

New Tools for Enhanced Proteome Characterization

by

Rachel M. Miller

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

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: 07/14/2022

The dissertation is approved by the following members of the Final Oral Committee:

Joshua J. Coon, Professor, Chemistry and Biomolecular Chemistry

Lingjun Li, Professor, Chemistry and Pharmacy

Douglas G. McNeel, Professor, Medicine

Lloyd M. Smith, Professor, Chemistry

© Copyright by Rachel M. Miller 2022

All Rights Reserved

To Matthew and My Family

ACKNOWLEDGMENTS

I have found working on my thesis to be an incredibly introspective experience. I find myself reflecting not only on my entire PhD experience, but everything in my life that has brought me to this point. I have been so incredibly lucky to have such an incredible support system who have supported me throughout my life and along this journey towards a PhD. I find myself struggling to find the right words to convey my gratitude and how much all those who have supported me mean to me. I will do my best in the upcoming paragraphs to try and capture my thoughts and feelings.

I want to start by thanking Dr. Lloyd Smith for being my PI and mentor during my tenure at UW-Madison. When I decided to go to graduate school, I was unsure exactly what area of research I would pursue. I knew that I wanted to be at the interface of chemistry and biology and that I was very intrigued about proteins, and how minor changes such as post-translational modifications could have such a profound impact. Beyond that I was just hoping to find a mentor whose research I felt a connection to. When I heard Lloyd speak about his research and proteoforms I was immediately sold. I knew at that moment that proteomics was where I belonged. Shortly after that I was able to meet with Lloyd, and I knew that he was who I wanted as my PI and mentor. His passion for science was contagious, and I loved his philosophy on fostering the growth of independent scientists and idea generators. Lloyd was kind enough to offer me the ability to begin working in his lab early over the summer, and I knew within my first week that I did not want to leave the Smith group, and that it was where I belonged. From day one Lloyd has been nothing but encouraging, pushing me to strive for more than I ever thought possible as an undergraduate. I knew I was capable of following directions, and executing experiments, but the thought of coming up with projects of my own was an intimidating feat early on in my tenure

in the Smith lab. I communicated this to Lloyd, and he helped encourage and coach me towards following my scientific intuition and not fearing failure. Lloyd helped me gain a confidence in myself as a scientist that I will be forever grateful for. He has supported me as I followed my scientific passion through many different diverse projects in the lab and encouraged me to do what makes me happy. It is difficult to put into words how much I have learned from Lloyd through our conversations and meetings, and how much they have meant to me. I can say without a doubt I would not be the scientist I am today if it were not for Lloyd's mentorship.

I am also very grateful for Dr. Michael Shortreed as he has been a constant and stable force for me throughout my graduate school experience. He was the first person in the Smith group I met after Lloyd. He instantly made me feel at home within the group, and that has never changed during my 5 years. I always know I can count on Shortreed when I need to talk either about science or life. He always seems to know exactly what to say and provides valuable perspective and insight. His kindness and compassion seem to know no bounds, and it is clear how much he cares about each one of us as students. Shortreed has always been there, either with the turn of a chair or just a slack video call away. I will be forever thankful for all of his support, and the time we shared as officemates.

I want to thank Dr. Brian Frey for being another incredible mentor within the Smith group. I will always appreciate his willingness to make time to talk science, whether it be helping me troubleshoot an experiment, or discussing an idea I have for a future project. Brian's insight was always invaluable. When I first started working in the lab, we shared an office, and Brian always made me feel so incredibly welcomed, and helped me navigate the first months of graduate school that are always so overwhelming. I will always be grateful for Brian's kindness, support and mentorship.

Thank you to Dr. Mark Scalf for all your support over the past 5 years. I always felt so lucky to have a bench space in your lab, because you were always just around the corner willing to help. I will always be grateful for your patience, and willingness to teach me about the mass spectrometers. You made instrumentation, something I always felt was so daunting, very approachable and fun. Your positive and relaxed energy in the lab was infectious. I feel very lucky to say that I have learned from and worked with you over the years.

I want to thank my undergraduate mentors Dr. Douglas Stack and Dr. John Conrad. To Dr. Stack, thank you for letting an overly eager undergraduate into your research lab before she had finished all of the necessary prerequisites. Thank you for supporting me as I learned how to do research and developed fundamental skill sets. Thank you for always being there and setting an incredibly high bar for what I would look for in a research mentor in graduate school. Dr. Conrad, thank you for being the professor that showed me that my passion lie at the interface of chemistry and biology. Thank you for your open-door policy to talk about science, life and cooking. I will always be incredibly grateful for your encouragement and guidance when it came to looking at graduate schools and programs. He encouraged me to apply for UW-Madison when I felt it was out of my reach. Without both of your encouragement and support, I would not have ever made it to UW-Madison to begin this crazy journey that is now coming to an end.

I now want to begin thanking my family. Family means the world to me, and without their support I would not be who I am today. This journey has not always been easy, but it is has been made easier knowing you all are behind me, supporting me. I want to begin by thanking my husband's family, my family-in-love, the Millers (Mitchells and Bishops). I am so incredibly lucky to be able to call you my family. Ever since Matthew and I began dating, you all have been incredibly loving and

supportive, welcoming me into your family with open arms. To Katie and Emily, you both are the best sisters I could have ever asked for. You both have always made me feel like part of the family and incredibly loved and supported. To Mike and Rhoda, you are the best second parents anyone could hope for. You have been a huge part of my life for over 12 years, and your unconditional love and support means the world to me. You both have always gone out of your way to make me know how loved, supported and valued I am.

To my immediate family, the Eastmans, what can I say. What words could come even close to describing how much your support, not just over the past 5 years, but over my entire life has meant to me. You are my original support system, and all the love and care that you have poured into me have helped make me who I am today. To my mother Jeanne, father Mark and brother Connor, you have all been my endless supporters, and I know you will be there for me no matter what. Having you all in my corner means the world to me. Connor, I always put a little bit extra pressure on myself during school growing up because I wanted to be a good role model for you. I wanted you to be proud of your big sister. The moments where you have told me you have looked up to me have been some of the proudest moments of my life. As you have grown and become a formidable scientist in your own right, I love the conversations we have surrounding work. It is amazing to be able to talk science with you discussing experiments that worked and the experiments that failed. I look forward to being a rock for you during your graduate school experience as you have been for me. To Mom and Dad, thank you for everything. Thank you for instilling in me the values of hard work and determination which have served me well. I will always be grateful that you always encouraged my curiosity and fostered my desire for learning both inside and outside of school. You both always pushed me to succeed and helped me set and achieve many goals throughout my life. Dad, thank you for

instilling in me a love of classic rock and for all of the wisdom and life advice you gave on our car rides to school, and to and from basketball all those years. You taught me to always be coachable, and open-minded which has served me well in my life, and particularly in graduate school. Thank you for always making me laugh and smile. Thank you for always being there when I need you and being able to keep a level head when I have dove off the deep end. Thank you for being my father and loving me unconditionally. Mom, you are not only my mother but also one of my best friends. Your love, kindness and compassion have shaped me into the woman I am today. Your passion for learning was instilled within me, and thank you for fostering its growth through many years. Thank you for always pushing me to excel, and for being there to comfort me when I fall. Thank you for your handwritten notes of encouragement and love which always come at exactly the right time, I will cherish them forever. Thank you for always being there when I need to talk, vent, or share great news. Thank you for always being willing to proof-read my writing, and for all the proteomics you have picked up over the years by doing so. I can say definitively, without a doubt, that I would not be who I am today if it were not for the constant love and support of my parents and family. Words cannot describe how much I love you all, and how much you all mean to me. Leaving you all as Matthew and I moved to Wisconsin for my graduate schooling was one of the most difficult experiences of my life, and without your unwavering support I do not know if we would have been able to make it through. I love you all to infinity and beyond.

To Matthew, my soulmate, best friend and partner in crime, there are truly no words to capture what your unwavering support and encouragement has meant to me while we embarked on this PhD journey together. Thank you for encouraging me to choose UW-Madison, and embracing the change of leaving home and going on this adventure together as newlyweds. Truly, this accomplishment is not one just of

my own, but of us. I would not have been able to do this without you. Matthew has been there for me every step of the way, celebrating every high, and comforting me at every low. He has been there encouraging me to keep going when I want to give up, and is always there to encourage me to take breaks, and remind me to eat, when I get in too deep. He has always eagerly attended Smith group gathering and participated in departmental recruiting events, fully embracing the graduate school experience as a team. We have carpooled to work over my entire graduate school career, and he reminds me every morning as he drops me off that "Science is fun". He has come and stayed with me late at night in the chemistry building, when experiments are going late, and the building gets creepy. He has driven up to chemistry in the middle of the night when I jolt awake realizing, I forgot to take my samples out of the SpeedVac. Matthew, in every conceivable way you have always been there for me, even when it's hard, and not fun at all. There is no one I would have rather gone on this journey with. I cannot wait for us to start the next chapter of our lives! I love you the mostest.

CONTENTS

Contents viii

List of Tables xii

List of Figures xiv

Abstract xviii

1 Introduction 1

1.1 *Overview of Mass Spectrometry-Based Proteomics* 2

1.2 *The Process of Protein Inference* 6

1.3 *The Value of Alternative Proteases* 9

1.4 *The Importance of Sample-Specific Protein Databases* 13

1.5 *The Analysis and Discovery of Post-Translational Modifications* 17

1.6 *References* 21

2 Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data 30

2.1 *Abstract* 31

2.2 *Introduction* 31

2.3 *Methods* 34

2.4 *Results and Discussion* 39

2.5 *Conclusion* 54

2.6 *References* 55

3 ProteaseGuru: A Tool for Protease Selection in Bottom-up Proteomics 59

3.1 *Abstract* 60

3.2	<i>Introduction</i>	60
3.3	<i>Methods</i>	63
3.4	<i>Results and Discussion</i>	64
3.5	<i>Conclusion</i>	75
3.6	<i>References</i>	75
4	Enhanced Protein Isoform Characterization Through Long-Read Proteogenomics	80
4.1	<i>Abstract</i>	81
4.2	<i>Background</i>	82
4.3	<i>Results</i>	85
4.4	<i>Discussion</i>	107
4.5	<i>Conclusion</i>	110
4.6	<i>Methods</i>	111
4.7	<i>Availability of data and materials</i>	119
4.8	<i>References</i>	121
5	Discovery of Dehydroamino Acid Residues in the Capsid and Matrix Structural Proteins of HIV-1	131
5.1	<i>Abstract</i>	132
5.2	<i>Introduction</i>	132
5.3	<i>Methods</i>	135
5.4	<i>Results</i>	139
5.5	<i>Conclusions</i>	149
5.6	<i>References</i>	151
6	Conclusion	157

- 6.1 *Summary* 158
- 6.2 *Leveraging Multi-Protease Data for Proteoform Inference* 160
- 6.3 *Expansion of ProteaseGuru to Include Peptide Detectability Estimates* 162
- 6.4 *Integration of PacBio Long-Read Sequencing with Top-Down Proteomics* 164
- 6.5 *Functional Investigation of Dehydroamino Acids in HIV* 166
- 6.6 *References* 169

- 7 Appendix I: Supporting Information for "Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data" 171
 - 7.1 *Supplementary Experimental Methods* 172
 - 7.2 *Supplementary Parameter Tables* 175
 - 7.3 *Supplementary Methods for Data Analysis* 180
 - 7.4 *Supplementary Information for MetaMorpheus' Separate and Integrated Multi-Protease Comparison* 182
 - 7.5 *Supplementary Figures* 184
 - 7.6 *Supplementary Tables* 188
 - 7.7 *References* 189

- 8 Appendix II: Supporting Information for "ProteaseGuru: A Tool for Protease Selection in Bottom-up Proteomics" 190
 - 8.1 *Supplementary Table* 191

- 9 Appendix III: Supporting Information for "Enhanced Protein Isoform Characterization Through Long-Read Proteogenomics" 198
 - 9.1 *Supplementary Figures* 199
 - 9.2 *Supplementary Note 1: Long- read transcriptome sequencing of a human cell line* 202

9.3	<i>Supplementary Note 2: ORF calling from long-read transcripts</i>	205
9.4	<i>Supplementary Note 3: Determination of the high confidence protein database space based on long-read RNA-seq coverage of peptides</i>	208
9.5	<i>Supplementary Note 4: Criteria for Novel Peptide Identification</i>	209
9.6	<i>Supplementary Note 5: Rescue & Resolve algorithm abundance threshold optimization</i>	212
9.7	<i>Supplementary Note 6: Multi-protease validation for Rescue & Resolve results</i>	214
9.8	<i>Supplementary Tables</i>	218
9.9	<i>References</i>	221
10	Appendix IV: Supporting Information for "Discovery of Dehydroamino Acid Residues in the Capsid and Matrix Structural Proteins of HIV-1"	224
10.1	<i>Supplementary Note 1: Determination of the Reproducibility of Peptide Identifications</i>	225
10.2	<i>Supplementary Note 2: Evaluation of Virion Sample Preparation</i>	225
10.3	<i>Supplementary Figures</i>	226
10.4	<i>Supplementary Tables</i>	235
	Colophon	253

LIST OF TABLES

2.1	Comparison of Results from the Separate and Integrated Multi-Protease Approaches	40
2.2	Comparison of Entrapment Results from the Separate and Integrated Multi-Protease Approaches	42
2.3	Comparison of Results from the Tryptic Digest and Integrated Multi-Protease Approach	43
2.4	Comparison of Entrapment Results from Fido and MetaMorpheus	46
2.5	Comparison of Entrapment Results from ProteinProphet and MetaMorpheus	49
2.6	Comparison of Protein Group Ambiguity between ProteinProphet and MetaMorpheus	49
2.7	Comparison of Entrapment Results from DTASelect2 and MetaMorpheus	51
2.8	Comparison of Protein Group Ambiguity between DTASelect2 and MetaMorpheus	51
3.1	Variant Protein Results	70
3.2	Comparison of In Silico Digestion Tool Features	74
5.1	Confirmed Dehydroamino Acids and Their Properties	144
5.2	Summary of HIV Site-Specific Mutagenesis	150
7.1	Protease-Specific Digestion Condition	173
7.2	MetaMorpheus Search Parameters	175
7.3	Modifications for GPTMD	176
7.4	Comet Search Parameters	177
7.5	TPP-Protein Prophet Parameters	178
7.6	ProLuCID Search Parameters	179

7.7	DTASelect2 Parameters	180
7.8	Comparison of Entrapment Results of the Top 7,472 Protein Groups from the Separate and Integrated Multi-Protease Approaches.	183
7.9	Comparison of Entrapment Results of the Top 7,716 Protein Groups from the Separate and Integrated Multi-Protease Approaches.	184
7.10	Comparison of Peptide Sequences Identified by MetaMorpheus and Comet at 1% FDR.	188
7.11	Comparison of Peptide Sequences Identified by MetaMorpheus and ProLuCID at 1% FDR.	188
8.1	Species for the Skin Microbiome Subset	191
9.1	Protein Classifications Based on SQANTI Protein	218
9.2	Number of Isoforms for Each Transcript and Protein Isoform Classifications Between SQANTI and SQANTI Protein	218
9.3	Summary of MetaMorpheus Search Results	219
9.4	Search Parameters for MetaMorpheus	220
10.1	Percent of Peptides Identified in Multiple Biological Replicates	225
10.2	Summary of Proteomic Search Results for Unlabeled and Glutathione-Labeled Samples	236
10.3	Relative Abundance of DHA/DHB containing PSMs in the 100 Most Abundant Human Proteins	237
10.4	HIV-1 Protein Sequences for Search Database	246
10.5	MetaMorpheus GPTMD Modifications	251
10.6	MetaMorpheus Search Task Settings	252

LIST OF FIGURES

1.1	Sources of proteome complexity	3
1.2	Experimental workflows for bottom-up and top-down proteomic approaches	5
1.3	Comparison of the theoretical and experimental length distributions of tryptic peptides	10
1.4	Comparison of short- and long-read sequencing for the reconstruction of transcript isoforms	16
2.1	Workflows for protein inference comparisons.	36
2.2	Comparison of protein group sizes between the separate or integrated multi-protease protein inference approaches	40
2.3	Comparison of protein groups identified between the separate or integrated multi-protease protein inference approaches	41
2.4	Comparison of protein group sizes between the tryptic digest or integrated multi-protease protein inference approaches	44
2.5	Comparison of protein groups identified between the tryptic digest or integrated multi-protease protein inference approaches	45
2.6	Comparison of multi-protease protein inference algorithms for false positive identifications.	47
2.7	Benefits of utilizing multiple proteases for the identification post-translational modifications	52
2.8	Breakdown of GPTMD identified post-translational modifications by protease	53
3.1	Comparison of percent unique peptide sequences for mouse and human databases	67

3.2	Comparison of the percent of shared peptide sequences for each protease between the Spritz proteogenomic database and the reference UniProt database	68
3.3	Number of variant proteins that can be identified by unique peptides . .	69
3.4	Sequence coverage map of variant containing protein (H3BQZ5_C25R) exported from ProteaseGuru	71
3.5	Histogram comparing the distribution of percent protein sequence coverage for the skin microbiome based on the protease used for in silico digestion	73
4.1	Challenges of protein isoform identification using MS-based proteomics	83
4.2	Long-read proteogenomic approach for enhanced sample-specific protein identification	87
4.3	Generation and characterization of a long-read RNA-seq derived protein database	89
4.4	Customized long-read-derived protein database for protein isoform detection	95
4.5	Discovery of novel peptides and full-length protein isoforms	99
4.6	Long-read-informed protein isoform detection	101
5.1	Schemas for dehydroamino acid generation and labeling	142
5.2	Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 86 in the HIV matrix protein	145
5.3	Tertiary structure of the HIV-1 matrix and capsid proteins	147
7.1	Distribution of target and decoy PSM scores resulting from digestion . .	181
7.2	Depiction of how the peptide sequence “FHSMASR” can be attributed to different protein groups based on which protease it originated from . . .	183
7.3	Investigation of protein groups unique to the separate approach	184

7.4	Investigation of protein groups unique to the protein inference results of the tryptic digest	185
7.5	Comparison of MetaMorpheus' and ProteinProphet's protein inference algorithms for false positive identifications	186
7.6	Comparison of MetaMorpheus' and DTASelect2's protein inference algorithms for false positive identifications	187
9.1	Detailed schematic of the Nextflow computational pipeline for long-read proteogenomics	199
9.2	Generation and characterization of candidate protein isoform sequences from long-read RNA-seq data	200
9.3	Comparison of MS-based proteomic coverage when using different protein databases for MS searching	201
9.5	Relationship between RNA and protein estimated abundances	202
9.6	Long-read transcriptome length and abundance distributions	203
9.7	Co-expression of multiple isoforms from the same gene	203
9.8	Abundance distribution of major versus minor transcript isoforms	204
9.9	Fraction of transcript isoforms in which the major isoform does not match the GENCODE principle APPRIS transcript isoforms	204
9.10	Breakdown of transcript isoforms by their novelty category	205
9.11	Comparison of ORF callers in predicting ORFs from full-length transcripts (PacBio)	206
9.12	Distribution of ORF scores from the CPAT algorithm	206
9.13	Evaluation of ORF plausibility and weighting	207
9.14	Fraction of ORFs predicted from the GENCODE transcriptome using the modified CPAT ORF calling pipeline	207

9.15	Characterizing the length and abundance biases that contribute to lower proteomic coverage from PacBio-derived databases	209
9.16	Evaluation of peptide identification recovery as a function of gene average transcript length and abundance	210
9.17	Comparison of novel and canonical peptide distributions for q -values and PEP values	211
9.18	Abundance threshold evaluation for the Rescue & Resolve algorithm	213
9.19	Relationship between transcriptional abundance and MS detectability	214
10.1	Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 56 in the HIV matrix protein	227
10.2	Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 16 in the HIV capsid protein	228
10.3	Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 119 in the HIV capsid protein	229
10.4	Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 188 in the HIV capsid protein	230
10.5	Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 198 in the HIV capsid protein	231
10.6	Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 210 in the HIV capsid protein	232
10.7	Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 216 in the HIV capsid protein	233
10.8	Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 218 in the HIV capsid protein	234

ABSTRACT

Proteins are the key biological actors within cells, driving many biological processes integral for both healthy and diseased states. Understanding the depth of complexity represented within the proteome is crucial to our scientific understanding of cellular biology and to provide disease specific insights for clinical applications. Mass spectrometry-based proteomics is the premier method for proteome analysis, with the ability to both identify and quantify proteins. Although proteomics continues to grow as a robust field of bioanalytical chemistry, advances are still necessary to enable a more comprehensive view of the proteome. In this thesis, several new tools for the improvement of proteome characterization are described, seeking to not only increase the depth of proteome characterization, but also the precision of the results obtained. In **Chapter 1**, an overview of mass spectrometry-based proteomics is provided including specific background information for the different areas of proteomic analysis addressed in the chapters of this thesis. **Chapter 2** introduces multi-protease protein inference and illustrates advantages of utilizing peptides from multiple proteolytic digests for protein inference. **Chapter 3** describes an in silico digestion software tool called ProteaseGuru, designed to aid in the consideration of alternative proteases for bottom-up experiments. **Chapter 4** establishes a software pipeline for the generation of sample-specific databases from PacBio long-read sequencing data. **Chapter 5** describes the discovery and validation of dehydroamino acid residues within the HIV-1 virus. These uncommon post-translational modifications (PTMs) were initially discovered using global PTM discovery (GPTMD), and then subsequently validated using a chemical labeling strategy. All the works described throughout this thesis are summarized and future directions for the improvement of proteome characterization are outlined in **Chapter 6**.

1 INTRODUCTION

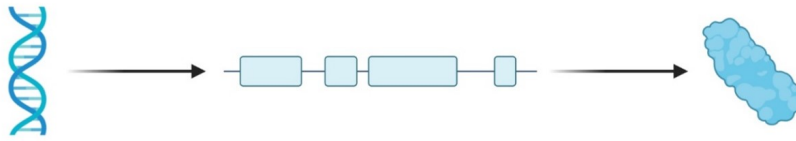
1.1 Overview of Mass Spectrometry-Based Proteomics

Proteins are central to nearly all major biological processes within the cell, acting as a bridge between genotype and phenotype.¹ Comprehensive characterization of the proteome would deepen our understanding of diseases and complex biological processes, and is an ongoing goal of mass spectrometry-based proteomics. However, this is not trivial, as the proteome is not only highly complex but also dynamic in nature. Therefore, proteomic analysis must not only seek to identify which proteins are present, but also their abundance and modification status.

Initially, it was thought that a single gene was transcribed to a single RNA transcript, which was then translated into a single protein. However, this single gene to single protein hypothesis has since been abandoned. Instead, it is now understood that the proteome is incredibly diverse, with numerous protein products, or proteoforms, coming from a single gene (see **Figure 1.1**). A proteoform is defined as the distinct molecular form of a protein, with a specific amino acid sequence and set of post-translational modifications.² The depth of proteoform complexity within the proteome is not yet fully understood, but this immense diversity can stem from numerous sources such as mutations at the gene-level, variants or alternative splicing at the transcript-level, and post-translational modifications or cleavage events at the protein-level.²⁻⁶ This depth of complexity further supports the importance of characterizing the proteome, because analysis of the genome and transcriptome alone cannot fully account for the complex phenotypes observed in healthy and disease states.

Mass spectrometry-based proteomics has quickly become the most high-throughput, reliable, and sensitive method for the characterization of the proteome.⁷⁻¹⁰ The principle of applying tandem mass spectrometry to the study of proteins is quite simple. In

Gene-Centric: One Gene = One Protein



Proteoform-Centric: Capture all diversity at the DNA, RNA and Protein level

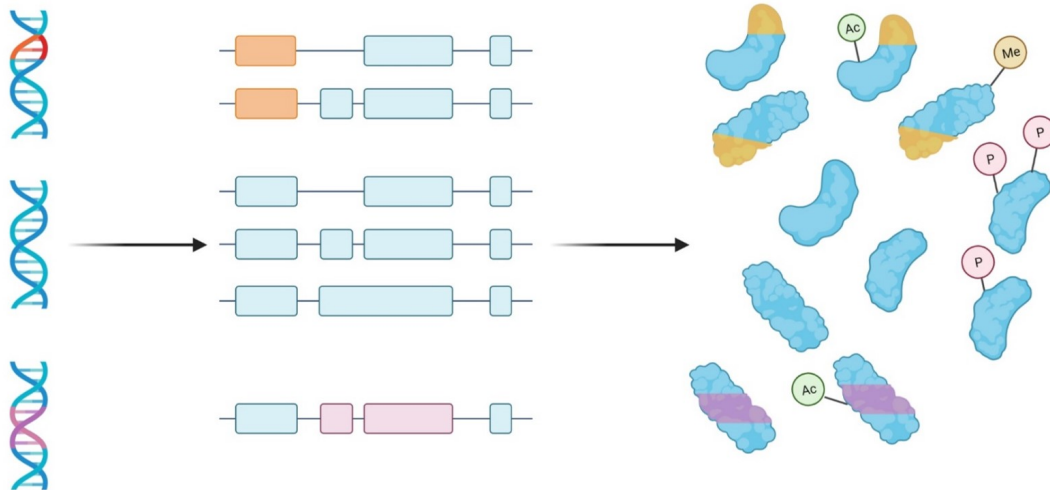


Figure 1.1: Sources of proteome complexity. Proteoforms provide a depth of complexity to the proteome which would not be possible if a gene only led to the production of a single protein product. Instead, mutations at the gene-level, variants or alternative splicing at the transcript-level, and post-translational modifications or cleavage events at the protein-level contribute to a still undefined number of proteoforms, which are the functional units of the proteome.

the initial MS1 spectra, the intact mass of a peptide or proteoform analyte is determined by measuring its mass-to-charge (m/z) ratio and using the observed charge state (z). In the subsequent MS2 scan, the intact peptide or proteoform is fragmented, generating product ions whose m/z values enable amino acid sequence determination leading to identifications. Beyond identification, mass spectrometry-based proteomics can also facilitate the quantification of peptide or proteoform analytes.

Mass spectrometry-based proteomics can be divided into two different approaches,

bottom-up and top-down. The key difference between these two approaches is the analyte, which is either a peptide or a proteoform, respectively (see **Figure 1.2**). The vast majority of proteomics experiments utilize the bottom-up approach. In bottom-up, or shotgun proteomics, proteins are digested into peptides which are then analyzed via LC-MS/MS.¹¹ Peptides are ideal analytes for mass spectrometry-based proteomics because they are easy to solubilize, separate and ionize. Since peptides are the observed unit in bottom-up proteomics, but protein-level information is still the desired outcome, peptides must act as proxies for their proteins or proteoforms of origin. All information regarding the presence and abundance of proteins in the sample are inferred from the peptides identified. The assumption that peptides are ideal proxies for the proteins or proteoforms in the sample is somewhat faulty. When proteoforms are digested into peptides, they lose their connectivity to their proteoforms of origin, which not only complicates the process of protein identification (see **Section 1.2**) but also prevents the determination of which proteoforms are present in the sample. When reconstructing proteins from peptides, it is impossible to completely reconstruct the complexity of the proteome at the proteoform-level.

In top-down proteomics, intact proteins/proteoforms are analyzed via tandem mass spectrometry.¹²⁻¹⁷ Here intact proteoforms are directly being observed, and the relationship between the base amino acid sequence and the post-translational modifications on the proteoform are preserved. Therefore, no proxies are required in top-down proteomics. However, top-down analysis is very complicated and there are many challenges that must be overcome including but not limited to the low abundance of many proteoforms, the low signal-to-noise ratio of large molecular weight proteoforms, and low solubility of intact proteoforms.¹⁸⁻²⁰ Currently, the sensitivity of top-down proteomics is quite restricted compared to that of bottom-up proteomics. Top-down proteomics is limited to those proteins with high abundance

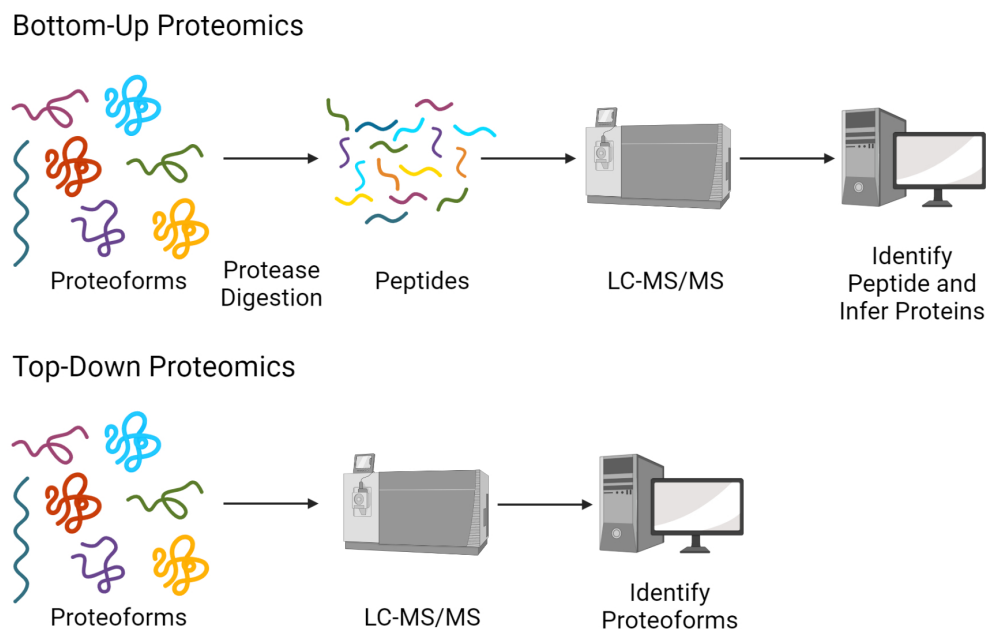


Figure 1.2: Experimental workflows for bottom-up and top-down proteomic approaches.

and low molecular weight, with sensitivity diminishing drastically for proteins with a mass above 30 kDa.¹⁸

This dissertation will focus on the development of tools to improve bottom-up proteomic characterization of the proteome. Although bottom-up proteomics is quite robust, there are many places within the conventional workflow where improvements can be made to enhance the characterization of the proteome. Here we will discuss four areas in which the development of new tools and methods can further improve proteome characterization via mass spectrometry-based proteomics: 1) the process of protein inference, 2) the use of alternative proteases, 3) the use of sample-specific databases and 4) the discovery and validation of post-translational modifications.

1.2 The Process of Protein Inference

In bottom-up proteomics, as discussed in the previous section, peptides are the analyte. Proteins within a sample are digested into peptides, which are then analyzed via tandem mass spectrometry. Although peptides are directly observed, more often than not, protein-level identifications and abundance measures are desired.²¹ Therefore, the observed peptides serve as an intermediate to the desired protein-level results, making it necessary to reconstruct the original proteins in the sample. This reconstruction process is called protein inference and is often quite complicated and imperfect. The process of protein inference is convoluted by the existence of “shared peptides”, which are peptide sequences that could result from the digestion of multiple proteins present in the sequence database.^{21,22} The identification of these peptides generates ambiguity in the protein-level results, because it is impossible to distinguish the peptide’s protein of origin. Conversely, there are “unique peptides”, which are peptides distinct to a single protein within the sequence database, and the identification of such a peptide can confidently identify a single protein.^{21,22} The more shared peptides identified, the more complicated the process of protein inference becomes. Shared peptides are increasingly prevalent in higher order eukaryotic organisms where there is a greater degree of sequence homology resulting from related protein families, paralogous genes and complex alternative splicing.^{21,23,24} Various models exist to address the protein inference problem, most of which differ from each other in their approach to handling the complications arising from shared peptides.

Algorithms for protein inference can be broadly grouped into three categories: 1) optimistic, 2) statistical and 3) parsimonious. In optimistic algorithms, all possible proteins which could exist, based on the peptides identified, are considered detected.

The underlying assumption made when utilizing this approach, is that the sample contains a large number of homologous proteins.²¹ Optimistic algorithms tend to be the simplest approach to protein inference, since there is no effort to reduce the ambiguity conferred by shared peptides. This also makes these algorithms the easiest to follow and comprehend for the end user. However, the increased ambiguity present in these algorithms is also why this model for protein inference is not widely utilized. Instead, statistical and parsimonious approaches have been and continue to be heavily favored. One example of optimistic inference is the original algorithm employed in DTASelect.²⁵

Statistical approaches assemble evidence from the peptide identifications to estimate the probability a given protein is present in the sample. Typically, these algorithms utilize peptide posterior error probability (PEP) values, or other peptide scoring metrics to calculate protein-level probabilities.^{10,21,22} Statistical protein inference algorithms can be further sub-classified into non-parametric or parametric models. Non-parametric, or distribution free methods, make few to no assumptions regarding the probability distributions of the data being assessed.²¹ Due to this, these methods are easier to use and are generally more robust. One of the most well-known and utilized non-parametric statistical protein inference algorithms is ProteinProphet.²⁶ Conversely, parametric models assume that the data used to generate the model comes from a probability distribution, and also makes assumptions regarding the parameters of said distribution.²¹ Due to the increased number of assumptions made in parametric models, they tend to produce more accurate protein probability estimates than non-parametric models, when the assumptions made are accurate. A major limitation to statistical approaches to protein inference is the inaccessibility of the logic underlying the algorithm. It can be unclear to the end user why certain proteins are weighted more heavily than others.

Parsimonious approaches to protein inference seek to apply the principle of Occam's razor, which states the simplest answer is most likely the correct answer, to handle the problem of shared peptides.²¹ The goal of these approaches is to establish the minimum set of proteins which can explain all the identified peptides. The complexity of parsimony is equivalent to the computationally prohibitive NP-hard set cover problem.²¹ Therefore, to be able to "solve" what the minimum set of proteins are in the sample, heuristics and assumptions must be established, enforcing the simplest answer is likely to be the correct answer. Several statistical approaches have principles of parsimonious algorithms at their core.^{10,27} The discarding of putative proteins when alternative protein identifications have more support is a major limitation of parsimonious approaches, because these removed proteins could be present in the sample.²² Additionally, the heuristics and assumptions that are central to the algorithm may not be clear to the end user, making it difficult to understand the end protein list, and why some proteins are absent.²⁷

The problem of protein inference and how to handle shared peptides is not yet solved, and new algorithms are still being developed.²⁷ One method for improving the quality of protein inference results, outside of continued algorithm development, is the curation of peptide identifications used as input for the inference algorithm. All assumptions regarding the presence or absence of a protein are based on the peptides used within the inference process. Increasing the depth and quality of the peptide identifications will in turn also increase the depth and quality of the inferred protein identifications. One approach to increasing the quality of inferred proteins is to increase the stringency of applied peptide filters.²⁸ If false positive peptide identifications are incorporated for inference, they can lead to identifications that are not reflective of the sample's proteome. However, being overly conservative can result in the loss of valuable true positives.²⁸ In the process of protein inference, the more

quality peptide identifications utilized, the better. Towards this end, several studies have shown protein inference results can be improved through the aggregation of peptide identifications across multiple search engines prior to inference.²⁹ Another approach to improve protein inference, that will be expanded upon in **Chapter 2**, is the use of peptide identifications from multiple orthogonal proteolytic digests.³⁰ Leveraging these peptide identifications from alternate proteases increases sequence coverage of the proteome, and the number of unique peptides identified, both of which have a positive impact on the accuracy of protein inference results.³⁰

1.3 The Value of Alternative Proteases

For bottom-up proteomics, the serine protease trypsin is used almost exclusively. Trypsin is robust, reliable, and affordable.³¹⁻³³ Cleaving after lysine or arginine residues, trypsin generates small peptides with a charged residue at the C-terminal position, ideal for collision-induced dissociation (CID) fragmentation.³¹⁻³³ However, the near ubiquitous utilization of trypsin provides a tunnel-like view of the proteome.^{31,34} Trypsin alone is incapable of producing peptide identifications sufficient for the comprehensive characterization of the proteome. One factor contributing to this lack of comprehension is the mismatch of the peptide length distribution between those produced by tryptic digest and those identified via mass spectrometry (see **Figure 1.3**). Most peptides identified by mass spectrometry are between 7-35 amino acids in length. Nearly one-third of the peptides theoretically produced by tryptic digestion of the human proteome are under 6 amino acids in length and are too small for MS/MS based identification. This can lead to regions of proteins which are intractable to tryptic peptides. There are also entire classes of proteins and post-translational modifications (PTMs) which are difficult to characterize with tryptic

digests.³⁴ One such class of proteins are membrane proteins, whose transmembrane domains are composed mainly of hydrophobic amino acids, with very few lysine or arginine residues. Digestion of these transmembrane proteins generate very long and very hydrophobic peptides which are difficult to solubilize and ionize for mass spectrometry-based proteomics.

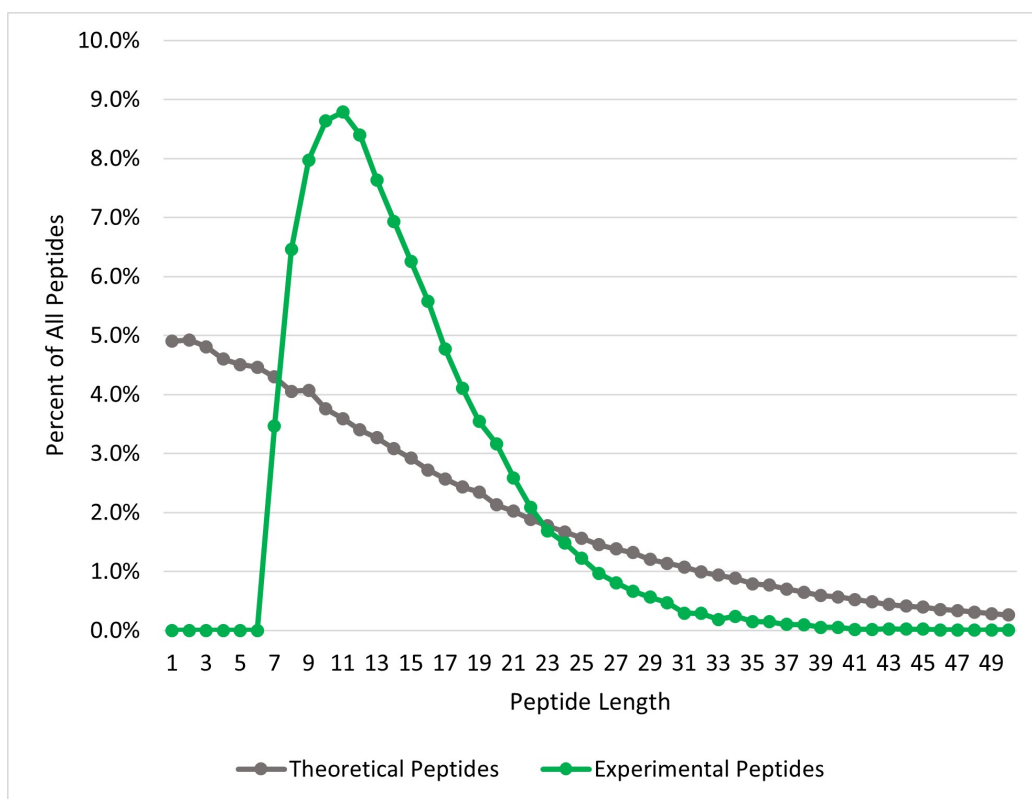


Figure 1.3: Comparison of the theoretical and experimental length distribution of tryptic peptides. The length distribution of in silico digested tryptic peptides (grey), as determined by ProteaseGuru, is compared to the length distribution of peptides experimentally identified from MetaMorpheus analysis of the tryptic data from **Chapter 2** (green).³⁰ Most experimentally identified peptides are between 7-35 amino acids in length, whereas the theoretical tryptic digest favors the generation of shorter peptides.

Additionally, tryptic digestion may elicit an inherent bias in the proteomic results obtained. Examples of this are 1) phosphoproteome analysis, 2) the identification of splice junction peptides and 3) quantitative proteomic experiments. In phospho-

proteome analysis, when negatively charged phosphorylated serine or threonine residues are adjacent to arginine or lysine residues, cleavage with trypsin can be inhibited. This results in longer peptides, with higher charge states that are not as amenable to identification with CID or higher-energy C-trap dissociation (HCD) fragmentation.^{31,33} This can result in biased phosphoproteome results, missing key phosphorylation sites, and lacking coverage in some of the most important regulatory regions throughout the proteome.^{34,35} For the identification of splice junction peptides, trypsin can also provide incomplete and therefore biased results. Surrounding exon boundaries, there are evolutionary preferred nucleotides which increase the occurrence of lysine and arginine coding triplets.³⁶ Due to this, most identifiable tryptic peptides flank splice junctions, and the peptides crossing the junction are too small to be identified. This is problematic for the characterization of proteome-wide alternative splicing, where identifying splice junction peptides are critical. The use of trypsin alone can also introduce bias in protein quantification results. Studies have shown that protein quantification values differ based on the protease used for analysis, and that the pooling of data from multiple proteases can provide the best estimate for accurate protein abundance values.^{37,38}

To overcome these pitfalls of trypsin, alternative proteases can be considered. The use of an alternative protease or multiple proteases has been shown to increase protein sequence coverage, the number of post-translational modifications identified, and the number of splice junctions covered.^{30-32,34,36,37,39} Different proteases have various strengths which may make them ideal for different proteomic applications.

Chymotrypsin, Glu-C and Lys-C, like trypsin, are all serine proteases and can be utilized for high-throughput proteomic analysis.³¹ Chymotrypsin cleaves after tyrosine, phenylalanine and tryptophan residues, and is favored for proteins with long stretches of hydrophobic amino acids. As an alternative protease, chymotrypsin

produces peptides which are generally considered to be the most orthogonal to those obtained by tryptic digests.³¹ Lys-C has strict specificity, cleaving only after lysine residues, and can produce longer peptides than trypsin.³¹ Lys-C is often paired with trypsin to improve the efficiency of cleavage after lysine residues. Glu-C which cleaves after glutamic acid, and also after aspartic acid when in phosphate buffers, is ideal for the digestion of heavily glycosylated proteins.³¹ Since the side chains of both glutamic and aspartic acid cannot be glycosylated, the modification will not inhibit cleavage of the proteins to peptides. Glu-C has also been heavily utilized for plasma proteomic applications.^{31,40}

There are also proteases which cleave N-terminally, or before their triggering amino acids. Asp-N cleaves before aspartic acid residues. One distinct advantage of Asp-N is its compatibility with detergents during the digestion process. Asp-N has been noted as an especially valuable alternative protease for sensitive targeted proteomic applications such as selected reaction monitoring (SRM) analyses.^{31,33,37} Lys-N, which cleaves before lysine residues, has high resistance to both denaturants and temperatures up to 70 °C.³¹ Peptide products of Lys-N digestion, when paired with electron transfer dissociation (ETD) fragmentation, can provide exceptional product ion coverage which in many cases could even enable facile de novo sequencing of the peptides.³¹

Arg-C, which cleaves after arginine residues, is another valuable alternative protease. Unlike with trypsin, the presence of a proline residue adjacent to an arginine residue does not prevent cleavage when using Arg-C.³¹ Arg-C, like Lys-C, produces longer peptides than what is achieved with trypsin. Arg-C is typically utilized alongside other proteases in a multi-protease approach to help characterize and map post-translational modifications as well as increase protein sequence coverage.

Protease discovery and optimization is an on-going area of research and interest.

One of the newer proteases is Proalanase, which cleaves after proline and alanine residues in highly acidic conditions.⁴¹ Proalanase enables the digestion of proline-rich proteins, such as collagen, and enables phospho-site profiling. It has been shown to be heavily orthogonal to tryptic digestion, providing valuable complementary coverage of the proteome.⁴¹

The use of multiple proteases, or alternative proteases, is crucial for the comprehensive characterization of the proteome. There are barriers that exist preventing widespread adoption of multiple, or alternative proteases. One such hurdle is the determination of which proteases are most beneficial to specific applications. This hurdle can be addressed using an *in silico* digestion tool to aid in experimental planning (see **Chapter 3**). Tools like this can be utilized to determine which proteases provide adequate or unique sequence coverage of target proteins, or sufficient PTM coverage. Another, more critical hurdle, is the increased time and sample requirements needed for multi-protease approaches. Towards addressing this concern, the Swaney group at University of California- San Francisco has developed a method which enables the pooling of peptides from multiple proteolytic digestions prior to data independent acquisition (DIA) analysis.⁴² Advances such as this are key to the future of comprehensive bottom-up proteomics leveraging multiple proteases.

1.4 The Importance of Sample-Specific Protein

Databases

Protein sequence databases are critical for high-throughput proteomic data analysis. Within search programs for bottom-up proteomics, protein sequence databases are digested *in silico* to generate a pool of candidate theoretical peptides. For each theoretical peptide, theoretical fragment ion m/z values are generated. These theoret-

ical peaks are then compared to those experimentally observed in the MS2 spectra to determine peptide identifications. Without protein databases, peptide identifications would necessitate the use of de novo sequencing, or more recently spectral library searching. De novo search approaches take significantly longer than database searching methods, and generally tend to have higher false positive rates.

Typically, for many model organisms, there are reference protein databases (UniProt, Ensembl, RefSeq) which can be utilized for proteomic analysis.^{43,44} These reference protein databases seek to broadly represent all proteins present. While these reference databases are useful starting points, it is known that even within the same species, protein sequences can vary between individuals, tissues, and cell lines. Therefore, reference databases may be incomplete and fail to represent each individual sample. If the protein database used for proteomic analysis is not concordant with the sample being analyzed, the accuracy of the proteomic results is detrimentally impacted, and the biological conclusions drawn from the results may be inaccurate. In many cases, the reference database may not only lack sequence variants, but may lack entire protein isoform sequences for a given gene. When the reference database is incomplete in this manner, peptides containing these variants, or that are unique to missing isoforms cannot be identified. Peptides shared between the missing isoforms and those present in the sequence database will be incorrectly parsed resulting in inaccurate protein inference results. It is also possible the sample may express a subset of the protein isoforms present in the reference database. In this case, protein-level results can have false positive identifications, or an inflated level of protein ambiguity.

