability. *Biochim. Biophys. Acta - Gene Regul. Mech.* **2016**, *1859* (1), 128–138.

(39) Tripathi, V.; Ellis, J. D.; Shen, Z.; Song, D. Y.; Pan, Q.; Watt, A. T.; Freier, S. M.; Bennett, C. F.; Sharma, A.; Bubulya, P. A.; et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **2010**, *39* (6), 925–938.

(40) Erde, J.; Loo, R. R. O.; Loo, J. A. Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. *J. Proteome Res.* **2014**, *13* (4), 1885–1895.

(41) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced global post-translational modification discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17* (5), 1844–1851.

(42) Li, Q.; Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Schaffer, L. V; Scalf, M.; Smith, L. M. Global post-translational modification discovery. *J. Proteome Res.* **2017**, *16* (4), 1383–1390.

(43) Dai, Y.; Buxton, K. E.; Schaffer, L. V; Miller, R. M.; Millikin, R. J.; Scalf, M.; Frey, B. L.; Shortreed, M. R.; Smith, L. M. Constructing human proteoform families using intact-mass and top-down proteomics with a multi-protease global post-translational modification discovery database. *J. Proteome Res.* **2019**, *18* (10), 3671–3680.

(44) Millikin, R. J.; Solntsev, S. K.; Shortreed, M. R.; Smith, L. M. Ultrafast peptide label-free quantification with FlashLFQ. *J. Proteome Res.* **2018**, *17* (1), 386–391.

(45) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **2014**, *13* (1), 228–240.

(46) Miller, R. M.; Millikin, R. J.; Hoffmann, C. V; Solntsev, S. K.; Sheynkman, G. M.; Shortreed, M. R.; Smith, L. M. Improved protein inference from multiple protease bottom-up mass spectrometry data. *J. Proteome Res.* **2019**, *18* (9), 3429–3438.

(47) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **2016**, *13* (9), 731–740.

(48) Wessel, D.; Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **1984**, *138* (1), 141–143.

(49) Tran, J. C.; Doucette, A. A. Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. *Anal. Chem.* **2008**, *80* (5), 1568–1573.

(50) Huang, J.-L.; Liu, W.; Tian, L.-H.; Chai, T.-T.; Liu, Y.; Zhang, F.; Fu, H.-Y.; Zhou, H.-R.; Shen, J.-Z. Upregulation of long non-coding RNA MALAT-1 confers poor prognosis and influences cell proliferation and apoptosis in acute monocytic leukemia. *Oncol. Rep.* **2017**, *38* (3), 1353–1362.

(51) Hu, N.; Chen, L.; Wang, C.; Zhao, H. MALAT1 knockdown inhibits proliferation and enhances cytarabine chemosensitivity by upregulating miR-96 in acute myeloid leukemia cells. *Biomed. Pharmacother.* **2019**, *112*, 108720.

(52) Ahmadi, A.; Kaviani, S.; Yaghmaie, M.; Pashaiefar, H.; Ahmadvand, M.; Jalili, M.; Alimoghaddam, K.; Eslamijouybari, M.; Ghavamzadeh, A. Altered expression of MALAT1 lncRNA in chronic lymphocytic leukemia patients, correlation with cytogenetic findings. *Blood Res.* **2018**, *53* (4), 320–324.

(53) Pouyanrad, S.; Rahgozar, S.; Ghodousi, E. S. Dysregulation of miR-335-3p, targeted by NEAT1 and MALAT1 long non-coding RNAs, is associated with poor prognosis in childhood acute lymphoblastic leukemia. *Gene* **2019**, *692*, 35–43.

(54) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.

(55) Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **2003**, *31* (13), 3406–3415.

(56) Kennedy-Darling, J.; Holden, M. T.; Shortreed, M. R.; Smith, L. M. Multiplexed programmable release of captured DNA. *ChemBioChem* **2014**, *15* (16), 2353–2356.

(57) Mi, H.; Muruganujan, A.; Huang, X.; Ebert, D.; Mills, C.; Guo, X.; Thomas, P. D. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* **2019**, *14* (3), 703–721.

(58) Mi, H.; Ebert, D.; Muruganujan, A.; Mills, C.; Albou, L.-P.; Mushayamaha, T.; Thomas, P. D. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer

regions and extensive API. *Nucleic Acids Res.* **2021**, *49* (D1), D394–D403.

(59) Gezer, U.; Özgür, E.; Cetinkaya, M.; Isin, M.; Dalay, N. Long non-coding RNAs with low expression levels in cells are enriched in secreted exosomes. *Cell Biol. Int.* **2014**, *38* (9), 1076–1079.

(60) Jackson, V. Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell* **1978**, *15* (3), 945–954.

(61) Kennedy-Darling, J.; Smith, L. M. Measuring the formaldehyde protein-DNA cross-link reversal rate. *Anal. Chem.* **2014**, *86* (12), 5678–5681.

(62) Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A. Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation. *J. Proteome Res.* **2010**, *9* (6), 2863–2870.

(63) Crowell, A. M. J.; Wall, M. J.; Doucette, A. A. Maximizing recovery of water-soluble proteins through acetone precipitation. *Anal. Chim. Acta* **2013**, *796*, 48–54.

(64) Koontz, L. TCA precipitation. *Methods Enzymol.* **2014**, *541*, 3–10.

(65) Yang, Z.; Shen, X.; Chen, D.; Sun, L. Toward a universal sample preparation method for denaturing top-down proteomics of complex proteomes. *J. Proteome Res.* **2020**, *19* (8), 3315–3325.