One approach to dealing with this database-sample discordance is the generation of sample-specific databases. This idea spawns from the sub-field of proteomics called proteogenomics, which seeks to integrate transcriptomic and proteomic data.^{45,46} For the specific application of sample-specific database generation, RNA-sequencing data

can be translated *in silico* to construct a protein sequence database. Since this database is based on the RNA transcripts which function as protein precursors, the generated database is likely more accurate to the proteins and protein isoforms present in the sample than the reference. However, these constructed sample-specific databases still are subject to several limitations such as the sensitivity and specificity of the RNA-sequencing technologies utilized. Also, not all transcripts carry equivalent coding potential, and select protein isoforms, although translated, may not be stable.

Initially utilizing proteogenomics, reference databases were supplemented with peptide sequences containing variants or alternative splice junctions, as identified from short-read RNA-sequencing technology.⁴⁷⁻⁵² These augmented databases represented the first attempts to generate a sample-specific search space. However these databases could become rather large, containing many sequences within reference proteins that were not relevant to the actual sample.^{46,53} To address this, tools such as Spritz were created to generate entire sample-specific protein databases by reconstructing full transcripts from short-read RNA-sequencing followed by *in silico* translation.⁵⁴ Short-read RNA-sequencing has many parallels to bottom-up proteomics, in that the transcripts within the samples are fragmented to form short RNA oligonucleotides. These RNA fragments are then sequenced and mapped back to a reference genome to reconstruct RNA transcripts, much in the way that peptides are mapped to proteins through the protein inference process. Just like protein inference is imperfect, the process of reconstructing full transcripts from short-read RNA-sequencing is also imperfect. Short-read RNA-sequencing excels at the identification of sequence variants but can fall short in the reconstruction of alternatively spliced transcripts, just as bottom-up proteomics cannot reliably identify proteoforms (see **Figure 1.4**).

To overcome complications arising from the inaccurate parsing of RNA fragments into full-length transcripts, long-read RNA-sequencing technologies can be utilized.

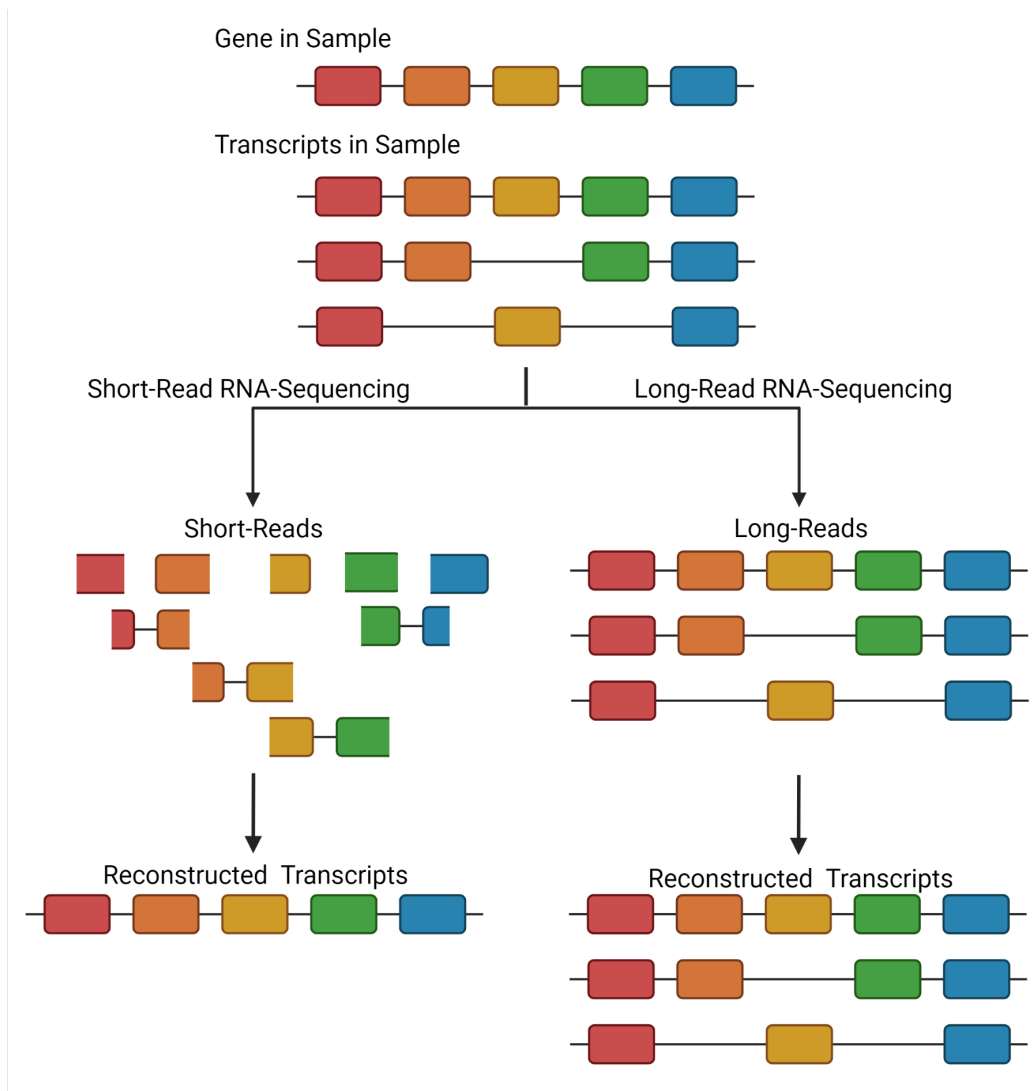


Figure 1.4: Comparison of short- and long-read sequencing for the reconstruction of transcript isoforms. In short-read RNA-sequencing approaches, RNA fragments are generated from which full-length transcripts must be reconstructed. Depending on the coverage of alternative splice junctions, incorrect transcript inference can be achieved. In this example, based on the fragments identified, a single transcript is reconstructed. Therefore, the two additional transcript isoforms are missed. In long-read RNA-sequencing, full-length transcripts are sequenced, and no reconstruction is required. Therefore, in the provided example, all three transcript isoforms expressed in the sample are identified.

In recent years, technology platforms from Pacific Biosciences and Oxford Nanopore, have become more prevalent in the transcriptomic community due to their ability to sequence full-length RNA transcripts with increasing accuracy.⁵⁵⁻⁵⁸ Specifically, for PacBio, technology has been developed to provide greater than 99% accuracy for the sequencing of single RNA transcripts.⁵⁹ Sequencing of intact, full-length RNA transcripts eliminates the read parsing issues of short-read sequencing approaches and enables a more comprehensive view of the transcript isoform landscape of the sample (see **Figure 1.4**). This can provide an even more precise sample-specific database than those constructed using short-read RNA-sequencing data, especially for protein isoforms. The use of PacBio long-read sequencing for sample-specific database generation, and its strengths will be discussed further in **Chapter 4**.

1.5 The Analysis and Discovery of Post-Translational Modifications

As powerful as proteogenomic approaches are for the generation of sample-specific databases, no transcriptional information can inform on the presence of post-translational modifications (PTMs). Post-translational modifications represent a critical layer of proteome diversity and are central to many important biological processes. The presence or absence of PTMs impact the function of proteoforms, contribute to signaling cascades and regulate diverse cellular functions.⁶⁰⁻⁶⁶ Mass spectrometry-based proteomics has quickly become the premier tool for the proteome-wide analysis of post-translational modifications. Using mass spectrometry-based proteomics for PTM mapping provides high sensitivity and throughput, as well as the ability to localize PTMs to a single amino acid residue. Unlike other PTM mapping approaches, such as antibody-based methods, proteomic analysis of PTMs is flexible

in terms of the PTMs being analyzed and is not limited to a single modification at a time.⁹

As the field of proteomics has evolved, the ability to characterize PTMs reliably and accurately has grown in its importance. Ignoring modified peptides or proteins leads to a vast under sampling of the proteome. Some modification sites are very well established and may be included as annotated modifications in the reference database (UniProt XML). However, these annotations are nowhere near complete, and proteins with unannotated PTMs are a large contributor to the dark proteome.⁶⁷⁻⁷¹

Methods for the discovery of PTMs not present in the protein database have evolved over time, giving greater PTM coverage. Initially, an approach called variable modification searching was applied.⁷² In this strategy, a selected PTM is allowed to occur on any amino acid residue fitting the modification motif in the search space. For example, for phosphorylation, theoretical peptides with phosphorylation at all serine, threonine or tyrosine residues are generated, as well as the unmodified theoretical peptides. This approach greatly expands the search space, increases search time, and introduces bias in the peptide-level false discovery rate (FDR) calculations.⁷¹ The bias in FDR calculations leads to a dramatic increase in the false positive rate for modified peptides. This approach is most valid when the variable modification being considered is widespread or enriched in the sample.⁷¹ Variable modification searching should only be applied for a small number of modifications at a time, as the negative repercussions of inflated false positive rates and increased database size compound with increasing numbers of modifications.⁷¹ These limitations make variable modification searching incompatible with reliable proteome-wide PTM discovery. To overcome many of the limitations of variable modification searching, Chick et. al. proposed a flexible method for PTM discovery and coined it “open search” or “open mass search”.⁶⁹ In open searching, a large precursor mass tolerance is per-

mitted. Therefore, the precursor mass of the experimental peptide can vary from the unmodified theoretical mass of the peptide and still be considered a match. The difference in mass observed can be accounted for by the mass(es) of unannotated PTMs. For the purpose of the Chick et. al foundational study, a mass difference up to 500 Da between the experimental and theoretical peptides was permitted.⁶⁹ In the open search approach, the product mass tolerance applied for the search remains narrow, requiring high-mass accuracy for fragment ion matches.⁶⁹ Therefore, a quality sequence tag can be utilized to identify the peptide's amino acid sequence in question, and the difference between the experimental precursor and theoretical peptide mass could be used to identify a PTM, or combination of PTMs. This process eliminates the database size issues of variable modification searching and maintains an accurate FDR rate for modified peptides. However, there are still several limitations with the open search approach, first of which is the high computational and time requirements necessary to complete this kind of search.^{67,68,71} Second, the difference in mass between the experimental and theoretical peptide may not always be easily identifiable as a PTM or combination of PTMs, leaving confusion and ambiguity. Third, this method fails to permit the identification of fragment ions containing the modified amino acid residue.⁶⁹ This becomes problematic if many of the potentially identifiable product ions contain the modified residue, making the modified peptide intractable to identification.

To further build on the open search approach, and address its downfalls, the Smith group invented global post-translational modification discovery (GPTMD), a multi-notch search approach for global discovery of PTMs.^{67,68} GPTMD searches for putative modifications found with an initial search using a multi-notch approach. This PTM discovery approach enables the identification of a large variety of PTMs while maintaining high confidence. The process of GPTMD has two main steps: 1)

a multi-notch initial search to augment the protein database with putative PTMs and 2) a narrow precursor mass search using the augmented database to confidently identify PTM modified peptides.⁶⁸ The multi-notch search is an extension of a narrow-precursor mass search enabling the inclusion of a variety of specific mass differences, or notches, between the precursor and theoretical masses. This approach improves upon the advantages of the open search approach, enabling the discovery of PTMs, without generating identifications with incomprehensible mass shifts.⁶⁸ GPTMD also reduces the search time and increases the accuracy of modified peptide identifications relative to open search approaches.⁶⁸ Using GPTMD, users define the mass notches they are willing to accept by selecting a list of modifications they are interested in discovering. A notch is generated for each mass shift associated with a PTM. Then, for each theoretical peptide, only experimental spectra with precursor masses that correspond to the unmodified peptide, or that differ by one of the defined notches are considered. These candidate spectra are then investigated for fragment ions matching the theoretical peptide. If a spectrum could correspond to a modified peptide, the corresponding PTM for the given notch is added to the augmented database. Once this augmented GPTMD database containing putative PTMs is generated, a final narrow-precursor mass search is completed to generate high confidence peptide identifications for both modified and unmodified peptides.⁶⁸ This approach can be used to consistently identify PTM modified peptides which are not present in the sequence database.⁶⁸ The application of GPTMD to the discovery of previously unknown PTMs in HIV-1 virions will be further discussed in **Chapter 5**.

1.6 References

- (1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, Type: Journal Article, 198–207.
- (2) Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nat Methods* **2013**, *10*, Type: Journal Article, 186–7.
- (3) Aebersold, R. et al. How many human proteoforms are there? *Nat Chem Biol* **2018**, *14*, Type: Journal Article, 206–214.
- (4) Smith, L. M.; Kelleher, N. L. Proteoforms as the next proteomics currency. *Science* **2018**, *359*, Type: Journal Article, 1106–1107.
- (5) Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347*, Type: Journal Article, 1260419.
- (6) Gaudet, P. et al. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res* **2017**, *45*, Type: Journal Article, D177–D182.
- (7) Han, X.; Aslanian, A.; Yates J. R., 3. Mass spectrometry for proteomics. *Curr Opin Chem Biol* **2008**, *12*, Type: Journal Article, 483–90.
- (8) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun* **2014**, *445*, Type: Journal Article, 683–93.
- (9) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates J. R., 3. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* **2013**, *113*, Type: Journal Article, 2343–94.

- (10) Serang, O.; Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat Interface* **2012**, *5*, Type: Journal Article, 3–20.
- (11) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates J. R., 3. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **1999**, *17*, Type: Journal Article, 676–82.
- (12) Siuti, N.; Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nat Methods* **2007**, *4*, Type: Journal Article, 817–21.
- (13) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal Chem* **2018**, *90*, Type: Journal Article, 110–127.
- (14) Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y. Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev Proteomics* **2016**, *13*, Type: Journal Article, 717–30.
- (15) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**, *9*, Type: Journal Article, 499–519.
- (16) Armirotti, A.; Damonte, G. Achievements and perspectives of top-down proteomics. *Proteomics* **2010**, *10*, Type: Journal Article, 3566–76.
- (17) Gregorich, Z. R.; Ge, Y. Top-down proteomics in health and disease: challenges and opportunities. *Proteomics* **2014**, *14*, Type: Journal Article, 1195–210.
- (18) 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*, Type: Journal Article, 6868–74.

- (19) Moore, S. M.; Hess, S. M.; Jorgenson, J. W. Extraction, Enrichment, Solubilization, and Digestion Techniques for Membrane Proteomics. *J Proteome Res* **2016**, *15*, Type: Journal Article, 1243–52.
- (20) Schaffer, L. V. et al. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19*, Type: Journal Article, e1800361.
- (21) Huang, T.; Wang, J.; Yu, W.; He, Z. Protein inference: a review. *Brief Bioinform* **2012**, *13*, Type: Journal Article, 586–614.
- (22) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005**, *4*, Type: Journal Article, 1419–40.
- (23) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem Sci* **2002**, *27*, Type: Journal Article, 74–8.
- (24) Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **2000**, *103*, Type: Journal Article, 367–70.
- (25) Tabb, D. L.; McDonald, W. H.; Yates J. R., 3. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002**, *1*, Type: Journal Article, 21–6.
- (26) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003**, *75*, Type: Journal Article, 4646–58.
- (27) Pfeuffer, J.; Sachsenberg, T.; Dijkstra, T. M. H.; Serang, O.; Reinert, K.; Kohlbacher, O. EPIFANY: A Method for Efficient High-Confidence Protein Inference. *J Proteome Res* **2020**, *19*, Type: Journal Article, 1060–1072.

- (28) Claassen, M.; Reiter, L.; Hengartner, M. O.; Buhmann, J. M.; Aebersold, R. Generic comparison of protein inference engines. *Mol Cell Proteomics* **2012**, *11*, Type: Journal Article, O110 007088.
- (29) Audain, E.; Uszkoreit, J.; Sachsenberg, T.; Pfeuffer, J.; Liang, X.; Hermjakob, H.; Sanchez, A.; Eisenacher, M.; Reinert, K.; Tabb, D. L.; Kohlbacher, O.; Perez-Riverol, Y. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J Proteomics* **2017**, *150*, Type: Journal Article, 170–182.
- (30) 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*, Type: Journal Article, 3429–3438.
- (31) Tsiatsiani, L.; Heck, A. J. Proteomics beyond trypsin. *FEBS J* **2015**, *282*, Type: Journal Article, 2612–26.
- (32) 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*, Type: Journal Article, 1323–9.
- (33) Vandermarliere, E.; Mueller, M.; Martens, L. Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom Rev* **2013**, *32*, Type: Journal Article, 453–65.
- (34) Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc* **2016**, *11*, Type: Journal Article, 993–1006.

- (35) Schlosser, A.; Vanselow, J. T.; Kramer, A. Mapping of phosphorylation sites by a multi-protease approach with specific phosphopeptide enrichment and NanoLC-MS/MS analysis. *Anal Chem* **2005**, *77*, Type: Journal Article, 5243–50.
- (36) Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* **2019**, *15*, Type: Journal Article, e8503.
- (37) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol Cell Proteomics* **2014**, *13*, Type: Journal Article, 1573–84.
- (38) Peng, M.; Taouatas, N.; Cappadona, S.; van Breukelen, B.; Mohammed, S.; Scholten, A.; Heck, A. J. Protease bias in absolute protein quantitation. *Nat Methods* **2012**, *9*, Type: Journal Article, 524–5.
- (39) Lanigan, L. T.; Mackie, M.; Feine, S.; Hublin, J. J.; Schmitz, R. W.; Wilcke, A.; Collins, M. J.; Cappellini, E.; Olsen, J. V.; Taurozzi, A. J.; Welker, F. Multi-protease analysis of Pleistocene bone proteomes. *J Proteomics* **2020**, *228*, Type: Journal Article, 103889.
- (40) Fossati, A.; Richards, A. L.; Chen, K. H.; Jaganath, D.; Cattamanchi, A.; Ernst, J. D.; Swaney, D. L. Toward Comprehensive Plasma Proteomics by Orthogonal Protease Digestion. *J Proteome Res* **2021**, *20*, Type: Journal Article, 4031–4040.
- (41) Samodova, D.; Hosfield, C. M.; Cramer, C. N.; Giuli, M. V.; Cappellini, E.; Franciosa, G.; Rosenblatt, M. M.; Kelstrup, C. D.; Olsen, J. V. ProAlanase is an Effective Alternative to Trypsin for Proteomics Applications and Disulfide Bond Mapping. *Mol Cell Proteomics* **2020**, *19*, Type: Journal Article, 2139–2157.
- (42) Richards, A. L.; Chen, K. H.; Wilburn, D. B.; Stevenson, E.; Polacco, B. J.; Searle, B. C.; Swaney, D. L. Data-Independent Acquisition Protease-Multiplexing En-

- ables Increased Proteome Sequence Coverage Across Multiple Fragmentation Modes. *J Proteome Res* **2022**, *21*, Type: Journal Article, 1124–1136.
- (43) The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **2017**, *45*, Type: Journal Article, D158–D169.
- (44) Aken, B. L. et al. Ensembl 2017. *Nucleic Acids Res* **2017**, *45*, Type: Journal Article, D635–D642.
- (45) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **2014**, *11*, Type: Journal Article, 1114–25.
- (46) Wang, X.; Liu, Q.; Zhang, B. Leveraging the complementary nature of RNA-Seq and shotgun proteomics data. *Proteomics* **2014**, *14*, Type: Journal Article, 2676–87.
- (47) 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*, Type: Journal Article, 2341–53.
- (48) 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*, Type: Journal Article, 228–40.
- (49) Sheynkman, G. M.; Johnson, J. E.; Jagtap, P. D.; Shortreed, M. R.; Onsongo, G.; Frey, B. L.; Griffin, T. J.; Smith, L. M. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **2014**, *15*, Type: Journal Article, 703.
- (50) Low, T. Y.; van Heesch, S.; van den Toorn, H.; Giansanti, P.; Cristobal, A.; Toonen, P.; Schafer, S.; Hubner, N.; van Breukelen, B.; Mohammed, S.; Cuppen, E.; Heck, A. J.; Guryev, V. Quantitative and qualitative proteome characteristics extracted

- from in-depth integrated genomics and proteomics analysis. *Cell Rep* **2013**, *5*, Type: Journal Article, 1469–78.
- (51) Ning, K.; Nesvizhskii, A. I. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics* **2010**, *11 Suppl 11*, Type: Journal Article, S14.
- (52) Evans, V. C.; Barker, G.; Heesom, K. J.; Fan, J.; Bessant, C.; Matthews, D. A. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat Methods* **2012**, *9*, Type: Journal Article, 1207–11.
- (53) Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification. *BMC Bioinformatics* **2012**, *13 Suppl 16*, Type: Journal Article, S2.
- (54) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *J Proteome Res* **2021**, *20*, Type: Journal Article, 1826–1834.
- (55) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A* **1996**, *93*, Type: Journal Article, 13770–3.
- (56) Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **2018**, *36*, Type: Journal Article, 338–345.
- (57) Van Dijk, E.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **2018**, *34*, Type: Journal Article, 15.
- (58) Sharon, D.; Tilgner, H.; Grubert, F.; Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **2013**, *31*, Type: Journal Article, 1009–14.

- (59) Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **2019**, *37*, Type: Journal Article, 1155–1162.
- (60) Doerr, A. Making PTMs a priority. *Nat Methods* **2012**, *9*, Type: Journal Article, 862–3.
- (61) Deribe, Y. L.; Pawson, T.; Dikic, I. Post-translational modifications in signal integration. *Nat Struct Mol Biol* **2010**, *17*, Type: Journal Article, 666–72.
- (62) Sirover, M. A. Subcellular dynamics of multifunctional protein regulation: mechanisms of GAPDH intracellular translocation. *J Cell Biochem* **2012**, *113*, Type: Journal Article, 2193–200.
- (63) Gould, N.; Doulias, P. T.; Tenopoulou, M.; Raju, K.; Ischiropoulos, H. Regulation of protein function and signaling by reversible cysteine S-nitrosylation. *J Biol Chem* **2013**, *288*, Type: Journal Article, 26473–9.
- (64) Cousin, C.; Derouiche, A.; Shi, L.; Pagot, Y.; Poncet, S.; Mijakovic, I. Protein-serine/threonine/tyrosine kinases in bacterial signaling and regulation. *FEMS Microbiol Lett* **2013**, *346*, Type: Journal Article, 11–9.
- (65) Doll, S.; Burlingame, A. L. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem Biol* **2015**, *10*, Type: Journal Article, 63–71.
- (66) Olsen, J. V.; Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics* **2013**, *12*, Type: Journal Article, 3444–52.
- (67) 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*, Type: Journal Article, 1383–1390.

- (68) 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*, Type: Journal Article, 1844–1851.
- (69) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **2015**, *33*, Type: Journal Article, 743–9.
- (70) Skinner, O. S.; Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat Biotechnol* **2015**, *33*, Type: Journal Article, 717–8.
- (71) Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Sheynkman, G. M.; Scalf, M.; Keller, M. P.; Attie, A. D.; Smith, L. M. Global Identification of Protein Post-translational Modifications in a Single-Pass Database Search. *J Proteome Res* **2015**, *14*, Type: Journal Article, 4714–20.
- (72) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **1994**, *5*, Type: Journal Article, 976–89.

2 IMPROVED PROTEIN INFERENCE FROM MULTIPLE PROTEASE

BOTTOM-UP MASS SPECTROMETRY DATA

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Millikin, R.J.; Hoffman, 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. *Journal of Proteome Research* **2019**, *18*(9), 3429–3438. <https://doi.org/10.1021/acs.jproteome.9b00330>.

Copyright © 2019 American Chemical Society.

2.1 Abstract

Peptides detected by tandem mass spectrometry (MS/MS) in bottom-up proteomics serve as proxies for the proteins expressed in the sample. Protein inference is a process routinely applied to these peptides to generate a plausible list of candidate protein identifications. The use of multiple proteases for parallel protein digestions expands sequence coverage, provides additional peptide identifications, and increases the probability of identifying peptides that are unique to a single protein, which are all valuable for protein inference. We have developed and implemented a multi-protease protein inference algorithm in MetaMorpheus, a bottom-up search software program, which incorporates the calculation of protease-specific q -values and preserves the association of peptide sequences and their protease of origin. This integrated multi-protease protein inference algorithm provides more accurate results than either the aggregation of results from the separate analysis of the peptide identifications produced by each protease (separate approach) in MetaMorpheus, or results that are obtained using Fido, ProteinProphet, or DTASelect2. MetaMorpheus' integrated multi-protease data analysis decreases the ambiguity of the protein group list, reduces the frequency of erroneous identifications, and increases the number of post-translational modifications identified, while combining multi-protease search and protein inference into a single software program.

2.2 Introduction

A frequent goal of proteomic studies is to accurately identify, characterize, and quantify all proteins expressed in a biological sample. The most prevalent strategy is referred to as "bottom-up" proteomics, wherein proteins present in the sample are digested into peptides and identified with liquid chromatography–mass spectrometry

(LC-MS/MS). These peptides serve as surrogate markers for the proteins from which they are derived. Peptides that can only result from the digestion of a single protein are referred to as “unique”, while peptides that can result from multiple different proteins are called “shared” and do not enable unambiguous identification of their protein of origin. This issue is addressed using a process referred to as “protein inference”, whereby the proteins most likely to be present in the sample are inferred from the observed peptides. When protein identifications are ambiguous, protein inference yields “protein groups”. Protein groups are collections of proteins that may be present in the sample and cannot be distinguished from one another on the basis of the peptides identified. Many different approaches have been described for protein inference,¹⁻³ but it remains a far from perfect process.

Combining the peptide identifications from orthogonal search programs or pooling peptide identifications from multiple replicates can expand proteome coverage, increase the number of confident peptide identifications used for protein inference, and enhance the quality of protein group identifications.⁴⁻⁹ The use of multiple proteolytic digestions in parallel has been shown to dramatically expand both the protein sequence coverage and the number of peptides identified from a given sample,^{10,11} indicating that combining multi-protease data for protein inference could have a strong positive impact on protein inference results.

We present here a multi-protease protein inference algorithm implemented in MetaMorpheus, a bottom-up search software program developed and maintained by our lab. MetaMorpheus allows for the parallel analysis of spectra files from multiple proteolytic digestions through the use of file-specific search settings. This advance facilitates the complete analysis of multi-protease data, from search to protein inference, in a single instance of MetaMorpheus without the need for any additional post-processing software programs. The multi-protease protein inference algorithm

also contains features which facilitate improved multi-protease protein inference. Specifically, protease-specific peptide false discovery rates (FDR) are calculated and employed in an attempt to maximize the number of high-quality peptides used for protein inference, and peptide identifications remain associated with their digesting protease to ensure proper determination of potential proteins of origin (further explanation of these improvements can be found in **Appendix I: Section 7.3**).

The performance of MetaMorpheus' integrated multi-protease protein inference algorithm was evaluated by comparing its results to those obtained by the manual aggregation of protein inference results from separate analysis of each protease's spectra files ("separate" approach) and to those obtained by analyzing only the files from a single protease (trypsin) digest. We also benchmarked the performance of the algorithm by comparing its results to those obtained by analyzing the data with three different protein inference software programs (Fido³, ProteinProphet¹², and DTASelect2⁹). An entrapment strategy, in which spectra files were searched against a concatenated human (*Homo sapiens*) and *Arabidopsis thaliana* protein database, was employed to facilitate intra- and intersoftware comparisons of protein inference results and assess their accuracy.¹³ Multi-protease data used for these studies was obtained via the analysis of highly fractionated aliquots of Jurkat cell lysate that were digested with one of six proteolytic enzymes (Arg-C, Asp-N, chymotrypsin, Glu-C, Lys-C, or trypsin).

We provide data below demonstrating that MetaMorpheus' multi-protease protein inference algorithm outperforms the separate, Fido, ProteinProphet, and DTASelect2 approaches. MetaMorpheus' multi-protease protein inference algorithm decreases the ambiguity of the protein group list and reduces the frequency of erroneous identifications. The MetaMorpheus implementation of the multi-protease protein inference algorithm is robust, user-friendly, and readily available to the community

as an open-source software program (<https://github.com/smith-chem-wisc/MetaMorpheus>).

2.3 Methods

The experimental procedures for the generation of the multi-protease data set were adapted from those reported in previous studies^{14,15} (see **Appendix I: Section 7.1**). A brief synopsis is as follows: six aliquots, one per protease (Arg-C, Asp-N, chymotrypsin, Glu-C, Lys-C, or trypsin), of approximately 10^7 Jurkat cells were lysed, and 150 μg of lysate was utilized for filter-aided sample preparation.¹⁶ Following digestion, peptides were fractionated off-line by high-pH reverse-phase liquid chromatography. Fractions were dried down and reconstituted in 5% acetonitrile and 1% formic acid prior to the LC-MS/MS analysis on a nanoACQUITY LC system (Waters, Milford, MA) interfaced with a Thermo Scientific LTQ Orbitrap Velos mass spectrometer. All mass spectrometry raw files are freely available on the MassIVE platform (<https://massive.ucsd.edu;ID:MSV000083304>).

Databases Used for Searches

The Swiss-Prot human FASTA (canonical and isoform) database containing 42,419 protein entries (downloaded from UniProt 3/22/19) was used for the comparison of protein inference results from the integrated and separate approaches with MetaMorpheus. All entrapment studies described used a concatenated human and *A. thaliana* FASTA database (canonical and isoforms, downloaded on 3/24/19) containing 60,391 protein entries. For the DTASelect2 analysis, a database containing both target (forward) and decoy (reverse) sequences was generated from the concatenated human and *A. thaliana* database.

Comparison of Integrated and Separate Protein Inference Methods Using MetaMorpheus

Spectra files for the six proteolytic digestions were calibrated, subjected to GPTMD (a post-translational modification (PTM)-discovery algorithm),^{17,18} and searched with MetaMorpheus (version 0.0.299). The final search was performed using either an integrated or separate approach to protein inference. In the integrated approach (**Figure 2.1A**), the “file-specific parameters” feature of MetaMorpheus was used to assign a protease to each spectra file. When using this setting, MetaMorpheus first identifies peptides via spectral matching, using the specified protease for each spectra file to generate the search space of theoretical peptides. After all searches have completed, peptide-spectral matches (PSMs) are grouped by their protease and the false discovery rate (FDR) is computed separately for each group. After FDRs have been calculated, PSMs from each protease are combined and filtered to an estimated 1% FDR prior to proteins being inferred.

In contrast to the integrated protein inference approach, the separate approach (**Figure 2.1B**) treats each set of spectra files for each protease as a separate analysis (i.e., all trypsin files are searched together, and a list of inferred proteins is output for that search). The resulting sets of protein lists were manually aggregated together into a single list while minimizing redundancy and ambiguity of the protein groups. For both the integrated and separate approaches, a 1% protein FDR threshold is estimated using q -values calculated from the standard target-decoy approach.

A video tutorial of how to use MetaMorpheus for the multi-protease analysis is available at <https://youtu.be/sk64xp5nfyI>, and a vignette containing the data used in the tutorial can be accessed at <https://uwmadison.box.com/s/dm6ezjbeyeahfe0xlc9hw9a5cl0i4j01>. Search settings and PTMs selected for discovery with GPTMD can

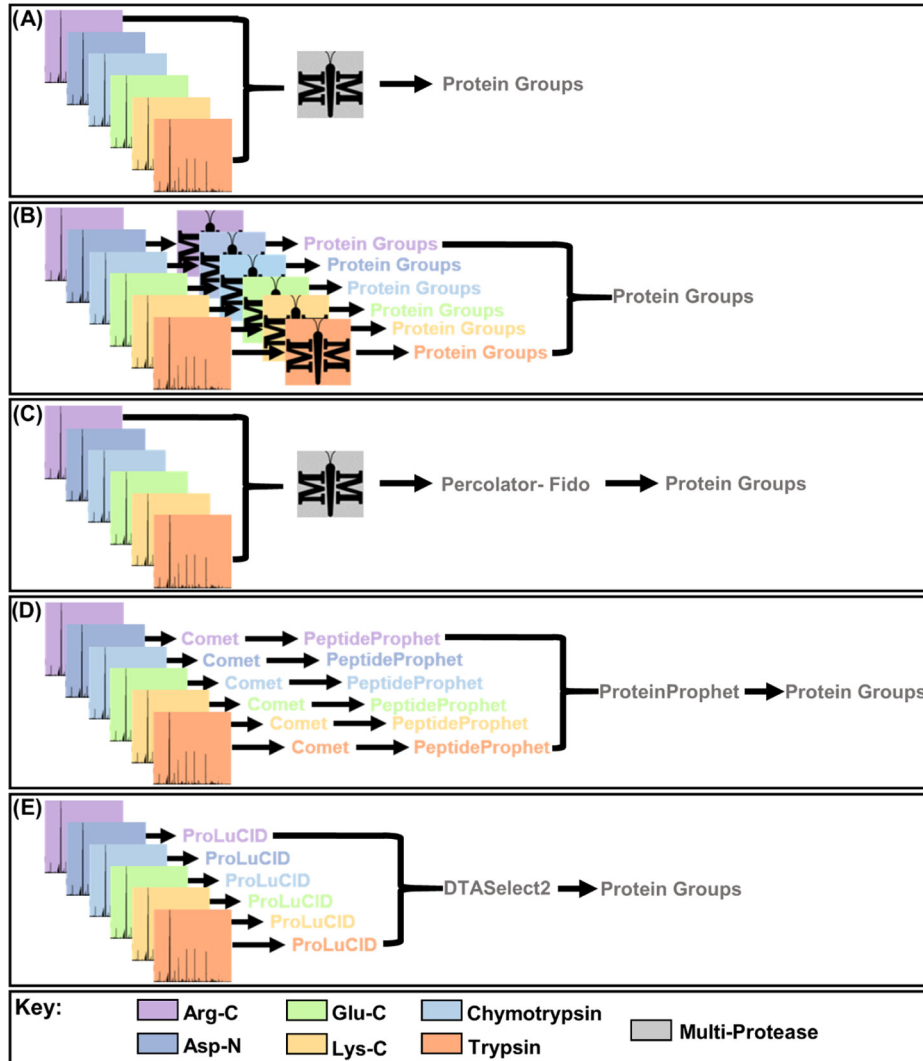


Figure 2.1: Workflows for protein inference comparisons. Workflows employed to compare results between (A) integrated multi-protease protein inference with MetaMorpheus, (B) separate multi-protease protein inference with MetaMorpheus, (C) multi-protease protein inference with Fido in Percolator, (D) multi-protease protein inference using ProteinProphet within trans-proteomic pipeline (TPP), and (E) multi-protease protein inference with DTASelect2.

be found in **Appendix I: Tables 7.2 and 7.3**, respectively.

Comparison of MetaMorpheus' Integrated Protein Inference to Percolator-Fido

Fido, a protein inference algorithm, is currently housed within Percolator (v.3-02), a software program whose purpose is to improve the number of confident peptide identifications through a semisupervised machine learning algorithm. MetaMorpheus writes search results that are compatible with Percolator's input requirements. We used MetaMorpheus' multi-protease search results as input to Percolator and used Fido for protein inference, comparing the results to MetaMorpheus' integrated protein inference approach (**Figure 2.1C**).

Comparison of MetaMorpheus' Integrated Protein Inference to ProteinProphet

ProteinProphet is housed within the trans-proteomic pipeline (TPP, v5.2.0). ProteinProphet can calculate protein-level probabilities but requires peptide-level probabilities as input, which MetaMorpheus does not calculate. This makes evaluating ProteinProphet's protein inference algorithm directly with MetaMorpheus' peptide-level search results infeasible. To compare MetaMorpheus' protein inference results to ProteinProphet's, all spectra files (grouped by protease) were searched by Comet¹⁹ (version 2018.01 rev.4) from within the trans-proteomic pipeline (search parameters can be found in **Appendix I: Table 7.4**). Search results for each protease were then input to PeptideProphet,²⁰ where the peptide-level probabilities necessary for ProteinProphet's algorithm were calculated (parameters for the ProteinProphet analysis can be found in **Appendix I: Table 7.5**). Peptide identifications that were not

present in the MetaMorpheus search results were removed (a comparison of the peptide identifications from MetaMorpheus and Comet-PeptideProphet can be found in **Appendix I: Table 7.10**). The trimmed PeptideProphet result files were loaded into ProteinProphet for protein inference (**Figure 2.1D**). The MetaMorpheus source code was altered to contain an exclusion list of all of the peptides whose sequences were not identified by Comet. Only peptide sequences that were not on this exclusion list were used for protein inference. This eliminated the effect of different peptide identifications on protein inference results.

Comparison of MetaMorpheus' Integrated Protein Inference to DTASelect2

To compare MetaMorpheus' protein inference results to DTASelect2's, all spectra files (grouped by protease) were searched by ProLuCID²¹ (version 1.3.5) (search parameters can be found in **Appendix I: Table 7.6**). Peptide identifications that were not present in the MetaMorpheus search results were removed (a comparison of the peptide identifications from MetaMorpheus and ProLuCID can be found in **Appendix I: Table 7.11**). The trimmed search results for each protease were then input to DTASelect2 (version 2.1.3) using "no enzyme" specificity to accommodate the multi-protease data (parameters can be found in **Appendix I: Table 7.7**) (**Figure 2.1E**). As with the ProteinProphet analysis, only peptide sequences that were identified using both search programs (ProLuCID and MetaMorpheus) were used for protein inference (**Appendix I: Table 7.11**).

2.4 Results and Discussion

Separate Versus Integrated Multi-Protease Analysis

The goal of protein inference is to, based on the peptides identified, generate a list of proteins that are in the sample and exclude those that are not. One approach to achieving this goal is to use Occam's razor, a "parsimonious" principle that seeks to generate the simplest set of proteins that can explain all identified peptides. We apply this principle via three key metrics that can be used to evaluate protein inference results of the integrated and separate multi-protease approaches: (a) minimize the number of protein groups identified, (b) maximize the percent of protein groups that contain a single protein accession, and (c) minimize the average number of protein accessions per group. **Table 2.1** gives these values for the separate and integrated multi-protease approaches and shows the improvement obtained in protein inference for each of the three metrics. The reduction in protein group ambiguity afforded by the integrated multi-protease approach is also illustrated in **Figure 2.2**. The reduction of protein group ambiguity indicates that the integrated multi-protease approach provides a simpler answer as to which proteins are projected to exist in the sample, indicating that it is more successful in achieving the goal of parsimonious protein inference. While the percent changes shown are modest in size, they are nonetheless notable improvements given the widespread importance of protein inference for protein studies.

The Venn diagram shown in **Figure 2.3** compares the protein group identifications obtained from the two protein inference approaches. While there is substantial overlap in the identifications (7,122 protein groups), there were 367 protein groups that were identified by only the integrated multi-protease approach (yellow region) and 864 protein groups by only the separate multi-protease approach (blue region). While a

Table 2.1: Comparison of Results from the Separate and Integrated Multi-Protease Approaches

	separate multi- protease approach	integrated multi- protease approach	percent change (%)
total number of protein groups	7,986	7,489	-6.2
percent of protein groups containing a single protein accession	66%	71%	+7.6
average number of protein accessions per protein group	1.58	1.48	-6.3

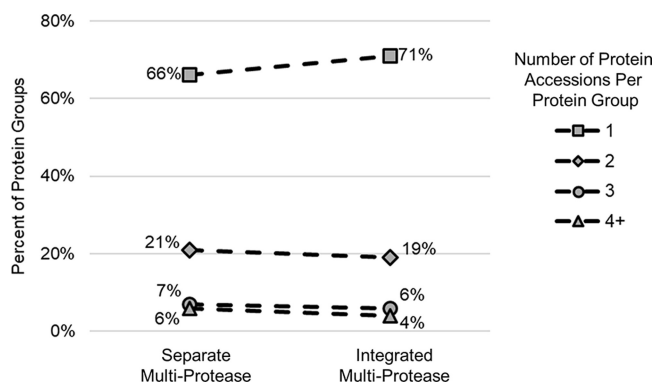


Figure 2.2: Comparison of protein group sizes between the separate or integrated multi-protease protein inference approaches. Plot of the percent of all protein groups produced by either the separate or integrated multi-protease protein inference approaches that contain between 1 and 4+ protein accessions. Reduction in the ambiguity of protein group identifications can be observed by comparing the integrated multi-protease protein inference results to those of the separate multi-protease approach.

smaller list of proteins may not at first seem like an improved result, in the case of parsimonious protein inference, a smaller protein list is actually preferred, given the same list of peptides as input (i.e., it is a simpler answer to explain the same data).

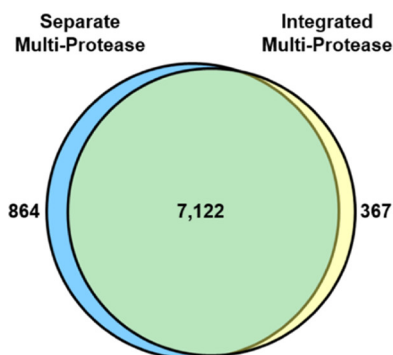


Figure 2.3: Comparison of protein groups identified between the separate or integrated multi-protease protein inference approaches. Venn diagram comparing protein groups identified at 1% FDR using the separate and integrated multi-protease protein inference approaches.

While the above results show a simpler answer, they are not necessarily more correct. To investigate if the results provided by the integrated protein inference method are not only simpler but also more accurate, the protein group differences between the two approaches were manually investigated. A majority of the protein groups unique to the separate multi-protease approach (519 of the 864) have their peptide identifications assigned to a protein group with reduced ambiguity in the integrated multi-protease approach: 349 became one or more protein groups containing only a single protein and 170 became a protein group with fewer protein members in the integrated multi-protease approach's results (**Appendix I: Figure 7.3**). This reduction in ambiguity further supports the claim that the integrated multi-protease approach provides improved protein inference results over the separate approach.

Further validation of the accuracy of the integrated multi-protease approach was provided by performing an entrapment study in which the spectra were searched

against a concatenated human and *A. thaliana* UniProt database, using either the separate or integrated approaches to protein inference. Since the spectra analyzed are from a human cell lysate, there should be only human proteins identified. If a protein group is identified as *A. thaliana*, it is necessarily a false positive. The accuracy of the protein inference algorithms can thus be evaluated by the number of *A. thaliana* protein groups identified and the corresponding false positive rate of the protein group results. The more accurate the algorithm, the lower the number of *A. thaliana* protein groups and the lower the corresponding false positive rate. The results of this analysis are summarized in **Table 2.2**. The number of *A. thaliana* protein groups identified and the corresponding false positive rate of protein group identifications decrease from the separate to integrated multi-protease approach, indicating that the integrated multi-protease strategy provides a more accurate list of protein groups. See the **Appendix I: Section 7.4** for additional analysis of the data.

Table 2.2: Comparison of Entrapment Results from the Separate and Integrated Multi-Protease Approaches

	separate multi- protease approach	integrated multi- protease approach	percent change (%)
number of human protein groups	7,400	7,255	-2.0
number of <i>A. thaliana</i> protein groups	316	217	-31.3
false positive rate	4.1%	2.9%	-29.3

MetaMorpheus Comparison: Tryptic Digest Versus Integrated Multi-Protease Analysis

The robust and reliable nature of tryptic digestions has made trypsin the primary proteolytic enzyme employed for bottom-up analyses. However, tryptic digests are far from perfect, as the majority of peptides produced are too short for MS-based sequence determination (56% of theoretical tryptic digestion products are six amino acids or less in length).^{10,22} The use of multiple proteolytic enzymes in parallel provides access to regions of the proteome that are not accessible when only using trypsin, resulting in more protein identifications, as well as improved protein inference. **Table 2.3** summarizes the protein inference results obtained for the tryptic digest and the integrated multi-protease approach. The reduction in protein group ambiguity afforded by the integrated multi-protease approach is also illustrated in **Figure 2.4**. The increase in the total number of protein groups for the integrated multi-protease approach compared to that for the tryptic digest does not violate Occam's razor because the number of peptide identifications used for protein inference differs between the two approaches.

Table 2.3: Comparison of Results from the Tryptic Digest and Integrated Multi-Protease Approach

	tryptic digest	integrated multi-protease approach	percent change (%)
total number of protein groups	5,173	7,489	+44.8
percent of protein groups containing a single protein accession	58%	71%	+22.4
average number of protein accessions per protein group	1.74%	1.48%	-14.9

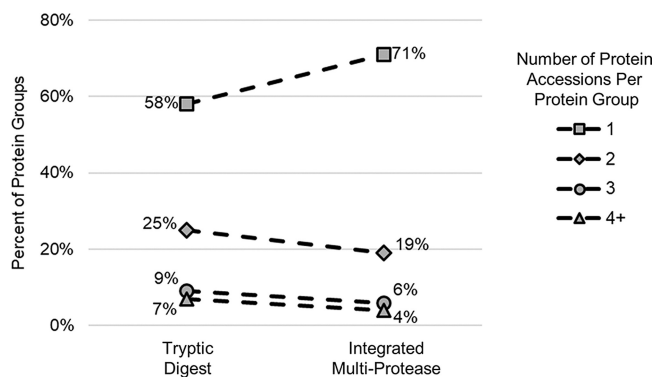


Figure 2.4: Comparison of protein group sizes between the tryptic digest or integrated multi-protease protein inference approaches. Plot of the percent of all protein groups produced by either the tryptic digest or integrated multi-protease protein inference approaches that contain between 1 and 4+ protein accessions. Reduction in the ambiguity of protein group identifications can be observed by comparing the integrated multi-protease protein inference results to those of the tryptic digest alone.

The Venn diagram shown in **Figure 2.5** compares the protein group identifications obtained from the two protein inference approaches. The large number of protein groups unique to the integrated multi-protease approach (3,518, yellow region) was expected based on the substantial increase in the total number of protein groups identified with the multi-protease data. There are 1,202 protein groups identified by the tryptic digest alone (red region). Further investigation of these protein groups showed that almost all of them (1,166 of the 1,202) have their peptide identifications assigned to a protein group with reduced ambiguity in the integrated multi-protease approach: reduction to one or more protein groups, each containing a single protein member (854 of the 1,166); reduction to a protein group with fewer protein members (288 of the 1,166); or reduction to one or more single protein groups and a protein group with fewer protein members (24 of the 1,166) (**Appendix I: Figure 7.4**). The majority (89%) of the protein groups that were disambiguated to a protein group containing a single protein member gained at least four additional peptide identifications from the other five proteolytic digestions (Arg-C, Asp-N, chymotrypsin, Glu-C,

and Lys-C). This indicates that the disambiguation of protein groups by a random peptide hit is highly unlikely.

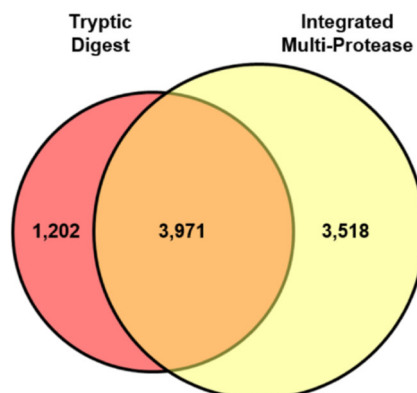


Figure 2.5: Comparison of protein groups identified between the tryptic digest or integrated multi-protease protein inference approaches. Venn diagram comparing the protein groups identified at 1% FDR from the tryptic digest and from the integrated multi-protease protein inference approach.

Intersoftware Comparisons: MetaMorpheus Integrated Multi-Protease Inference Compared to ProteinProphet, Fido, and DTASelect2

For all comparisons between software programs, entrapment studies were performed. Spectra were searched against the concatenated human and *A. thaliana* UniProt database. Protein groups that were identified as human were true positives, whereas protein groups that were identified as *A. thaliana* were false positives. The lists of proteins were ranked by either q -value or probability, depending on the software's output. Curves plotting false positives (x -axis) versus true positives (y -axis) were then used to evaluate each protein inference algorithm. Perfect protein inference would produce a vertical line that passes through the upper-left corner of the graph (only true positives identified). The closer the curve is to the upper-left corner of

the graph, the higher the overall accuracy of the protein inference algorithm. In addition to these curves, the number of *A. thaliana* protein groups identified and the corresponding false positive rate are also used to assess the performance of the algorithms.

For the Fido vs MetaMorpheus comparison, MetaMorpheus peptide identifications were directly imported into Percolator/Fido. The results for the Fido and MetaMorpheus comparison are summarized in **Table 2.4** and **Figure 2.6a**. The curve representing MetaMorpheus' protein inference results is closer to the upper left-hand corner than the curve representing Fido's results, indicating that MetaMorpheus' integrated multi-protease protein inference is more accurate in differentiating true and false positives. MetaMorpheus reports fewer total proteins than Fido, given the same peptide input; according to Occam's razor, this simpler result is more likely to be correct. Indeed, MetaMorpheus' integrated multi-protease protein inference algorithm showed 38.3% fewer false positive protein group identifications (162 to 100) and a 39.1% decrease in the false positive rate (2.3% to 1.4%) compared to Fido's results.

Table 2.4: Comparison of Entrapment Results from Fido and MetaMorpheus

	Fido	MetaMorpheus	percent change (%)
number of human protein groups	6,994	6,832	-2.3
number of <i>A. thaliana</i> protein groups	162	100	-38.3
false positive rate	2.3%	1.4%	-39.1

For the ProteinProphet and MetaMorpheus comparison, additional steps were required to normalize peptide identifications used for protein inference because MetaMorpheus' peptide output is not compatible with ProteinProphet. Consequently, the

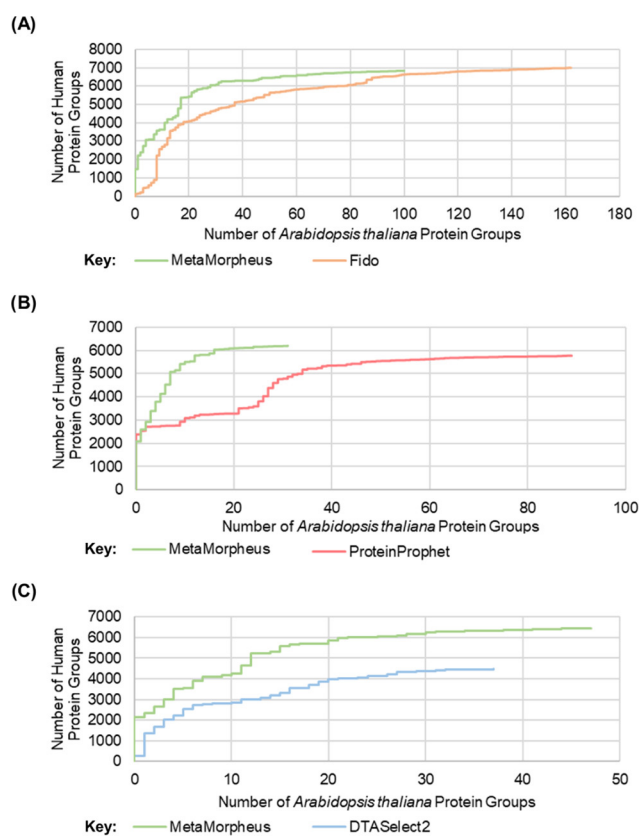


Figure 2.6: Comparison of multi-protease protein inference algorithms for false positive identifications. Curves comparing the ability of multi-protease protein inference algorithms to distinguish between human protein groups (true positives) and *A. thaliana* protein groups (false positives) for (A) protein group identifications from the MetaMorpheus and Fido comparison, (B) protein group identifications generated from the MetaMorpheus and ProteinProphet comparison, and (C) protein group identifications from the MetaMorpheus and DTASelect2 comparison. In all of the comparisons, MetaMorpheus produces a curve that is closer to the upper-left corner of the graph, indicating that the integrated multi-protease approach is more accurate.

Comet search engine within the trans-proteomic pipeline (TPP) was used for peptide identification. Although many proteomic software programs employ the same general strategy for peptide identification, the details of each algorithm are distinct and therefore they provide different results.⁷⁻⁹ Many of the high-confidence peptide identifications are found by multiple different algorithms, but others are unique. The peptide identifications unique to one search program can have a large impact on protein inference results (**Appendix I: Figure 7.5**). To provide a fair comparison of protein inference results, only peptides that were identified by both Comet and MetaMorpheus were used in the protein inference process. Additionally, ProteinProphet does not provide q -values for FDR threshold estimation. Therefore, to determine which ProteinProphet protein groups fall within the 1% protein FDR threshold, protein posterior error probabilities were summed and divided by the number of protein groups at or above the rank to calculate that protein group's q -value.²³ The results for the ProteinProphet and MetaMorpheus comparison are summarized in **Table 2.5** and **Figure 2.6b**. Overall, the curve representing MetaMorpheus' protein inference results is closer to the upper left-hand corner than the curve representing ProteinProphet's results, indicating that MetaMorpheus' integrated multi-protease protein inference is more accurate in differentiating true and false positives. MetaMorpheus' integrated multi-protease protein inference algorithm provides 5.7% more protein group identifications (5,877 to 6,214), an 80.5% decrease in the percent of *A. thaliana* protein group identifications (159 to 31) and an 80.8% decrease in the false positive rate (2.6% to 0.5%) as compared to ProteinProphet. ProteinProphet reports fewer total protein groups, making it appear to be the simpler answer, but MetaMorpheus has fewer *A. thaliana* identifications and a lower false positive rate, indicating that its multi-protease protein inference results are more accurate. To investigate the results further, the percent of protein groups containing a single protein and the

average number of proteins per group were determined for each protein inference algorithm (results are summarized in **Table 2.6**). Multi-protease protein inference with MetaMorpheus results in an 8.6% increase in the percent of protein groups that contain a single protein (58.3% to 63.3%) and a 10.5% decrease in the average number of proteins per group (1.81 to 1.62). This data further supports the premise that MetaMorpheus provides more accurate and less ambiguous multi-protease protein inference results compared to ProteinProphet.

Table 2.5: Comparison of Entrapment Results from ProteinProphet and MetaMorpheus

	ProteinProphet	MetaMorpheus	percent change (%)
number of human protein groups	5,877	6,214	+5.7
number of <i>A. thaliana</i> protein groups	159	31	-80.5
false positive rate	2.6%	0.5%	-80.8

Table 2.6: Comparison of Protein Group Ambiguity between ProteinProphet and MetaMorpheus

	ProteinProphet	MetaMorpheus	percent change (%)
percent of protein groups containing a single protein accession	58.3%	63.3%	+8.6
average number of protein accessions per protein	1.81	1.62	-10.5

For the DTASelect2 and MetaMorpheus comparison, the ProLuCID search engine was used for peptide identification, due to its compatibility with DTASelect2. To provide a fair comparison of MetaMorpheus' and DTASelect2's protein inference algorithms, only peptides that were identified by both ProLuCID and MetaMorpheus

were used in each protein inference process. The effect of this peptide filtering on the protein inference results can be observed in **Appendix I: Figure 7.6**. The results for the DTASelect2 and MetaMorpheus comparisons are summarized in **Tables 2.7, 2.8** and **Figure 2.6c**. DTASelect2's protein group output had a protein false discovery rate of 0.4%. MetaMorpheus protein group results at this same FDR were compared to those of DTASelect2. Overall, DTASelect2 provided a simpler result than MetaMorpheus. MetaMorpheus' integrated multi-protease results provide 43.6% more human protein group identifications (4,479 to 6,435), a 29.2% decrease in the percent of protein groups that contain a single protein (93.3% to 66.1%) and a 41.1% increase in the average number of proteins per group (1.12 to 1.58). This data indicates that DTASelect2 results not only have fewer protein group identifications but that the protein groups identified also have lower ambiguity. When evaluating the accuracy of the protein inference approaches, MetaMorpheus, despite having the more complex answer, outperforms DTASelect2. In **Figure 2.6c**, the curve representing MetaMorpheus' protein inference results is closer to the upper left-hand corner than the curve representing DTASelect2's results, indicating that MetaMorpheus more accurately differentiates true and false positives. Additionally, MetaMorpheus' integrated multi-protease protein inference algorithm provides a 12.0% decrease in the false positive rate (0.83% to 0.73%) as compared to DTASelect2. These results indicate that MetaMorpheus provides more accurate protein inference results compared to DTASelect2, despite violating Occam's razor.

Table 2.7: Comparison of Entrapment Results from DTASelect2 and MetaMorpheus

	DTASelect2	MetaMorpheus	percent change (%)
number of human protein groups	4,479	6,435	+43.6
number of <i>A. thaliana</i> protein groups	37	47	+27.0
false positive rate	0.83%	0.73%	-12.0

Table 2.8: Comparison of Protein Group Ambiguity between DTASelect2 and MetaMorpheus

	DTASelect2	MetaMorpheus	percent change (%)
percent of protein groups containing a single protein accession	93.3%	66.1%	-29.2
average number of protein accessions per protein	1.12	1.58	+41.1

Post-Translational Modification (PTM) and Localization with Multi-Protease Protein Inference

The GPTMD strategy within MetaMorpheus was used to discover 8,488 localized PTMs of biological origin (see **Appendix I: Table 7.3** for a list of PTMs) in the multi-protease data. Approximately 77% of the modifications identified (6,515 of the 8,488) were unique to a single protease (**Figure 2.7A**). The use of only one protease yielded 3- to 16-fold fewer biologically relevant PTMs, depending on the protease (**Figure 2.7B**), demonstrating the orthogonality of each protease's ability to identify PTMs. **Figure 2.8** shows the PTMs considered as common biological modifications that were observed in the six different proteolytic digestions.

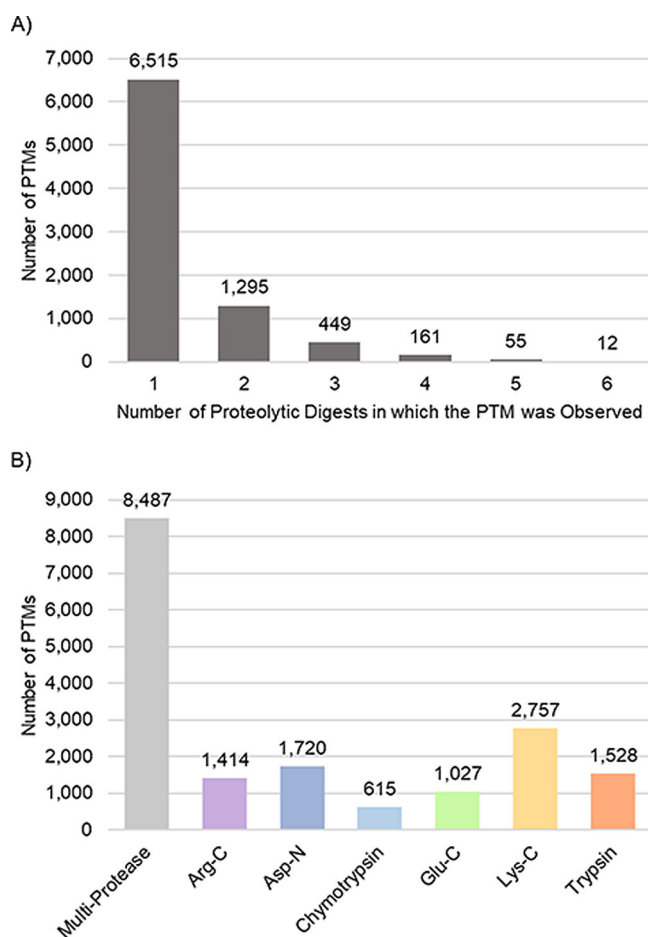


Figure 2.7: Benefits of utilizing multiple proteases for the identification post-translational modifications. Bar graphs detailing PTM results for the integrated multi-protease approach and the single proteolytic digestions. (A) Number of PTMs identified in the integrated multi-protease search observed in just one to all six of the proteolytic digestions. (B) Number of PTMs of biological origin (**Appendix I: Table 7.3**) identified from the integrated multi-protease approach compared to the number of modifications identified by each of the six individual proteases.

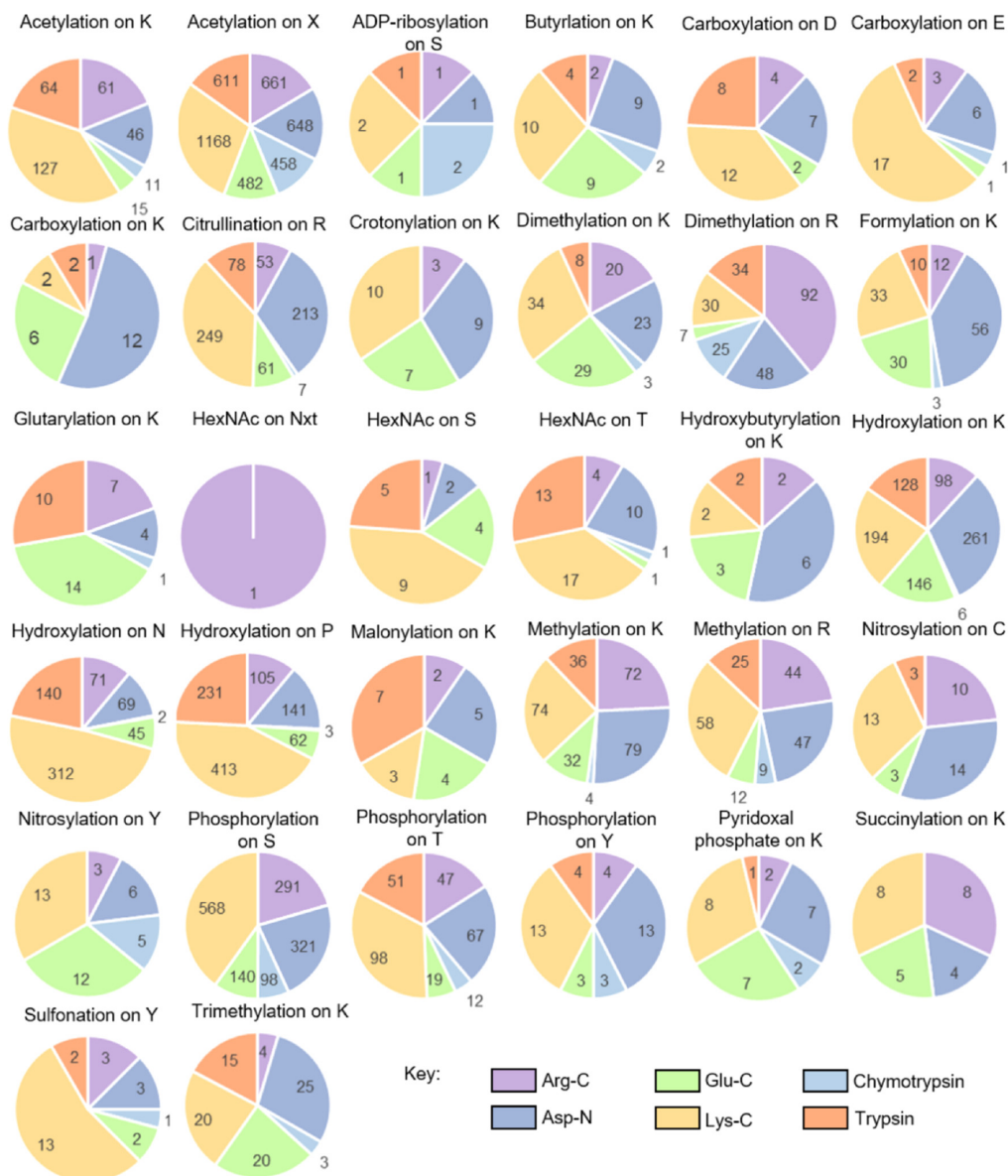


Figure 2.8: Breakdown of GPTMD identified post-translational modifications by protease. Pie charts for each common biological modification searched for with GPTMD showing the number of modifications identified with each protease.

2.5 Conclusion

Protein inference from bottom-up peptide identifications is an intriguing challenge. Protein group identifications that result from protein inference are used to draw biological and chemical conclusions, and therefore their proper identification is of the utmost importance. We present here a novel multi-protease protein inference algorithm implemented in the open-source bottom-up database search software MetaMorpheus.

We show that this integrated multi-protease protein inference algorithm provides simpler and more accurate protein inference results than the separate approach. The increase in simplicity and reduction in the ambiguity of the integrated approach can be seen from the 6.2% decrease in the number of protein group identifications (7,986 to 7,489), a 7.6% increase in the percent of single protein identifications (66% to 71%) and a 6.3% decrease in the average number of proteins per group (1.58 to 1.48). An entrapment study performed to evaluate the accuracy of the integrated approach showed a 31.3% decrease in the number of false positive *A. thaliana* protein group identifications (316 to 217), as well as a 29.3% decrease in the corresponding false positive rate (4.1% to 2.9%) compared to that of the separate approach. A decrease in the ambiguity of protein inference results was also observed when comparing the integrated multi-protease results to those obtained from the widely used trypsin-only digestion strategy. The number of single protein identifications increased 22.4% (58% to 71%), and the average number of proteins per group decreased 4.9% (1.74 to 1.48).

Entrapment studies were employed to benchmark the accuracy of the MetaMorpheus' integrated multi-protease approach against three well-known and well-regarded protein inference tools: Fido, ProteinProphet, and DTASelect2. MetaMorpheus' integrated multi-protease approach reduced false positive *A. thaliana* protein

identifications by 38.3% (162 to 100), and the corresponding false positive rate of the results decreased 39.1% (2.3% to 1.4%) compared to Fido. An increase in accuracy was also observed in MetaMorpheus' comparison with ProteinProphet, which showed an 80.5% decrease in the number of *A. thaliana* protein identifications (159 to 31) and an 80.8% decrease in the corresponding false positive rate (2.6% to 0.5%) of the protein inference results. Compared to DTASelect2, the MetaMorpheus' integrated multi-protease approach reduced the false positive rate of *A. thaliana* protein identifications by 12.0% (0.83% to 0.73%), indicating an increased accuracy. The integrated multi-protease approach in MetaMorpheus thus reduces the percent of erroneous protein group identifications, compared to Fido, ProteinProphet, and DTASelect2 while providing search and protein inference all in one software program.

An additional advantage conferred by the multi-protease digestion compared to conventional single protease digestion was the 3- to 16-fold increase in the identification of PTMs (depending on the single protease selected). The MetaMorpheus implementation of this multi-protease protein inference algorithm facilitates the ability to perform multi-protease search and protein inference all within a single software program, making it more user-friendly than those currently available that require the transfer of results from a search software to a distinct protein inference software.

2.6 References

- (1) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005**, *4*, Type: Journal Article, 1419–40.

- (2) Uszkoreit, J.; Perez-Riverol, Y.; Eggers, B.; Marcus, K.; Eisenacher, M. Protein Inference Using PIA Workflows and PSI Standard File Formats. *J Proteome Res* **2019**, *18*, Type: Journal Article, 741–747.
- (3) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J Proteome Res* **2010**, *9*, Type: Journal Article, 5346–57.
- (4) Huang, T.; Wang, J.; Yu, W.; He, Z. Protein inference: a review. *Brief Bioinform* **2012**, *13*, Type: Journal Article, 586–614.
- (5) Zhang, Y.; Xu, T.; Shan, B.; Hart, J.; Aslanian, A.; Han, X.; Zong, N.; Li, H.; Choi, H.; Wang, D.; Acharya, L.; Du, L.; Vogt, P. K.; Ping, P.; Yates J. R., 3. ProteinInferencer: Confident protein identification and multiple experiment comparison for large scale proteomics projects. *J Proteomics* **2015**, *129*, Type: Journal Article, 25–32.
- (6) Edwards, N.; Wu, X.; Tseng, C. An Unsupervised, Model-Free, Machine-Learning Combiner for Peptide Identifications from Tandem Mass Spectra. *Clinical Proteomics* **2009**, *5*, Type: Journal Article, DOI: 10.1007/s12014-009-9024-5.
- (7) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* **2011**, *10*, Type: Journal Article, M111 007690.
- (8) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* **2008**, *7*, Type: Journal Article, 245–53.

- (9) Tabb, D. L.; McDonald, W. H.; Yates J. R., 3. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002**, *1*, Type: Journal Article, 21–6.
- (10) 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*, Type: Journal Article, 1323–9.
- (11) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol Cell Proteomics* **2014**, *13*, Type: Journal Article, 1573–84.
- (12) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003**, *75*, Type: Journal Article, 4646–58.
- (13) Feng, X. D.; Li, L. W.; Zhang, J. H.; Zhu, Y. P.; Chang, C.; Shu, K. X.; Ma, J. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics* **2017**, *18*, Type: Journal Article, 143.
- (14) 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*, Type: Journal Article, 2341–53.
- (15) 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*, Type: Journal Article, 228–40.

- (16) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **2009**, *6*, Type: Journal Article, 359–62.
- (17) 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*, Type: Journal Article, 1383–1390.
- (18) 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*, Type: Journal Article, 1844–1851.
- (19) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, Type: Journal Article, 22–4.
- (20) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002**, *74*, Type: Journal Article, 5383–92.
- (21) Xu, T. et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J Proteomics* **2015**, *129*, Type: Journal Article, 16–24.
- (22) Tsiatsiani, L.; Heck, A. J. Proteomics beyond trypsin. *FEBS J* **2015**, *282*, Type: Journal Article, 2612–26.
- (23) Storey, J. D.; Akey, J. M.; Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **2005**, *3*, Type: Journal Article, e267.

3 PROTEASEGURU: A TOOL FOR PROTEASE SELECTION IN BOTTOM-UP PROTEOMICS

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Ibrahim, K.; Smith, L. M. ProteaseGuru: A Tool for Protease Selection in Bottom-Up Proteomics. *Journal of Proteome Research* **2021**, *20*(4), 1936–1942. <https://doi.org/10.1021/acs.jproteome.0c00954>.

Copyright © 2021 American Chemical Society.

3.1 Abstract

Bottom-up proteomics is currently the dominant strategy for proteome analysis. It relies critically upon the use of a protease to digest proteins into peptides, which are then identified by liquid chromatography–mass spectrometry (LC-MS). The choice of protease(s) has a substantial impact upon the utility of the bottom-up results obtained. Protease selection determines the nature of the peptides produced, which in turn affects the ability to infer the presence and quantity of the parent proteins and post-translational modifications in the sample. We present here the software tool ProteaseGuru, which provides *in silico* digestions by candidate proteases, allowing evaluation of their utility for bottom-up proteomic experiments. This information is useful for both studies focused on a single or small number of proteins, and for analysis of entire complex proteomes. ProteaseGuru provides a convenient user interface, valuable peptide information, and data visualizations enabling the comparison of digestion results of different proteases. The information provided includes data tables of theoretical peptide sequences and their biophysical properties, results summaries outlining the numbers of shared and unique peptides per protease, histograms facilitating the comparison of proteome-wide proteolytic data, protein-specific summaries, and sequence coverage maps. Examples are provided of its use to inform analysis of variant-containing proteins in the human proteome, as well as for studies requiring the use of multiple proteomic databases such as a human:mouse xenograft model, and microbiome metaproteomics.

3.2 Introduction

Bottom-up proteomics is the principal approach employed for the analysis of complex proteomes. In bottom-up proteomics, proteins are digested into peptides

prior to chromatographic separation and tandem mass spectrometric analysis.¹ Their identification and quantification aids in inference of the proteins present in the sample and provides valuable information on their abundances.¹ Bottom-up proteomics has evolved into a widespread and high-throughput approach providing sensitive and in-depth characterization of thousands of proteins in complex proteomes.²

This peptide-centric approach is entirely reliant on proteases, and their ability to generate predictable proteolytic peptides that span the proteome and are detectable by the mass spectrometer. Often, a single protease, trypsin, is used for digestion. Trypsin is robust, reproducible, and its cleavage motif at the carboxyl side of the amino acids lysine or arginine generates peptides that ionize well.^{3,4} In some cases, a protease other than trypsin can yield improved results identifying more critical peptides for the identification of select proteins, PTMs, or sequence variations of interest.^{3,5} Furthermore, we and others have shown that the use of multiple proteases in parallel produces superior results, increasing the number of proteins and post-translational modifications identified through increased proteome coverage.⁶⁻⁹

However, it is often not straightforward to determine which protease or combination of proteases is best suited for a given experiment. Due to cost, time, and sample limitations, it is frequently infeasible to employ a trial and error approach, digesting samples with all commonly used proteases to determine which worked the best. Selection of a protease or combination of proteases for sample digestion relies on the ability to determine which proteolytic digestions will produce peptides that are the most likely to be observed via mass spectrometry (based on their biophysical properties such as length and hydrophobicity), provide adequate protein sequence coverage, and generate sufficient numbers of unique peptides to identify specific proteins, or a large portion of the proteome. The ability to identify unique peptides is always important in bottom-up proteomics, but becomes even more critical

when samples include proteins from multiple species, as is the case for xenograft or microbiome samples.¹⁰⁻¹² Peptides can not only be ambiguous between proteins within a species, but also between proteins from different species, compromising the ability to draw biological conclusions from the proteomic results. This creates a need for experimental planning, in which theoretical peptides produced by potential proteolytic digests are generated and the proteases can be compared for their efficacy prior to initiating laboratory work.

We have developed a free and open-source software tool, ProteaseGuru, to enable the comparison of candidate proteases through *in silico* digestion of protein databases. We designed ProteaseGuru with the goal of making it the easiest to use and most versatile *in silico* digestion tool to date. Users can select as many proteases as desired to digest the elements of one or more protein databases generating a pool of theoretical peptide sequences. After *in silico* digestion, ProteaseGuru determines several biophysical characteristics of the theoretical peptide sequences which can help to assess their uniqueness and utility for bottom-up proteomic analysis. Digestion result summaries are provided for each *in silico* digested database, giving the number of shared and unique peptides. When more than one database is utilized, as for the xenograft and microbiome applications mentioned above, an additional analysis is performed to determine which peptides are unique to a single protein and which are distinct to a single species. Such peptides are valuable for the identification and quantification of select proteins in complex proteomic backgrounds. ProteaseGuru provides graphical visualizations, such as histograms and protein sequence coverage maps, that aid the user in evaluation of candidate proteolytic digestions of either select proteins or on a whole proteome level. Specific examples demonstrating ProteaseGuru's utility are shown for different experiment types, including proteogenomics, xenograft analysis, and microbiome metaproteomics.

3.3 Methods

The Tool

ProteaseGuru is a windows GUI application written in C# for the in silico digestion of protein databases. ProteaseGuru includes both MzLib (v1.0.485), a mass spectrometry code library (<https://github.com/smith-chem-wisc/mzLib>), and OxyPlot (v2.0.0), for data visualization, as Nuget packages. The application and its source code are available for download on GitHub (<https://github.com/smith-chem-wisc/ProteaseGuru>). The prediction of both a peptide sequence's hydrophobicity and electrophoretic mobility are incorporated into ProteaseGuru. The hydrophobicity of unmodified peptide sequences is predicted using the SSRCalc algorithm described by Krokhin et al.¹³ and electrophoretic mobility of peptide sequences, including PTMs, is calculated based on a modified Cifuentes's model¹⁴ as described in Chen et al.¹⁵

ProteaseGuru accepts, as input, UniProt formatted XML and FASTA databases. Post-translational modifications annotated in the UniProt XML database are loaded into ProteaseGuru, displayed within protein sequence coverage maps, annotated in the full sequence of theoretical peptides, and contribute toward the total molecular weight of the theoretical peptide. Additionally, users can choose, as part of the digestion parameters, to include carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification.

Data Analysis

The utility of ProteaseGuru was evaluated for three different applications: (1) human:mouse xenografts, (2) identification of sequence variant-containing proteins, and (3) a subset of the human skin microbiome. Analysis was performed using ProteaseGuru version 0.0.22 with the following digestion conditions: proteases =

[Arg-C, Asp-N, chymotrypsin (do not cleave before proline), Glu-C (with asp), Lys-C (do not cleave before proline), and trypsin (do not cleave before proline)]; max number of missed cleavages = 2; min peptide length = 7; and max peptide length = 50; and treat modified peptides as different = False.

For the xenograft application, human and mouse reference databases were downloaded from UniProt in .xml format. Only reviewed Swiss-Prot entries were included in the databases.

For the variant analysis, a proteogenomic database generated by Spritz¹⁶ (version 0.1.3) was utilized. The RNA-Seq data used as input for Spritz is publicly available and can be downloaded from the GEO Sequence Read Archives with the following identifier GSE45428.

For the skin microbiome analysis, a subset of the entire microbiome was analyzed. In a review by Byrd et al. concerning the human skin microbiome, a table was provided outlining the top 10 most abundant bacterial, eukaryotic, and viral species present in four different physiological sites (dry skin, moist skin, sebaceous skin, and foot skin).¹⁷ With duplicate species removed, a total of 59 remained. Of those 59 species, 57 are present on UniProt and the corresponding protein databases in .fasta format were downloaded (see **Appendix II: Table 8.1** for the specific species included, and download information).

3.4 Results and Discussion

The utility of ProteaseGuru as an experimental planning and protease comparison tool will be demonstrated through three different case studies, representative of three distinct bottom-up proteomic applications: (1) proteomics on xenograft samples, (2) variant proteomics, and (3) microbiome analyses. We will also evaluate the relative

ease of use and versatility of ProteaseGuru by benchmarking its features against those of existing tools.

Analysis of Patient-Derived Xenografts

Proteomic samples are sometimes more complex than a single species' proteome. Patient-derived xenografts (PDXs) are human tumor samples that have been transplanted into an immune-compromised, or humanized mouse. PDXs are a widely used model system for the study of cancer.¹⁸⁻²¹ ProteaseGuru is applied here for PDX proteomics, performing *in silico* digestion and analysis of both the human and mouse UniProt databases to guide experimental design.

As part of its postdigestion processing, ProteaseGuru determines a peptide's "uniqueness" for three different categories: (1) "Unique in database", a peptide is unique if it is the proteolytic product of a single protein within a database; (2) "Unique in all databases", a peptide is unique if it is the digestion product of a single protein within all of the databases analyzed; and (3) "Exclusive to this database", a peptide's sequence (regardless of its shared or unique peptide status) is only found in one protein database. This categorization enables the identification of theoretical peptide sequences that can distinguish proteins and species in complex mixtures. For all uniqueness categorizations, isoleucine and leucine are treated as distinct amino acids. All three "uniqueness" values are displayed in the ProteaseGuru peptide output files, and are included in result summaries, histograms, and sequence coverage maps. This feature of ProteaseGuru is critical for the combined analysis of the human and mouse databases since their proteomes have high sequence homology. Once the proteomes are digested, it can be difficult to determine which peptides belong to the human tumor, and which belong to the mouse. The ability to distinguish human and mouse proteins is critical to the success of many PDX studies, and their ability to inform

future functional or clinical research.

The extent to which homology between the two species complicates proteomic analysis was evaluated using the count of shared and unique peptides for the combined and individual database analyses provided in the ProteaseGuru summary file (**Figure 3.1**). The average percent unique peptide sequences for all of the *in silico* digestions is 97.6% for human and 98.5% for mouse (category 1). If there was no sequence homology between the two species, all peptides that were unique in the separate human and mouse analyses would remain unique peptides when the two proteomes are analyzed together, yielding a percent unique peptide value of approximately 98.01%. However, it is well documented that there is homology between the human and mouse proteomes with the average degree of protein sequence conservation for orthologous human and mouse genes being approximately 85%.²² When comparing the combined theoretical peptides from the human and mouse proteomes, the percent unique peptide sequences (category 2) observed was 81.5%, indicating the high homology of the human and mouse proteomes has a strong impact on the ability to identify peptides unique to a single protein.

Analysis of Sequence Variant-Containing Proteins

Proteomic experiments can be focused on the entire proteome, or can be focused on capturing a particular class of proteins, a specific protein, or a specific post-translational modification. ProteaseGuru allows selection of the protease or combination of proteases that will be most effective in achieving the goal of such proteomic experiments. Here we demonstrate this functionality by applying ProteaseGuru to a proteogenomic database generated by the software tool Spritz.¹⁶ Spritz utilizes RNA-sequencing data and a reference genome to generate a protein XML database containing sequence variations present in the sample's transcriptome.

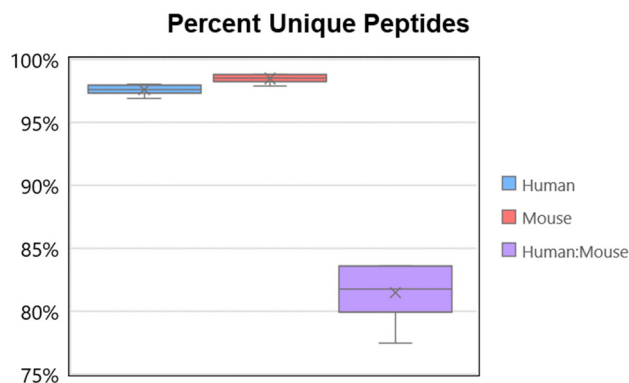


Figure 3.1: Comparison of percent unique peptide sequences for mouse and human databases. Box plot generated from the percent unique peptide sequences for all 6 proteolytic digests (Arg-C, Asp-N, Chym, Glu-C, Lys-C, and Tryp) when the human and mouse databases are analyzed either separately (category 1) or together (category 2). The high sequence homology between the human and mouse proteomes creates a significant decrease in the percent of unique peptides, and a corresponding increase in the percent of shared peptides.

Proteins translated from variant transcripts may have zero, minor or very substantive amino acid sequence differences, depending on the nature of the nucleic acid variation(s) present. A proteogenomic database, generated from these transcripts, will include greater proteomic complexity than the standard UniProt database because the translation products derived from both alleles are represented. These homologous alleles, producing related transcripts, will give rise to translation products which also have high homology, and accordingly a greater prevalence of peptides are shared between those homologous proteins (**Figure 3.2**). The average increase in the percent of shared peptide sequences across proteases is 61.8% when comparing Spritz and UniProt database results. The dramatic increase in the percent of shared peptide sequences underscores the importance of identifying protease(s) capable of producing unique peptides for the confident identification of variant-containing proteins, as well as the importance of utilizing proteogenomic databases in general.

Using the peptide output files from ProteaseGuru, the following information

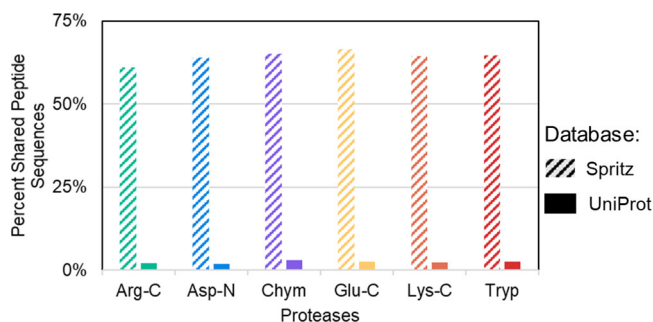


Figure 3.2: Comparison of the percent of shared peptide sequences for each protease between the Spritz proteogenomic database and the reference UniProt database.

is readily determined for each proteolytic digestion of the Spritz database: (a) the number of unique peptides for variant proteins (category 1), (b) the number of variant proteins which have unique peptide evidence, and (c) the number of variant proteins that can only be confidently identified by theoretical peptides from this digest (see **Table 3.1**). On the basis of these results, it is clear certain proteolytic digests have the capability of producing more variant protein identifications (e.g., Trypsin, Chymotrypsin and Glu-C), and in order to maximize the number of variant proteins identified, a combination of proteases must be used. The maximum number of variant proteins (5,355) can only be achieved when all proteolytic digests are performed since there are variant protein identifications unique to each digest. However, it is not always feasible to perform that many parallel digests, and it is prudent to determine, based on the number of digestions to be performed, the combination of proteases that captures the largest population of variant proteins. In **Figure 3.3**, the number of variant proteins that can be identified via a unique peptide sequence are determined for all individual proteases and all combinations of proteases.

ProteaseGuru also enables the investigation of individual proteins through the generation of protein-specific digestion result summaries and protein sequence coverage maps. Sequence coverage maps enable the visualization of theoretical peptide

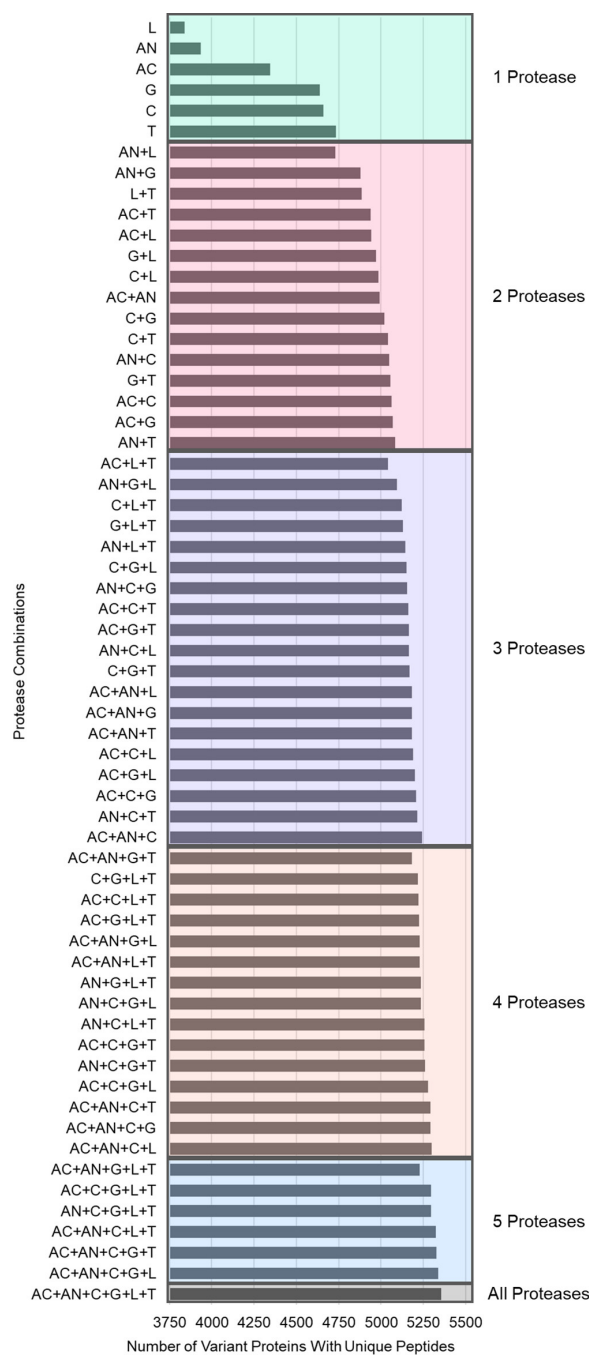


Figure 3.3: Number of variant proteins that can be identified by unique peptides. Plot for each protease and combination of proteases showing the number of variant proteins that can be confidently identified by unique peptides. This shows which protease or combination of proteases provides the best coverage of variant proteins within the proteome. (Arg-C: AC, Asp-N: AN, Chym: C, Glu-C: G, Lys-C: L, and Tryp: T).

Table 3.1: Variant Protein Results

protease	number of unique peptides for variant proteins (percent of total unique peptides)	number of variant protein with unique peptides	number of variant proteins with unique peptides exclusive to a protease
Arg-C	52,453 (9.5%)	4,346	46
Asp-N	39,981 (9.9%)	3,934	58
Chymotrypsin	127,011 (9.7%)	4,660	84
Glu-C	101,760 (10.3%)	4,639	38
Lys-C	45,753 (10.0%)	3,840	25
Trypsin	95,781 (9.9%)	4,733	15

coverage, for all proteolytic digests, for a given protein, and for its database-annotated PTMs and variants. This feature is valuable for more focused experiments because it allows the user to visualize which protease provides optimal coverage of proteins of interest, and which protease(s) can produce peptides that cross PTMs or variant sites. The sequence coverage map of UniProt protein H3BQZ5, with a single amino acid variant at residue 25 from cysteine to arginine, is shown in **Figure 3.4**. This coverage map highlights that only one of the six proteases evaluated, Arg-C, produces theoretical peptide sequences unique to this protein, and only one of those sequences crosses the variant site. This variant-crossing peptide is particularly valuable in that it confirms the presence of the variant.

Analysis of Skin Microbiome

Metaproteomics encompasses the study of incredibly complex and diverse multispecies proteomes such as those for microbial communities and microbiomes.²³ ProteaseGuru is able to perform *in silico* digestions on more than 2 proteomes at once, a functionality absent from existing *in silico* digestion tools. It is important to note the computational requirements for these analyses scales with the number and size of



Figure 3.4: Sequence coverage map of variant containing protein (H3BQZ5_C25R) exported from ProteaseGuru. Theoretical peptide sequences are mapped to the protein highlighting its coverage by shared and unique peptides for all proteolytic digests. Unique peptide sequences are bold colored, where shared peptide sequences are translucent. Peptides are ordered by their starting residue. Since peptides with up to two missed cleavages are allowed, multiple peptides from the same protease can overlap but will either start or end at different residues. Multiple amino acid gaps between peptide lines correspond to regions of the proteome that are not covered by any peptide sequence due to the constraints placed on acceptable peptide length, and number of missed cleavages. For peptides that span more than one row, the line extends beyond the margin before wrapping around to the next row down. Peptide sequences unique to this specific variant protein were only obtained in the Arg-C digest, and only a single theoretical unique Arg-C peptide crosses the variant site (the upper of the two bold lines).

the proteomes being analyzed. Shown here is the use of ProteaseGuru on 57 protein databases which compose a subset of the human skin microbiome, as described in **Section 3.3**.

ProteaseGuru generates various histograms within the graphical user interface (GUI) to enable the comparison of proteolytic digests. These histograms and the data tables used to generate the histograms can be exported. **Figure 3.5** shows a “Percent Protein Sequence Coverage” histogram generated by Excel for the microbiome analysis using the exported data table from ProteaseGuru. Often, peptides with fewer than seven amino acids are difficult to confidently identify via mass spectrometry.^{3,7} Therefore, setting a minimum peptide length of seven for in silico digestion enables the generation of theoretical peptides that, based on length, are likely to be identified. Specifying peptide length digestion criteria will result in regions of proteins without theoretical peptide sequence coverage, approximating a lack in identifiable coverage in actual digestion results. It is desirable to select proteases that provide the greatest proteome coverage overall, and on a protein by protein basis. As may be seen in **Figure 3.5**, the ProteaseGuru results show in silico digestion with Trypsin, Chymotrypsin and Glu-C produce peptide sequences that would provide the most comprehensive coverage of the proteome, whereas Lys-C digestion would provide the least proteome coverage.

Comparison of ProteaseGuru to Existing Tools

ProteaseGuru includes numerous features providing versatility for a wide-range of bottom-up proteomic experiments. A comparison of features between ProteaseGuru and other existing in silico digestion tools is provided in **Table 3.2**. iHDPM²⁴ is a great tool, with many wonderful visualization features, but lacks customizability. iHDPM is limited to analysis of the human proteome, with a predetermined set of

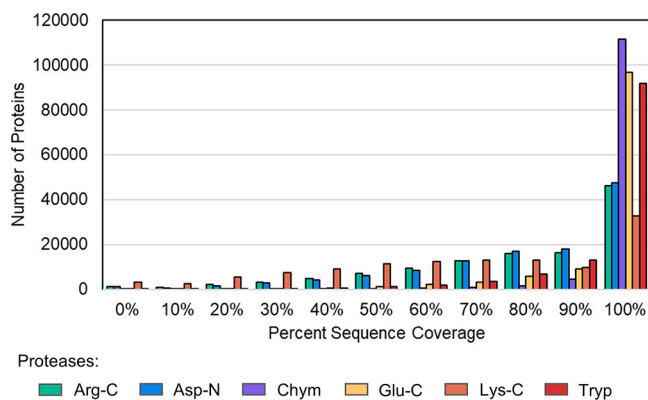


Figure 3.5: Histogram comparing the distribution of percent protein sequence coverage for the skin microbiome based on the protease used for in silico digestion.

proteolytic digests. In contrast, ProteaseGuru allows the user to supply as many of their own protein databases as necessary, providing the user with more control over their analysis. ProteaseGuru is one of two tools that permits the analysis of more than one database at a time, and is the only tool which allows for the analysis of more than two databases. ProteaseGuru also does not limit the user to digestion with the default proteases provided, only two other in silico digestion tools offer that same level of flexibility. ProteaseGuru makes the process of custom protease generation easy by allowing the user to add a custom protease within the GUI—simply requiring a protease name and cleavage motif. ProteaseGuru is also one of three tools that provides result visualizations. In silico digestion results can be visualized as histograms and protein sequence coverage maps. Both histograms and sequence coverage maps can be exported for publication. An additional feature unique to ProteaseGuru is the ability to export the data tables underlying each histogram which facilitates easy recreation of the plots in the user’s software of choice. ProteaseGuru provides a combination of features and a level of user-friendliness that provides an increased degree of versatility compared to existing in silico digestion tools.