(66) Wiśniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6* (5), 359–362.

(67) Zougman, A.; Selby, P. J.; Banks, R. E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **2014**, *14* (9), 1006–1010.

(68) HaileMariam, M.; Eguez, R. V.; Singh, H.; Bekele, S.; Ameni, G.; Pieper, R.; Yu, Y. S-Trap, an ultrafast sample-preparation approach for shotgun proteomics. *J. Proteome Res.* **2018**, *17* (9), 2917–2924.

(69) Ludwig, K. R.; Schroll, M. M.; Hummon, A. B. Comparison of in-solution, FASP, and S-Trap based digestion methods for bottom-up proteomic studies. *J. Proteome Res.* **2018**, *17* (7), 2480–2490.

(70) *S-Trap*. https://protifi.com/pages/s-trap (accessed 2021-11-10).

(71) Doucette, A. A.; Vieira, D. B.; Orton, D. J.; Wall, M. J. Resolubilization of precipitated intact membrane proteins with cold formic acid for analysis by mass spectrometry. *J. Proteome Res.* **2014**, *13* (12), 6001–6012.

(72) Zheng, S.; Doucette, A. A. Preventing N- and O-formylation of proteins when incubated in concentrated formic acid. *Proteomics* **2016**, *16* (7), 1059–1068.

(73) Wilusz, J. E.; Freier, S. M.; Spector, D. L. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **2008**, *135* (5), 919–932.

(74) Seelenfreund, E.; Robinson, W. A.; Amato, C. M.; Tan, A.-C.; Kim, J.; Robinson, S. E. Long term storage of dry versus frozen RNA for next generation molecular studies. *PLoS One* **2014**, *9* (11), e111827.

(75) Albihlal, W. S.; Gerber, A. P. Unconventional RNA-binding proteins: an uncharted zone in RNA biology. *FEBS Lett.* **2018**, *592* (17), 2917–2931.

(76) Soboleva, T. A.; Parker, B. J.; Nekrasov, M.; Hart-Smith, G.; Tay, Y. J.; Tng, W.-Q.; Wilkins, M.; Ryan, D.; Tremethick, D. J. A new link between transcriptional initiation and pre-mRNA splicing: the RNA binding histone variant H2A.B. *PLoS Genet.* **2017**, *13* (2), e1006633.

(77) Damavandi, F.; Wang, W.; Shen, W.-Z.; Cetinel, S.; Jordan, T.; Jovel, J.; Montemagno, C.; Wong, G. K.-S. Enrichment of low abundance DNA/RNA by oligonucleotide-clicked iron oxide nanoparticles. *Sci. Rep.* **2021**, *11* (1), 13053.

(78) Tayri-Wilk, T.; Slavin, M.; Zamel, J.; Blass, A.; Cohen, S.; Motzik, A.; Sun, X.; Shalev, D. E.; Ram, O.; Kalisman, N. Mass spectrometry reveals the chemistry of formaldehyde cross-linking in structured proteins. *Nat. Commun.* **2020**, *11* (1), 3128.

(79) Nickerson, J. L.; Baghalabadi, V.; Rajendran, S. R. C. K.; Jakubec, P. J.; Said, H.; McMillen, T. S.; Dang, Z.; Doucette, A. A. Recent advances in top-down proteome sample processing ahead of MS analysis. *Mass Spectrom. Rev.* **2021**, DOI: 10.1002/mas.21706.

(80) Wang, Z.; Ma, H.; Smith, K.; Wu, S. Two-dimensional separation using high-pH and low-pH reversed phase liquid chromatography for top-down proteomics. *Int. J. Mass Spectrom.* **2018**, *427*, 43–51.

(81) Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y. Top-down proteomics of large proteins up to 223 kDa enabled by serial size exclusion chromatography strategy. *Anal. Chem.* **2017**, *89* (10), 5467–5475.

(82) Tucholski, T.; Knott, S. J.; Chen, B.; Pistono, P.; Lin, Z.; Ge, Y. A top-down proteomics platform coupling serial size exclusion chromatography and Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **2019**, *91* (6), 3835–3844.

(83) Seckler, H. D. S.; Fornelli, L.; Mutharasan, R. K.; Thaxton, C. S.; Fellers, R.; Daviglus, M.; Sniderman, A.; Rader, D.; Kelleher, N. L.; Lloyd-Jones, D. M.; et al. A targeted, differential top-down proteomic methodology for comparison of ApoA-I proteoforms in individuals with high

and low HDL efflux capacity. *J. Proteome Res.* **2018**, *17* (6), 2156–2164.

(84) Hodge, K.; Have, S. Ten; Hutton, L.; Lamond, A. I. Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS. *J. Proteomics* **2013**, *88*, 92–103.

(85) van de Waterbeemd, M.; Fort, K. L.; Boll, D.; Reinhardt-Szyba, M.; Routh, A.; Makarov, A.; Heck, A. J. R. High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. *Nat. Methods* **2017**, *14* (3), 283–286.

(86) Schachner, L. F.; Ives, A. N.; McGee, J. P.; Melani, R. D.; Kafader, J. O.; Compton, P. D.; Patrie, S. M.; Kelleher, N. L. Standard proteoforms and their complexes for native mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (7), 1190–1198.

(87) Schroeder, A.; Mueller, O.; Stocker, S.; Salowsky, R.; Leiber, M.; Gassmann, M.; Lightfoot, S.; Menzel, W.; Granzow, M.; Ragg, T. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **2006**, *7*, 3.