Table 3.2: Comparison of In Silico Digestion Tool Features

tool name	digests whole proteome	user supplied database(s)	can digest multiple databases	runs parallel protease digestions	enables custom protease generation	determines uniqueness of peptides	provides data visualization	includes PTM annotations	cross-platform web interface	provides theoretical peptide fragmentation data
ProteaseGuru	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No (C#)	No
iHDPM ²⁴	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	No
IPEP ²⁵	No	No	No	Yes	No	No	Yes	No	Yes	No
MS-Digest	No	No	No	No	No	No	No	Yes	Yes	Yes
pepServe ²⁶	No	No	No	No	No	Yes	No	Yes	Yes	No
PeptideCutter ²⁷	No	No	No	No	No	No	No	No	Yes	No
PeptideManager ¹⁰	Yes	Yes	Yes (max:2)	No	No	Yes	No	Yes	No (C#)	No
PeptideMass ²⁸	No	No	No	No	No	No	No	Yes	Yes	No
Protein Digest	No	No	No	No	No	No	No	No	Yes	No
Proteogest ²⁹	Yes	Yes	No	No	Yes	Yes	No	Yes	No (Perl)	No
Rapid Peptides Generator ³⁰	Yes	Yes	No	Yes	Yes	No	No	No	No (Python)	No

3.5 Conclusion

ProteaseGuru is a software tool designed to aid in the selection of proteases for bottom-up proteomic experiments. The in silico digestion and subsequent analyses performed within ProteaseGuru provide result files and data visualizations that empower users to make informed choices on which proteases to select for bottom-up proteomic experiments. This eliminates the need for a trial and error approach, which is costly with respect to time, samples, and money. ProteaseGuru is the most broadly applicable in silico digestion software to date, enabling its use for proteomics experiments focused on PTM or variant identification, as well as for proteome-wide experiments analyzing samples composed of one or more species. ProteaseGuru provides not only the peptide sequences that result from in silico digestions, but also a wide variety of information about each peptide to enable customized analyses based on the users' needs, such as peptide's modification status, length, protein of origin, position within the protein of origin, hydrophobicity, electrophoretic mobility, and uniqueness. ProteaseGuru also generates several histograms to aid in the comparison of proteolytic digests, as well as providing the ability to investigate the in silico digestion of specific proteins. ProteaseGuru provides numerous features, along with a user-friendly experience to facilitate experimental planning for a wide-variety of bottom-up proteomic experiments.

3.6 References

- (1) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates J. R., 3. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* **2013**, *113*, Type: Journal Article, 2343–94.

- (2) 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 (Palo Alto Calif)* **2016**, *9*, Type: Journal Article, 449–72.
- (3) Tsiatsiani, L.; Heck, A. J. Proteomics beyond trypsin. *FEBS J* **2015**, *282*, Type: Journal Article, 2612–26.
- (4) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal Chem* **2005**, *77*, Type: Journal Article, 5800–13.
- (5) Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc* **2016**, *11*, Type: Journal Article, 993–1006.
- (6) 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*, Type: Journal Article, 3429–3438.
- (7) 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*, Type: Journal Article, 1323–9.
- (8) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. Confetti: a multiprotease map of the HeLa proteome for comprehensive proteomics. *Mol Cell Proteomics* **2014**, *13*, Type: Journal Article, 1573–84.
- (9) Biringer, R. G.; Amato, H.; Harrington, M. G.; Fonteh, A. N.; Riggins, J. N.; Huhmer, A. F. Enhanced sequence coverage of proteins in human cerebrospinal

- fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS. *Brief Funct Genomic Proteomic* **2006**, *5*, Type: Journal Article, 144–53.
- (10) Demeure, K.; Duriez, E.; Domon, B.; Niclou, S. P. PeptideManager: a peptide selection tool for targeted proteomic studies involving mixed samples from different species. *Front Genet* **2014**, *5*, Type: Journal Article, 305.
- (11) Saltzman, A. B.; Leng, M.; Bhatt, B.; Singh, P.; Chan, D. W.; Dobrolecki, L.; Chandrasekaran, H.; Choi, J. M.; Jain, A.; Jung, S. Y.; Lewis, M. T.; Ellis, M. J.; Malovannaya, A. gpGrouper: A Peptide Grouping Algorithm for Gene-Centric Inference and Quantitation of Bottom-Up Proteomics Data. *Mol Cell Proteomics* **2018**, *17*, Type: Journal Article, 2270–2283.
- (12) Heyer, R.; Schallert, K.; Budel, A.; Zoun, R.; Dorl, S.; Behne, A.; Kohrs, F.; Puttker, S.; Siewert, C.; Muth, T.; Saake, G.; Reichl, U.; Benndorf, D. A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 h Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer. *Front Microbiol* **2019**, *10*, Type: Journal Article, 1883.
- (13) Krokhin, O. V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-Å pore size C18 sorbents. *Anal Chem* **2006**, *78*, Type: Journal Article, 7785–95.
- (14) Cifuentes, A.; Poppe, H. Simulation and optimization of peptide separation by capillary electrophoresis. *J Chromatogr A* **1994**, *680*, Type: Journal Article, 321–40.
- (15) Chen, D.; Lubeckyj, R. A.; Yang, Z.; McCool, E. N.; Shen, X.; Wang, Q.; Xu, T.; Sun, L. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. *Anal Chem* **2020**, *92*, Type: Journal Article, 3503–3507.

- (16) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *J Proteome Res* **2021**, *20*, Type: Journal Article, 1826–1834.
- (17) Byrd, A. L.; Belkaid, Y.; Segre, J. A. The human skin microbiome. *Nat Rev Microbiol* **2018**, *16*, Type: Journal Article, 143–155.
- (18) Tentler, J. J.; Tan, A. C.; Weekes, C. D.; Jimeno, A.; Leong, S.; Pitts, T. M.; Arcaroli, J. J.; Messersmith, W. A.; Eckhardt, S. G. Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol* **2012**, *9*, Type: Journal Article, 338–50.
- (19) Hidalgo, M.; Amant, F.; Biankin, A. V.; Budinska, E.; Byrne, A. T.; Caldas, C.; Clarke, R. B.; de Jong, S.; Jonkers, J.; Maelandsmo, G. M.; Roman-Roman, S.; Seoane, J.; Trusolino, L.; Villanueva, A. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov* **2014**, *4*, Type: Journal Article, 998–1013.
- (20) Ntai, I. et al. Integrated Bottom-Up and Top-Down Proteomics of Patient-Derived Breast Tumor Xenografts. *Mol Cell Proteomics* **2016**, *15*, Type: Journal Article, 45–56.
- (21) Bhimani, J.; Ball, K.; Stebbing, J. Patient-derived xenograft models-the future of personalised cancer treatment. *Br J Cancer* **2020**, *122*, Type: Journal Article, 601–602.
- (22) Makalowski, W.; Zhang, J.; Boguski, M. S. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res* **1996**, *6*, Type: Journal Article, 846–57.

- (23) Heyer, R.; Schallert, K.; Zoun, R.; Becher, B.; Saake, G.; Benndorf, D. Challenges and perspectives of metaproteomic data analysis. *J Biotechnol* **2017**, *261*, Type: Journal Article, 24–36.
- (24) Choong, W. K.; Chen, C. T.; Wang, J. H.; Sung, T. Y. iHPDM: In Silico Human Proteome Digestion Map with Proteolytic Peptide Analysis and Graphical Visualizations. *J Proteome Res* **2019**, *18*, Type: Journal Article, 4124–4132.
- (25) Lu, D.; Liu, R. Z.; Izumi, V.; Fenstermacher, D.; Haura, E. B.; Koomen, J.; Eschrich, S. A. IPEP: an in silico tool to examine proteolytic peptides for mass spectrometry. *Bioinformatics* **2008**, *24*, Type: Journal Article, 2801–2.
- (26) Alexandridou, A.; Dovrolis, N.; Tsangaris, G. T.; Nikita, K.; Spyrou, G. PepServe: a web server for peptide analysis, clustering and visualization. *Nucleic Acids Res* **2011**, *39*, Type: Journal Article, W381–4.
- (27) Wilkins, M. R.; Gasteiger, E.; Bairoch, A.; Sanchez, J. C.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* **1999**, *112*, Type: Journal Article, 531–52.
- (28) Wilkins, M. R.; Lindskog, I.; Gasteiger, E.; Bairoch, A.; Sanchez, J. C.; Hochstrasser, D. F.; Appel, R. D. Detailed peptide characterization using PEPTIDEMASS—a World-Wide-Web-accessible tool. *Electrophoresis* **1997**, *18*, Type: Journal Article, 403–8.
- (29) Cagney, G.; Amiri, S.; Premawaradena, T.; Lindo, M.; Emili, A. In silico proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci* **2003**, *1*, Type: Journal Article, 5.
- (30) Maillet, N. Rapid Peptides Generator: fast and efficient in silico protein digestion. *NAR Genom Bioinform* **2020**, *2*, Type: Journal Article, lqz004.

4 ENHANCED PROTEIN ISOFORM CHARACTERIZATION THROUGH LONG-READ PROTEOGENOMICS

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Jordan, B.T.; Mehlferber, M.M.; Jeffery, E.D.; Chatzipantsiou, C.; Kaur, S. Millikin, R.J.; Dai, Y.; Tiberi, S.; Castaldi, P.J.; Shortreed, M.R.; Luckey, C.J.; Conesa, A.; Smith, L.M.; Deslattes-Mays, A.; Sheynkman, G.M. Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biology* **2022**, *23*(69). <https://doi.org/10.1186/s13059-022-02624-y>.

Copyright © 2022 Springer Nature.

4.1 Abstract

Background

The detection of physiologically relevant protein isoforms encoded by the human genome is critical to biomedicine. Mass spectrometry (MS)-based proteomics is the preeminent method for protein detection, but isoform-resolved proteomic analysis relies on accurate reference databases that match the sample; neither a subset nor a superset database is ideal. Long-read RNA sequencing (e.g., PacBio or Oxford Nanopore) provides full-length transcripts which can be used to predict full-length protein isoforms.

Results

We describe here a long-read proteogenomics approach for integrating sample-matched long-read RNA-seq and MS-based proteomics data to enhance isoform characterization. We introduce a classification scheme for protein isoforms, discover novel protein isoforms, and present the first protein inference algorithm for the direct incorporation of long-read transcriptome data to enable detection of protein isoforms previously intractable to MS-based detection. We have released an open-source Nextflow pipeline that integrates long-read sequencing in a proteomic workflow for isoform-resolved analysis.

Conclusions

Our work suggests that the incorporation of long-read sequencing and proteomic data can facilitate improved characterization of human protein isoform diversity. Our first-generation pipeline provides a strong foundation for future development of

long-read proteogenomics and its adoption for both basic and translational research.

4.2 Background

A comprehensive understanding of the proteome in healthy and diseased states is vital for nearly every area of biomedical research.¹ Multiple protein isoforms, containing distinct amino acid (AA) sequences, can arise from the same gene through mechanisms such as alternative promoter usage or splicing² and can exhibit different stabilities, molecular binding capabilities, and functional effects^{3,4}. Many protein isoforms have been implicated in diseases from neurodegeneration to cancer.⁵ It has been estimated, through transcriptome measurements, that over 300,000 human protein isoforms may exist.⁶ However, few experimental approaches readily detect proteins at isoform resolution, leaving open the question of the extent to which transcript isoform complexity propagates to the proteome.^{7,8}

Mass spectrometry (MS)-based proteomics has become the preeminent method for the comprehensive and sensitive characterization of the proteome.¹ Typically, the proteome is proteolytically digested into peptides that are analyzed via liquid chromatography (LC) and MS. The mass spectra are compared to theoretical peptides, generated from a protein database, to obtain peptide identifications. These peptide identifications are mapped back to their potential proteins of origin to obtain protein identifications (i.e., protein inference).⁹ Protein inference is complicated by shared peptides, which are peptides that map to two or more protein isoforms in the database. The presence of shared peptides can result in ambiguous protein identifications wherein multiple proteins are indistinguishable based on the peptide evidence. In these cases, a “protein group” (**Figure 4.1a**) is formed, signifying either all or some subset of proteins in the group may be present in the sample.

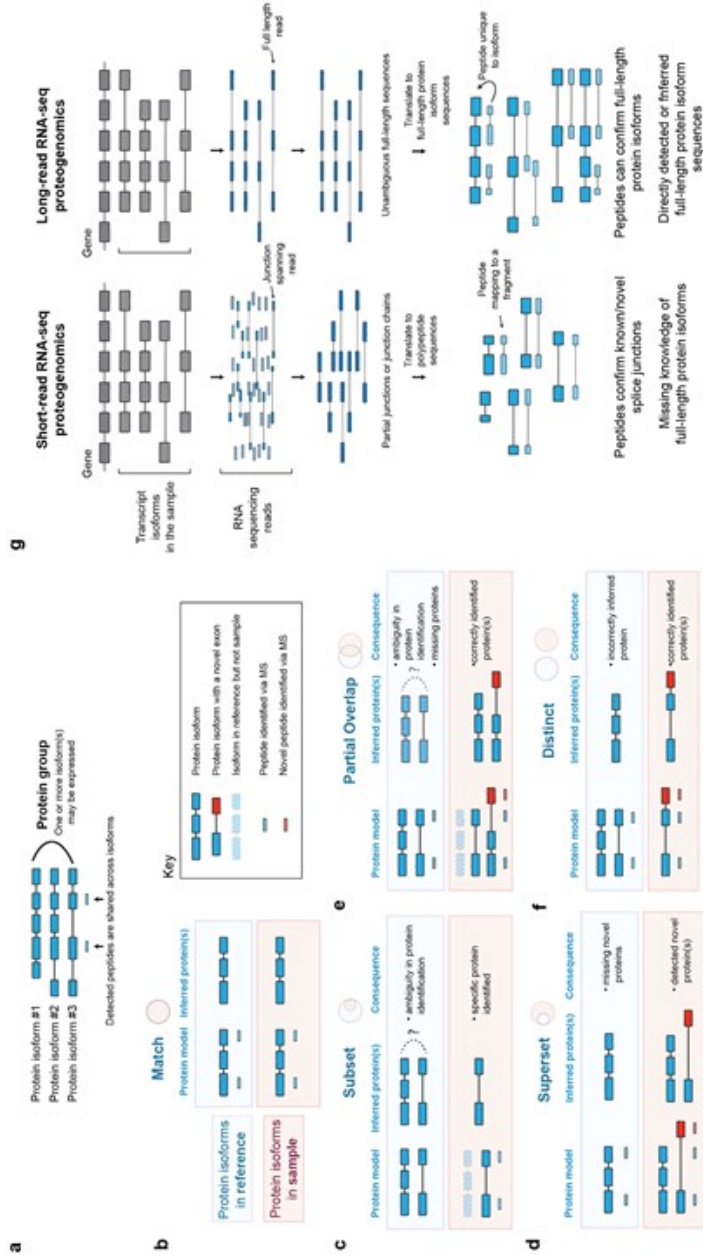


Figure 4.1: Challenges of protein isoform identification using MS-based proteomics. **a**, Many peptides detected in MS-based proteomics map to multiple protein isoforms. Indistinguishable protein isoforms are represented as protein groups. This ambiguity limits the utility of MS-based proteomics for isoform detection. **b**, The assumption of MS-based proteomics search algorithms is that the reference and sample isoforms match. Isoforms in the light blue boxes represent those annotated in a reference database. Isoforms in the light pink boxes represent isoforms that are actually expressed in a sample (which is unknowable using current technologies). When the reference and sample isoform are concordant (“Match”), protein identification can be accurate. **c-f**, Reference-sample discordances can result in inferred proteins that are ambiguous or incorrect. Schematic of a case in which a sample contains a subset of isoforms in the reference database (“Subset”, **c**), additional isoforms (i.e., novel) not found in the reference database (“Superset”, **d**), a subset of isoforms but also additional novel isoforms (“Partial Overlap”, **e**), or only additional novel isoforms (“Distinct”, **f**). **g**, Comparison of short versus long reads for proteogenomics analysis. Short-read RNA-Seq provides fragmented evidence of transcript isoforms, whereas long-read RNA-Seq provides full-length transcript sequences that can be used to predict full-length protein isoforms.

The peptide identification and protein inference processes are heavily reliant on the composition of the protein database used for analysis. Reference protein databases broadly represent an organism's proteome, but may fail to capture the proteomic variation across tissues, developmental and disease states, and individuals.¹⁰ Discordances between a database and a sample can have a direct impact on proteomic search results. Ideally, the protein isoform sequences annotated in the reference for a gene would exactly match those expressed in a sample ("Match," **Figure 4.1b**). In practice, however, perfect matches are rare. The protein isoforms from a sample could differ from those in the reference by either lacking isoforms ("Subset," **Figure 4.1c**) and/or possessing a surplus of isoforms ("Superset," "Distinct," "Partial Overlap," **Figure 4.1d–f**). Overall, reference-sample discordances lead to (1) ambiguity in identifying protein isoforms; (2) incorrectly identified protein isoforms; or (3) failure to identify known or novel relevant protein isoforms (such as those associated to disease and treatment).

Transcript sequencing can be used to generate a sample-specific candidate protein database, which is more reflective of the isoform diversity in the sample than the reference database, but still has limitations due to the sensitivity and specificity of sequencing technologies. Presently, such efforts to generate sample-specific databases have been dominated by using short-read RNA-seq^{11–20} which suffers from the inability to sequence full-length transcripts and can only deliver partial protein models^{21,22} (**Figure 4.1g**). Long-read sequencing technologies, such as those from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can delineate full-length transcriptomes with high fidelity.²³ These technologies can readily reveal thousands of novel isoforms based on full-length transcript reads.²⁴ Such developments present an opportunity to leverage transcript expression—a prerequisite and correlate of protein expression²⁵—to enhance isoform-resolved proteomics.

Here, we present a workflow for long-read proteogenomics that achieves enhanced characterization of protein isoform diversity through paired long-read RNA-seq and MS-based proteomics of the same sample. This approach is enabled by a computational pipeline that generates full-length protein databases constructed de novo from long-read RNA-seq data. Using this database, we demonstrate MS-based discovery of novel protein isoforms arising from mechanisms such as retained introns and skipped exons. With full-length protein predictions, we introduce a new classification system, SQANTI Protein, to characterize novel protein isoforms. Finally, we introduce a new heuristic-based protein inference algorithm, called “Rescue & Resolve,” that incorporates long-read transcript abundance into the protein inference process, which enables detection of protein isoforms typically discarded during parsimonious protein inference due to insufficient peptide support. The entire pipeline and workflow is freely available as an open-source and extensible computational resource, using the community-based workflow language, Nextflow. This first-generation long-read proteogenomics pipeline provides a strong foundation for the integration of long-read sequencing into proteomic workflows, advancing the characterization of human protein isoform diversity.

4.3 Results

We developed a long-read proteogenomics pipeline for protein isoform detection through integrated analysis of sample-matched long-read RNA-seq and MS-based proteomics data. A Nextflow pipeline processes PacBio data, converts full-length transcripts into a protein database, and performs proteomics database searching (**Figure 4.2, Appendix III: Figure 9.1**). We demonstrate the utility of our pipeline using transcriptomic and proteomic data from the same cell line, Jurkat T-lymphocyte.

Below we describe the following: (1) analysis of PacBio sequencing to reveal high-quality full-length transcript sequences; (2) open reading frame (ORF) prediction; (3) a novel protein isoform classification system called SQANTI Protein; (4) generation of a sample-specific, full-length protein database using both PacBio and GENCODE reference isoform models; and (5) creation of a novel protein inference algorithm that increases the number of protein isoform identifications through the direct incorporation of PacBio transcript abundance values.

Long-read RNA-seq reveals widespread isoform diversity that differs from the GENCODE reference set

We characterized the landscape of full-length transcripts in a human cell line through long-read RNA sequencing on the PacBio platform (see **Appendix III: Section 9.2**). Transcript isoforms were compared to GENCODE²⁶ reference transcripts (v35), and their novelty status classified using SQANTI3 (Structural and Quality Annotation of Novel Transcript Isoforms)²⁷. Among the transcript isoforms identified, 43,865 contained an exact match to GENCODE (“full splice matches,” FSMs) and 75,491 were novel. Of the novel cases, 43,075 transcripts contained novel combinations of known splice sites and/or junctions (“novel in catalog,” NICs), and 32,416 transcripts contained an entirely new splice site or exon (“novel not in catalog,” NNCs). On average, novel transcripts exhibit lower abundances than their known counterparts, despite exhibiting a broad range of abundances overall (**Appendix III: Figure 9.2a**). In 13.93% (1,274) of genes, the most abundant transcript isoform is novel. To determine the sampling sensitivity of the transcriptome, we generated saturation-discovery curves and confirmed that the number of unique genes and isoforms detected reaches a plateau (**Appendix III: Figure 9.2b**). Overall, these results

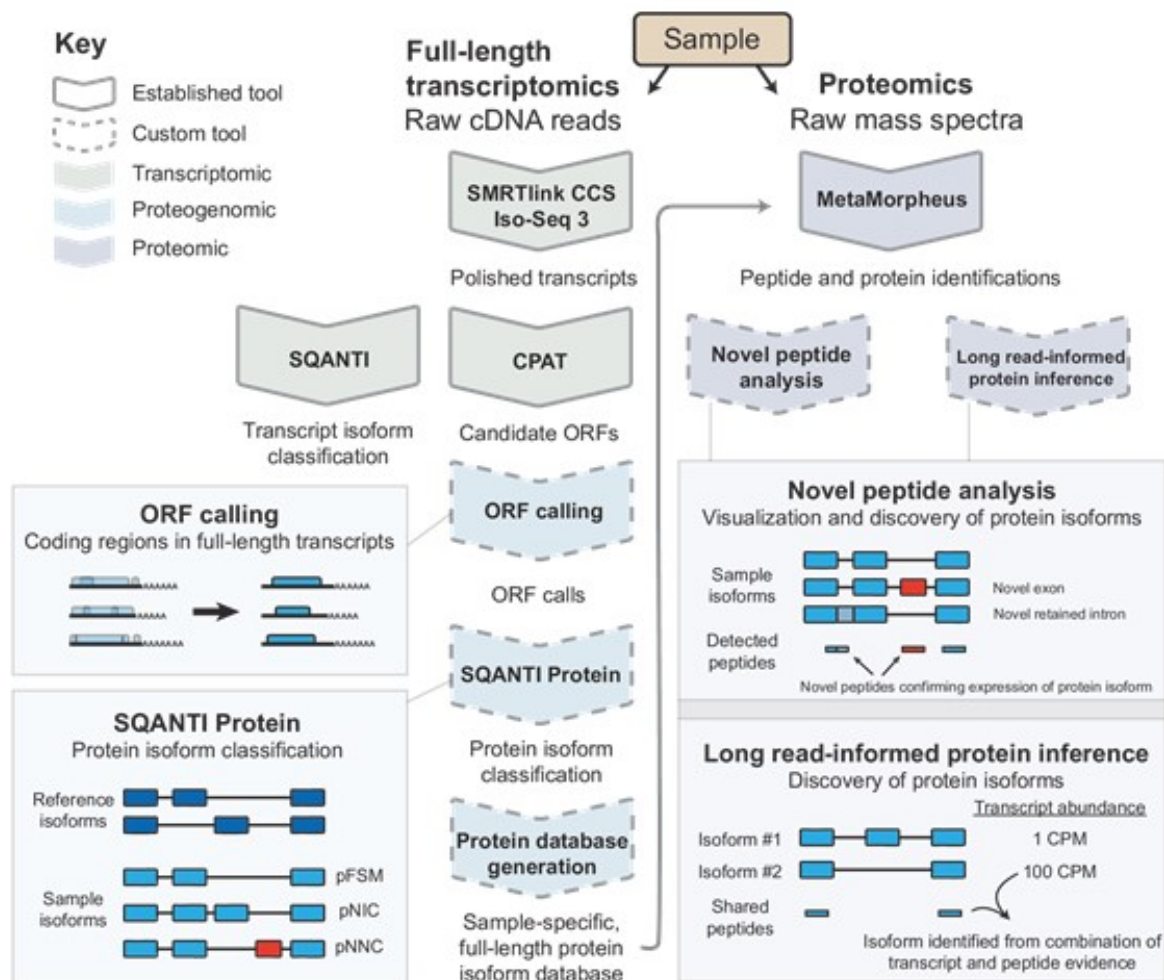


Figure 4.2: Long-read proteogenomic approach for enhanced sample-specific protein identification. Schematic of the long-read proteogenomics pipeline for improved protein isoform characterization. The pipeline includes approaches for ORF calling from long transcript reads, automated protein isoform classification (SQANTI Protein), novel protein isoform detection, and a long-read-informed protein inference algorithm. CPM - full-length read counts per million.

illustrate the widespread nature of alternative splicing and the need for empirically driven methods to characterize isoform diversity in human samples.

Note that for this study, transcript nucleotide sequences were derived from the reference genome (genome-corrected mode in SQANTI3); therefore, genetic variations are not captured in the current version of our pipeline (see **Section 4.4**).

A sample-specific, full-length protein isoform database derived from long-read RNA-seq data

ORF prediction from long-read RNA-seq data

We created a workflow to discern the most biologically plausible open reading frame (ORF) for each full-length transcript isoform. We considered multiple candidate ORFs for each transcript as defined by the Coding-Potential Assessment Tool (CPAT).²⁸ For most of the transcripts (91%), one ORF stands out as the most plausible protein-coding product based on its coding score; however, a sizable number of transcripts (12,787 or 9% of all transcripts) have two or more relatively high scoring ORFs (CPAT coding score above 0.9), in which the best ORF is unclear (**Appendix III: Figure 9.2c**). Therefore, for all ORFs, we incorporated additional metrics in the ORF ranking process, such as the GENCODE annotation status of the ATG start codon and the start codon's position relative to the 5' end of the transcript (see "ORF calling" in **Section 4.6** and see **Appendix III: Section 9.3**). After determining the ORF prediction for each transcript, we clustered transcripts containing identical ORF predictions (**Figure 4.3a**). Transcripts that differed only in their noncoding regions were assigned to the same protein entry in the database.

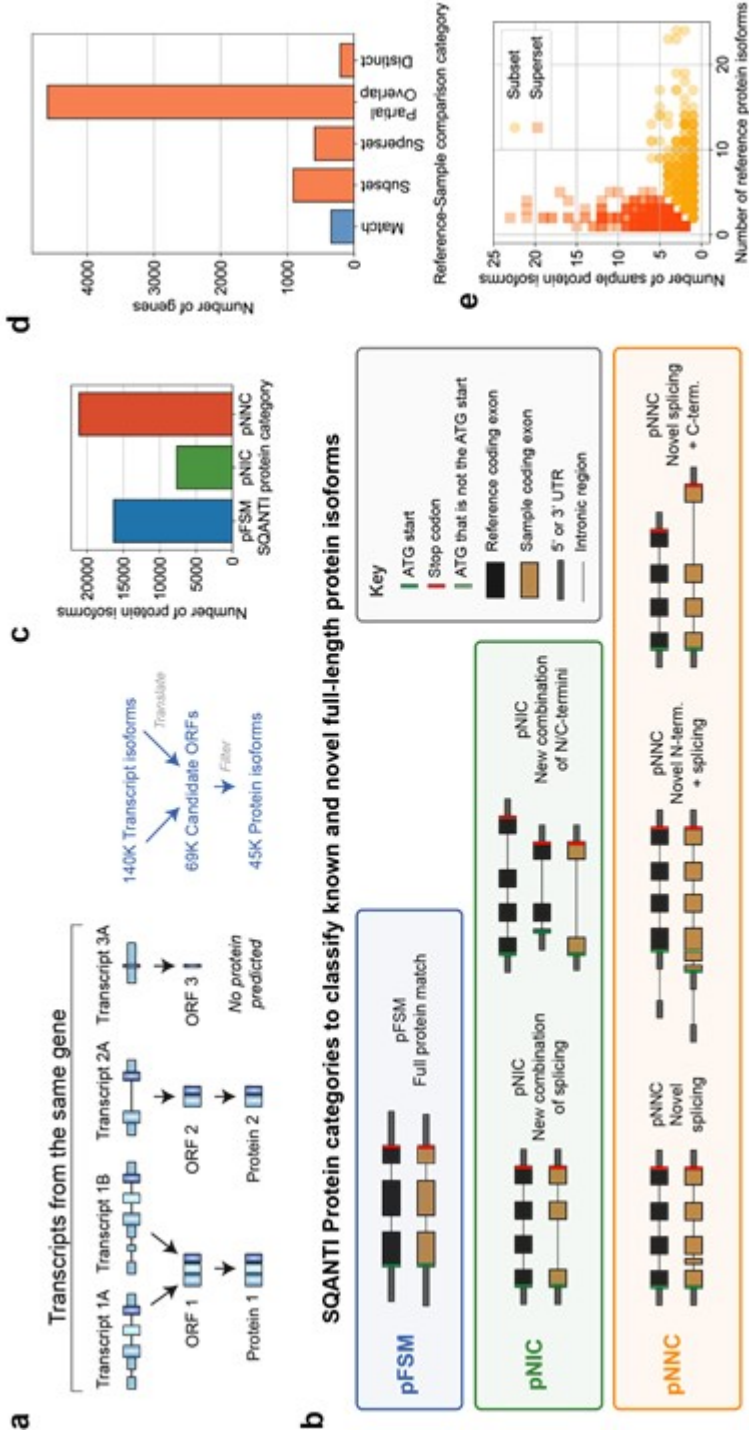


Figure 4.3: Generation and characterization of a long-read RNA-seq derived protein database. a, Schematic demonstrating grouping of transcript isoforms to protein isoform database entries. Some distinct transcript isoforms may have identical coding regions, producing the same theoretical protein isoform product. b, Schematic of the SQANTI Protein classification to compare long-read RNA-seq derived protein isoforms to those annotated in the reference proteome. c, Bar chart showing the frequency of protein isoform classifications for the protein database. d, Number of genes in each category described in **Figure 4.1b-f**, classified by the relationship between reference isoforms and predicted sample protein isoforms (genes in high confidence space). e, Comparison of the number of sample versus reference isoforms for Subset and Superset isoform comparison scenarios. pFSM, protein full splice match; pNIC, protein novel in catalog; pNNC, protein novel not in catalog.

SQANTI Protein: new classification scheme for full-length protein isoforms

We derived protein isoform models from long-read RNA sequencing data for each gene and found that many genes may concurrently express multiple protein isoforms (**Appendix III: Figure 9.2d**). To systematically characterize these full-length protein isoforms, we created a new protein isoform classification scheme, SQANTI Protein, to describe the relationship between the predicted protein isoforms and those annotated in GENCODE. SQANTI Protein extends SQANTI3 transcript-centric classifications to the protein isoform level, considering how three key protein sequence elements—the N-terminus, the identified splice junctions, and the C-terminus—compare to reference protein isoforms (**Figure 4.3b**). SQANTI Protein considers the full-length predicted protein sequence, detectable only by long-read RNA-seq, which differentiates it from previously proposed protein isoform classification schemas that have focused on “local” events, such as splice junctions or novel exons detected by microarrays or short-read RNA-seq.^{29,30}

We loosely follow the nomenclature first developed for transcript isoform classification in SQANTI. Major isoform categories for SQANTI Protein include pFSM, pNIC, pNNC, and pISM (**Figure 4.3b**). A “protein full splice match” (pFSM) represents a protein isoform where all elements exactly match at least one protein isoform in the reference. For a “novel in catalog” (pNIC) protein isoform, all protein sequence elements—such as the N-terminus, splice junctions, or C-terminus—are known (i.e., annotated in the reference), but the combination of elements is novel. A “novel not in catalog” (pNNC) protein isoform contains at least one novel element, such as a novel N-terminus or splice junction. Protein isoforms classified as an “incomplete splice match” (pISM) are cases in which the predicted protein isoform is a suspected artifact. For example, the originating transcript isoform could be degraded at the 5'end, resulting in a translation product missing the true ATG start codon. More

detailed protein isoform sub-classifications are provided in the “sqanti_protein” and “protein_classification” modules of the Nextflow pipeline.

Among the ORFs predicted from the long-read data, 16,331 (24%) have an exact GENCODE match and are deemed pFSMs (**Figure 4.3c**). We found 28,737 (41%) potentially novel protein isoforms, with 7,642 (11%) pNICs and 21,095 (30%) pNNCs. A more detailed breakdown of categorizations can be found in **Appendix III: Table 9.1**. The remaining sequences were classified as pISM or were putative translation products of transcripts unlikely to be protein coding, such as intergenic transcripts.

It is notable that transcript-level classification does not always translate directly to the protein-level classification (**Appendix III: Table 9.2**). For example, 371 transcript-level ISMs (ISMs) are actually protein-level FSMs (pFSMs). This occurs when part of the 5' untranslated region (UTR) of a reference transcript is missing, but the ATG start codon is preserved. As another example, for 4,086 known protein isoforms (pFSMs, 25% of total pFSMs), the originating transcript was novel (NIC or NNC) with novel splicing events exclusively occurring in the UTRs.

Predicted protein isoforms that are novel make up a substantial part of the database. For the majority of genes (75%), at least one pNIC or pNNC protein isoform was uncovered (**Appendix III: Figure 9.2e**). Furthermore, for a third of all genes with observed transcripts, the most abundant protein isoform by transcript abundance did not correspond to the “reference” isoform (i.e., GENCODE APPRIS principal reference isoform³¹, **Appendix III: Figure 9.2f**), and 42.5% (1215) of those isoforms were entirely novel.

After annotation with SQANTI Protein, 45,068 protein isoforms (pFSM, pNIC, and pNNC protein isoforms) from 10,348 genes were considered for database generation.

Defining a high-confidence PacBio-derived protein database

We generated a high-quality database for proteomic analysis with the following filtering criteria. Within our PacBio dataset, we found that genes producing transcripts with extreme lengths (e.g., less than 1 kb, longer than 4 kb), low abundance (e.g., below ~ 3 CPM, or full-length read counts per million), or without 3'polyadenylation were not fully covered due to technical limitations (**Appendix III: Section 9.4**). Therefore, we used these criteria to select genes in which we were confident in the sampling of protein-coding transcripts. By extension, we are confident that the protein isoform models for these genes are reasonably complete. A total of 6,653 genes meet our filtering criteria and are within the “high-confidence” space (HC space). For all other genes, we populated the protein database with GENCODE entries, generating a hybrid database to maintain integrity of downstream proteomic analysis. This hybrid database of PacBio-derived and GENCODE entries, called PacBio-Hybrid, is composed of 35,119 PacBio-derived protein entries from 6,653 genes, and 48,413 GENCODE protein entries for the remaining 13,276 protein-coding genes (**Appendix III: Figure 9.3a**).

PacBio-derived protein isoform models for most genes differ from the reference

As described in the **Section 4.2**, differences between what is expressed in the sample and the reference database (see **Figure 4.1b–f**; Match, Subset, Superset, Partial Overlap, Distinct) can have striking consequences on the protein isoforms inferred by MS analysis. Within the HC space, we found less than 5% of genes have PacBio-derived isoform models that exactly match the reference database (**Figure 4.3d**). The most frequent database-sample discordance observed at a rate of 69% is “Partial Overlap,” in which the PacBio-derived database contains one or more reference-matched isoforms, but also contains additional novel isoforms. A total of 19,838

novel isoforms belong to genes in the “Partial Overlap” category. The other database-sample discordance categories which contain novel PacBio isoforms, “Superset” and “Distinct,” account for 8.9% and 3.1% of the genes in the database, respectively. Overall, the number of predicted protein isoforms for a given gene can diverge greatly between the sample-specific and reference database (**Figure 4.3e**).

MS-based proteomics analysis with a PacBio-derived protein database

The PacBio-derived proteome differs substantially from the reference proteome. Since the database used for proteomic analysis serves not only as a model for identification but also for protein inference, its isoform composition directly impacts protein identifications. To assess such impacts, MS data from the Jurkat cell line was obtained and used for proteomic analysis with either the PacBio-Hybrid or GENCODE database. The MS spectra for analysis was generated via liquid chromatography-MS (LC-MS)/MS data-dependent analysis (DDA) of 28 fractions from high-pH reverse-phase liquid chromatography (RPLC) of a Jurkat tryptic digest. Acquired spectra were searched using the software tool MetaMorpheus¹⁶ to obtain peptide- and protein-level identifications at a 1% false discovery rate (FDR) (**Appendix III: Table 9.3**, Additional file 6: Table S4).

PacBio-derived protein database recovers peptides identified with the reference database

Notably, the proteomic results using the PacBio-Hybrid database recovered 99% of peptide and 99% of gene identifications found in the GENCODE reference database search results (1% FDR cut-off, **Figure 4.4a,b**). Similar trends of results were observed when considering data from only the HC space, as well as when comparing PacBio-Hybrid results to search results obtained when using the UniProt reference database

(**Appendix III: Figure 9.3b-g**). Additionally, the overlap between identified peptides and genes for the PacBio-Hybrid and reference database search results is comparable with the overlap found between the search results of the two reference databases (GENCODE vs. UniProt, **Appendix III: Figure 9.3h-i**) demonstrating that the PacBio-derived database is appropriately covering the protein space in the sample.

PacBio-derived isoform models lead to dramatically different protein isoform identification and can resolve ambiguities

MS-based identification of protein isoforms is challenging due to the uncertainty in assigning shared (multi-mapping) peptides to their isoform(s) of origin. The protein database utilized for analysis should represent the protein isoforms in the sample, but differences between isoforms in the database versus the sample can impact the accuracy and precision of the inferred protein groups (see **Figure 4.1**).⁹

We found that although the peptide and gene-level identifications between the PacBio-Hybrid and GENCODE MS search results were nearly 100% concordant (**Figure 4.4a,b**), indicating that the peptide set for protein inference is nearly identical, there were major differences in the protein isoform identifications obtained (**Figure 4.4c**). Only 41% (4,503) of the protein isoform groups from both PacBio-Hybrid and GENCODE results were identical. Similar results were observed for comparisons of protein groups in the HC space, against the protein groups from the UniProt reference database search, and between the protein groups obtained from the two reference database searches (**Appendix III: Figure 9.3j-m**). This low overlap of protein inference results, across all comparisons, indicate that differences in protein identifications are primarily caused by differences in protein isoform composition of the databases.

The PacBio-derived database provides transcript-backed evidence of protein iso-

form expression that, when combined with peptide evidence, can lead to enhanced protein isoform identification. We found 3,199 PacBio-Hybrid protein groups that are different from those protein groups inferred through the GENCODE reference search. Of these protein group differences, 673 cases (21%) result in increased specificity of protein isoform identification when using the sample-derived PacBio-Hybrid database. An illustration of this can be found in **Figure 4.4d**. Based purely on MS peptide evidence, there is ambiguity in terms of whether the isoform LNPK-201 or LNPK-212 is expressed, but the PacBio transcript evidence indicates LNPK-201 is the main isoform likely to be expressed in the cell line. Another common scenario, accounting for 873 cases (27%), is that of partially overlapping protein isoform groups between the PacBio-Hybrid and reference results, as illustrated by isoforms of *MECP2* (**Figure 4.4e**). Using the GENCODE database as reference, MECP2-205 and MECP2-201 form a single protein isoform group and are indistinguishable based on the peptide evidence. However, when using the PacBio-Hybrid database, there was no transcriptional support for MECP2-201. Instead, MECP2-205 forms a protein isoform group with the novel PacBio-derived isoform PB.16836.37. A third scenario, accounting for 382 cases (12%), occurs when all of the protein isoforms for a protein group in the PacBio-Hybrid analysis are absent from any protein groups within the GENCODE reference database analysis. This results in a protein group that is entirely distinct to the PacBio-Hybrid protein inference results. An example of this can be found in **Figure 4.4f**, where the PacBio-derived database lists a single isoform which is not found in the reference database, representing a case of an entirely distinct isoform model.

For many of these cases, peptides were not detected in the isoform-specific regions, leading to a high dependence of protein isoform inference on the isoforms represented in the database. The isoform composition of a database has an outsized impact on

the protein inference results obtained, and we believe that sample-specific databases improve the accuracy of protein isoform detection.

Characterization novel *RUNX1* isoforms relevant to thymocyte biology

Within our data, we uncovered an excellent example of biologically relevant protein isoforms from *RUNX1* using full-length PacBio sequencing. *RUNX1* expresses a key transcription factor that regulates early thymocyte development.^{32,33} Rearrangements or mutations of *RUNX1* are associated with multiple hematopoietic neoplasms.^{34,35} Interestingly, recent evidence indicates germline mutations in *RUNX1* are associated with an increased risk of acute lymphoblastic leukemia (ALL) and that these mutations result in the generation of dominant negative isoforms of *RUNX1*.³⁶ The Jurkat cell line, analyzed here, is derived from a 14-year-old male patient with ALL.³⁷ Therefore, understanding the isoform landscape of *RUNX1* in our sample is highly relevant. Overall, we predicted 11 novel full-length protein isoforms of *RUNX1* (**Appendix III: Figure 9.4**). Eight of these predicted protein isoforms contain the complete DNA binding Runt homology domain (RHD) sequence expressed in-frame with novel downstream sequences (PB.15792.9, PB.15792.10, PB.15792.15, PB.15792.17, PB.15792.18, PB.15792.32, PB.15792.33, PB.15792.40). Additionally, five of these predicted isoforms (PB.15792.17, PB.15792.18, PB.15792.32, PB.15792.33, PB.15792.40) lack the transactivation domain (TAD) found in the longer *RUNX1* protein isoforms. The TAD recruits multiple cofactors (P300, CREBBP, TLE1) to *RUNX1*-binding sites, and thus each novel protein isoform has the potential to represent a functional dominant negative isoform capable of binding *RUNX1* sites but unable to recruit relevant cofactors that mediate gene activation or repression.^{35,38} Since full-length *RUNX1* is

known to generally activate T-cell differentiation genes and suppress multipotent hematopoietic genes³³, expression of these newly predicted dominant negative isoforms is consistent with supporting leukemogenic potential in Jurkat T-ALL. Peptide identifications provide support for the presence of three protein isoforms in two distinct protein groups. The two isoforms PB.15792.10 and PB.15792.15, containing both the RHD and TAD, are inferred as an indistinguishable protein group. Interestingly, PB.15792.40, one of the predicted dominant negative isoforms, is identified with a uniquely mapping peptide.

Long-read, sample-specific database leads to discovery of novel protein isoforms

The MS search with the PacBio-Hybrid database revealed novel peptide sequences which were absent from both the GENCODE and UniProt reference databases. Stringent validation criteria were applied for novel peptide identifications and are described in more depth in **Appendix III: Section 9.5**. We manually examined candidate mass spectra and confidently identified 14 novel peptides, each corresponding to a distinct event (Additional file 6: Table S4). Such events arose from a diversity of mechanisms, including upstream ATG start site usage, translation of a retained intronic region, and novel exons (**Figure 4.5a–c**).

Notably, 6 of the 14 novel detected peptides each map to a single isoform and therefore provide direct evidence for expression of the corresponding full-length protein isoform. Such a direct link from peptide to full-length protein is only available with knowledge of full-length transcripts expressed in the sample.³⁹ An example of this is illustrated for the peptide, abbreviated as ESD, which confirms the novel terminal exon in *RABGAP1L*, but also unambiguously maps to the full-length PacBio-

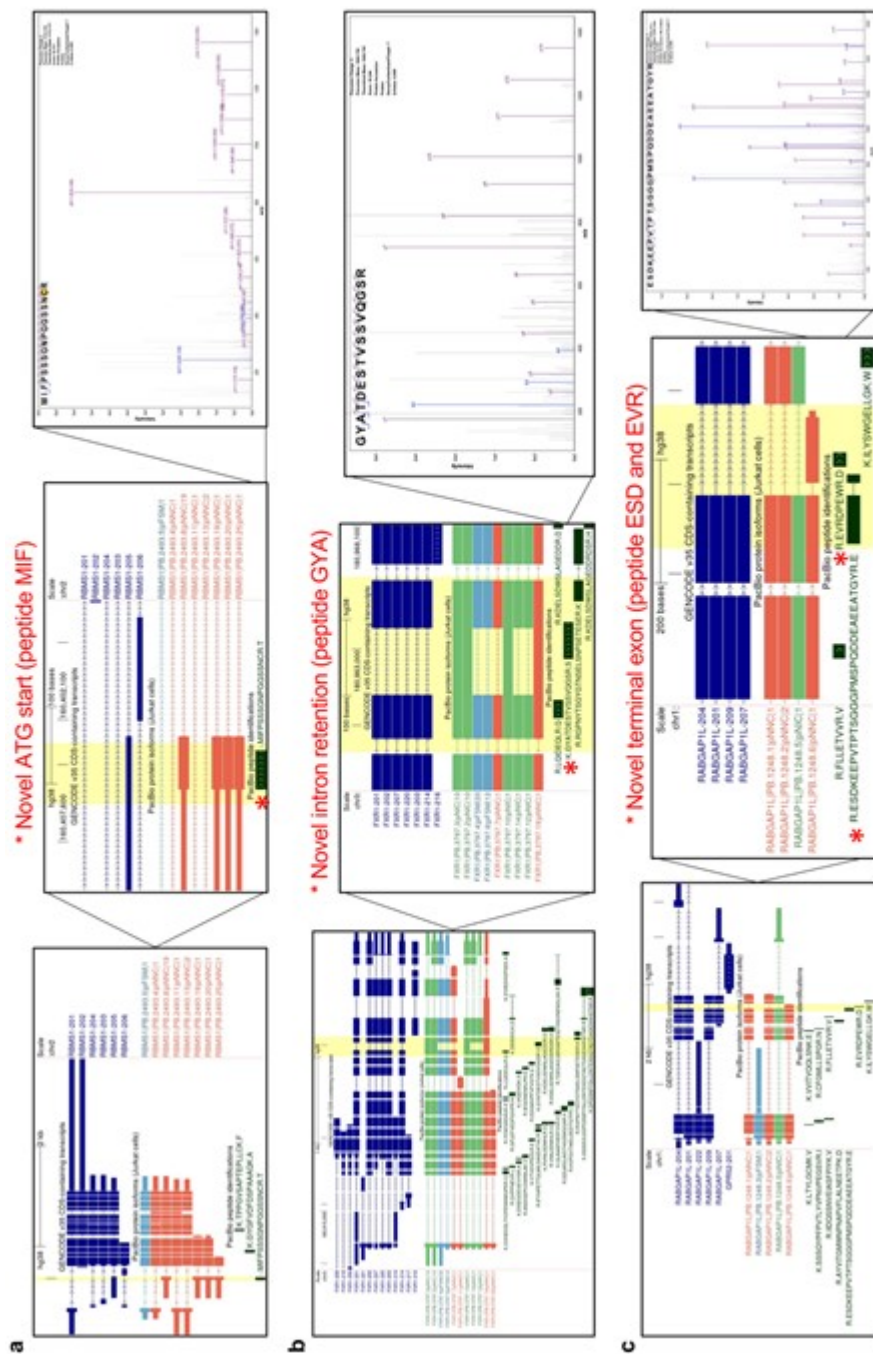


Figure 4-5: Discovery of novel peptides and full-length protein isoforms. a, Novel peptide MIF confirms translation of an ATG start for RBMS1. b, Novel peptide GYA confirms translation of a novel retained intron for FXR1. c, Novel peptides ESD and EVR confirms the translation of a novel terminal exon for *RABGAP1L*. In this case, since the novel peptide maps exclusively to PB.1248.6, the corresponding full-length protein isoform is likely translated. Note that only ESD passed strict manual annotation, but EVR, which passed a 1% FDR in the global MS search, supports the expression of the same terminal exon.

derived protein isoform PB.1248.6 (**Figure 4.5c**). Only a small fraction of all potential novel protein isoforms are identified directly by a novel peptide. This is unsurprising based on previous reports regarding the detectability of isoform-specific tryptic peptides. The low peptide coverage of alternative isoforms could be technical in origin^{40,41}, and the debate is ongoing regarding the extent to which novel transcript isoforms are translated into proteins^{7,8}.

Long-read RNA-seq-informed protein isoform identification

In order to infer the presence of protein isoforms, most protein inference algorithms employ a probabilistic or parsimonious approach. Probabilistic protein inference algorithms seek to estimate the probability that a given protein isoform is in the sample on the basis of the peptides observed.^{42–45} Parsimonious protein inference algorithms are more heuristically driven and follow Occam's razor, which attempts to define the smallest number of protein isoforms that "covers" the set of identified peptides.^{9,45–49}

Parsimonious algorithms are commonly used in the MS proteomics field as part of search software platforms like Andromeda/MaxQuant and MetaMorpheus. However, this approach can lead to elimination of bona fide protein isoforms that lack sufficient peptide support relative to other isoforms (**Figure 4.6a**).⁵⁰ Alternative isoforms are particularly susceptible, because their isoform-specific regions comprise a small fraction of the proteome and suffer from a negative detection bias in traditional MS-based proteomics workflows using tryptic digestion.⁵¹

In our tryptic dataset, the peptides observed at 1% FDR could be the digestion products of up to 26,931 different PacBio-derived protein isoforms in the high-confidence space. When traditional, parsimonious protein inference is applied to this peptide set, the number of PacBio-derived protein isoforms present in inferred protein groups

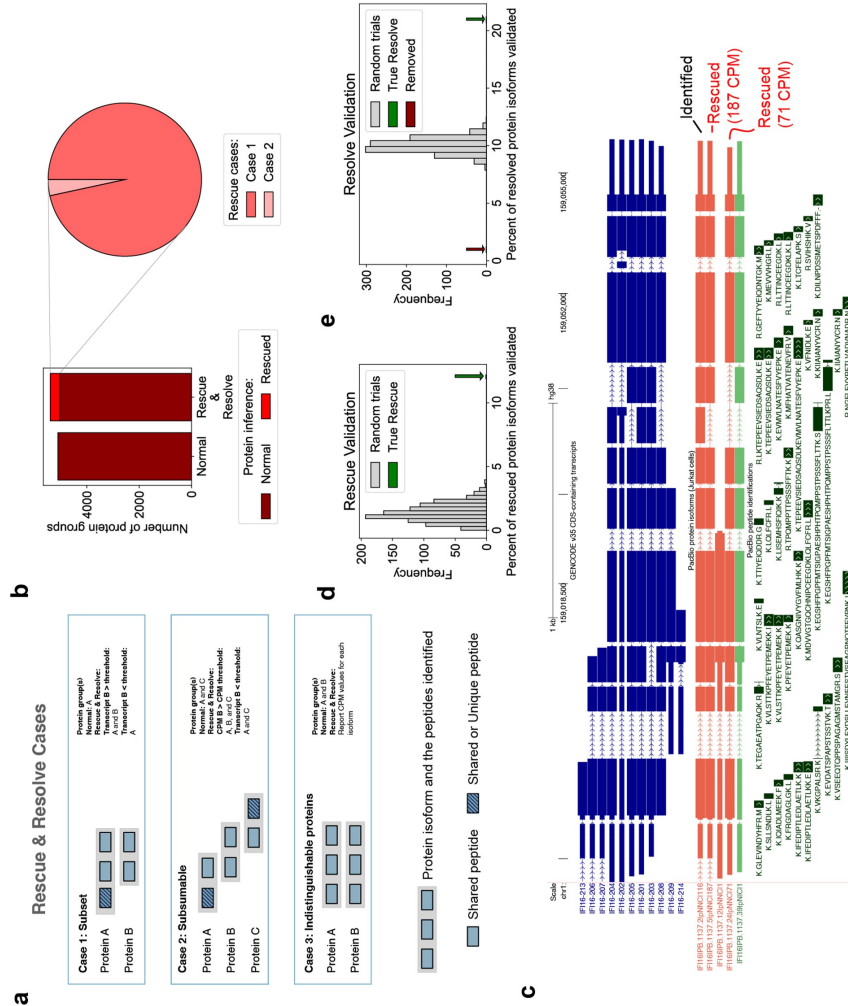


Figure 4.6: Long-read-informed protein isoform detection. a, Scenarios in which long-read transcriptomic abundance values can be used to “Rescue and Resolve” protein group identifications for improved protein isoform detection. Note that peptide locations are not drawn to scale. b, The number of protein groups containing PacBio-seq data to rescue isoforms identified in a traditional (normal) search versus one that incorporates long-read RNA-seq data to rescue isoforms (Rescue & Resolve algorithm). c, Example of two rescued isoforms from the gene *IFI16*. PB.1137.5 and PB.1137.24 both have high transcriptional abundance, and contain novel exon skipping events, but both lack unique sequence regions containing an identified peptide. CPM, full-length copies per million reads. d, e, The percent of rescued (d) and resolved or removed (e) protein isoforms validated compared to the derived null distribution from randomly rescued isoforms.

drops to 11,231, eliminating 15,700 potential protein isoforms due to lack of sufficient peptide support. We hypothesize that a fraction of these eliminated protein isoforms may actually exist in the sample, and their elimination reduces the precision and accuracy of the protein inference results obtained.

Rescue & Resolve: direct incorporation of long-read data into protein inference

To overcome limitations of incomplete peptide coverage for protein isoform detection, we reasoned that the incorporation of long-read transcript isoform data directly in the protein inference process could help inform on the presence of a protein isoform. For this purpose, we developed a heuristic-based protein inference algorithm called “Rescue & Resolve” (R&R), which is implemented within a custom version of MetaMorpheus (see **Section 4.6**). To our knowledge, this is the first protein inference algorithm that incorporates long-read transcriptional abundance as an orthogonal data source. As previously mentioned, the parsimonious protein inference process makes decisions throughout the algorithm to discard, or eliminate, protein isoforms from consideration for identification, because they lack the same level of peptide evidence that competing isoforms possess. During this process, protein isoforms that are actually present in the sample could be eliminated, generating false negatives. The “rescue” portion of our “R&R” algorithm defines two cases in which a protein isoform could be “rescued” from elimination (**Figure 4.6a**). The first case occurs when a protein isoform’s mapped peptides are a subset of the peptides mapped to another protein isoform (Case 1, **Figure 4.6a**). In this scenario, the parsimonious algorithm would determine that the protein isoform which accounts for the most peptides is the simplest answer, and therefore more likely to be correct by the principle of Occam’s razor. The protein isoform that accounts only for a subset of the peptides observed is eliminated from consideration for identification. The second

case occurs when a protein isoform's mapped peptides are subsumable to (i.e., can be explained by) two or more protein isoforms which have additional peptide evidence (Case 2, **Figure 4.6a**). In this scenario, there is a protein isoform for which all of its peptide evidence can be explained by the existence of multiple protein isoforms that all have more peptide identifications supporting their existence. Again, as in Case 1, the parsimonious approach dictates that it is simpler, and therefore more likely, that the protein isoforms with additional peptide support are the sole contributors to the peptides being identified. The subsumable protein isoform is then eliminated from consideration for identification. In the "rescue" portion of our R&R algorithm, during the parsimonious process, protein isoforms that were eliminated due to scenarios such as Case 1 and Case 2, are identified, and set aside as potential false negatives that can be "rescued" from elimination. To determine whether or not a protein isoform should be "rescued" or eliminated, the long-read transcriptional abundance information obtained for each isoform is leveraged as an additional source of data. Since RNA abundance is at least moderately correlated with protein expression^{25,52} (R-squared = 0.65, **Appendix III: Figure 9.5a**), a high abundance transcript would have a higher probability, than a low abundance transcript, of generating the corresponding protein which was observed in our dataset (**Appendix III: Figure 9.5b**). In the R&R algorithm, protein isoforms are only rescued from elimination if their transcriptional abundance is greater than a user-specified abundance threshold. We selected a conservative transcript abundance threshold of 25 CPM (see **Appendix III: Section 9.6** for parameter optimization details). The impact of the "rescue" portion of the "Rescue & Resolve" algorithm on the protein inference results obtained were compared to those obtained with the traditional parsimonious protein inference algorithm within MetaMorpheus (details regarding MetaMorpheus's inference algorithm can be found at <https://github.com/smith-chem-wisc/MetaMorpheus/wiki/Protei>

n-Parsimony-&-Grouping-(Protein-Inference)).

We rescued 355 protein groups, of which 343 (96.6%) are Case 1 and 12 (3.4%) are Case 2 (**Figure 4.6b**). A common example, Case 1, is shown in **Figure 4.6c** for isoforms of *IFI16*, in which the dominant isoforms (PB.1137.5 and PB.1137.24) are not the isoform that contains the longest sequence (PB.1137.2). Notably, these isoforms are entirely novel, as compared to isoforms found in GENCODE. Collectively, the “rescued” protein isoforms represented a 6.5% increase in the number of PacBio-derived protein isoforms identified at 1% FDR, compared to what is obtained without the “R&R” algorithm, using MetaMorpheus’ traditional parsimonious approach. Validation of protein inference approaches is exceedingly difficult, in that we do not know the true composition of the sample, and standard protein mixtures lack the complexity necessary to model the human proteome. This is especially true in the case of modeling human isoform diversity where the “Rescue & Resolve” algorithm is most beneficial. To validate the accuracy of the “rescued” protein isoform identifications, we used an independent multi-protease MS dataset to generate a “ground truth” of protein isoform presence, enabling us to calculate the rate of validation of the “rescued” protein isoforms within the high coverage multi-protease dataset, as compared to the validation rate of a random control (see **Appendix III: Section 9.7**). We observed that 12.2% of protein groups that were “rescued” were confirmed to be expressed in the multi-protease data, which is much greater than the average fraction of “rescued” protein isoforms validated from the distribution of the randomized control at 1.4% (N = 1,000 permutations, p -value < 0.0001, **Figure 4.6d**). Details on the construction of the randomized control permutations can be found in **Appendix III: Section 9.7**. Therefore, these results indicate that many true protein isoforms are rescued based on the incorporation of long-read sequencing knowledge.

The “resolve” portion of the R&R algorithm addresses a third scenario which

can arise during protein inference (Case 3, **Figure 4.6a**), where the parsimonious process generates ambiguity through a protein group which contains two or more indistinguishable protein isoforms (based on equivalent peptide evidence). Ambiguous protein groups can be composed of three different classes of isoforms categorized by their relative transcriptional abundance: (1) dominant (a “resolved” isoform), (2) minor, or (3) co-expressed. The “resolve” portion of the algorithm provides the opportunity to “resolve” these ambiguous protein groups to a single, dominant isoform, or provides support for the co-expression of multiple protein isoforms based on relative transcriptional abundance of each isoform within the group. For instances of Case 3, the relative transcriptional abundances underlying the predicted protein isoforms could indicate likelihood of expression.

We found 2,600 cases (Case 3, **Figure 4.6a**) of indistinguishable protein isoform groups in the high-confidence space, in which one or more protein isoforms are indistinguishable by peptide evidence alone. Our algorithm provides the relative transcript abundance measures for protein isoforms within a group, enabling the opportunity to resolve isoform identifications based on underlying transcript support, which is fully at the discretion of the user (**Appendix III: Figure 9.5c**). We found that in 1,434 cases, one isoform comprises more than 90% of the transcript abundance, suggesting that a single dominant isoform could comprise the group. For these dominant isoform-containing protein groups, the ambiguity of which protein isoform is present within the sample was resolved, and a single protein isoform was considered to be identified, increasing the precision of the protein inference results obtained. Notably, not all protein groups can or should be resolved to a single isoform. There are cases where multiple protein isoforms are co-expressed and the peptide evidence is not comprehensive enough to be able to sufficiently distinguish them. It is important to maintain protein group ambiguity when necessary and valid. We discovered 295

protein isoform groups in which multiple protein isoforms may be co-expressed at appreciable levels (2+ isoforms with relative abundance > 30%), indicating that a single representative isoform cannot be assumed for these cases. We validated the accuracy of the “resolved” protein isoform identifications by applying the same multi-protease validation strategy used for “rescued” protein isoforms (see **Appendix III: Section 9.7**). We observed that 21.2% of the “resolved” protein isoforms were confirmed to be expressed in the multi-protease data, which is much greater than the average fraction of “resolved” proteins validated from the distribution of the randomized control, 10.0% (N = 1,000 permutations, p -value < 0.0001, **Figure 4.6e**). Details on the construction of the randomized control permutations can be found in **Appendix III: Section 9.7**. We also investigated the validation rate of the protein isoforms that were removed from the protein groups, to determine if their removal was justified. We observed that only 0.7% of the removed isoforms were confirmed to be expressed in the multi-protease data. This is much less than the average fraction of “resolved” proteins validated from the distribution of the randomized control and the validation rate of the experimentally “resolved” protein isoforms (**Figure 4.6e**). Although the majority of the “resolved” protein isoforms (73%) are incapable of producing a detectable unique peptide (7 to 50 amino acids) in any of the six protease digests (Arg-C, Asp-N, Chym, Glu-C, Tryp, and Lys-C), 86 of the 387 (22%) “resolved” isoforms capable of producing a theoretical unique peptide were confirmed by the identification of a unique peptide identified in the multi-protease dataset. All “rescued” and “resolved” groups may be found in Additional file 7: Table S5.

These results indicate that the incorporation of long-read transcriptional abundance values into the protein inference process reveals protein isoforms that were difficult to identify solely with MS peptide data.

4.4 Discussion

The comprehensive characterization of the cellular proteome is a major goal in proteomics to understand the molecular underpinnings of normal and disease states. One factor impeding progress towards this goal is the lack of experimental approaches that can easily identify proteins at isoform resolution. Current efforts employ short-read RNA-seq approaches which cannot characterize full-length isoforms.²² Long-read sequencing provides the ability to obtain full-length transcript reads²³, allowing the delineation of transcript isoforms and, therefore, potential full-length protein isoforms for MS analysis^{39,53,54}.

To our knowledge, this is the first long-read based proteogenomics pipeline that integrates full-length transcripts with MS data for full-length protein isoform characterization. We show that the availability of long-read-derived, sample-specific protein isoform models is critical to enhance protein isoform detection. Our pipeline produces sample-specific, full-length protein isoform databases which enables novel peptide discovery, and outputs genome browser tracks for visualization of reference- and sample-derived isoforms as well as peptide identifications. The pipeline also includes the first protein inference algorithm to directly incorporate long-read sequencing data to detect protein isoforms heretofore intractable to MS analysis (“Rescue & Resolve”).

Integrating long-read sequencing and proteomic data presented new challenges, which we addressed through the development of new components in the pipeline. We defined for each full-length transcript the most likely canonical ORF based on a modified output of CPAT. Further, we created a new protein isoform classification system, SQANTI Protein, based on the transcript isoform classification tool SQANTI3. Finally, the “Rescue & Resolve” algorithm, through incorporation of long-read transcript isoform expression data into the protein inference process, enables the “rescue”

of protein isoforms that have significant transcriptional support but are nonetheless difficult to identify in MS due to high sequence overlap. The algorithm also enables the user to “resolve” ambiguous protein isoforms that are indistinguishable based on peptide evidence alone, by leveraging the relative transcriptional abundance for such isoforms.

Our workflow identified 45,068 distinct candidate protein isoforms from a human cell line (Jurkat cells), 22,807 of which were novel. These long-read sequencing-derived protein isoforms were filtered, and a sample-specific PacBio-Hybrid database containing 35,119 PacBio-derived protein isoform entries was generated. Proteomic analysis of this database revealed 14 novel peptide identifications and 5,100 protein isoform groups within the high-confidence space identifications at 1% FDR. Notably, one of the novel peptides confirmed the translation of a transcript with a retained intron, which highlights the utility of an empirical approach to uncover the translation of transcripts not commonly thought to be translated. The implementation of the heuristic-based Rescue & Resolve protein inference algorithm increased the number of PacBio-derived protein isoform groups identified by 355, and resolved 1434 ambiguous protein isoform groups to a single protein isoform identification. The resolve approach also highlighted the existence of 295 protein isoform groups in which multiple protein isoforms appeared to be co-expressed at appreciable levels (2+ isoforms with relative abundance $> 30\%$), demonstrating it is not always appropriate to assume a single isoform is expressed.¹⁴ Although the Rescue & Resolve algorithm was developed for use with long-read sequencing information, the algorithm could also be applied to proteogenomic databases and transcriptional abundance information derived from short-read sequencing approaches.

The results and concepts described here provide a foundation for future development of long-read proteogenomics. The pipeline’s flexible and modular nature lends

itself to adaptation. For example, the proteomic analysis portion of the pipeline could be expanded to include a semisupervised learning post-search program such as percolator⁵⁵ or mokapot⁵⁶. In the future, we plan to expand the custom ORF prediction algorithm to include the discovery of noncanonical ORFs, such as those with cognate start sites (e.g., CTG) or short upstream ORFs commonly found in the 5'UTR.⁵⁷⁻⁵⁹ Another improvement to the pipeline will be an evolution of the heuristically driven “Rescue & Resolve” approach. We plan to develop a probabilistic protein inference algorithm in which transcriptional abundance values are incorporated into a rigorous statistical framework for the inference of protein isoforms.^{43,60} The applications of our computational pipeline could also include the analysis of novel genes or genetic variations that are detectable in long-read data or separately available from previous genotyping, use of ONT (i.e., nanopore) cDNA or direct RNA sequencing data⁵⁴, the analysis of single-cell RNA-seq, use of targeted long-read datasets⁶¹, or the use of top-down proteomics data for the analysis of proteoform diversity⁶².

Though long-read proteogenomics and its application hold promise, limitations remain. First, for the “Rescue and Resolve” approach, we assume at least a moderate degree of RNA-protein correlation. Although isoforms from the same gene should not greatly differ in their transcript-protein correlation, several studies have reported isoform-specific mRNA translation^{63,64} suggesting that alternative splicing can generate transcripts with distinct cis-regulatory landscapes. Therefore, caution must be taken for any given protein isoform, including follow-up confirmation of expression *in vivo*. Second, as with any RNA-Seq-based dataset, even though a majority of the isoform diversity detected from long-read RNA-seq approaches are likely due to co- and post-transcriptional processing mechanisms, it is possible that genetic translocations, deletions, or other mutations may give rise to what is ostensibly transcript isoform variations that are actually genetic in origin. We used Jurkat cells as a model

system, which is tetraploid, and may contain some isoform variations due to cancer-related or natural genetic variants.⁶⁵ Third, the pipeline results are dependent on the quality of long-read RNA sequencing. Limitations in quality of the extracted RNA or artifacts generated during the sample handling and library preparation process (e.g., PCR artifacts) can detrimentally impact accuracy of predicted protein models. The sampling of full-length transcripts is known to be incomplete—ultra-long transcripts or those transcripts lacking a polyA tail may be under sampled—and can impede the ability to derive the entire proteome from transcript data alone. However, as both ONT and PacBio sequencing improves in both coverage and sensitivity, an entire long-read-derived proteome should be able to be generated *de novo* from sample-specific transcriptomes. Furthermore, rigorous benchmarking studies, such as those being conducted by The Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) Consortium, will reveal strength and limitations of these methods for the community.⁶⁶

Overall, the incorporation of long-read sequencing into proteogenomic workflows represents a tremendous opportunity for isoform-resolved investigations in basic and translational research. As long-read sequencing continues to evolve in throughput, accuracy, and accessibility, long-read proteogenomics will be adopted by researchers and clinicians and become a routine practice in the context of precision medicine.

4.5 Conclusion

We show that sample-specific protein isoform models derived from long-read RNA-seq can lead to enhanced protein isoform detection. Our pipeline enables novel peptide discovery and outputs genome browser tracks for visualization of reference- and sample-derived isoforms as well as peptide identifications. We introduce the first

protein inference algorithm that directly incorporates long-read sequencing data to detect protein isoforms heretofore intractable to MS analysis (“Rescue & Resolve”). This work represents a foundation for subsequent studies that integrate long-read RNA-seq with proteomics for protein isoform characterization.

4.6 Methods

PacBio long-read RNA-seq

PacBio (Iso-Seq) data was collected on the Jurkat T-lymphocyte cell line. Jurkat RNA was procured from Ambion (Thermo, PN AM7858). The RNA was analyzed on a Thermo Nanodrop UV-Vis and an Agilent Bioanalyzer to confirm the nominal concentration and ensure RNA integrity. We observed a RIN value of 9.9. From the RNA, cDNA was synthesized using the NEB Single Cell/Low Input cDNA Synthesis and Amplification Module (New England Biolabs).

Approximately 300 ng of Jurkat cDNA was converted into a SMRTbell library using the Iso-Seq Express Kit SMRT Bell Express Template prep kit 2.0 (Pacific Biosciences). This protocol employs bead-based size selection to remove low mass cDNA, specifically using an 86:100 bead-to-sample ratio (Pronex Beads, Promega). Library preparations were performed in technical duplicate. We sequenced each library on a SMRT cell on the Sequel II system using polymerase v2.1 with a loading concentration of 85pM. A 2-h extension and 30-h movie collection time was used for data collection. The “ccs” command from the PacBio SMRTLink suite (SMRTLink version 9) was used to convert Raw reads (~ 6 million, over 349 Gbps) into Circular Consensus Sequence (CCS) reads. CCS reads with a minimum of three full passes and a 99% minimum predicted accuracy (QV20) were kept for further analysis.

Jurkat RNA-Seq data download and analysis

Jurkat RNA-Seq data was previously collected on an Illumina HiSeq2000, generating ~ 38.8 million paired-end 150 bp reads.⁶⁷ The data was downloaded from GEO (GSE45428).

To obtain estimated gene and isoform-level abundances, Kallisto (version 0.44.0) was used, with raw reads and the GENCODE reference transcriptome (version 35, GTF file of the comprehensive set, protein-coding genes only) as input.

Mass Spectrometry data collection

Bottom-up proteomic data was previously collected for the multi-protease and trypsin-only data sets.^{48,67} Briefly, cells were cultured and processed with aliquots of approximately 10^7 cells each (6 aliquots for multi-protease digest and 1 aliquot for trypsin digest). Aliquots were lysed in SDT buffer (4% SDS, 500 mM Tris-HCl (pH 7.4) and 180 mM DTT) and approximately 150 μ g of lysate was used for filter-aided sample preparation.⁶⁸ Each aliquot for the multi-protease dataset was digested with a different protease (Arg-C, Asp-N, chymotrypsin, Glu-C, Lys-C, or trypsin), and the trypsin-only aliquot was digested using trypsin. Following digestion, peptides were fractionated off-line by high-pH reverse-phase liquid chromatography (trypsin-only: 28 fractions, multi-protease: 11 fractions–10 fractions for the second trypsin sample) and dried down. Fractions were then reconstituted in 5% acetonitrile and 1% formic acid prior to LC-MS/MS analysis on a nanoACQUITY LC system (Waters, Milford, MA) interfaced with a Thermo Scientific LTQ Orbitrap Velos mass spectrometer. All mass spectrometry raw files are freely available online (multi-protease: <https://massive.ucsd.edu/MSV000083304/>; 28 fraction trypsin: https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View?da

tasetPassword=RE4343upo&identifier=PASS00215\$).

PacBio Iso-Seq data analysis

Raw reads obtained from PacBio Sequel II sequencing were processed into high fidelity (HiFi/CCS) reads using the `ccs` command in SMRTLink. Following CCS read generation, the `lima` command was run to generate full-length reads containing both the 5' and 3' primer. The 5' primer consists of the NEB cDNA sequence (sequence: GCAATGAAGTCGCAGGGTTGGG). The 3' primer consists of the Clontech SMARTer cDNA primer (sequence: GTACTCTGCGTTGATACTACTGCTT). Following `isoseq3 refine` processing, polyA tail sequences are removed. Then, `isoseq3 cluster` is run in order to cluster full-length reads that correspond to the same transcript isoform. This process allows for generation of full-length, non-concatamer (FLNC) reads, which are subjected to further downstream processing, as described below.

The high-quality, polished transcript sequences were mapped to hg38 using `minimap`⁶⁹ (`pbbmm2`, version 1.4.0) with the following parameters

`--preset ISOSEQ -sort`. Finally, `isoseq3 collapse` was run in order to combine redundant reads which were not properly clustered in the `isoseq3 cluster` step.

We recovered the relative abundance of each of the final isoforms in each sample by extracting the number of full-length reads supporting each polished isoform. Full-length counts per million (CPM) were derived by dividing the number of full-length non-chimeric reads aligning to a transcript isoform (i.e., the read became part of the transcript during the isoform clustering step) by the total number of reads and multiplying by a factor of 1,000,000. Only transcripts above one CPM were subjected to further analysis in this study.

Transcript isoform classification and filtering

SQANTI is a computational tool for classification and quality assessment of full-length isoforms sequenced on long-read platforms.²⁸ We used SQANTI3 version 1.3 to annotate the polished transcript isoforms obtained from the Iso-Seq analysis. We used default parameters. Note that this includes the option to use genome-derived sequences for the isoform output; therefore, transcriptional variations (alternative N-termini, alternative splicing, etc.), but not genetic variations, will be captured in the current version of our pipeline.

The inputs for SQANTI3 analysis include the GENCODE version 35 annotations (i.e., GTF file) and the human reference genome (GRCh38, only canonical chromosomes chr1-22, X, Y). The SQANTI3 outputs—isoform and junction classification files—were subjected to additional analysis using custom python scripts, which are part of the Nextflow pipeline.

After running SQANTI3, we filtered out any transcript that was (1) classified as a RT-template switching artifact by SQANTI3, (2) had 95% or higher Adenosine (i.e., polyA) content in 20 nt of the genome immediately downstream of the aligned 3' end of the transcript, indicating a possible dT intra-priming artifact, or (3) did not align to a GENCODE-annotated protein-coding gene (while SQANTI3 does not exclude transcripts based on coding potential, for the purpose of this study, we have excluded them). Finally, we employed a modified version of Cupcake `filter_away_subset.py` (https://github.com/Magdoll/cDNA_Cupcake) to remove 5' transcript degradation products.

Generation of a full-length protein isoform database from long-read RNA-seq

ORF prediction

After deriving a high-confidence set of full-length transcript isoforms, we developed a pipeline for selection of the most biologically plausible canonical ORF for each Iso-Seq transcript (`orf_calling` module in the Nextflow pipeline).

The Coding-Potential Assessment Tool (CPAT) was used to find all candidate open reading frames (ORFs), allowing up to 50 candidate ORFs of 50 nt or longer. The metrics in the CPAT result output (e.g., coding score, which incorporates a hexamer score, ORF length and other metrics) were used for subsequent derivation of a final score for each candidate ORF. Additional information on ATG start codon status was used to generate this final score. For each candidate ORF, the ATG start codon was determined and compared to the GENCODE-annotated ATG start codon. It is difficult to predict the exact ATG start *ab initio* due to lack of a strong consensus sequence for translational initiation sites genome-wide, but the identity of at least some of these sites has been manually curated where literature evidence exists (e.g., HAVANA group, GENCODE). Therefore, any ORF containing an ATG start previously annotated by GENCODE was selected in all cases. In the case that there are multiple ORFs corresponding to two or more GENCODE proteins, we selected the upstream-most ORF. Otherwise, the number of ATGs found upstream of the candidate ORF start site was determined for incorporation into the final scoring metric. Note that this final score employed heavy weighting for ORFs with ATG start sites closer to the 5' end of the PacBio transcript.

Protein database compilation

To generate a PacBio-derived protein database for MS searching, we grouped transcripts that produce ORFs (i.e., proteins) of the same sequence (`refine_orf_database` module in the Nextflow pipeline). Within each transcript grouping, a representative or base PacBio accession was chosen based on alphanumeric sorting. The total transcript abundance for each grouping is the sum of all CPM values for member transcripts.

A FASTA file was generated containing in the accession line the base Iso-Seq accession and gene name. In addition to the FASTA file, a metadata table (`jurkat_orf_refined.tsv`) was generated containing information on the base Iso-Seq accession, all other accession(s) in the same protein sequence group, the gene to which the isoform mapped, and the aggregated CPM.

GENCODE reference protein database

The GENCODE protein database used in this study was created by downloading the protein-coding translation FASTA and grouping entries with the same protein sequence for each gene (see `make_gencode_database` module in the Nextflow pipeline). There are many cases in which one or more GENCODE transcripts from the same gene lead to the same protein sequence. We grouped such cases and defined a representative protein accession as the first alphanumeric GENCODE protein accession, by transcript name (e.g., GAPDH-201).

Cross-mapping of protein isoforms across databases

To compare protein isoform entries across the sample-specific (PacBio-derived) and reference (GENCODE, UniProt) databases, we performed a standard sequence-

alignment-based mapping (see `accession_mapping` module in the Nextflow pipeline). Specifically, a pairwise alignment of all proteins between databases is conducted, tolerating up to two AA mismatches. Up to two AA differences are tolerated since the three databases originate from different sources of genomic or transcript nucleotide sequence. For example, GENCODE protein sequences are derived from the human reference genome, while many UniProt sequences were derived from cDNA sequences. The mapping was done in an iterative manner, in which perfect alignments (i.e., end-to-end match, no AA differences) were first sought and any remaining unmapped entries were compared to the other databases allow for first a single AA and then (if still unmapped) two AA mismatches. Any entries with differing protein lengths or with more than two AA mismatches were considered distinct entries.

Mass spectrometry searching against the PacBio-derived and GENCODE database

Standard proteomic analysis of the tryptic and multi-protease datasets was performed using the free and open-source search software program MetaMorpheus.⁷⁰ A custom branch and docker image of MetaMorpheus was created (GitHub: <https://github.com/smith-chem-wisc/MetaMorpheus/tree/LongReadProteogenomics>, Docker: https://hub.docker.com/r/smithchemwisc/metamorpheus/tags?page=1&ordering=last_updated tag: `lrproteogenomics`) based on MetaMorpheus version 0.0.316 which includes a novel protein inference algorithm termed "Rescue & Resolve". Analysis was performed using either the sample-specific hybrid (PacBio+GENCODE, called "PacBio-Hybrid") database (83,532 protein entries from 19,929 genes; in which the subset of PacBio-derived entries are 35,119 protein entries from 6,653 genes), the GENCODE human database (version 35; 87,729 protein entries from 19,929 genes),

or the UniProt reviewed human database with isoforms (downloaded November 1st, 2020; 42,358 protein entries from 20,292 genes). All searches were conducted with a contaminants database, included in MetaMorpheus, which contains 264 common contaminant proteins frequently found in MS samples.

All RAW spectra files were first converted to MzML format with MSConvert (centroid mode) prior to analysis with MetaMorpheus (see `mass_spec_raw_convert` module in the Nextflow pipeline). For the MetaMorpheus MS search, the settings used for all search tasks can be found in **Appendix III: Table 9.4**. MetaMorpheus produces peptide spectral match (PSM), peptide and protein group result files, which we analyzed in downstream custom modules. Peptide identifications constitute not only the base amino acid sequence but also any post-translational modifications. Two separate peptide identifications may be present for the same base sequence, but exist as the modified and unmodified form. All peptide and protein results reported employ a 1% false discovery rate (FDR) threshold after target-decoy searching.⁷¹

Computational pipeline with Nextflow

We implemented the long-read proteogenomic pipeline in Nextflow, a domain-specific language allowing for the highly flexible development of bioinformatic pipelines capable of being deployed on local machines, servers, or cloud environments.⁷² The ability to create distinct modules for different analyses through containerization (e.g., Docker) is a key benefit of this framework, enabling both the seamless integration of long-read RNA-seq and mass spectrometry analysis workflows and the flexibility to collaborate across research groups. These processes are automatically parallelized for optimal efficiency of compute resources.

We developed a Nextflow pipeline to process PacBio data, convert resulting transcripts into a protein database, and perform proteomics database searching. The

workflow, including all source code, is publicly available on GitHub at <https://github.com/sheynkman-lab/Long-Read-Proteogenomics>. All docker images may be found in the Docker Hub (<https://hub.docker.com/>) under the repository `gsheynkmanlab`.

The analyses were performed on the Lifebit CloudOS platform (link: <https://lifebit.ai/>), and the analysis page is available with the shareable link <https://cloudos.lifebit.ai/public/jobs/60bcb29b303ee601a69d8c74>. The pipeline structure, including details for each module, is included in **Appendix III: Figure 9.2**. Modules can represent a previously established program, a modified program, or a customized script for either processing or analysis. The full details may be found on the Long-Read-Proteogenomics GitHub Wiki page <https://github.com/sheynkman-lab/Long-Read-Proteogenomics/wiki>.

Data analysis and plot generation

All downstream data analyses were performed through custom Python and/or C# scripts. Data analysis scripts used for figure generation may be found in the following GitHub repository: <https://github.com/sheynkman-lab/Long-Read-Proteogenomics-Analysis>.

4.7 Availability of data and materials

All materials, including data used, workflows and analysis notebooks are available in full accordance with the NIH Grants Policy Statement and the Principles and Guidelines for Recipients of NIH Research Grants and Contracts (<https://grants.nih.gov/policy/sharing.htm>). Third-party datasets used in this manuscript include short-read Jurkat RNA-seq data (Gene Expression Omnibus GSE45428) and bottom-up mass spectrometry data for Jurkat cells (PeptideAtlas: PASS00215, ProteomeExchange:

PXD012272). Raw long-read RNA-seq data collected on the PacBio platform are available from the Sequence Read Archive (PRJNA783347, corresponding to accessions SRX13222302 and SRX13222303).

Data generated by both mass spectrometry and long-read RNA sequencing used in the execution of results for this work are available on Zenodo (10.5281/zenodo.5703754). The long-read proteogenomics workflow results generated using the mass spectrometry and long-read RNA-sequencing data are available on Zenodo (<https://doi.org/10.5281/zenodo.5987905>).

The open-source software produced in the making of this work is freely available under the MIT license found in the GitHub repository (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics>). The workflow language used in the generation of the results was Nextflow (<http://nextflow.io>) and the long-read proteogenomics workflow may be found in the repository (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics/main.nf>). A README (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics/blob/main/README.md>) is located in the repository, guiding the user to the Wiki (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics/wiki>) describing each of the pipeline processes (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics/wiki/Pipeline-Processes>) and provides for pipeline vignette (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics/wiki/Pipeline-Vignette>). Test data used in the pipeline vignette and with the GitHub actions run to ensure workflow integrity through continuous testing are available on Zenodo (10.5281/zenodo.5234651).

Code used to generate the main figures and tables in this manuscript can be found in the GitHub repository (<https://github.com/sheynkman-lab/Long-Read-Proteogenomics-Analysis>).

All containers used in the workflow are located in Dockerhub (<https://hub.dock>

er.com/r/sheynkmanlab/long-read-proteogenomics).

4.8 References

- (1) Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* **2013**, *49*, Type: Journal Article, 583–90.
- (2) Tapial, J. et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res* **2017**, *27*, Type: Journal Article, 1759–1768.
- (3) Kelemen, O.; Convertini, P.; Zhang, Z.; Wen, Y.; Shen, M.; Falaleeva, M.; Stamm, S. Function of alternative splicing. *Gene* **2013**, *514*, Type: Journal Article, 1–30.
- (4) Yang, X. et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **2016**, *164*, Type: Journal Article, 805–17.
- (5) Cooper, T. A.; Wan, L.; Dreyfuss, G. RNA and disease. *Cell* **2009**, *136*, Type: Journal Article, 777–93.
- (6) Deveson, I. W.; Brunck, M. E.; Blackburn, J.; Tseng, E.; Hon, T.; Clark, T. A.; Clark, M. B.; Crawford, J.; Dinger, M. E.; Nielsen, L. K.; Mattick, J. S.; Mercer, T. R. Universal Alternative Splicing of Noncoding Exons. *Cell Syst* **2018**, *6*, Type: Journal Article, 245–255 e5.
- (7) Tress, M. L.; Abascal, F.; Valencia, A. Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem Sci* **2017**, *42*, Type: Journal Article, 408–410.
- (8) Blencowe, B. J. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem Sci* **2017**, *42*, Type: Journal Article, 407–408.

- (9) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005**, *4*, Type: Journal Article, 1419–40.
- (10) Mudge, J. M.; Harrow, J. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **2016**, *17*, Type: Journal Article, 758–772.
- (11) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **2014**, *11*, Type: Journal Article, 1114–25.
- (12) Sheynkman, G. M.; Shortreed, M. R.; Cesnik, A. J.; Smith, L. M. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**, *9*, Type: Journal Article, 521–45.
- (13) Carlyle, B. C.; Kitchen, R. R.; Zhang, J.; Wilson, R. S.; Lam, T. T.; Rozowsky, J. S.; Williams, K. R.; Sestan, N.; Gerstein, M. B.; Nairn, A. C. Isoform-Level Interpretation of High-Throughput Proteomics Data Enabled by Deep Integration with RNA-seq. *J Proteome Res* **2018**, *17*, Type: Journal Article, 3431–3444.
- (14) Salowska, B.; Zhu, H.; Gandhi, T.; Frank, M.; Li, W.; Rosenberger, G.; Wu, C.; Germain, P. L.; Zhou, H.; Hodny, Z.; Reiter, L.; Liu, Y. Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation. *Mol Syst Biol* **2020**, *16*, Type: Journal Article, e9170.
- (15) Liu, Y.; Gonzalez-Porta, M.; Santos, S.; Brazma, A.; Marioni, J. C.; Aebersold, R.; Venkitaraman, A. R.; Wickramasinghe, V. O. Impact of Alternative Splicing on the Human Proteome. *Cell Rep* **2017**, *20*, Type: Journal Article, 1229–1241.
- (16) Shanmugam, A. K.; Yocum, A. K.; Nesvizhskii, A. I. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein

- identification by tandem MS. *J Proteome Res* **2014**, *13*, Type: Journal Article, 4113–9.
- (17) Wang, X.; Slebos, R. J.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J Proteome Res* **2012**, *11*, Type: Journal Article, 1009–17.
- (18) Jeong, S. K.; Kim, C. Y.; Paik, Y. K. ASV-ID, a Proteogenomic Workflow To Predict Candidate Protein Isoforms on the Basis of Transcript Evidence. *J Proteome Res* **2018**, *17*, Type: Journal Article, 4235–4242.
- (19) Agosto, L. M.; Gazzara, M. R.; Radens, C. M.; Sidoli, S.; Baeza, J.; Garcia, B. A.; Lynch, K. W. Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing. *Genome Res* **2019**, *29*, Type: Journal Article, 2046–2055.
- (20) Lau, E.; Han, Y.; Williams, D. R.; Thomas, C. T.; Shrestha, R.; Wu, J. C.; Lam, M. P. Y. Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. *Cell Rep* **2019**, *29*, Type: Journal Article, 3751–3765 e5.
- (21) Kannan, S.; Hui, J.; Mazooji, K.; Pachter, L.; Tse, D. Shannon: An Information-Optimal de Novo RNA-Seq Assembler. *BioRxiv* **2016**, Type: Journal Article, DOI: <https://doi.org/10.1101/039230>.
- (22) Steijger, T.; Abril, J. F.; Engstrom, P. G.; Kokocinski, F.; Consortium, R.; Hubbard, T. J.; Guigo, R.; Harrow, J.; Bertone, P. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **2013**, *10*, Type: Journal Article, 1177–84.
- (23) Van Dijk, E.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **2018**, *34*, Type: Journal Article, 15.

- (24) Sharon, D.; Tilgner, H.; Grubert, F.; Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **2013**, *31*, Type: Journal Article, 1009–14.
- (25) Liu, Y.; Beyer, A.; Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **2016**, *165*, Type: Journal Article, 535–50.
- (26) Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **2019**, *47*, Type: Journal Article, D766–D773.
- (27) Tardaguila, M. et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **2018**, Type: Journal Article, DOI: 10.1101/gr.222976.117.
- (28) Wang, L.; Park, H. J.; Dasari, S.; Wang, S.; Kocher, J. P.; Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **2013**, *41*, Type: Journal Article, e74.
- (29) Sammeth, M.; Foissac, S.; Guigo, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol* **2008**, *4*, Type: Journal Article, e1000147.
- (30) Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtkova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. Alternative isoform regulation in human tissue transcriptomes. *Nature* **2008**, *456*, Type: Journal Article, 470–6.
- (31) Rodriguez, J. M.; Maietta, P.; Ezkurdia, I.; Pietrelli, A.; Wesselink, J. J.; Lopez, G.; Valencia, A.; Tress, M. L. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res* **2013**, *41*, Type: Journal Article, D110–7.
- (32) Hosokawa, H.; Rothenberg, E. V. How transcription factors drive choice of the T cell fate. *Nat Rev Immunol* **2021**, *21*, Type: Journal Article, 162–176.

- (33) Shin, B.; Hosokawa, H.; Romero-Wolf, M.; Zhou, W.; Masuhara, K.; Tobin, V. R.; Levanon, D.; Groner, Y.; Rothenberg, E. V. Runx1 and Runx3 drive progenitor to T-lineage transcriptome conversion in mouse T cell commitment via dynamic genomic site switching. *Proc Natl Acad Sci U S A* **2021**, *118*, Type: Journal Article, DOI: 10.1073/pnas.2019655118.
- (34) Blyth, K.; Cameron, E. R.; Neil, J. C. The RUNX genes: gain or loss of function in cancer. *Nat Rev Cancer* **2005**, *5*, Type: Journal Article, 376–87.
- (35) Sood, R.; Kamikubo, Y.; Liu, P. Role of RUNX1 in hematological malignancies. *Blood* **2017**, *129*, Type: Journal Article, 2070–2082.
- (36) Li, Y. et al. Germline RUNX1 variation and predisposition to childhood acute lymphoblastic leukemia. *J Clin Invest* **2021**, Type: Journal Article, DOI: 10.1172/JCI147898.
- (37) 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*, Type: Journal Article, 621–6.
- (38) Bonifer, C.; Levantini, E.; Kouskoff, V.; Lacaud, G. Runx1 Structure and Function in Blood Cell Development. *Adv Exp Med Biol* **2017**, *962*, Type: Journal Article, 65–81.
- (39) Deslattes Mays, A.; Schmidt, M.; Graham, G.; Tseng, E.; Baybayan, P.; Sebra, R.; Sanda, M.; Mazarati, J. B.; Riegel, A.; Wellstein, A. Single-Molecule Real-Time (SMRT) Full-Length RNA-Sequencing Reveals Novel and Distinct mRNA Isoforms in Human Bone Marrow Cell Subpopulations. *Genes (Basel)* **2019**, *10*, Type: Journal Article, DOI: 10.3390/genes10040253.

- (40) Weatheritt, R. J.; Sterne-Weiler, T.; Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nat Struct Mol Biol* **2016**, *23*, Type: Journal Article, 1117–1123.
- (41) Blakeley, P.; Siepen, J. A.; Lawless, C.; Hubbard, S. J. Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* **2010**, *10*, Type: Journal Article, 1127–40.
- (42) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003**, *75*, Type: Journal Article, 4646–58.
- (43) Pfeuffer, J.; Sachsenberg, T.; Dijkstra, T. M. H.; Serang, O.; Reinert, K.; Kohlbacher, O. EPIFANY: A Method for Efficient High-Confidence Protein Inference. *J Proteome Res* **2020**, *19*, Type: Journal Article, 1060–1072.
- (44) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J Proteome Res* **2010**, *9*, Type: Journal Article, 5346–57.
- (45) Huang, T.; Wang, J.; Yu, W.; He, Z. Protein inference: a review. *Brief Bioinform* **2012**, *13*, Type: Journal Article, 586–614.
- (46) Yang, X.; Dondeti, V.; Dezube, R.; Maynard, D. M.; Geer, L. Y.; Epstein, J.; Chen, X.; Markey, S. P.; Kowalak, J. A. DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* **2004**, *3*, Type: Journal Article, 1002–8.
- (47) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **2011**, *10*, Type: Journal Article, 1794–1805.

- (48) 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*, Type: Journal Article, 3429–3438.
- (49) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res* **2007**, *6*, Type: Journal Article, 3549–57.
- (50) Searle, B. C. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **2010**, *10*, Type: Journal Article, 1265–9.
- (51) Wang, X.; Codreanu, S. G.; Wen, B.; Li, K.; Chambers, M. C.; Liebler, D. C.; Zhang, B. Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. *Mol Cell Proteomics* **2018**, *17*, Type: Journal Article, 422–430.
- (52) Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* **2019**, *15*, Type: Journal Article, e8503.
- (53) Komor, M. A.; Pham, T. V.; Hiemstra, A. C.; Piersma, S. R.; Bolijn, A. S.; Schelfhorst, T.; Delis-van Diemen, P. M.; Tijssen, M.; Sebra, R. P.; Ashby, M.; Meijer, G. A.; Jimenez, C. R.; Fijneman, R. J. A. Identification of Differentially Expressed Splice Variants by the Proteogenomic Pipeline Splicify. *Mol Cell Proteomics* **2017**, *16*, Type: Journal Article, 1850–1863.
- (54) Verbruggen, S.; Gessulat, S.; Gabriels, R.; Matsaroki, A.; Van de Voorde, H.; Kuster, B.; Degroeve, S.; Martens, L.; Van Criekinge, W.; Wilhelm, M.; Menschaeert, G. Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics. *Mol Cell Proteomics* **2021**, *20*, Type: Journal Article, 100076.

- (55) The, M.; MacCoss, M. J.; Noble, W. S.; Kall, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom* **2016**, *27*, Type: Journal Article, 1719–1727.
- (56) Fondrie, W. E.; Noble, W. S. mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *J Proteome Res* **2021**, *20*, Type: Journal Article, 1966–1971.
- (57) Brunet, M. A.; Leblanc, S.; Roucou, X. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. *Exp Cell Res* **2020**, *393*, Type: Journal Article, 112057.
- (58) Chen, J.; Brunner, A. D.; Cogan, J. Z.; Nunez, J. K.; Fields, A. P.; Adamson, B.; Itzhak, D. N.; Li, J. Y.; Mann, M.; Leonetti, M. D.; Weissman, J. S. Pervasive functional translation of noncanonical human open reading frames. *Science* **2020**, *367*, Type: Journal Article, 1140–1146.
- (59) Calviello, L.; Mukherjee, N.; Wyler, E.; Zauber, H.; Hirsekorn, A.; Selbach, M.; Landthaler, M.; Obermayer, B.; Ohler, U. Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods* **2016**, *13*, Type: Journal Article, 165–70.
- (60) Serang, O.; Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat Interface* **2012**, *5*, Type: Journal Article, 3–20.
- (61) Sheynkman, G. M.; Tuttle, K. S.; Laval, F.; Tseng, E.; Underwood, J. G.; Yu, L.; Dong, D.; Smith, M. L.; Sebra, R.; Willems, L.; Hao, T.; Calderwood, M. A.; Hill, D. E.; Vidal, M. ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms. *Nat Commun* **2020**, *11*, Type: Journal Article, 2326.

- (62) Schaffer, L. V. et al. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19*, Type: Journal Article, e1800361.
- (63) Sterne-Weiler, T.; Martinez-Nunez, R. T.; Howard, J. M.; Cvitovik, I.; Katzman, S.; Tariq, M. A.; Pourmand, N.; Sanford, J. R. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res* **2013**, *23*, Type: Journal Article, 1615–23.
- (64) Floor, S. N.; Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *Elife* **2016**, *5*, Type: Journal Article, DOI: 10.7554/eLife.10921.
- (65) Gioia, L.; Siddique, A.; Head, S. R.; Salomon, D. R.; Su, A. I. A genome-wide survey of mutations in the Jurkat cell line. *BMC Genomics* **2018**, *19*, Type: Journal Article, 334.
- (66) Pardo-Palacios, Francisco et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. **2021**, Type: Generic, DOI: 10.21203/rs.3.rs-777702/v1.
- (67) 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*, Type: Journal Article, 2341–53.
- (68) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **2009**, *6*, Type: Journal Article, 359–62.
- (69) Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, Type: Journal Article, 3094–3100.
- (70) 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*, Type: Journal Article, 1844–1851.

- (71) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007**, *4*, Type: Journal Article, 207–14.
- (72) Di Tommaso, P.; Chatzou, M.; Floden, E. W.; Barja, P. P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **2017**, *35*, Type: Journal Article, 316–319.

5 DISCOVERY OF DEHYDROAMINO ACID RESIDUES IN THE CAPSID AND MATRIX STRUCTURAL PROTEINS OF HIV-1

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Knoener, R.A; Benner, B.E.; Frey, B.L.; Shortreed, M.R.; Sherer, N.M.; Smith, L. M. Discovery of Dehydroamino Acid Residues in the Capsid and Matrix Structural Proteins of HIV-1. *Journal of Proteome Research* **2022**, *21*(4), 993–1001. <https://doi.org/10.1021/acs.jproteome.1c00867>.

Copyright © 2022 American Chemical Society.

5.1 Abstract

Human immunodeficiency virus type 1 (HIV-1) remains a deadly infectious disease despite existing antiretroviral therapies. A comprehensive understanding of the specific mechanisms of viral infectivity remains elusive and currently limits the development of new and effective therapies. Through in-depth proteomic analysis of HIV-1 virions, we discovered the novel post-translational modification of highly conserved residues within the viral matrix and capsid proteins to the dehydroamino acids, dehydroalanine and dehydrobutyrine. We further confirmed their presence by labeling the reactive alkene, characteristic of dehydroamino acids, with glutathione via Michael addition. Dehydroamino acids are rare, understudied, and have been observed mainly in select bacterial and fungal species. Until now, they have not been observed in HIV proteins. We hypothesize that these residues are important in viral particle maturation and could provide valuable insight into HIV infectivity mechanisms.

5.2 Introduction

Human immunodeficiency virus type 1 (HIV-1) remains a contagious and deadly infectious disease around the globe despite being heavily studied.^{1,2} Advances in antiretroviral therapies have enabled the control of viremia, limiting disease progression and host transmission, but have been unable to completely eradicate latent viral reservoirs.³ Research enhancing the understanding of the viral mechanisms underlying HIV pathogenicity, latency, and persistence is critical for the development of new and effective therapeutics.

Matrix and capsid proteins play critical roles in virion assembly, maturation, and infectivity, making them important therapeutic targets. Immature HIV virions contain

two copies of the HIV RNA genome surrounded by the structural viral polyproteins Gag and Gag-Pol.^{1,4-6} Efficient maturation of viral particles, enabling infectivity, is dependent on the cleavage of these polyproteins into their protein products, including the matrix and capsid proteins.^{4,7} The matrix structural protein is critical to most stages of the viral life cycle and is integral for targeting of the viral genome to the plasma membrane.^{1,5,8} Most matrix protein molecules are associated with the membrane, forming a protective and stabilizing shell, while others associate with the viral capsid core.^{5,8} The capsid structural protein forms the viral core.^{4,9,10} Approximately 1,500 capsid structural protein monomers assemble into 240 hexameric and 12 pentameric subunits, generating a fullerene cone that encapsulates the viral genome and other associated proteins necessary for viral replication in the host.^{4,9-13} Proper assembly and disassembly of the capsid core are vital for effective viral replication and infectivity.^{5,9} Characterization of matrix and capsid post-translational modifications (PTMs) is critical to understanding viral mechanisms of action.

Dehydroamino acids are noncanonical amino acids installed in peptides and proteins post-translationally, either enzymatically or nonenzymatically.¹⁴ Dehydroalanine (DHA) and dehydrobutyrine (DHB) are the most frequently observed dehydroamino acids and are most commonly generated via (a) the dehydration of serine or threonine, (b) the loss of hydrogen sulfide from cysteine, or (c) the elimination of phosphate from serine and threonine, or thiophosphate from cysteine (**Figure 5.1A**).^{14,15} Dehydroalanine can also be generated from the noncanonical amino acid selenocysteine via the elimination of H₂Se.¹⁶ Some research groups synthetically install dehydroamino acids into peptides and proteins to leverage their chemical properties for further derivatization.¹⁷⁻¹⁹ Dehydroamino acids are characterized by their α,β -unsaturated carbonyl structure, which contains a highly reactive, electrophilic alkene capable of undergoing numerous reactions such as Michael addition, hydrogenation, and

cycloaddition.^{14,15,20,21} The conformational and chemical reactivity properties of dehydroamino acids can influence the bioactivity of the peptides or proteins in which they are found.¹⁴ These residues can form irreversible inter- or intramolecular cross-links with lysine, cysteine, or histidine, generating aggregates or cyclic peptides.^{14,20-23}

Dehydroamino acids are most common in peptide natural products of bacteria or fungal species¹⁴ but have also been observed in some human proteins such as human serum albumin,²⁴ thyroglobulin,^{17,25,26} and the lens proteins of the eye.²⁰⁻²² Bacterial enzymes such as phospholyases or dehydratases, such as *Shigella* type III effector OspF, can lead to the generation of dehydroamino acids via an enzymatic pathway.¹⁸ However, in humans, there are no obvious orthologs to known phospholyases or dehydratases.²⁷ Because of this, within the human proteome, dehydroamino acids are largely considered to be formed via nonenzymatic pathways. For example, the lens proteins within the human eye are some of the most long-lived proteins in the human body, and the formation of dehydroamino acid residues within these proteins is suspected to occur via hydroxide ion-induced β -elimination reactions that can occur under physiological conditions over time or may be induced by chemical stress caused by UV-light exposure.²⁰ Overall, the exact origin of many dehydroamino acids in proteins is uncertain. Dehydroamino acid residues have not been widely observed in viruses,²⁸⁻³⁰ are largely unstudied, and to the best of our knowledge, have not previously been observed as modifications to HIV viral proteins.

Here, we report the discovery of DHA and DHB residues in the HIV-1 matrix and capsid proteins, including validation of their presence by chemical derivatization. We analyzed HIV-1 virions with mass spectrometry-based proteomics, both with and without glutathione labeling. Glutathione reacts with DHA- or DHB-containing peptides via Michael addition with the DHA or DHB alkene moieties. The glutathione-labeling reaction yielded nine glutathione-modified residues: confirming

the existence of two and three DHA residues within the matrix and capsid proteins, respectively, and four DHB residues within the capsid protein. The sites of the DHA and DHB modifications are highly conserved in various strains of HIV-1; therefore, we hypothesize that their modification to DHA or DHB is important in viral particle maturation and infectivity.

5.3 Methods

Plasmids and Cell Culture

Human embryonic kidney (HEK) 293T cells were obtained from American Type Culture Collection (ATCC, Manassas, VA). Cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum, 1% l-glutamine, and 1% penicillin–streptomycin. All cells were maintained at 37 °C, 50% humidity, and 5% CO₂. The full-length HIV-1 proviral plasmid, wild type (WT), was derived from the pNL4-3 molecular clone, where both *env* and *vpr* reading frames contain inactivating point mutations in addition to the expression of a GFP reporter from the *nef* reading frame.

Retroviral Assembly and Virion Preparation

HEK293T cells were plated as a monolayer into 10 cm tissue culture plates (Genesee Scientific, San Diego, CA) prior to transfection. For each plate, 10 µg of plasmid DNA encoding the WT virus was transfected into cells using polyethylenimine ((PEI) Polysciences Inc., Warrington, PA) and Gibco Opti-MEM (ThermoFisher Scientific, Waltham, MA). Twenty-four hours post transfection, each well was supplemented with fresh medium. Forty-eight hours post transfection, supernatants were harvested

and passed through a 0.45 μm filter. Virions were concentrated and pelleted through a 20% (wt/vol) sucrose cushion and subjected to centrifugation for 2 h at 21,130g and at 4 °C. Pelleted virions were washed 3x with 10 mL phosphate-buffered saline (PBS), followed by centrifugation for 10 min at 21,130g and at 4 °C. Pelleted virions were flash-frozen in liquid nitrogen prior to downstream processing for mass spectrometry.

Proteomic Sample Preparation

Pelleted virions were lysed in 100 μL of lysis buffer (0.1% RapiGest SF and 8M Urea) prior to sonication on ice for a total of 2 min, cycling between 30 s of sonication and 30 s of rest. Samples were reduced by the addition of 5 μL of reducing buffer (200 mM TCEP in 25 mM ammonium bicarbonate (ABC)) and incubated at room temperature for 1 h. Alkylation of reduced cysteine residues was performed by the addition of 10 μL of alkylation buffer (200 mM IAA in 25 mM ABC) to the sample and incubation in the dark for 1 h. The lysate was diluted with 9 volumes (900 μL) of 50 mM ABC (pH.8.5) to reduce urea concentration to a level compatible with tryptic digestion. The sample was then divided into two 500 μL aliquots (glutathione (GSH)-labeled and unlabeled). To the GSH-labeled sample, 100 μL of 600 mM GSH in 50 mM ABC (pH 8.5) was added, making the final concentration of GSH 100 mM in solution. To the unlabeled sample, 100 μL of 50 mM ABC was added. Both samples were allowed to incubate at 37 °C for 3 h prior to the addition of 1 μg trypsin. Digestion was allowed to proceed overnight (17 h) at 37 °C.

Following digestion, both labeled and unlabeled samples were acidified with TFA to reach a final concentration of 0.5% TFA and an approximate pH of 1 to facilitate cleavage of the RapiGest SF detergent molecule into an insoluble fraction and MS-compatible fraction. Acidified samples were heated at 37 °C for 30 min followed by 10 min of centrifugation at 13,000 rpm. The supernatant was transferred to a new

microcentrifuge tube and dried using a SpeedVac concentrator. Samples were reconstituted in 100 μL of 95:5 H_2O :ACN 0.1% TFA for desalting using a C18 solid-phase extraction pipette tip (OMIX C18, 100 μL , Agilent Technologies). Desalted samples were dried using the SpeedVac concentrator and reconstituted in 95:5 H_2O :ACN 0.1% TFA for mass spectrometric analysis.

Mass Spectrometry of Peptides

Samples were analyzed using a UPLC-MS/MS system consisting of an Easy-nLC 1200 ultra-high-pressure liquid chromatography system and an Orbitrap Fusion Lumos mass spectrometer (ThermoFisher Scientific). Peptides were loaded in buffer A (H_2O , 0.2% formic acid) at a pressure of 300 Bar onto a 20 cm long fused silica capillary nanocolumn packed with C18 beads (1.7 μm diameter, 130 Angstrom pore size from Waters). Peptides eluted over 120 min at a flow rate of 350 nL/min with the following gradient, where buffers consist of A (H_2O , 0.2% formic acid) and B (80% acetonitrile, 0.2% formic acid): time 1 min, 5% buffer B; time 52 min, 30% buffer B; time 80 min, 42% buffer B; time 90 min, 64% ACN; time 95–100 min, 85% buffer B; time 101–120 min, equilibrate at 0% buffer B. The nanocolumn was held at 60 $^\circ\text{C}$ using a column heater (in-house-constructed).

The nanospray source voltage was set to 2,200 V. Full-mass profile scans were performed in the orbitrap between 375 and 1,500 m/z at a resolution of 120,000, followed by MS/MS HCD scans in the orbitrap of the highest intensity parent ions in a 3 second cycle time at a 30% relative collision energy and a 15,000 resolution, with a 2.5 m/z isolation window. Charge states 2–6 were included, and dynamic exclusion was enabled with a repeat count of one over a duration of 30 s and a 10 ppm exclusion width, both low and high. The AGC target was set to “standard”, the maximum inject time was set to “auto”, and 1 microscan was collected for the MS/MS orbitrap HCD

scans.

Spectra files for all four replicates, labeled and unlabeled, can be accessed on MassIVE with the following identifier (MSV000088220).

Data Analysis

Spectral files were analyzed with the free and open-source search software program MetaMorpheus (version 0.0.319, <https://github.com/smith-chem-wisc/MetaMorpheus>). Labeled and unlabeled samples were searched separately but using the same conditions. Since HIV virions contain both human and viral proteins, multiple search databases were utilized. The Swiss-Prot human XML (canonical) database containing 20,380 protein entries (downloaded from UniProt 4/12/2021) was utilized. For HIV, the human immunodeficiency virus type 1 group M subtype B (isolate HXB2) database was downloaded from UniProt and amino acid differences between the HXB2 strain and the strain used for analysis were made manually. Additional entries were generated for the protein cleavage products of Gag and Gal-pol. All viral protein sequences can be found in **Appendix IV: Table 10.4**.

Global post-translational modification discovery (GPTMD), within MetaMorpheus, was used to identify post-translational modifications that are not annotated in the reference database and may constitute novel PTM identifications. GPTMD enables numerous modifications to be searched for concurrently, without the same deleterious false discovery rate (FDR) implications that would occur if a variable modification search strategy was employed.^{31,32} A custom class of modifications was added to GPTMD to enable the identification of peptides containing glutathione-labeled dehydroamino acids. Since the discovered dehydroamino acid residues are not present in the reference database, the mass shift searched for by GPTMD for the labeled corresponds to the total mass shift that would occur, going from an unmodified

serine, cysteine, or threonine residue to a glutathione-labeled dehydroalanine or dehydrobutyrine residue (+289.073, +273.096, and +289.073 Da, respectively). GPTMD was performed separately for the HIV and human database (all modifications selected for GPTMD can be found in **Appendix IV: Table 10.5**). Raw mass spectral files were then searched against the GPTMD HIV and GPTMD human databases, as well as the internal contaminant database included with MetaMorpheus. MetaMorpheus combines the three databases and searches them together in a single search (search task settings can be found in **Appendix IV: Table 10.6**), with carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as a variable modification. All of the protein sequences from the searched databases were reversed to form the decoy database used to calculate FDR. Search results can be accessed with Zenodo (<https://doi.org/10.5281/zenodo.5838422>).

5.4 Results

Our lab has developed a proteomic search software program, MetaMorpheus, which enables the highly confident identification of post-translational modifications from bottom-up mass spectrometry (MS) data (<https://github.com/smith-chem-wisc/MetaMorpheus>). Global post-translational modification discovery (GPTMD) analysis, within MetaMorpheus, can reveal PTMs previously unannotated in the search database by identifying peptide spectral matches (PSMs) whose precursor mass differs from its most likely theoretical peptide match by that of a PTM.^{31,32} DHA and DHB from serine and threonine would both have a mass shift of -18.01 Da, and DHA from cysteine would have a mass shift of -33.987 Da. When these mass shifts, corresponding to potential DHA and DHB modifications, are observed in an initial search of a peptide mixture, the potentially modified candidate peptides are then

added to the search database, which is employed in a second and final search. The use of this search tool for the analysis of the HIV-1 viral proteome revealed the presence of possible dehydroamino acid residues within the capsid and matrix structural proteins, which had not previously been identified in HIV viral proteins.

Bottom-up mass spectrometry-based proteomic analysis was performed on four biological replicates of HIV-1 virions. These virions were isolated from HEK293T cells transfected with a HIV-1 proviral plasmid derived from the pNL4-3 molecular clone, rendered biosafe due to inactivating point mutations in both the *env* and *vpr* reading frames. Proteomic data analysis, including GPTMD followed by basic database search, was carried out using MetaMorpheus (see the Methods section for details). Search results yielded 169,860 peptide spectral matches (PSMs) and 3,742 protein identifications from both the human host and HIV-1 viral proteomes at a 1% false discovery rate (FDR) (see **Appendix IV: Table 10.2**). A total of 1,953 PSMs from the host and viral proteomes were identified as containing possible dehydroalanine or dehydrobutyrine residues; these PSMs were filtered to both a 1% FDR and a 1% posterior error probability (PEP) q -value. The capsid and matrix viral structural proteins accounted for 54% of all PSMs containing putative dehydroamino acids. Additionally, the capsid and matrix proteins exhibited a greater relative abundance of dehydroamino acid-containing PSMs compared to most host proteins, as determined by dividing the number of dehydroamino acid-containing PSMs by the total number of PSMs for that protein. Specifically, 4% of all capsid PSMs and 11% of all matrix PSMs contained a putative dehydroamino acid residue, compared to an average of 1.8% for the 100 most abundant host proteins (by PSM count) containing putative dehydroamino acid residues (see **Appendix IV: Table 10.3**). These results indicate the enrichment of dehydroamino acids in the viral capsid and matrix proteins relative to the rest of the human proteome, suggesting that these residues may constitute an

important HIV-1 post-translational modification.

The reproducibility of the identification and localization of the putative dehydroamino acid residues within the capsid and matrix proteins is very important to the confidence of their discovery. Four biological replicates of HIV virions were analyzed, and a total of 1,058 dehydroamino acid-containing PSMs that map to either the capsid or matrix protein were identified. These PSMs were attributed to 133 distinct peptide identifications (peptides with no ambiguity and a unique combination of PTMs and base amino acid sequence (1% FDR, 0% group FDR, and 1% PEP q -value)). Eighty-one of these modified peptides were found in at least 2 biological replicates and are therefore considered reproducible (see **Appendix IV: Section 10.1**). Although a peptide may only be identified in two biological replicates, the dehydroamino acid residue it identified can also be found in other peptides. These 81 peptides support the putative existence of 29 dehydroamino acid residues (16 DHA and 13 DHB), with 28 of the 29 being identified in 3 or more biological replicates.

The existence of the candidate post-translationally installed dehydroamino acid residues was confirmed by the derivatization of lysate aliquots, from each of the four biological replicates, with glutathione. Accessible dehydroamino acids, containing the characteristic electrophilic double bond of the α,β -unsaturated carbonyl, will undergo Michael addition with glutathione, leading to labeled peptides with a large distinctive mass shift of +307.32 Da from the dehydroamino acid (**Figure 5.1B**). Confirmation of the existence of the putative dehydroamino acid residues was required for two main reasons. First, proteomic searches for infrequent PTMs in complex samples are susceptible to false positive identifications. False positive identifications are always a concern in proteomic data analysis, and their likelihood increases when previously unannotated PTMs are investigated.³² Specific labeling with glutathione of the reactive alkene present in dehydroamino acids would confirm

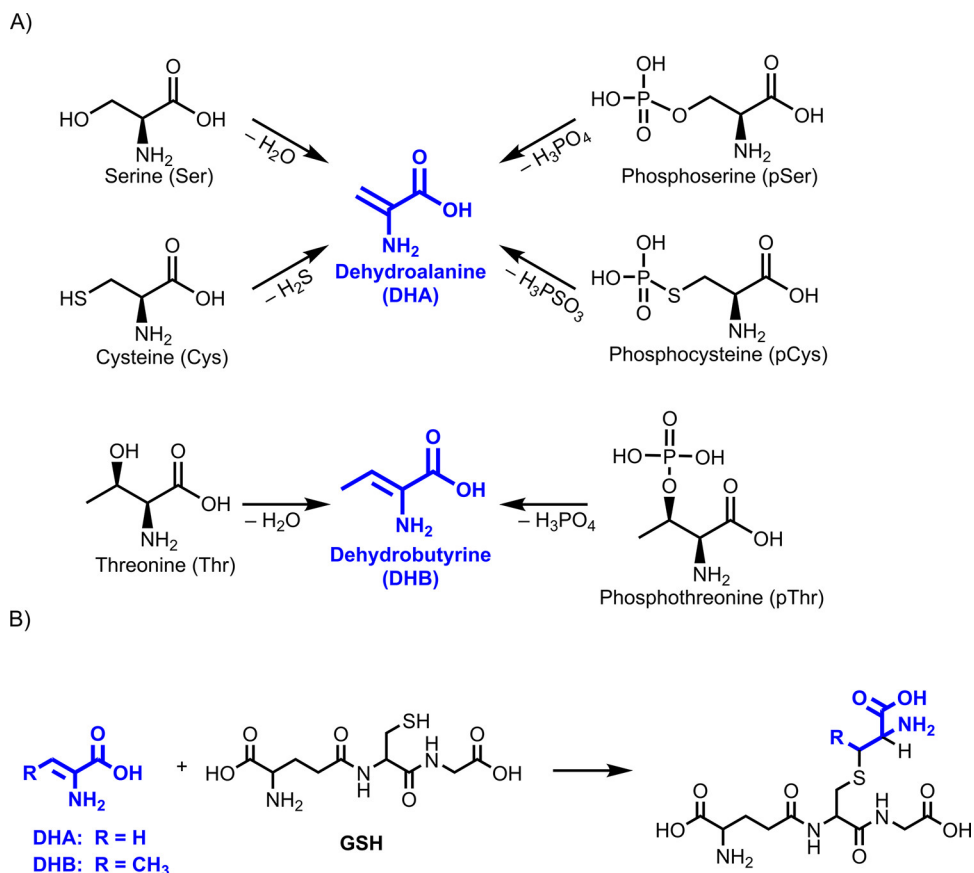


Figure 5.1: Schemas for dehydroamino acid generation and labeling. (A) Most common routes to dehydroalanine (DHA) and dehydrobutyryne (DHB) formation. Dehydroalanine (DHA) can be generated via the dehydration of serine, the elimination of hydrogen sulfide from cysteine, or the elimination of phosphate/thiophosphate from either phosphoserine or phosphocysteine. Dehydrobutyryne (DHB) can be generated via either the dehydration of threonine or the elimination of phosphate from phosphothreonine. (B) Schema of labeling DHA and DHB with glutathione for a distinct mass shift of +307.32 Da from the dehydroamino acid. Labeling of dehydroalanine from serine or dehydrobutyryne from threonine gives a net mass shift of +289.07 Da relative to the originating serine or threonine residue. Labeling of dehydroalanine from cysteine gives a net mass shift of +273.096 Da relative to the originating cysteine residue.

their identities. Second, the -18 Da mass shift indicative for DHA and DHB from serine and threonine, respectively, correlates generally to the loss of water, which could arise from multiple sources such as other water losses on the peptide either already present in the sample itself or induced during the preparation process. Labeling with glutathione not only confirms that the characteristic alkene of dehydroamino acids was present but also that the mass shift of -18 Da did not occur through an alternative mechanism. MetaMorpheus search results of glutathione-treated samples identified 33 glutathione-labeled capsid and matrix peptides (1% FDR, 0% group FDR, and 1% PEP q -value) with no ambiguity regarding their base amino acid sequence or the localization of post-translational modifications, 18 of which were present in two or more biological replicates. These 18 labeled peptides confirmed 9 dehydroamino acid residues in total, 2 in the matrix protein and 7 in the capsid protein (see **Table 5.1**). Representative annotated spectra for peptide identifications confirming the DHA residue at cysteine 86 in the matrix protein are shown in **Figure 5.2**. Annotated spectra for the PSM with the most matching fragment ions for the other dehydroamino acid residues can be found in **Appendix IV: Figures 10.1 to 10.8**. Since the elimination of phosphate from phosphorylated amino acids is one potential route to DHA/DHB formation (see **Figure 5.1A**), it is interesting to note that for all seven of the DHA/DHB residues confirmed in the capsid protein, the phosphorylated form of the precursor amino acid at that position was also identified in at least two biological replicates (**Table 5.1**). All nine confirmed dehydroamino acid modifications occur at highly conserved serine, threonine, or cysteine residues across various HIV strains, indicating their potential functional significance (**Table 5.1**).³³

We believe the confirmed dehydroamino acid residues to be of biological origin, not generated during the sample preparation process. During the entire sample preparation process, conditions were avoided which could lead to the artifactual

Table 5.1: Confirmed Dehydroamino Acids and Their Properties

protein	dehydroamino acid	residue # in cleaved protein	residue # in Gag	original residue	percent conserved (%)	phosphorylated residue identified
matrix	DHA	56	57	cysteine	97.1	no
matrix	DHA	86	87	cysteine	85.9	no
capsid	DHA	16	148	serine	94.7	yes
capsid	DHB	119	251	threonine	99.4	yes
capsid	DHB	188	320	threonine	98.8	yes
capsid	DHA	198	330	cysteine	98.8	yes
capsid	DHB	210	342	threonine	79.4	yes
capsid	DHB	216	348	threonine	94.1	yes
capsid	DHA	218	350	cysteine	98.8	yes

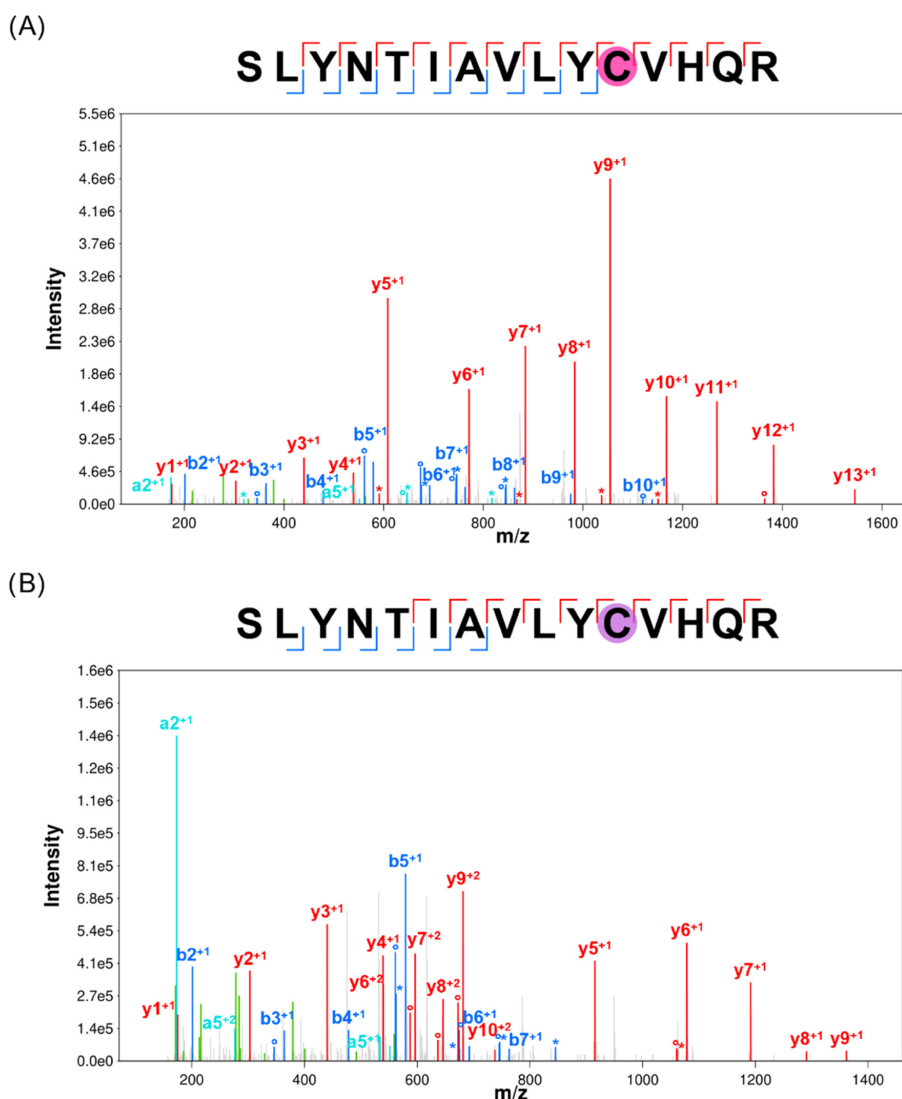


Figure 5.2: Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 86 in the HIV matrix protein. (A) DHA and (B) glutathione-labeled DHA, which is normally a cysteine in an unmodified peptide. Identified y-ions are in red, b-ions are in blue, a-ions are in cyan, and internal ions are in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHA modification is highlighted by the pink circle, and the site of the glutathione-labeled DHA modification is highlighted by the purple circle. In panel A, y-ions 5–13 all confirm the DHA modification and are shifted by -33.987 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, y-ions 5–10 all confirm the glutathione-labeled DHA modification and are shifted by $+273.096$ Da relative to the theoretical m/z of the peaks of an unmodified peptide and by $+307.32$ Da relative to those of the DHA-modified peptide.

generation of dehydroamino acids, such as heavily basic conditions ($\text{pH} > 10$) or prolonged high heat. Post sample preparation, it is possible for dehydroamino acids to be generated in the source of the mass spectrometer. We know this is not the case for our confirmed dehydroamino acids for two reasons. First, if the dehydroamino acid were generated in the source, labeling with glutathione would not have occurred. Second, the normal and dehydroamino acid-modified peptides would have the same retention time, which is not the case (the retention times for all DHA/DHB-modified PSMs and their unmodified counterparts can be found in Supporting File Table S2).

The matrix and capsid proteins have been heavily studied due to their critical importance to the HIV virus. We investigated the existing literature to determine if the sites of any of the confirmed dehydroamino acid residues had been shown to have an impact on virion production or viral infectivity. We found, for the nine confirmed sites of dehydroamino acid modifications, studies examining the effects of site-specific mutagenesis.^{6,34-41} For all but one of the modified residues, mutagenesis had a detrimental effect on virion production or infectivity.

Both the HIV matrix and capsid proteins are composed of an N-terminal domain (NTD) and a C-terminal domain (CTD). The HIV matrix protein includes five α -helices, a 3_{10} helical stretch, and a three-strand mixed β -sheet.⁴² The N-terminal domain (NTD) of the matrix protein is composed of α -helices 1-4 and the 3_{10} helix, which are tightly packed together forming a globular structure.⁴² The C-terminal domain (CTD) is composed of the final α -helix and the β -sheet, which protrudes out from the NTD.⁴² The DHA residues present in the matrix protein at residues 56 and 86 are within the NTD in helices 3 and 4, respectively (**Figure 5.3A**).³⁴ Site-specific mutagenesis studies targeting these sites did show that the mutation of these cysteine residues had an impact on HIV-1 infectivity.^{34,35} The mutation of cysteine 56 to serine resulted in no virion production and no infectivity, likely creating a defect during

either assembly or release of the virions.^{34,35} This mutated form of the matrix protein, when purified, was shown to be completely unfolded, indicating the importance of the cysteine 56 residue to the matrix protein tertiary structure.³⁴ This phenomenon is of particular interest, as dehydroalanine residues can provide structural integrity through intramolecular cross-links. In contrast to the effect of mutating cysteine 56, the mutation of cysteine 86 to serine does not have as significant of an impact on the HIV life cycle and the mutated product supports normal virion production levels. However, it does exhibit slowed viral replication and reduced infectivity.^{34,35}

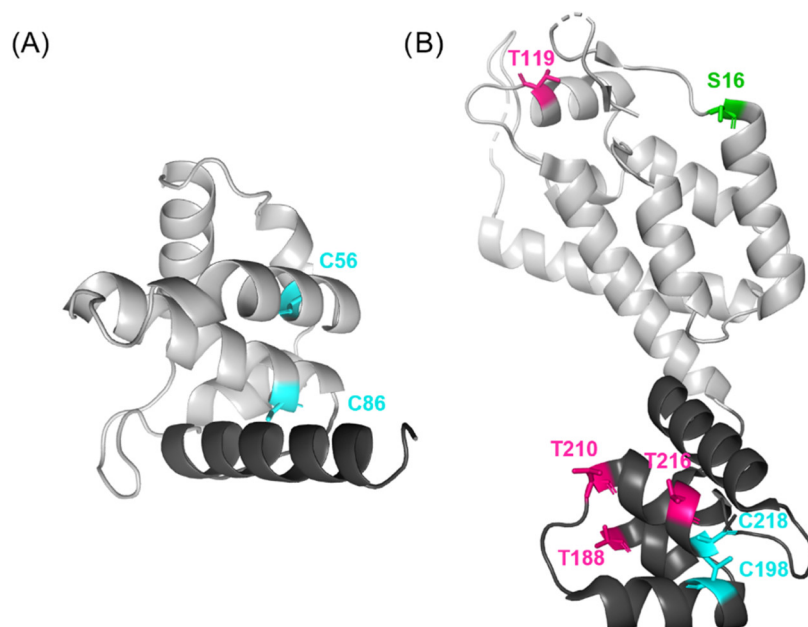


Figure 5.3: Tertiary structure of the HIV-1 matrix and capsid proteins. (A) Structure of the matrix protein monomer (PDB: 1HIW). (B) Structure of the capsid protein monomer (PDB: 3H47). N-terminal domains of both proteins are in light gray, and C-terminal domains are in black. Residues with confirmed dehydroamino acid modifications are highlighted in color, with DHA from serine and cysteine in green and cyan, respectively, and dehydrobutyrine from threonine in pink.

The NTD of the capsid protein is composed of an N-terminal β -hairpin, seven α -helices, and an extended loop.³⁶ The NTD is connected to the CTD, which is composed of four smaller α -helices, by a flexible linker region.³⁶ The CTD of the HIV capsid

protein is known to dimerize and further multimerize to drive the formation of the capsid protein hexamer and pentamer structures that compose the fullerene cone-shaped viral core.³⁶ The first confirmed DHA site within the capsid protein sequence is located at residue 16 on a loop between the β -hairpin and the first α -helix. Phosphorylation of the originating serine residue is well established and is thought to play a crucial role in the capsid core disassembly process through interaction with the host peptidyl-prolyl isomerase, Pin1.³⁷ Pin1 mediates the disassembly of the capsid core through an unknown mechanism when phosphorylation of serine 16 is enabled and can interact with serine 16 and the neighboring proline residue.³⁷ Since phosphorylation is a precursor to the formation of dehydroamino acids, it is intriguing to speculate that the generation of DHA from the phosphoserine residue may play a role in capsid core disassembly. The mutation of serine 16 to proline results in nonviable virions, and mutation from serine to alanine results in decreased infectivity relative to the wild-type virion.^{37,38}

The DHB on residue 119 resides in helix 6 of the capsid protein, and previous experiments mutating the originating threonine residue to alanine did not show any drastic effects on infectivity or viral core formation.^{6,36} The other five confirmed dehydroamino acids occur in the CTD of the capsid protein (**Figure 5.3B**). The DHB at residue 188 is part of helix 9. The originating threonine residue is thought to participate in capsid–capsid association, and its mutagenesis to alanine has a significant impact on the dimerization of capsid proteins.³⁹ It is interesting to consider the possibility that these contacts are stabilized by intermolecular cross-links with the DHB residue.

Capsid cysteine residues 198 (helix 10) and 218 (immediately after helix 11) each have a confirmed DHA modification and influence the binding of host protein cyclophilin A (CypA). CypA is known to bind to the capsid core at the loop structure

between helices 4 and 5 and to induce a large-scale conformational shift in the capsid protein.⁴⁰ The mutation of cysteine 198 to alanine prevents the CypA-mediated conformational shift from occurring and is associated with the inefficient disassembly of the viral core.⁴⁰ The mutation of the cysteine to alanine at residue 218 greatly reduces CypA binding affinity and is associated with the inefficient assembly of the viral core.⁴⁰ Thus, both residues have been shown to play important roles in viral core formation and stability. The DHB at residue 210 is located within a loop between helices 10 and 11, which appears to be an interaction point for the host protein lysyl-tRNA synthetase (LysRS).⁴¹ The interaction of LysRS with the capsid protein is critical for the specific packaging of tRNA^{Lys3}, which serves as the primer for HIV-1 reverse transcription.⁴¹ Mutation of threonine 210 to alanine drastically decreases the binding affinity of LysRS to the capsid protein.⁴¹ The last confirmed DHB at residue 216 is part of helix 11, and mutation of this threonine residue to alanine decreases the infectivity of the virus and also severely impacts the spreading fitness of the virus relative to the wild type^{38,41} (**Table 5.2**).

In summary, it is evident that the majority of the residues, where dehydroamino acid modifications were confirmed in this study, play critical roles in the viral life cycle. Future work will seek to determine possible mechanistic roles played by these interesting protein modifications.

5.5 Conclusions

We have discovered the presence of multiple DHA and DHB residues in the matrix and capsid HIV viral proteins and confirmed their presence by means of chemical derivatization with glutathione and MS analysis. Many of the residues, where DHA and DHB modifications are present, have been shown to play critical roles

Table 5.2: Summary of HIV Site-Specific Mutagenesis

protein	residue no.	original residue	mutated residue	impact of mutagenesis on HIV
matrix	56	cysteine	serine	results in no virion production and no infectivity. ^{34,35}
matrix	86	cysteine	serine	exhibits slowed viral replication and reduced infectivity. ^{34,35}
capsid	16	serine	proline	produces nonviable virions. ^{37,38}
capsid	16	serine	alanine	results in decreased infectivity. ^{37,38}
capsid	119	threonine	alanine	has no significant impact on infectivity or viral core formation. ^{6,36}
capsid	188	threonine	alanine	impacts dimerization of capsid proteins. ³⁹
capsid	198	cysteine	alanine	prevents conformational shift induced by CypA binding and results in inefficient viral core disassembly. ⁴⁰
capsid	210	threonine	alanine	decreases lysRS building to the viral core. ⁴¹
capsid	216	threonine	alanine	decreases infectivity and spreading fitness. ^{37,41}
capsid	218	cysteine	alanine	results in inefficient viral core assembly and reduces CypA binding affinity. ⁴⁰

in the HIV viral life cycle and infectivity. In future work, we will seek to determine if these modifications give rise to intra- or intermolecular cross-links and explore their possible functional roles within the HIV-1 virion. Additionally, we plan to investigate the mechanism of generation for these modifications within the matrix and capsid proteins. We hope to explore whether they are enzymatically installed, via a phospholyase-like enzyme, or nonenzymatically generated. Once the enzymatic or nonenzymatic pathway generating these modifications has been determined, we will seek to ascertain whether the proteins are modified in the host cell, in the immature HIV-1 virion or in the mature HIV-1 virion.

5.6 References

- (1) Fiorentini, S.; Marini, E.; Caracciolo, S.; Caruso, A. Functions of the HIV-1 matrix protein p17. *New Microbiol* **2006**, *29*, Type: Journal Article, 1–10.
- (2) Li, M. Proteomics in the investigation of HIV-1 interactions with host proteins. *Proteomics Clin Appl* **2015**, *9*, Type: Journal Article, 221–34.
- (3) Donnelly, M. R.; Ciborowski, P. Proteomics, biomarkers, and HIV-1: A current perspective. *Proteomics Clin Appl* **2016**, *10*, Type: Journal Article, 110–25.
- (4) Briggs, J. A.; Simon, M. N.; Gross, I.; Krausslich, H. G.; Fuller, S. D.; Vogt, V. M.; Johnson, M. C. The stoichiometry of Gag protein in HIV-1. *Nat Struct Mol Biol* **2004**, *11*, Type: Journal Article, 672–5.
- (5) Hikichi, Y.; Takeda, E.; Fujino, M.; Nakayama, E.; Matano, T.; Murakami, T. HIV-1 matrix mutations that alter gag membrane binding modulate mature core formation and post-entry events. *Virology* **2019**, *532*, Type: Journal Article, 97–107.

- (6) Forshey, B. M.; von Schwedler, U.; Sundquist, W. I.; Aiken, C. Formation of a human immunodeficiency virus type 1 core of optimal stability is crucial for viral replication. *J Virol* **2002**, *76*, Type: Journal Article, 5667–77.
- (7) Von Schwedler, U. K.; Stemmler, T. L.; Klishko, V. Y.; Li, S.; Albertine, K. H.; Davis, D. R.; Sundquist, W. I. Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO J* **1998**, *17*, Type: Journal Article, 1555–68.
- (8) Bukrinskaya, A. HIV-1 matrix protein: a mysterious regulator of the viral life cycle. *Virus Res* **2007**, *124*, Type: Journal Article, 1–11.
- (9) Campbell, E. M.; Hope, T. J. HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nat Rev Microbiol* **2015**, *13*, Type: Journal Article, 471–83.
- (10) Ganser, B. K.; Li, S.; Klishko, V. Y.; Finch, J. T.; Sundquist, W. I. Assembly and analysis of conical models for the HIV-1 core. *Science* **1999**, *283*, Type: Journal Article, 80–3.
- (11) Pornillos, O.; Ganser-Pornillos, B. K.; Yeager, M. Atomic-level modelling of the HIV capsid. *Nature* **2011**, *469*, Type: Journal Article, 424–7.
- (12) Cardone, G.; Purdy, J. G.; Cheng, N.; Craven, R. C.; Steven, A. C. Visualization of a missing link in retrovirus capsid assembly. *Nature* **2009**, *457*, Type: Journal Article, 694–8.
- (13) Li, S.; Hill, C. P.; Sundquist, W. I.; Finch, J. T. Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* **2000**, *407*, Type: Journal Article, 409–13.
- (14) Siodlak, D. alpha,beta-Dehydroamino acids in naturally occurring peptides. *Amino Acids* **2015**, *47*, Type: Journal Article, 1–17.

- (15) Chambers, K. A.; Abularrage, N. S.; Hill, C. J.; Khan, I. H.; Scheck, R. A. A Chemical Probe for Dehydrobutyrine. *Angew Chem Int Ed Engl* **2020**, *59*, Type: Journal Article, 7350–7355.
- (16) Ma, S.; Caprioli, R. M.; Hill, K. E.; Burk, R. F. Loss of selenium from selenoproteins: conversion of selenocysteine to dehydroalanine in vitro. *J Am Soc Mass Spectrom* **2003**, *14*, Type: Journal Article, 593–600.
- (17) Chalker, J. M.; Gunnoo, S. B.; Boutureira, O.; Gerstberger, S. C.; Fernández-González, M.; Bernardes, G. J. L.; Griffin, L.; Hailu, H.; Schofield, C. J.; Davis, B. G. Methods for converting cysteine to dehydroalanine on peptides and proteins. *Chemical Science* **2011**, *2*, Type: Journal Article, DOI: 10.1039/c1sc00185j.
- (18) Jones, L. H. Dehydroamino acid chemical biology: an example of functional group interconversion on proteins. *RSC Chemical Biology* **2020**, *1*, Type: Journal Article, 298–304.
- (19) Chandrasekar, J.; Wylder, A. C.; Silverman, S. K. Phosphoserine Lyase Deoxyribozymes: DNA-Catalyzed Formation of Dehydroalanine Residues in Peptides. *J Am Chem Soc* **2015**, *137*, Type: Journal Article, 9575–8.
- (20) Wang, Z.; Lyons, B.; Truscott, R. J.; Schey, K. L. Human protein aging: modification and crosslinking through dehydroalanine and dehydrobutyrine intermediates. *Aging Cell* **2014**, *13*, Type: Journal Article, 226–34.
- (21) Wang, Z.; Schey, K. L. Quantification of thioether-linked glutathione modifications in human lens proteins. *Exp Eye Res* **2018**, *175*, Type: Journal Article, 83–89.
- (22) Friedrich, M. G.; Wang, Z.; Oakley, A. J.; Schey, K. L.; Truscott, R. J. W. Hotspots of age-related protein degradation: the importance of neighboring residues

- for the formation of non-disulfide crosslinks derived from cysteine. *Biochem J* **2017**, *474*, Type: Journal Article, 2475–2487.
- (23) Seebeck, F. P.; Szostak, J. W. Ribosomal synthesis of dehydroalanine-containing peptides. *J Am Chem Soc* **2006**, *128*, Type: Journal Article, 7150–1.
- (24) Bar-Or, R.; Rael, L. T.; Bar-Or, D. Dehydroalanine derived from cysteine is a common post-translational modification in human serum albumin. *Rapid Commun Mass Spectrom* **2008**, *22*, Type: Journal Article, 711–6.
- (25) Gavaret, J. M.; Cahnmann, H. J.; Nunez, J. Thyroid hormone synthesis in thyroglobulin. The mechanism of the coupling reaction. *J Biol Chem* **1981**, *256*, Type: Journal Article, 9167–73.
- (26) Gavaret, J. M.; Nunez, J.; Cahnmann, H. J. Formation of dehydroalanine residues during thyroid hormone synthesis in thyroglobulin. *J Biol Chem* **1980**, *255*, Type: Journal Article, 5281–5.
- (27) Lai, K. Y. et al. LanCLs add glutathione to dehydroamino acids generated at phosphorylated sites in the proteome. *Cell* **2021**, *184*, Type: Journal Article, 2680–2695 e26.
- (28) Chen, Y. R.; Lees-Miller, S. P.; Tegtmeyer, P.; Anderson, C. W. The human DNA-activated protein kinase phosphorylates simian virus 40 T antigen at amino- and carboxy-terminal sites. *J Virol* **1991**, *65*, Type: Journal Article, 5131–40.
- (29) Pepinsky, R. B.; Papayannopoulos, I. A.; Campbell, S.; Vogt, V. M. Analysis of Rous sarcoma virus Gag protein by mass spectrometry indicates trimming by host exopeptidase. *J Virol* **1996**, *70*, Type: Journal Article, 3313–8.
- (30) Perlman, D. H.; Berg, E. A.; O'Connor, P. B.; Costello, C. E.; Hu, J. Reverse transcription-associated dephosphorylation of hepadnavirus nucleocapsids. *Proc Natl Acad Sci U S A* **2005**, *102*, Type: Journal Article, 9020–5.

- (31) 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*, Type: Journal Article, 1844–1851.
- (32) 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*, Type: Journal Article, 1383–1390.
- (33) Davey, N. E.; Satagopam, V. P.; Santiago-Mozos, S.; Villacorta-Martin, C.; Bharat, T. A.; Schneider, R.; Briggs, J. A. The HIV mutation browser: a resource for human immunodeficiency virus mutagenesis and polymorphism data. *PLoS Comput Biol* **2014**, *10*, Type: Journal Article, e1003951.
- (34) Cannon, P. M.; Matthews, S.; Clark, N.; Byles, E. D.; Iourin, O.; Hockley, D. J.; Kingsman, S. M.; Kingsman, A. J. Structure-function studies of the human immunodeficiency virus type 1 matrix protein, p17. *J Virol* **1997**, *71*, Type: Journal Article, 3474–83.
- (35) Freed, E. O.; Orenstein, J. M.; Buckler-White, A. J.; Martin, M. A. Single amino acid changes in the human immunodeficiency virus type 1 matrix protein block virus particle production. *J Virol* **1994**, *68*, Type: Journal Article, 5311–20.
- (36) Von Schwedler, U. K.; Stray, K. M.; Garrus, J. E.; Sundquist, W. I. Functional surfaces of the human immunodeficiency virus type 1 capsid protein. *J Virol* **2003**, *77*, Type: Journal Article, 5439–50.
- (37) Dochi, T.; Nakano, T.; Inoue, M.; Takamune, N.; Shoji, S.; Sano, K.; Misumi, S. Phosphorylation of human immunodeficiency virus type 1 capsid protein at serine 16, required for peptidyl-prolyl isomerase-dependent uncoating, is mediated by virion-incorporated extracellular signal-regulated kinase 2. *J Gen Virol* **2014**, *95*, Type: Journal Article, 1156–1166.

- (38) Rihn, S. J.; Wilson, S. J.; Loman, N. J.; Alim, M.; Bakker, S. E.; Bhella, D.; Gifford, R. J.; Rixon, F. J.; Bieniasz, P. D. Extreme genetic fragility of the HIV-1 capsid. *PLoS Pathog* **2013**, *9*, Type: Journal Article, e1003461.
- (39) Del Alamo, M.; Neira, J. L.; Mateu, M. G. Thermodynamic dissection of a low affinity protein-protein interface involved in human immunodeficiency virus assembly. *J Biol Chem* **2003**, *278*, Type: Journal Article, 27923–9.
- (40) Bon Homme, M.; Carter, C.; Scarlata, S. The cysteine residues of HIV-1 capsid regulate oligomerization and cyclophilin A-induced changes. *Biophys J* **2005**, *88*, Type: Journal Article, 2078–88.
- (41) Kovaleski, B. J.; Kennedy, R.; Khorchid, A.; Kleiman, L.; Matsuo, H.; Musier-Forsyth, K. Critical role of helix 4 of HIV-1 capsid C-terminal domain in interactions with human lysyl-tRNA synthetase. *J Biol Chem* **2007**, *282*, Type: Journal Article, 32274–9.
- (42) Massiah, M. A.; Starich, M. R.; Paschall, C.; Summers, M. F.; Christensen, A. M.; Sundquist, W. I. Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. *J Mol Biol* **1994**, *244*, Type: Journal Article, 198–223.

6 CONCLUSION

6.1 Summary

The focus of this thesis is the development of new tools and methods to improve characterization of the proteome through the incorporation of additional data. Towards this goal, the projects described here highlight the expansion of bottom-up proteomic experiments and data analysis workflows to include additional data from either alternative proteases, long-read RNA-sequencing, or from post-translational modification discovery.

Both **Chapters 2 and 3** focus on the incorporation of alternative proteases to improve proteome characterization. Orthogonal proteases can provide a great deal of additional information by reaching into regions of the proteome considered intractable to tryptic digestions, increasing sequence coverage and decreasing sampling bias. In **Chapter 2**, we discuss our development and implementation of a novel protein inference algorithm which leverages multi-protease data to provide more precise and accurate results relative to what can be achieved with either trypsin alone, or with other protein inference algorithms not optimized for multi-protease data. The accuracy of protein inference results is of critical importance because the protein or protein group identifications made are subsequently used downstream to draw biological conclusions. Our approach helps maximize the benefits of collecting multi-protease data for bottom-up proteomic experiments. In **Chapter 3**, we outline the development and utility of our *in silico* digestion tool, ProteaseGuru, which seeks to facilitate the evaluation of alternative proteases for specific experimental applications. ProteaseGuru provides theoretical peptides, their physiochemical properties, information regarding uniqueness, as well as data visualization to empower users to make informed decisions regarding the incorporation of alternative proteases into their experimental workflows. ProteaseGuru is the most broadly applicable *in silico*

digestion tool because it can help facilitate both very targeted experiments such as those focused on discovering post-translational modifications or amino acid variants, as well as experiments with a broader scope focused on profiling the entire proteome of a single species, or a sample containing multiple species.

In **Chapter 4** we highlight the importance of incorporating transcriptomic data into proteomic analysis for the purpose of generating a sample-specific database. More specifically, this project focused on the benefits of utilizing PacBio long-read sequencing data to create an isoform-centric protein database. We developed a software pipeline which generates sample-specific databases from PacBio long-read sequencing data, facilitates proteomic analysis and provides genome browser tracks for isoform visualization. Utilizing this isoform-centric sample-specific database led to the identification of novel peptides and enhanced overall protein isoform characterization. This work provides a foundation for future studies seeking to integrate long-read transcriptomics with proteomic data for protein isoform identification.

The routine inclusion of post-translational modification discovery in proteomic data analysis workflows is very important because PTMs are widely considered to be one of the largest contributing factors to the dark proteome. Ignoring the presence of PTMs not only limits the coverage and comprehensive nature of the results obtained, but also their biological relevance as PTMs can be critical to understanding complex biological processes. In **Chapter 5**, we describe a very intriguing application of global post-translational modification discovery, which resulted in the identification of multiple dehydroamino acids within the matrix and capsid proteins of HIV. Following their initial discovery, we were able to further confirm their presence through the development of a chemical labeling strategy in which glutathione covalently attaches to the modified residue, creating a distinct and identifiable mass shift. The impact of these modifications on the virus itself is not yet understood, but the residues on

which they were found are known to be critical to the viral life cycle and to infectivity. Without robust PTM discovery, or a validation schema, these interesting and rare PTMs would still be part of the HIV dark proteome.

The new tools and approaches for the integration of additional data for enhanced proteomic analysis described here are widely applicable and can serve as a foundation for others seeking to maximize the information obtained from their bottom-up proteomic experiments. In the remainder of this chapter, I will discuss future ideas extending each one of the projects highlighted in this thesis.

6.2 Leveraging Multi-Protease Data for Proteoform

Inference

On the path to comprehensive characterization of the proteome, a shift towards the identification of proteoforms over proteins and peptides alone is inevitable and critical. However, top-down proteomics is not currently in a position to be able to characterize the entire proteome due to its mass range and sensitivity limitations.¹ Bottom-up proteomics, while unable to identify intact proteoforms directly, has the depth of sensitivity required to characterize an entire proteome more comprehensively. Integration of top-down and bottom-up data can help bridge the gap between what is necessary for proteoform-level characterization of the proteome, and what is achievable by top-down proteomics today.

There are many different manners in which top-down and bottom-up data can be integrated to benefit characterization of the proteome. Schaffer et. al. outlined several of these approaches including the use of bottom-up data to help localize post-translational modifications within proteoforms and utilizing proteoform identifications to help improve bottom-up protein inference.² One area of bottom-up and

top-down integration that I am particularly interested in extending from Schaffer et. al. is the use of bottom-up peptide identifications to infer the presence of proteoforms.² Inferring the presence of proteoforms from peptides of any single protease would be incredibly difficult and likely inaccurate. However, due to the increased sequence coverage and presence of overlapping peptides, multi-protease data can feasibly be used to more accurately infer the presence of proteoforms. The process of developing multi-protease proteoform inference would not be trivial, and many of the same complications that plague protein inference would also be problematic for proteoform inference. Due to the inherent uncertainty conferred by the inference of proteoforms from peptide identifications, inferred proteoforms should be considered more as putative proteoform identifications and not as confident as those obtained directly from top-down or intact mass data. These putative identifications can be utilized to generate a database of theoretical proteoforms present in the sample which can help serve to inform future top-down proteomic experiments and could function as a starting point for the generation of inclusion lists in targeted proteoform analyses.

Development of a multi-protease proteoform inference algorithm would further extend work demonstrated in **Chapter 2** of this thesis. As in **Chapter 2**, the developed algorithm would maximize the utility of the multi-protease data for the inference process. For example, overlapping peptides spanning multiple PTMs can indicate the presence of particular proteoforms. Many different approaches could be taken in the development of the algorithm, mirroring the approaches taken for bottom-up protein inference such as optimistic, parsimonious or probabilistic. An optimistic approach would be the most anti-conservative, in which peptide identifications would be used to infer the maximal set of proteoforms possible. Conversely, the parsimonious approach would be the most conservative, using the set of peptides to infer the minimal set of proteoforms capable of explaining the identified peptides.

Finally, probabilistic modeling would provide likelihood estimates for the existence of different proteoforms given the peptide identifications. I would seek to first develop a parsimonious approach, as it would inherit the most logic from our existing multi-protease protein inference algorithm. Additionally, the use of a conservative approach limits the number of false positive inferences. Once a parsimonious approach has been established, I would then move towards development of a probabilistic model. Using machine learning, a model for proteoform inference could be established through training on current paired bottom-up and top-down data sets.

6.3 Expansion of ProteaseGuru to Include Peptide Detectability Estimates

It is well understood that not all theoretical peptides for a given protein, or proteome will end up being detected via mass spectrometry. There are a large number of highly complex features that go into detection of a peptide, including but not limited to the ionizability of the peptide and the stochastic nature of peptide selection by the mass spectrometer. For a long time, the idea of determining a peptide's detectability, or probability that a peptide will be identified via an LC-MS/MS experiment, has been of interest and highly desirable. A good deal of this interest has resulted from its many potential applications in downstream data analysis processes such as protein inference and protein quantification.³ I believe the inclusion of peptide detectability estimates as part of the experimental planning process can be equally advantageous and could serve as another factor to aid in the determination of which protease or proteases should be utilized based on the sample and goal of the experiment. Expansion of ProteaseGuru to include peptide detectability estimates would greatly expand its functionality. With this expansion, users would not only be able to determine

which proteases span a specific region or PTM of interest for a given protein, but also the likelihood that the specific peptide would be detectable in a complex proteomic mixture.

The determination of peptide detectability is very complex and has a long history. Initially, standard protein mixtures were used to determine “standard” detectability of a peptide. Although these algorithms were novel first steps, these approaches failed to reliably model detectability for peptides outside of the standard mixtures and fell short when applied to digests of complex lysates.³ What is widely desired out of peptide detectability algorithms is an estimate of “effective” detectability.⁴ This means when analyzing a real-world complex mixture of proteins with varying abundances, how likely is it that the peptide of interest will be identified. This is a very difficult question to address due to the multitude of features that contribute to the detectability of a peptide which expands beyond its basic physiochemical properties. This question also takes into account the experimental protocol used, the mass spectrometry platform, the abundance of the peptide and the software used for peptide identification, amongst many other variables.³ It is near impossible to generate an algorithm which can accurately account for all of these factors, but many modern tools attempt to satisfy this by using machine learning on a great deal of publicly available data from multiple organisms, platforms and preparation protocols to generate a model for providing detectability estimates. Many of these tools, such as DeepMSPeptide, have utilized this approach to achieve higher precision and accuracy than what had been possible in the past.⁵

Ideally, we would like to implement an already existing open-source peptide detectability algorithm such as DeepMSPeptide into ProteaseGuru. However, this may not be entirely possible because almost all of the available tools have been trained either exclusively on tryptic data, or have used data from publicly available

repositories, of which the overwhelming majority would be tryptic.^{3,5} This means that the algorithms are heavily biased towards tryptic peptides, and may fail to give accurate peptide detectability values for non-tryptic peptides. This is problematic because one of the critical features of ProteaseGuru is its emphasis on alternative proteases. Having an algorithm which does not function reliably on peptides from these alternative proteases would be entirely unacceptable. To address this, we would first attempt to adapt the existing algorithm to accommodate peptides from alternative proteases. This likely would require re-training the core model using data from alternative proteases, with each protease as equally weighted as possible. Because of the lack of publicly available multi-protease data, generation of additional quality data sets would likely be required. If existing algorithms such as DeepMSPeptide could not be sufficiently adapted, it would be necessary to generate our own peptide detectability algorithm to suit our needs, likely inheriting components from existing tools.

6.4 Integration of PacBio Long-Read Sequencing with Top-Down Proteomics

The composition of the protein database used for analysis is just as important for top-down proteomics as it is for bottom-up proteomics. In top-down proteomics, with the measurable unit being intact proteoforms, there is very little room for error regarding sample and database discrepancies. The entire sequence of the database entry must exactly reflect that of the proteoform in the sample in order to be identified by matching intact masses.

We have previously demonstrated the utility of sample-specific databases for proteoform identifications in Cesnik et. al. in which short-read RNA-seq data was

utilized to make a sample-specific database informed with sequence variants.⁶ We were able to identify several proteoforms containing variants that were not represented in the reference database, and therefore would have gone unidentified.⁶ However, this approach did not consider protein isoforms. As discussed in **Chapter 4**, when it comes to evaluating isoform diversity of a sample, the use of long-read RNA-sequencing enables the identification of intact transcript isoforms eliminating the need for transcript isoform reconstruction required with short-read RNA-sequencing.

It is clear that top-down proteomics and long-read RNA-sequencing are synergistic technologies, both committed to the characterization of intact biomolecules. Extension of our pipeline to include proteoform analysis using PacBio isoform-centric databases is a clear next step in expanding its utility. This integration would enable the identification of proteoforms translated from novel transcript isoforms. These proteoforms would go unidentified when utilizing conventional reference databases, and contribute to the dark proteome. Unfortunately, the limitations of top-down proteomics might restrict the number of novel proteoforms we are able to identify using a PacBio informed protein database. The molecular weight range limitations of top-down proteomics will likely play a role in limiting the number of proteoforms identified from novel spliced transcripts.¹⁷ For smaller proteins amenable for proteoform analysis, less than 30 kDa, alternative splicing is very well characterized.⁸ This means it is likely there are fewer truly novel proteoforms that are identifiable relative to those present in mass ranges not currently amenable to proteoform analysis of complex mixtures. Additionally, the restricted sensitivity of top-down proteomics would further complicate our ability to identify novel proteoforms.¹ As mentioned previously, top-down proteomics can only identify the most abundant proteoforms present in complex mixtures. This is problematic because many of the novel transcript isoforms discovered are not the major isoform for a given gene and are likely lower

abundance on the protein-level.⁹ Although fractionation approaches could be utilized to increase the depth of proteoform coverage, it is likely that many of the novel proteoforms of interest would still be below the abundance threshold necessary for identification. To address this, a more targeted approach to proteoform identifications could be employed. Based on the isoform-centric database, an inclusion list of novel proteoforms could be generated.

6.5 Functional Investigation of Dehydroamino Acids in HIV

In **Chapter 5** of this thesis we described the identification and confirmation of nine dehydroamino acids within the HIV proteome, but their functional significance remains unresolved. Here I will outline what I believe to be first steps towards determining whether or not our discovered dehydroamino acids play an important functional role in HIV.

As part of **Chapter 5**, we looked to previously existing literature surrounding site-specific mutagenesis of our residues of interest, in order to get a sense of whether or not they were believed to be important to the viral life cycle. The information that can be extracted from a results section of a manuscript, written with a different scientific hypothesis in mind, is limited relative to what can be gained from performing the experiment first-hand. Towards this end, the first step on the journey to functional characterization would be to perform our own site-specific mutagenesis experiments with more targeted functional assays. Since four of the nine dehydroamino acids originate from cysteine residues, it would be very important to try and distinguish between any loss of function caused by the loss of the thiol group, and that caused by the loss of the dehydroalanine residue. In an effort to try and parse out these

functional implications, I propose a step-wise site-specific mutagenesis platform. First, I would mutate the cysteine residues to serine residues, since serine is another amino acid precursors for dehydroalanine. Ideally, if the source of dehydroamino acid formation is adaptable to this alteration, the dehydroalanine site would remain intact. We would perform proteomics experiments on the mutants to confirm the retention of the dehydroalanine residues of interest. If the dehydroalanine residue is conserved, we may be able to attribute any functional defect observed to the loss of the thiol functional group. Subsequently, cysteine to alanine mutants would be generated which would represent the loss of both the thiol functional group and the loss of the dehydroamino acid modifications. By comparing the functional differences of these two mutants, we should be able to extract out attributes unique to the loss of the dehydroamino acid residues. In general, the results of our own site-specific mutagenesis experiments would provide strong direction towards what experiments should be pursued further to functionally characterize dehydroamino acids within the HIV proteome.

Alongside our site-specific mutagenesis work, we could also perform a few very targeted proteomic experiments to evaluate the viability of some of our current hypotheses surrounding the function of dehydroamino acids within the capsid and matrix proteins. Based on their chemical properties and elevated reactivity, we believe it is plausible that dehydroamino acids are forming inter or intra-molecular crosslinks with nucleophilic amino acids residues to act as protein-protein interaction stabilizers, and to facilitate structural rigidity of the capsid and matrix proteins. Both the capsid and matrix proteins are known to have very interesting structural biology, but not as much is known as to what drives these interesting formations. For example, capsid protein monomers assemble into 240 hexamers and only 12 pentamers.¹⁰⁻¹⁴ The formation of the 12 pentamers is absolutely critical to enabling the closure of the viral core's

distinct fullerene cone shape. Without the pentamers, capsid monomers form open tubes or small spheres as the viral core. However, it is not understood what drives the formation of these pentamers relative to the much more common hexameric structure. We hypothesize that at least some of the dehydroamino acids identified within the capsid proteins could form intermolecular crosslinks with other capsid protein monomers to generate the pentamers observed in the viral core. We also believe it is feasible that intermolecular crosslinks facilitated by dehydroamino acids could stabilize interactions between hexameric units and between hexameric and pentameric units. We also have a similar hypothesis surrounding the dehydroamino acids found in the matrix protein. Matrix protein monomers assemble into trimers which then aggregate into a hexameric lattice structure.¹⁵ Here, the orientation of the lattice structure is known to change during the maturation process of the virus, moving from a very open structure with large pores, to a more closed structure with small pores.¹⁵ We believe that the formation of intermolecular crosslinks between matrix proteins during the viral maturation process could be the driving force altering the orientation of the lattice. To evaluate whether these hypotheses have any foundation, we can investigate proteomic data in attempts to identify crosslinked peptide pairs that would result from the Michael addition of a nucleophilic amino acid residue from one matrix or capsid protein across the electrophilic double bond of a dehydroamino acid on another matrix or capsid protein, respectively. Identification of crosslinked peptides would not confirm our hypothesis, but would represent a strong first step that could be further investigated using structural biology techniques such as cryoEM with EM-active probes designed to target the crosslinked peptides.

6.6 References

- (1) Schaffer, L. V. et al. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19*, Type: Journal Article, e1800361.
- (2) 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. *Journal of Proteome Research* **2020**, *19*, 3510–3517.
- (3) Li, Y. F.; Arnold, R. J.; Tang, H.; Radivojac, P. The Importance of Peptide Detectability for Protein Identification, Quantification, and Experiment Design in MS/MS Proteomics. *Journal of Proteome Research* **2010**, *9*, 6288–6297.
- (4) Li, Y. F.; Arnold, R. J.; Li, Y.; Radivojac, P.; Sheng, Q.; Tang, H. A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. *Journal of Computational Biology* **2009**, *16*, 1183–1193.
- (5) Serrano, G.; Guruceaga, E.; Segura, V. DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics* **2019**, ed. by Hancock, J., btz708.
- (6) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *J Proteome Res* **2021**, *20*, Type: Journal Article, 1826–1834.
- (7) 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*, Type: Journal Article, 6868–74.
- (8) Melani, R. D. et al. The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science* **2022**, *375*, 411–418.
- (9) Miller, R. M. et al. Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biology* **2022**, *23*, 69.

- (10) Briggs, J. A.; Simon, M. N.; Gross, I.; Krausslich, H. G.; Fuller, S. D.; Vogt, V. M.; Johnson, M. C. The stoichiometry of Gag protein in HIV-1. *Nat Struct Mol Biol* **2004**, *11*, Type: Journal Article, 672–5.
- (11) Campbell, E. M.; Hope, T. J. HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nat Rev Microbiol* **2015**, *13*, Type: Journal Article, 471–83.
- (12) Ganser, B. K.; Li, S.; Klishko, V. Y.; Finch, J. T.; Sundquist, W. I. Assembly and analysis of conical models for the HIV-1 core. *Science* **1999**, *283*, Type: Journal Article, 80–3.
- (13) Pornillos, O.; Ganser-Pornillos, B. K.; Yeager, M. Atomic-level modelling of the HIV capsid. *Nature* **2011**, *469*, Type: Journal Article, 424–7.
- (14) Cardone, G.; Purdy, J. G.; Cheng, N.; Craven, R. C.; Steven, A. C. Visualization of a missing link in retrovirus capsid assembly. *Nature* **2009**, *457*, Type: Journal Article, 694–8.
- (15) Qu, K.; Ke, Z.; Zila, V.; Anders-Össwein, M.; Glass, B.; Mücksch, F.; Müller, R.; Schultz, C.; Müller, B.; Kräusslich, H.-G.; Briggs, J. A. G. Maturation of the matrix and viral membrane of HIV-1. *Science* **2021**, *373*, 700–704.

7 APPENDIX I: SUPPORTING INFORMATION FOR "IMPROVED
PROTEIN INFERENCE FROM MULTIPLE PROTEASE BOTTOM-UP MASS
SPECTROMETRY DATA"

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Millikin, R.J.; Hoffman, 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. *Journal of Proteome Research* **2019**, *18*(9), 3429–3438. <https://doi.org/10.1021/acs.jproteome.9b00330>.

Copyright © 2019 American Chemical Society.

7.1 Supplementary Experimental Methods

Cell Culture

The Jurkat cell line (TIB-152) chosen for this study was obtained from the American Type Culture Collection (ATCC, Manassas, VA). Cells were cultured in 10% Fetal Bovine Serum (FBS) and 90% RPMI medium at 37 °C and grown to a concentration of approximately 1.3×10^6 cells/mL. Six cell aliquots of approximately 3.2×10^7 cells each were centrifuged at $180 \times g$ and 4 °C for 10 minutes. The cell pellets were washed twice with ice-cold PBS buffer. The final cell pellets were flash frozen and stored at -80 °C until needed.

Protein Extraction

Flash frozen Jurkat cell pellets were thawed on ice. Cells were lysed by pipetting the pellet repeatedly with SDT lysis buffer ([4% SDS, 500 mM Tris-HCl (pH 7.4)] and 180 mM dithiothreitol (DTT), added in a 5:1 volume ratio to that of the cell pellet) followed by a 5-minute incubation at 95 °C. Lysate was probe sonicated on ice for 3 to 5 minutes, cycling between 30 seconds sonication and 30 seconds rest.

Filter-Aided Sample Preparation

Approximately 150 µg of protein from each aliquot of lysate was transferred to a 100K Amicon Ultra filter (Millipore, Billerica, MA). A modified FASP protocol was utilized to accommodate differing protease digestion conditions. The original FASP protocol was employed until the final wash with ammonium bicarbonate (pH 7.8)¹; then each filter was washed two additional times with the buffer system appropriate for its specific protease. Protease aliquots were then added to their respective filters.

Optimized digestion conditions for each protease are given in **Table 7.1**. Following digestion, each filter was centrifuged at 14,000xg for 15 minutes to recover digested peptides. The amount of peptide recovered was quantified via Pierce BCA assay (ThermoFisher Scientific).

Table 7.1: Protease-Specific Digestion Condition

Protease	Protein:Enzyme Ratio	Buffer System	Temperature of Digest (°C)	Length of Digest (hrs)
Arg-C	100:1	Incubation Buffer: 50 mM Tris-HCl (pH 7.6), 5 mM CaCl ₂ , and 2 mM EDTA Activation Buffer: 50 mM Tris-HCl (pH 7.6), 50 mM DTT, and 2 mM EDTA Combined Incubation and Activation buffer in 9:1 ratio	37	16
Asp-N	100:1	50 mM sodium phosphate (pH 8.0)	25	16
Chymotrypsin	100:1	100 mM Tris-HCl (pH 8.0) and 10 mM CaCl ₂	25	12
Glu-C	100:1	25 mM Ammonium Bicarbonate (pH 7.8)	25	16
Lys-C	100:1	25 mM Tris-HCl (pH 8.5), 1 mM EDTA, and 4M urea	37	16
Trypsin	50:1	50 mM Ammonium Bicarbonate (pH 7.8)	37	16

Peptide Fractionation

At least 100 µg of each peptide digest was fractionated at high pH on a Shimadzu HPLC using a Phenomenex C18 Gemini 3 µ, 110Å, 3.0 × 150 mm column. The buffers used for separation were 20 mM ammonium formate (pH 10) in water (mobile phase

A, MA), 20 mM ammonium formate (pH 10) in 70% acetonitrile (mobile phase B, MB). The flow rate was 0.5 mL/min and the binary gradient was: 0% MB for 15 minutes, linear ramp to 100% MB over 45 minutes, hold at 100% MB for 5 minutes, linear descent to 0% B over 2 minutes followed by equilibration at 0% MB for 20 minutes. Eleven 1 mL fractions of peptides were collected for all proteases with the exception of the tryptic digest where only 10 fractions were obtained. Fractions were lyophilized via SpeedVac and stored at -80 °C.

LC-MS/MS Analysis

Each lyophilized fraction was reconstituted in 5% acetonitrile and 1% formic acid, followed by chromatography on a nanoACQUITY LC system (Waters, Milford, MA) interfaced with a Thermo Scientific LTQ Orbitrap Velos mass spectrometer (Thermo Fisher, Waltham, MA) using a 20 cm reverse-phase capillary column packed with 3 μm C18 beads. Buffers used were 0.2% formic acid in water (mobile phase A, MA) and 0.2% formic acid in acetonitrile (mobile phase B, MB). Full scans from 300-1,500 m/z were collected at a resolution of 60,000. These MS1 scans were followed by top 10 precursor HCD fragmentation to produce spectra at a resolution of 7,500. Precursor fragmentation repeat count was set to two, and dynamic exclusion was set to 60 s.

7.2 Supplementary Parameter Tables

Table 7.2: MetaMorpheus Search Parameters

Global Search Parameters
<i>Search Mode</i>
Classic Search
<i>In silico Digestion Parameters</i>
Generate target proteins
Generate decoy proteins
Generate reversed decoy proteins
Max Missed Cleavages :2
Initiator Methionine: Variable
Max Modification Isoforms: 1024
Min Peptide Length: 7 (5 for ProteinProphet Comparison)
Max Peptide Length: none
Max mods per peptide: 2
<i>Fragment Ion Search Parameters</i>
Dissociation Type: HCD
Max Threads: 39
Max Fragment Mass (Da): 30000
N-Terminal Ions
C-Terminal Ions
<i>Mass Difference Acceptors</i>
1 Missed Monoisotopic Peak
<i>Ambiguity Parameters</i>
Report PSM ambiguity
<i>Scoring Options</i>
Minimum score allowed: 5
<i>Post-Search Analysis</i>
Apply protein parsimony and construct protein groups

Table 7.3: Modifications for GPTMD

GPTMD Modifications	
<i>Common Biological</i>	
Acetylation on K	HexNAc on T
Acetylation on X (Prot N-Term)	Hydroxybutyrylation on K
ADP-ribosylation on S	Hydroxylation on K
Butyrylation on K	Hydroxylation on N
Carboxylation on D	Hydroxylation on P
Carboxylation on E	Malonylation on K
Carboxylation on K	Methylation on K
Citrullination on R	Methylation on R
Crotonylation on K	Nitrosylation on C
Dimethylation on K	Nitrosylation on Y
Dimethylation on R	Phosphorylation on S
Formylation on K	Phosphorylation on T
Glu to PyroGlu on Q (Prot N-Term)	Pyridoxal phosphate on K
Glutarylation on K	Succinylation on K
HexNAc on Nxs	Sulfonation on Y
HexNAc on S	Trimethylation on K
HexNAc on Nxt	
<i>Common Artifact</i>	
Ammonia loss on C (Pep N-Term)	Carbamyl on R
Ammonia loss on N	Carbamyl on X (Pep N-Term)
Carbamyl on C	Deamidation on N
Carbamyl on K	Deamidation on Q
Carbamyl on M	Water Loss on E (Pep N-Term)
<i>Metal</i>	
Calcium on D	Magnesium on D
Calcium on E	Magnesium on E
Cu[I] on D	Potassium on D
Cu[I] on E	Potassium on E
Fe[II] on D	Sodium on D
Fe[II] on E	Sodium on E
Fe[III] on D	Zinc on D
Fe[III] on E	Zinc on E

Table 7.4: Comet Search Parameters

<i>Search</i>		
Decoy search:1	Peff format:1	Num threads: -1
<i>Masses</i>		
Peptide mass tolerance: 20.00	Mass type fragment: 1	Peptide mass units: 2
Precursor tolerance type: 1	Mass type parent: 1	Isotope error: 1
<i>Search Enzyme</i>		
1- Trypsin, 3- Lys-C, 5- Arg-C, 6- Asp-N, 8- Glu-C & 10- Chymotrypsin		
Num enzyme termini:2 2	Allowed missed cleavage: 2	
<i>Modifications</i>		
variable mod01: 15.9949 M 0 2 -1 0 0	max variable mods in pep- tide: 5	Require variable mod: 0
<i>Fragment Ions</i>		
Fragment bin tol: 0.4	Use C ions: 0	Fragment bin offset: 0.4
Theoretical fragment ions: 1	Use X ions: 0	Use Y ions: 1
Use A ions: 0	Use Z ions: 0	Use B ions: 1
Use NL ions: 0		
<i>Misc. Parameters</i>		
Digest mass range: 600.0-5000.0	Nucleotide reading frame: 0	Num results: 1000
Clip nterm methionine: 0	Skip researching: 1	Spectrum batch size: 0
Max fragment charge: 3 Equals I and L: 1	Decoy prefix: DECOY	Max precursor charge: 6
<i>Spectral Processing</i>		
Minimum peaks: 10	Remove precursor tolerance: 1.5	Minimum intensity: 0
Clear mz range: 0.0 0.0	Remove precursor peak: 0	
<i>Additional Modifications</i>		
Add Cterm peptide: 0.0	Add Q: 0.0000	Add Nterm peptide: 0.0
Add K: 0.0000	Add Cterm protein: 0.0	Add E: 0.0000
Add Nterm protein: 0.0	Add M: 0.0000	Add G: 0.0000
Add O: 0.0000	Add A: 0.0000	Add H: 0.0000
Add S: 0.0000	Add F: 0.0000	Add P: 0.0000
Add U: 0.0000	Add V: 0.0000	Add R: 0.0000
Add T: 0.0000	Add Y: 0.0000	Add C: 57.021464
Add W: 0.0000	Add L: 0.0000	Add N: 0.0000
Add I: 0.0000	Add D: 0.0000	Add B user AA: 0.0000
Add X user AA: 0.0000	Add J user AA: 0.0000	Add Z user AA: 0.0000

Table 7.5: TPP-Protein Prophet Parameters

Input is from iProphet	false
Import XPRESS protein ratios	false
Import ASPARatio protein ratios and pvalues	false
Import Libra protein ratios	false
Do not include zero probability protein entries in output	true
Do not report protein length	false
Report(calculated) protein molecular weight	false
Icat data	false
N-glycosylation data	false
Delude (do not look up ALL proteins corresponding to shared peps)	false
Do not use Occam's razor for shared peps	false
Do not assemble protein groups	false
Normalize NSP using protein length	false
Use expected number of ion instance to adjust the peptide probabilities prior to NSP adjustment	false
Check peptide's Protein Weight against the threshold	false

Table 7.6: ProLuCID Search Parameters

<i>Search Mode</i>	
Primary score type	1-XCorr
Secondary score type	2-zScore
Locus type	0-accession
Charge disambiguation	0
Atomic enrichment	0-no labeling
Min match	5
Peak rank threshold	200
Candidate peptide threshold	500
Num output	5
Is decharged	0
Fragmentation method	CID
Pre process	1-do XCorr like preprocessing
<i>Isotopes</i>	
Precursor	Mono
Fragment	Mono
Num peaks	0
<i>Tolerance</i>	
Precursor high	4500
Precursor low	4500
Precursor mass accuracy	5
Fragment ion mass accuracy	20
<i>Precursor mass limits</i>	
Minimum	600
Maximum	1600
<i>Peptide Length limits</i>	
Minimum	7
<i>Num peak limits</i>	
Minimum	25
Maximum	5000
Max num diff mods	0
<i>Modifications</i>	
Static Mods	C & U mass shift: 57.02146
Diff Mods	M mass shift: 15.9949146
<i>Enzyme Info</i>	
Specificity	2-both ends
Man num internal mis cleavage	2
Name, type and residues	Depends on protease

Table 7.7: DTASelect2 Parameters

Enzyme number	0
Diff search options	16 M
Include peptides regardless of cleavage status	-y0
Peptide FDR	-fp 0.01
Protein FDR	-pfp 0.01
Decoy identifier	-decoy DECOY

7.3 Supplementary Methods for Data Analysis

The implementation of the “integrated” multi-protease protein inference algorithm in MetaMorpheus required that peptide confidence (q -values) be calculated separately for each protease, and that the peptide identification be associated with their protease of origin. These two modifications to MetaMorpheus’ conventional protein inference are explained below.

Peptide q -Values

MetaMorpheus uses q -values as an assessment of confidence in a given identification, describing the minimum FDR threshold at which the peptide would exist within the dataset. Peptide spectral match (PSM) q -values are calculated by ranking the identifications by score, then calculating the ratio of cumulative decoy to targets identifications for each PSM. The length of the peptide is directly correlated to the number of fragment matching opportunities a target or decoy spectra has to the theoretical database, and each protease produces peptides with differing length distributions. In MetaMorpheus, the number of fragment ion matches is used to determine the score of the PSM, therefore target and decoy score distributions differ by protease (**Figure7.1**). PSM scores are subsequently used for ranking prior to q -value calculation. High scoring decoys from one protease should not penalize the q -value of target PSMs from

another protease, hence the need to calculate peptide confidence levels separately for each protease in order to maintain the integrity of the value.

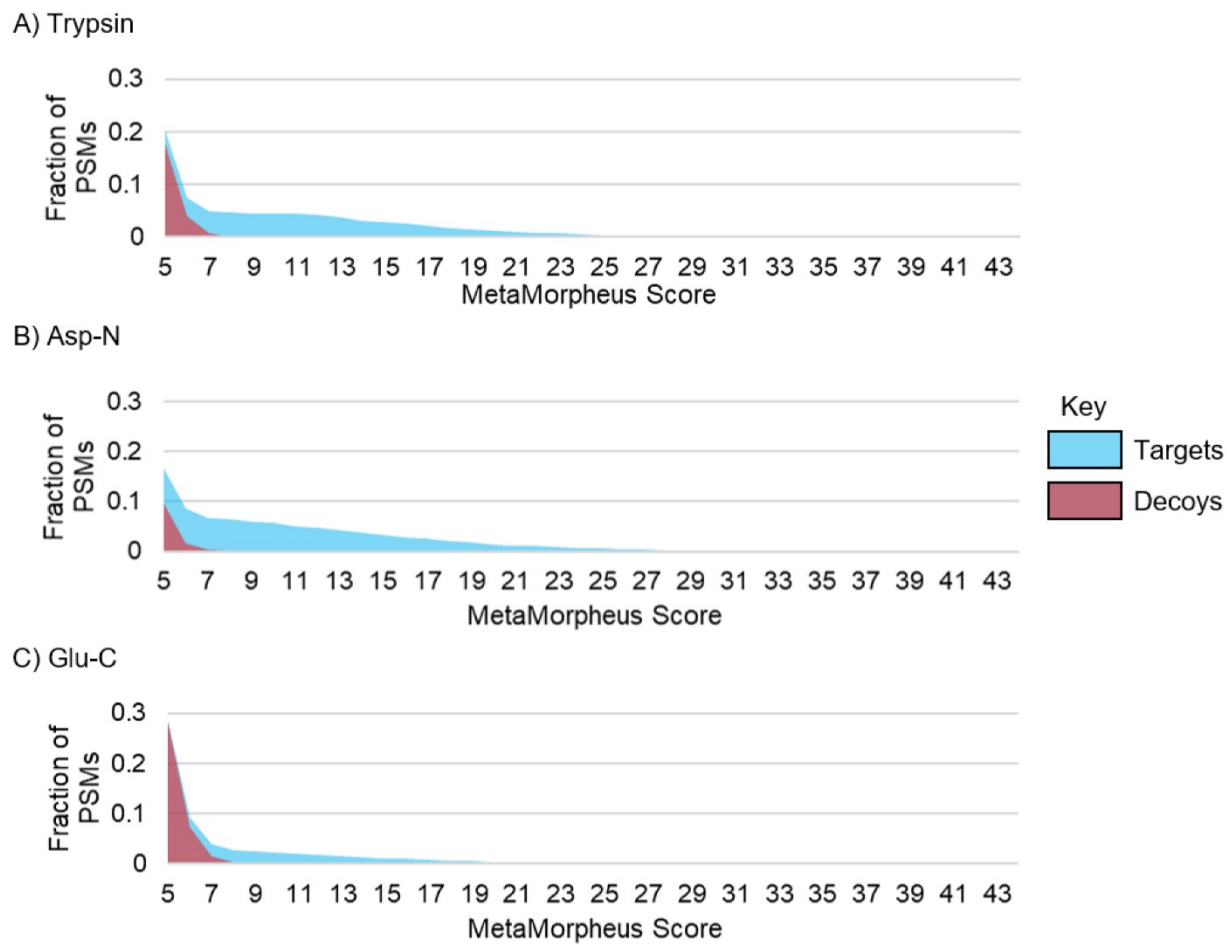


Figure 7.1: Distribution of target and decoy PSM scores resulting from digestion. A) Trypsin, B) Asp-N and C) Glu-C.

Associating Peptides with Their Protease of Origin

Ignoring the sequence-protease relationship can add unnecessary ambiguity to multi-protease protein inference results. The integrated multi-protease protein inference algorithm maintains this relationship, enabling accurate determination of

whether a peptide should be classified as shared or unique, and which proteins the peptide could have potentially originated. A sequence of amino acids could be unique within the proteome for digestion with one protease but shared for another, or be shared among different proteins based on its originating protease (**Figure 7.2**). These situations arise for 5.3% and 12% of the 42,419 proteins present in the UniProt Human Canonical and Isoform reviewed database, respectively. For example, if the peptide “FHSMASR”, from **Figure 7.2**, is identified from spectra resulting from Lys-C digestion, it could have resulted from the digestion of Q86UV7 or Q86UV6. If the protease that resulted in the production of this peptide is unknown, then the number of possible parent proteins increases from 2 to 5 (Q86T4 or Q86XT4-2 or Q86UV6 or Q86UV6-2 or Q86UV7).

7.4 Supplementary Information for MetaMorpheus’ Separate and Integrated Multi-Protease Comparison

The total number of protein groups identified with the separate multi-protease approach is greater than that of the integrated multi-protease approach at a 1% protein FDR threshold (**Table 2.2**). To ensure that this difference in the total number of protein groups does not bias the analysis of the accuracy of the multi-protease protein inference approaches, the results were compared at the total number of protein groups present at the 1% FDR threshold for both the integrated and separate multi-protease approaches (7,472 protein groups and 7,716 protein groups, respectively). The results of this analysis are summarized in **Tables 7.8 and 7.9**. A decrease in the number of *Arabidopsis thaliana* identifications and the corresponding false positive rate was observed for the integrated approach, further indicating that the integrated approach provides more accurate protein group results than the separate approach.

Peptide Sequence: FHSMASR				
Protease	Arg-C	Chymotrypsin	Lys-C	Trypsin
Possible Protein Accessions	Q86XT4-2	Q86UV6-2	Q86UV7 Q86UV6	Q86XT4 Q86XT4-2 Q86UV7 Q86UV6

Key Unique Peptide Shared Peptide

Figure 7.2: Depiction of how the peptide sequence “FHSMASR” can be attributed to different protein groups based on which protease it originated from. Sections in blue indicate that the peptide sequence is unique to a single protein accession for the assigned protease whereas sections in green indicate the same peptide sequence is shared for the assigned protease. This shows how the association of a peptide sequence with its protease of origin can eliminate unnecessary protein group ambiguity.

Table 7.8: Comparison of Entrapment Results of the Top 7,472 Protein Groups from the Separate and Integrated Multi-Protease Approaches.


	Separate Multi- Protease Approach	Integrated Multi- Protease Approach	Percent Change
Number of Human Protein Groups	7,253	7,255	+0.03%
Number of <i>Arabidopsis thaliana</i> Protein Groups	219	217	-0.93%
False Positive Rate	2.93%	2.90%	-1.02%


Table 7.9: Comparison of Entrapment Results of the Top 7,716 Protein Groups from the Separate and Integrated Multi-Protease Approaches.

	Separate Multi- Protease Approach	Integrated Multi- Protease Approach	Percent Change
Number of Human Protein Groups	7,400	7,407	+0.09%
Number of <i>Arabidopsis thaliana</i> Protein Groups	316	309	-2.22%
False Positive Rate	4.10%	4.00%	-2.44%

7.5 Supplementary Figures

Protein Group Reassignment

 The protein group identified was disambiguated to one or more protein groups with only a single protein accession.
Ex. A|B|C → A

 Protein group ambiguity was reduced leading to a protein group with fewer protein accessions.
Ex. A|B|C → B|C

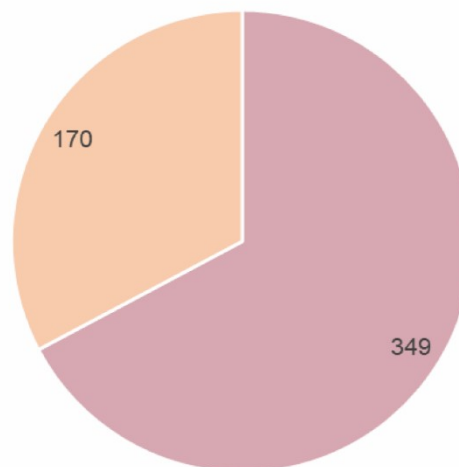


Figure 7.3: Investigation of protein groups unique to the separate approach. A majority of the protein groups unique to the separate approach (519 of 864) were disambiguated into simpler protein groups in the results of the integrated multi-protease approach and can be assigned to two distinct categories: disambiguation to one or more protein groups with a single protein member, or disambiguation to a protein group with fewer protein members. A pie chart representing the distribution of the 519 protein groups into these categories is shown.

Protein Group Reassignment

■ The protein group identified was disambiguated to one or more protein groups with only a single protein accession.
Ex. A|B|C → A

■ Protein group ambiguity was reduced leading to a protein group with fewer protein accessions.
Ex. A|B|C → B|C

■ The protein group identified was disambiguated to a protein group with a single accession as well as to a protein group with fewer protein accessions.
Ex. A|B|C → A and B|C

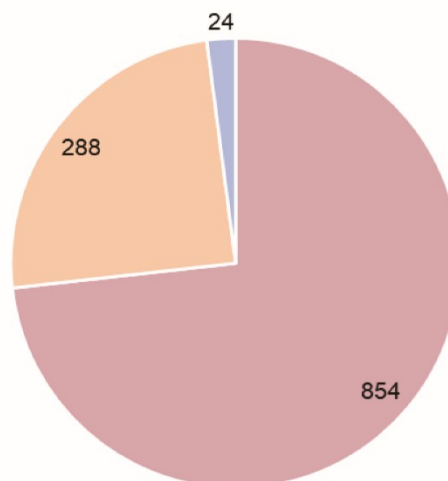


Figure 7.4: Investigation of protein groups unique to the protein inference results of the tryptic digest. Almost all of the protein groups unique to the tryptic digest (1,166 of 1,202) were disambiguated into simpler protein groups in the results of the integrated multi-protease approach could be assigned to three distinct categories: disambiguation to one or more protein groups with a single protein member, disambiguation to a protein group with fewer protein members, or disambiguation to both a protein groups containing a single member and a protein group containing fewer protein members. A pie chart representing the distribution of the 1,166 protein groups into these categories is shown.

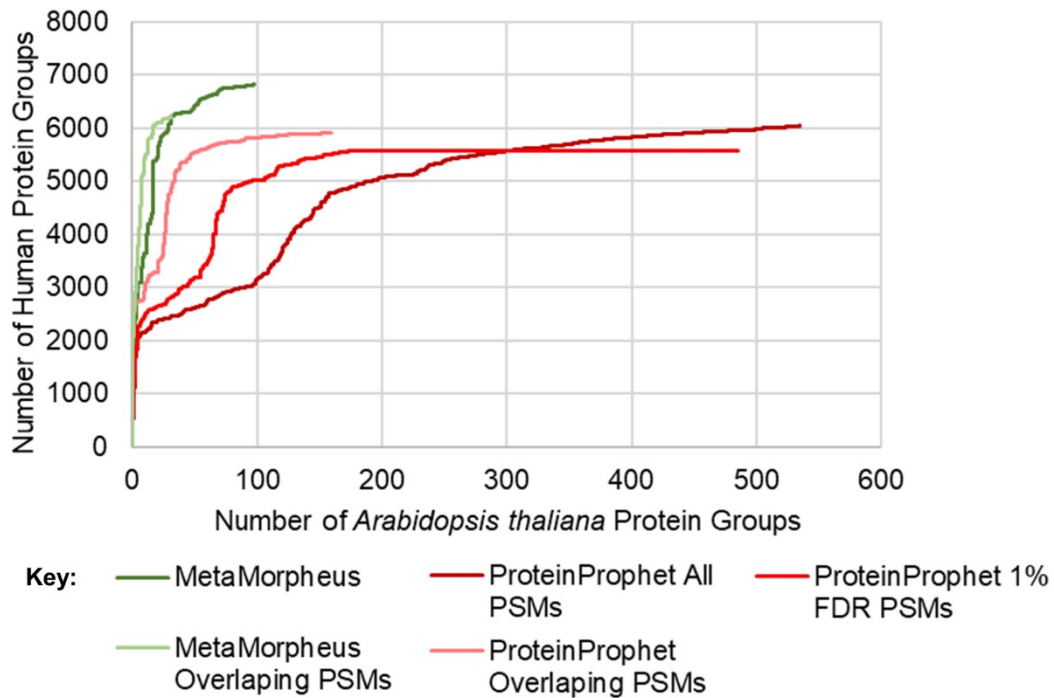


Figure 7.5: Comparison of MetaMorpheus' and ProteinProphet's protein inference algorithms for false positive identifications. Curves comparing the ability of MetaMorpheus' and ProteinProphet's multi-protease protein inference algorithms to distinguish between human protein groups (true positives) and *Arabidopsis thaliana* protein groups (false positives) based on the PSMs used for protein inference. Limiting the PSMs used for protein inference to those that were identified in both MetaMorpheus and Comet searches provided increased accuracy for both algorithms compared to their un-filtered counterparts and provided a more unbiased comparison.

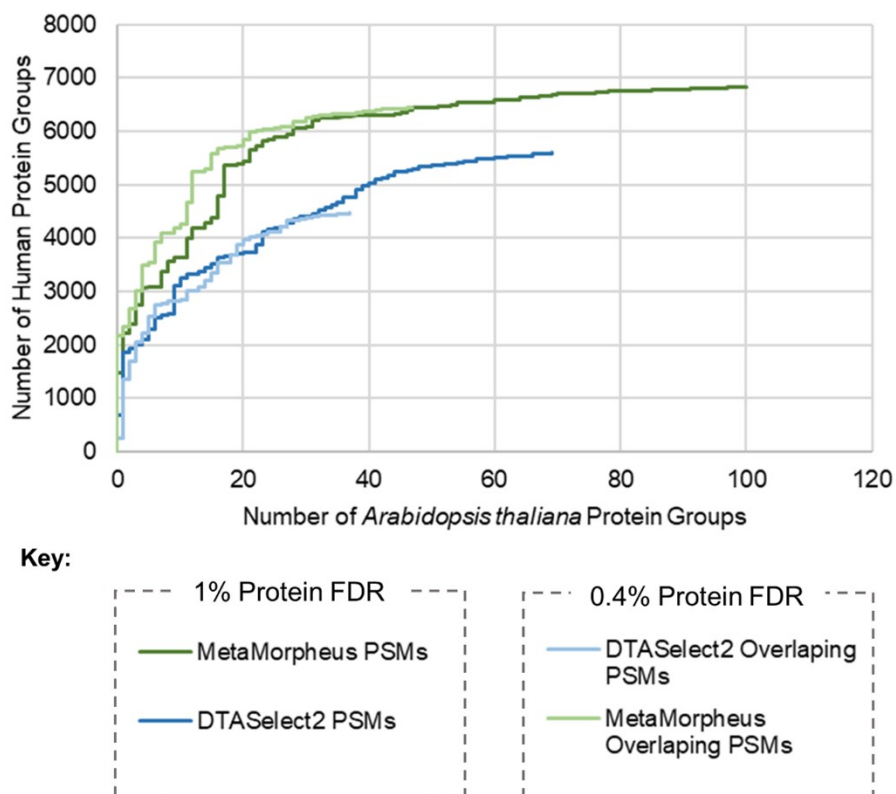


Figure 7.6: Comparison of MetaMorpheus' and DTASelect2's protein inference algorithms for false positive identifications. Curves comparing the ability of MetaMorpheus' and DTASelect2's multi-protease protein inference algorithms to distinguish between human protein groups (true positives) and *Arabidopsis thaliana* protein groups (false positives) based on the PSMs used for protein inference. Limiting the PSMs used for protein inference to those that were identified in both MetaMorpheus and ProLuCID searches provided increased accuracy for both algorithms compared to their un-filtered counterparts and provided a more unbiased comparison.

7.6 Supplementary Tables

Table 7.10: Comparison of Peptide Sequences Identified by MetaMorpheus and Comet at 1% FDR.

	Comet	MetaMorpheus	Overlap
Number of Arg-C Peptide Identifications	11,148	17,270	10,109
Number of Asp-N Peptide Identifications	17,326	12,065	9,565
Number of Chymotrypsin Peptide Identifications	18,073	10,831	6,218
Number of Glu-C Peptide Identifications	14,828	10,956	8,666
Number of Lys-C Peptide Identifications	25,569	31,962	24,321
Number of Trypsin Peptide Identifications	24,873	29,497	23,058

Table 7.11: Comparison of Peptide Sequences Identified by MetaMorpheus and ProLuCID at 1% FDR.

	ProLuCID	MetaMorpheus	Overlap
Number of Arg-C Peptide Identifications	16,696	17,064	13,122
Number of Asp-N Peptide Identifications	16,794	13,912	11,005
Number of Chymotrypsin Peptide Identifications	3,203	12,157	3,072
Number of Glu-C Peptide Identifications	9,629	13,423	8,316
Number of Lys-C Peptide Identifications	38,156	37,125	28,649
Number of Trypsin Peptide Identifications	26,440	41,935	26,137

7.7 References

- (1) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **2009**, *6*, Type: Journal Article, 359–62.

8 APPENDIX II: SUPPORTING INFORMATION FOR "PROTEASEGURU:
A TOOL FOR PROTEASE SELECTION IN BOTTOM-UP PROTEOMICS"

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Ibrahim, K.; Smith, L. M. ProteaseGuru: A Tool for Protease Selection in Bottom-Up Proteomics. *Journal of Proteome Research* **2021**, *20*(4), 1936–1942. <https://doi.org/10.1021/acs.jproteome.0c00954>.

Copyright © 2021 American Chemical Society.

8.1 Supplementary Table

Table 8.1: Species for the Skin Microbiome Subset

Begin Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Propionibacterium acnes	Cutibacterium acnes (strain DSM 16379 / KPA171202)	UP000000603	2,294
Corynebacterium tuberculostearicum	Corynebacterium tuberculostearicum SK141	UP000004384	2,209
Streptococcus mitis	Streptococcus mitis SK597	UP000003316	1,870
Streptococcus oralis	Streptococcus oralis subsp. oralis	UP000033716	1,810
Streptococcus pseudopneumoniae	Streptococcus pseudopneumoniae 5247	UP000018724	2,016
Streptococcus sanguinis	Streptococcus sanguinis (strain SK36)	UP000002148	2269

Continuation of Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Micrococcus luteus	Micrococcus luteus (strain ATCC 4698 / DSM 20030 / JCM 1464 / NBRC 3333 / NCIMB 9278 / NCTC 2665 / VKM Ac-2230)	UP000000738	2,207
Staphylococcus epidermidis	Staphylococcus epidermidis (strain ATCC 35984 / RP62A)	UP000000531	2,492
Staphylococcus capitis	Staphylococcus capitis subsp. capitis	UP000236440	2,232
Veillonella parvula	Veillonella parvula (strain ATCC 10790 / DSM 2008 / JCM 12972 / Te3)	UP000007968	1,843
Staphylococcus hominis	Staphylococcus hominis	UP000315294	2,602
Corynebacterium fastidiosum	NOT IN UNIPROT	X	X
Corynebacterium afermentans	Corynebacterium afermentans	UP000185547	2,165

Continuation of Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Enhydrobacter aerosaccus	Enhydrobacter aerosaccus	UP000190092	6,507
Corynebacterium simulans	Corynebacterium simulans	UP000074804	2,519
Corynebacterium aurimucosum	Corynebacterium aurimucosum (strain ATCC 700975 / DSM 44827 / CN-1)	UP000002077	2,528
Corynebacterium kroppenstedtii	Corynebacterium kroppenstedtii (strain DSM 44385 / JCM 11950 / CIP 105744 / CCUG 35717)	UP000001473	2,018
Corynebacterium amycolatum	Corynebacterium amycolatum SK46	UP000003275	2,103
Staphylococcus warneri	Staphylococcus warneri	UP000292953	2,831
Staphylococcus haemolyticus	Staphylococcus haemolyticus (strain JCSC1435)	UP000000543	2,640

Continuation of Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Corynebacterium resistens	Corynebacterium resistens (strain DSM 45100 / JCM 12819 / GTC 2026 / SICGH 158)	UP000000492	2,160
Malassezia restricta	Malassezia restricta CBS 7877	UP000269793	4096
Malassezia globosa	Malassezia globosa (strain ATCC MYA-4612 / CBS 7966)	UP000008837	4,274
Aspergillus tubingenis	Aspergillus tubingenis (strain CBS 134.48)	UP000184304	12,319
Candida parapsilosis	Candida parapsilosis (strain CDC 317 / ATCC MYA-4646)	UP000005221	5,777
Zymoseptoria tritici	Zymoseptoria tritici (strain CBS 115943 / IPO323)	UP000008062	10,972
Malassezia sympodialis	Malassezia sympodialis (strain ATCC 42132)	UP000186303	4,501

Continuation of Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Epidermophyton floccosum	Epidermophyton floccosum	N/A	79
Pyramimonas parkeae	Pyramimonas parkeae	N/A	239
Nannizzia nana	Nannizzia nana	N/A	33
Parachlorella kessleri	Parachlorella kessleri (Green alga) (Chlorella kessleri)	N/A	147
Tilletia walkeri	Tilletia walkeri	UP000078113	7,968
Nephroselmis olivacea	Nephroselmis olivacea (Green alga)	N/A	179
Cyanophora paradoxa	Cyanophora paradoxa	N/A	475
Aureoumbra lagunensis	Aureoumbra lagunensis	N/A	123
Pycnococcus solii	Pycnococcus solii	N/A	94
Gracilaria tenuistipitata	Gracilaria tenuistipitata var. liui (Red alga)	N/A	251
Leucocytozoon majoris	Leucocytozoon majoris	N/A	21

Continuation of Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Trichophyton rubrum	Trichophyton rubrum (strain ATCC MYA-4607 / CBS 118892)	UP000008864	10,006
Trichophyton mentagrophytes	Arthroderma hamiae (strain ATCC MYA-4681 / CBS 112371)	UP000008866	7,976
Molluscum contagiosum virus	Molluscum contagiosum virus subtype 1	UP000000869	163
Propionibacterium phage	Propionibacterium phage PAS50	UP000008740	46
Merkel cell polyomavirus	Merkel cell polyomavirus	UP000154903	4
Polyomavirus HPyV7	Human polyomavirus 7	N/A	68
Acheta domestica densovirus	Acheta domestica densovirus	UP000121107	5
Human papillomavirus (β)	Human papillomavirus type 5	UP000009252	9
Actinomyces phage	Actinomyces phage xhp1	UP000241342	54
Simian virus	Simian virus 41	UP000108270	7

Continuation of Table 8.1			
Species	UniProt Taxonomy	UniProt Proteome Key	Protein Count
Streptococcus phage	Streptococcus phage SPQS1	UP000014703	104
Stenotrophomonas phage	Stenotrophomonas phage SMA7	UP000014423	10
Polyomavirus HPyV6	Human polyomavirus 6	UP000119412	5
Human papillomavirus (γ)	Human papillomavirus type 103	UP000100457	6
Staphylococcus phage	Staphylococcus phage 44AHJD	UP000007462	21
Gamma papillomavirus HPV127	NOT IN UNIPROT	X	X
Enterobacteria phage	Enterobacteria phage MX1	UP000001832	4
Alphapapillomavirus	Alphapapillomavirus 9	UP000136316	8
Human papillomavirus (μ)	Human papillomavirus type 16	UP000009251	9
Pseudomonas phage	Pseudomonas phage D3	UP000009085	98
RD114 retrovirus	RD114 retrovirus	UP000172387	2
End of Table 8.1			

9 APPENDIX III: SUPPORTING INFORMATION FOR "ENHANCED PROTEIN ISOFORM CHARACTERIZATION THROUGH LONG-READ PROTEOGENOMICS"

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Jordan, B.T.; Mehlferber, M.M.; Jeffery, E.D.; Chatzipantsiou, C.; Kaur, S. Millikin, R.J.; Dai, Y.; Tiberi, S.; Castaldi, P.J.; Shortreed, M.R.; Luckey, C.J; Conesa, A.; Smith, L.M.; Deslattes-Mays, A.; Sheynkman, G.M. Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biology* **2022**, *23*(69). <https://doi.org/10.1186/s13059-022-02624-y>.

Copyright © 2022 Springer Nature.

9.1 Supplementary Figures

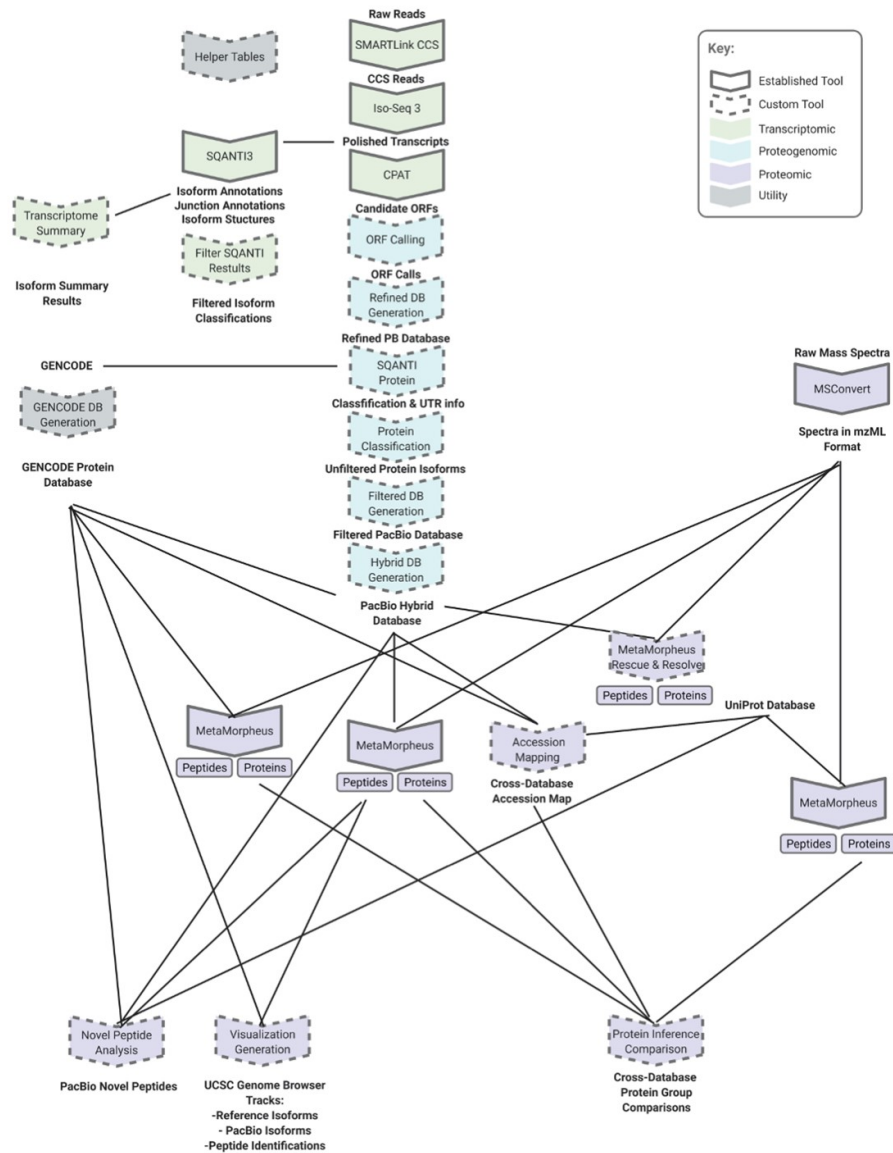


Figure 9.1: Detailed schematic of the Nextflow computational pipeline for long-read proteogenomics. Computational pipeline for full-length protein database generation, database searching, and downstream data analysis and visualization. Complete details of the pipeline may be found at <https://github.com/sheynkman-lab/Long-Read-Proteogenomics>.

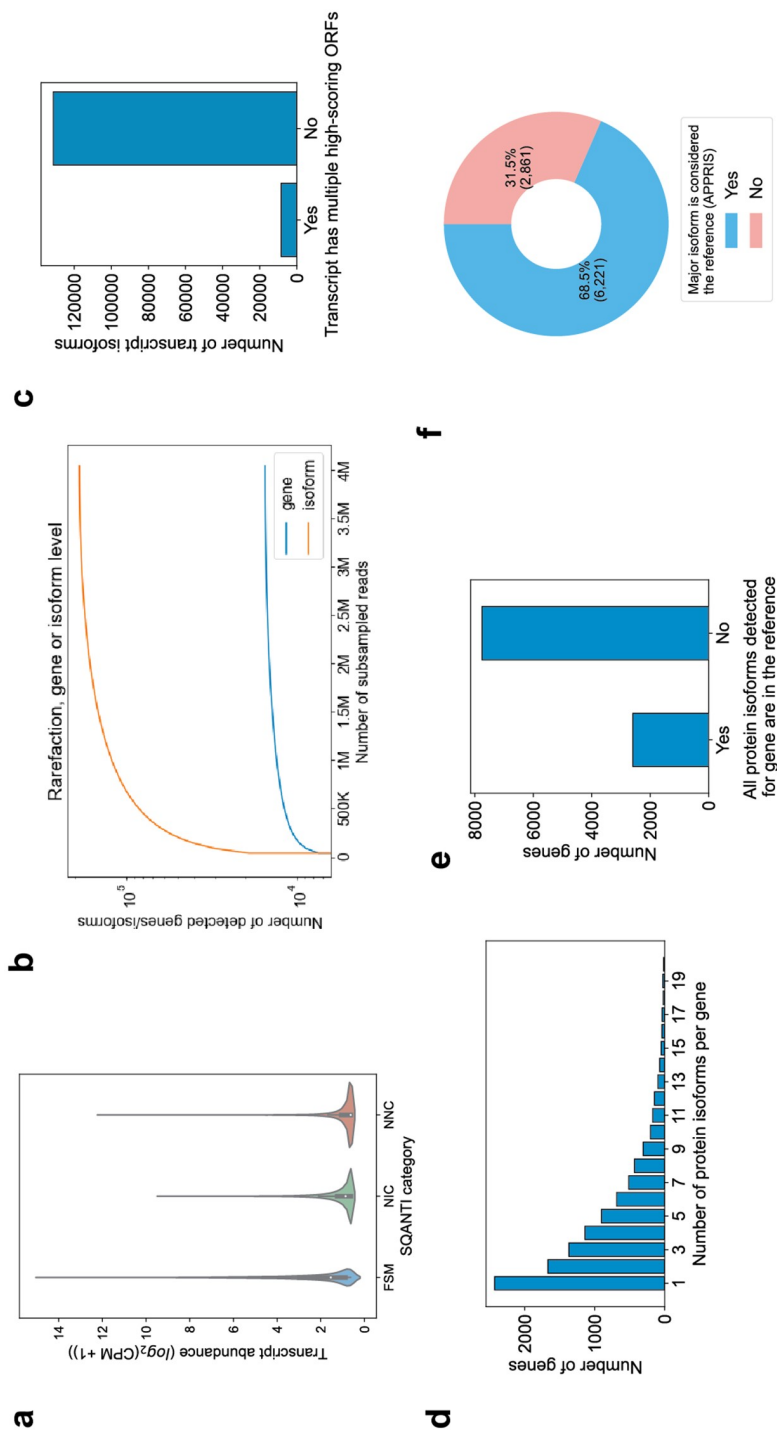


Figure 9.2: Generation and characterization of candidate protein isoform sequences from long-read RNA-seq data. a, Transcript abundance distributions for known (FSM) versus novel transcript isoforms (NIC, NNC). b, Saturation-discovery curve to determine the relationship between number of full-length reads subsampled and total number of genes and isoforms detected. c, Count of transcript isoforms with more than one high scoring ORF (defined as CPAT score above 0.9). d, Distribution of the number of protein isoforms per gene. e, Bar chart showing the number of genes that contain all known protein isoforms. Data based on the filtered PacBio database (45K protein isoforms). f, Fraction of genes for which the most abundant transcript underlying the protein isoform does not correspond to the APPRIS reference. ORF, open reading frame; CPAT, Coding-Potential Assessment Tool; FSM, full splice match; NIC, novel in catalog; NNC, novel not in catalog.

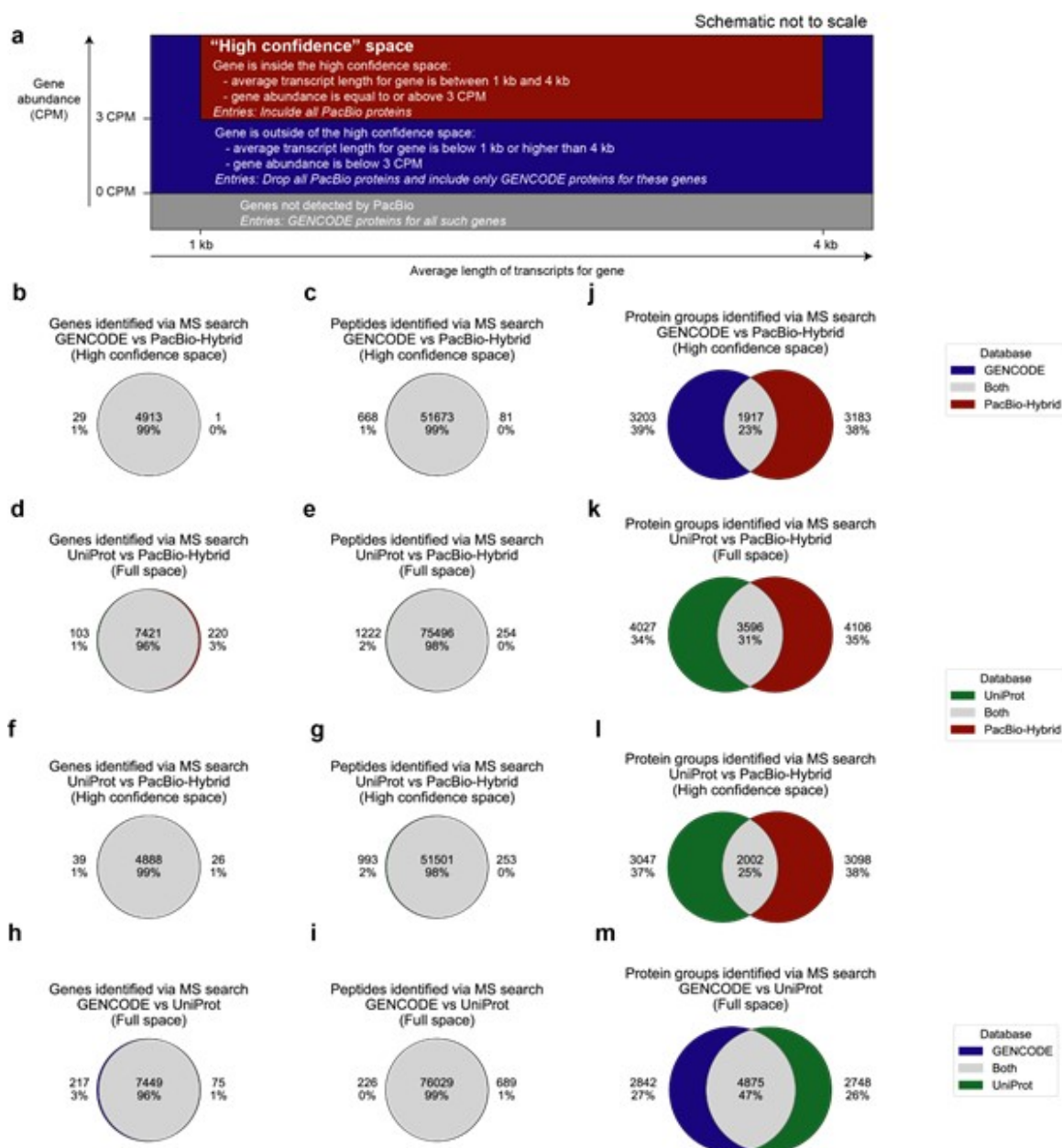


Figure 9.3: Comparison of MS-based proteomic coverage when using different protein databases for MS searching. *a*, Schematic of the contents of the PacBio-Hybrid database (not to scale). *b-m* Overlap of gene, peptide, and protein group identifications when comparing GENCODE versus PacBio-Hybrid in the high confidence space (*b, c, j*), UniProt versus PacBio-Hybrid in the full gene space (*d, e, k*), UniProt versus PacBio-Hybrid in the high confidence gene space (*f, g, l*), and GENCODE versus UniProt in the full gene space (*h, i, m*).

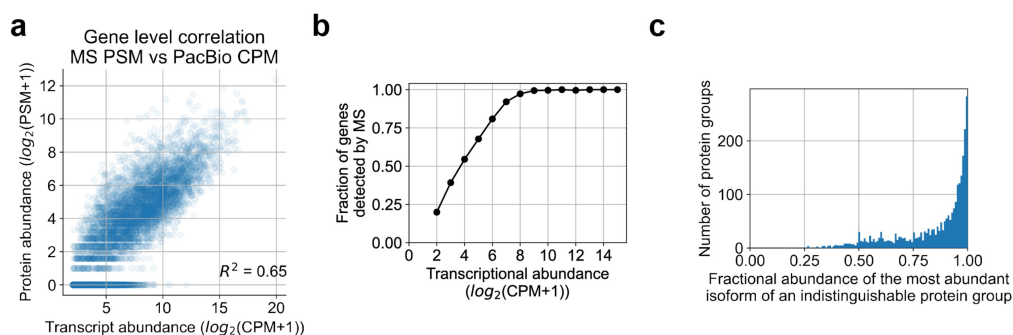


Figure 9.5: Relationship between RNA and protein estimated abundances. a, Correlation between long-read transcriptional abundance and protein abundance. Results are grouped by gene. b, Fraction of genes detected by MS as a function of transcript abundance. c, Distribution of the fractional relative abundance of the most abundant protein isoform in each indistinguishable protein group.

9.2 Supplementary Note 1: Long-read transcriptome sequencing of a human cell line

We sequenced two cDNA libraries of the human Jurkat T-lymphocyte cell line each with SMRT Cell 8M on the PacBio Sequel II system and obtained a total of 5 million HiFi (CCS) reads with an average read length of 2.1 kbp. Following a standard Iso-Seq bioinformatics workflow (see **Section 4.6**, <https://github.com/sheynkman-lab/Long-Read-Proteogenomics>), we classified and filtered the full-length transcript sequences, removing potential library artifacts.

Transcript isoform diversity is widespread in the sample. We identified 139,743 transcripts from 11,186 protein coding genes that exhibit a wide range of lengths and abundances (**Figure 9.6**). Many genes express multiple isoforms, with some genes co-expressing up to a dozen or more isoforms, and 84% (8,367) of the genes exhibiting co-expression of more than two isoforms (**Figure 9.7**).

For genes expressing multiple isoforms, we classified the corresponding isoforms

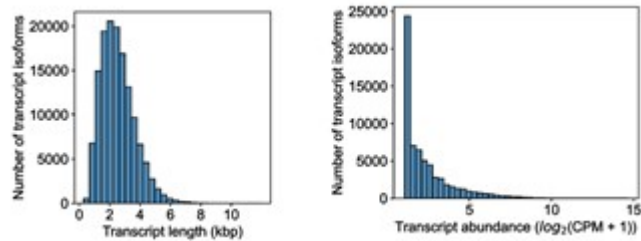


Figure 9.6: Long-read transcriptome length and abundance distributions. (Left) Distribution of transcript isoform lengths. (Right) Distribution of transcript isoform abundances. CPM, full-length read counts per million.

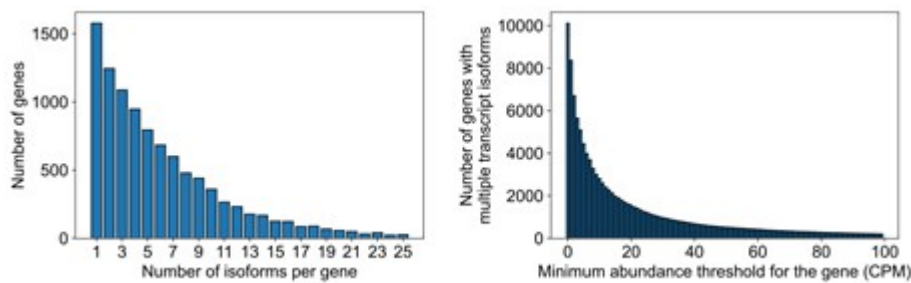


Figure 9.7: Co-expression of multiple isoforms from the same gene. (Left) Histogram of the number of distinct transcript isoforms per gene. (Right) Frequency of genes containing multiple isoforms, at different CPM abundance cut-offs. Only transcripts greater than 1 CPM were used for the data in these plots. CPM, full-length read counts per million.

as either major (i.e., most abundant isoform for a gene) or minor. Overall, minor isoforms tend to have lower abundance, but certain minor isoforms can still have robust expression and make up a large fraction of total gene expression (**Figure 9.8**). For a substantial fraction of genes expressing multiple isoforms (38%, 3,888), the major isoform expressed in Jurkat cells was not the “reference” isoform (GENCODE APPRIS principal isoform¹, **Figure 9.9**). Collectively, these results illustrate the widespread nature of alternative splicing and the need for empirically driven methods to characterize isoform diversity.

Approximately 86% of the transcript isoforms were classified as “full-splice match” (FSM) or a novel category NIC or NNC (see **Chapter 4**). The remaining 9,130 (14%)

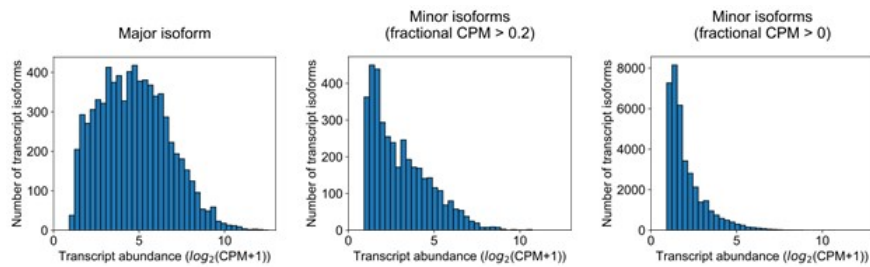


Figure 9.8: Abundance distribution of major versus minor transcript isoforms. (Left) Distribution of transcriptional abundance for major transcript isoforms. (Middle) Distribution of transcriptional abundance for minor transcript isoforms with a fractional abundance of more than 0.2. (Right) Histogram of transcriptional abundance for all minor transcript isoforms. Only transcripts greater than 1 CPM were used for the data in these plots. CPM, full-length read counts per million.

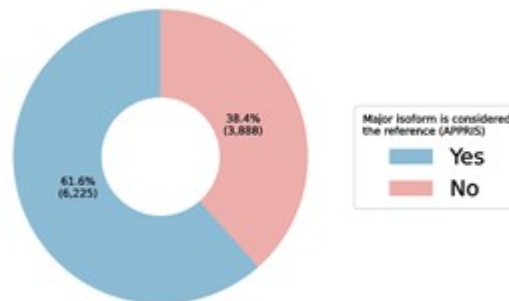


Figure 9.9: Fraction of transcript isoforms in which the major isoform (highest isoform expressed for a gene, based on CPM values) does not match the GENCODE principle APPRIS transcript isoforms.

transcripts were classified as “incomplete splice match” (ISM) cases, which can result from partially degraded transcripts generated during sample and library preparation or, alternatively, represent *bona fide* novel alternative promoter or polyadenylation sites. As expected, minor isoforms tend to be novel at a higher rate than major isoforms (**Figure 9.10**).

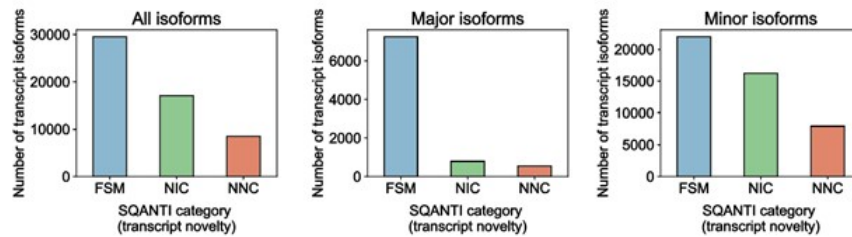


Figure 9.10: Breakdown of transcript isoforms by their novelty category. (Left) Number of transcript isoforms in each novelty classification category. (Middle) Breakdown of major (highest expressed for gene) transcript isoforms by novelty category. (Right) Breakdown of minor transcript isoforms by novelty category. Only transcripts greater than 1 CPM were used for the data in these plots.

9.3 Supplementary Note 2: ORF calling from long-read transcripts

For calling of ORFs from full-length transcript isoforms, we used the CPAT algorithm. Several ORF callers are available. We compared the identity of ORFs called using CPAT versus TransDecoder² and GMST³. To run TransDecoder (version 5.5.0), a minimum ORF size was set to 50 nucleotides to mimic the parameters we used for CPAT. The single best ORF for each isoform was selected, per the ‘Transdecoder.Predict’ parameter. To run GMST the same parameters as used in the SQANTI pipeline (GMST version 5.1) were used. We found that in a majority of cases, the same ORF is predicted; however, there are some differences in ORFs called.

CPAT returns a coding score for each candidate ORF. Overall, the scores of the candidate ORFs form a bimodal distribution, and there is a clear distinction between high and low scoring ORFs, overall (**Figure 9.12**).

In some cases, there are two or more ORFs that have a high coding score (from the CPAT algorithm). Generally speaking, the upstream-most ORF, containing an ATG closer to the 5’, was deemed more credible given the ribosomal scanning model of

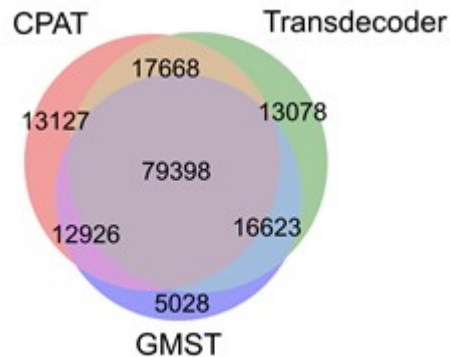


Figure 9.11: Comparison of ORF callers in predicting ORFs from full-length transcripts (PacBio).

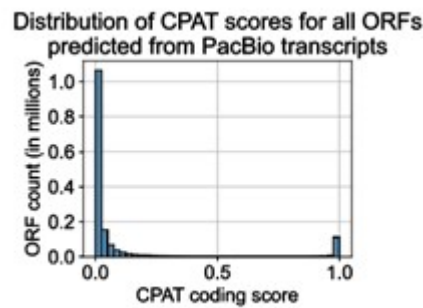


Figure 9.12: Distribution of ORF scores from the CPAT algorithm.

translation⁴. Therefore, we use this model as an assumption for ORF calling, in which higher weights are given to ORFs containing ATGs that are closer to the 5' end of the transcript (**Figure 9.13**). Implementation details about the ORF calling algorithm can be found in the `orf_calling` module in the Nextflow pipeline.

In order to determine whether our ORF calling algorithm is able to recover ORFs annotated in GENCODE, we compared the GENCODE reference ORFs (ENSPs, i.e., GENCODE proteins) versus the *ab initio* CPAT-predicted ORFs. We found that of the 58,860 GENCODE transcripts, the CPAT-predicted ORF matched the ORF annotated by GENCODE in 55,324 or 94% of cases. For 3,536 or 6% of cases, the ORF was not an exact match (**Figure 9.14**). A majority of the ORFs that differ from the reference differ

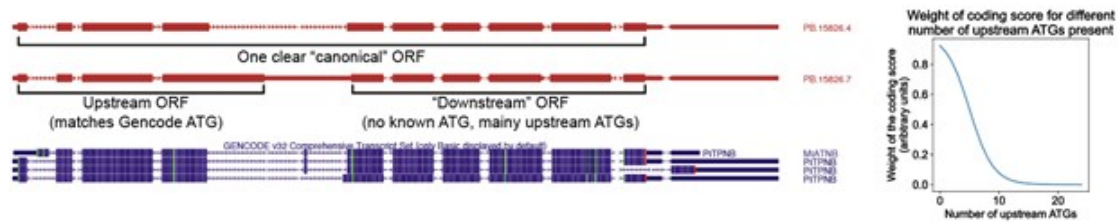


Figure 9.13: Evaluation of ORF plausibility and weighting. (Left) Example of two highly scoring ORFs from the same transcript called by CPAT for gene PITPNB. The top isoform PB.15826.4 contains one clear best ORF. The second isoform PB.15826.7 contains an extension in the 4th exon, leading to a frameshift and premature termination codon. In such cases, the upstream-most ORF is the most plausible translated region. (Right) ORF score weighting based on the number of upstream ATGs.

due to differences in the N-terminus, with 55% differing only in the ATG start location. Note that in our proteogenomics pipeline, as part of our ORF calling procedure, we heavily weigh the presence of a GENCODE ATG; therefore, a majority of these cases would be reverted to the GENCODE ORF within our pipeline. For 60 cases (0.1% of the dataset), only the C-terminus did not match, and a majority of such cases could be explained by selenocysteine re-coding events, in which the stop codon is re-coded to selenocysteine, thereby extending the protein C-terminus.

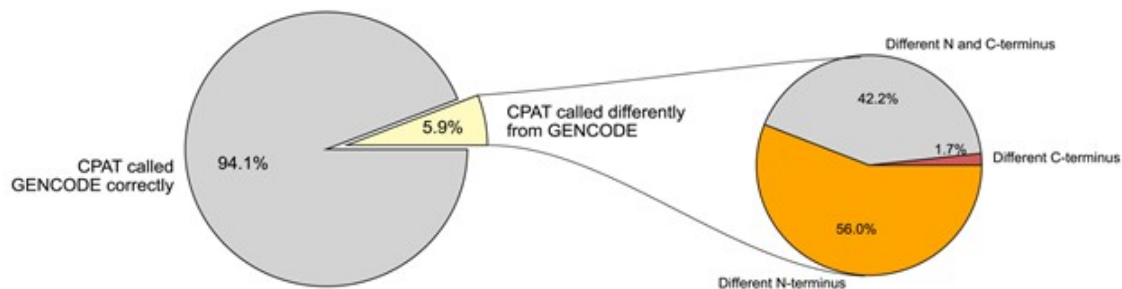


Figure 9.14: Fraction of ORFs predicted from the GENCODE transcriptome using the modified CPAT ORF calling pipeline. For predicted ORFs that did not exactly match GENCODE annotated ORFs, a breakdown of the part of the protein that differs (e.g., N-terminus) is indicated.

9.4 Supplementary Note 3: Determination of the high confidence protein database space based on long-read RNA-seq coverage of peptides

To determine factors underlying incomplete proteomic coverage using the PacBio database, we examined the properties of genes in which there were fewer peptides recovered when using the PacBio database than when using the GENCODE database. When searching the PacBio database (PacBio only), we detected 70,761 peptides and 7,068 genes, which corresponds to 90.7% of peptides and 92% of genes detected using the GENCODE database. Overall, lower peptide recovery was observed for genes with extreme transcript lengths (e.g., less than 1 kb, longer than 4 kb) or very low abundance (e.g., below 3 CPM) (**Figure 9.15**). The extremely short transcripts may not be sampled because the PacBio library preparation included a bead clean-up which removes short cDNAs. The longer transcripts are, in general, more difficult to convert to cDNA and sample for sequencing, although newer sequencing platforms may demonstrate less bias against lengths. Overall, as expected, the protein content of lower abundance genes or genes with extremes in lengths are not fully sampled in the long-read transcriptome dataset.

A majority of genes detected using the GENCODE database also returned 100% peptide coverage when using the PacBio database, suggesting long-read datasets are reaching a critical threshold of coverage capturing the full sequence content of protein-coding mRNA. In other words, given reasonable constraints (length 1-4 kb, abundance 3 CPM+), we achieved nearly 99% coverage of the known peptide sequence space (**Figure 9.16**). For this set of high confidence genes, it is likely that all expressed protein isoforms are represented in the long-read data, including novel

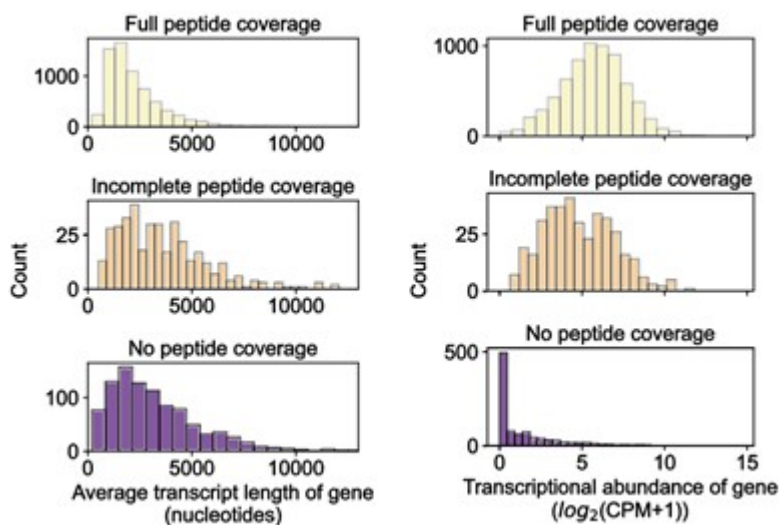


Figure 9.15: Characterizing the length and abundance biases that contribute to lower proteomic coverage from PacBio-derived databases (Left) Distribution of transcript lengths for genes with different extents of peptide coverage. (Right) Distribution of transcript abundances for genes with different extents of peptide coverage.

isoforms not represented in the reference database, which is advantageous for protein inference.

9.5 Supplementary Note 4: Criteria for Novel Peptide Identification

The identification of novel peptides requires rigorous validation and stringent filtering criteria to ensure that the spectrum does in fact represent the novel peptide sequence. First, all novel peptide identifications must meet the basic filtering criteria for all peptide identifications, having a minimum MetaMorpheus score of 5 and being present at a global 1% FDR cutoff. The distribution of the FDR, or q -values, for the novel peptide identifications was compared to the FDR, or q -value, distribution of the canonical peptides (Figure 9.17). The median q -value for the novel peptides

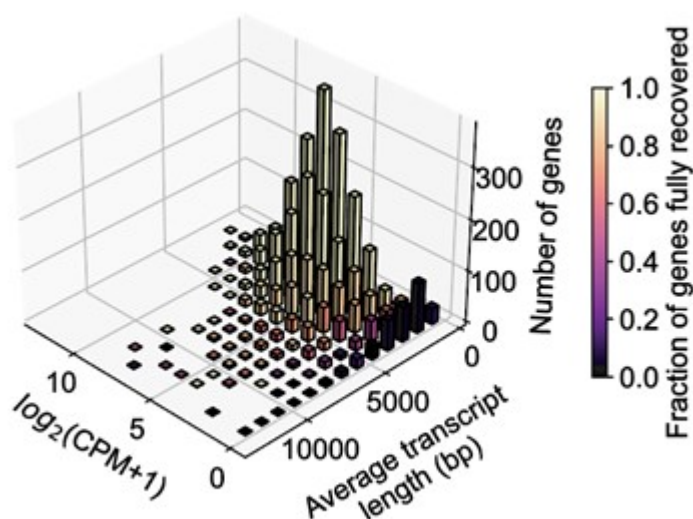


Figure 9.16: Evaluation of peptide identification recovery as a function of gene average transcript length and abundance. Three-dimensional bar plot that displays the fraction of genes in which all peptides are recovered in the PacBio database, as a function of gene average transcript length and abundance. CPM, full-length read counts per million.

was lower than the median q -value for the canonical peptides with values of 0 and 2.1×10^{-5} , respectively. Since the FDR, or q -value, is a global assessment of confidence for the entire set of peptide identifications, not a confidence metric for the specific peptide identification, the posterior error probability (PEP) of each novel peptide was also considered. All novel peptides had a PEP value less than 0.005. The distribution of novel peptide PEP values was compared to the distribution of PEP values for all canonical peptide identifications (**Figure 9.17**). The median PEP value for the novel peptides and canonical peptides were nearly identical, demonstrating the novel peptide identifications are as confident as the canonical peptide identifications.

In addition to the preliminary filtering and confidence evaluations of the novel peptide identifications, the Human Proteome Project MS data guidelines⁵ for manual validation of data-dependent acquisition spectra were applied. These criteria include:

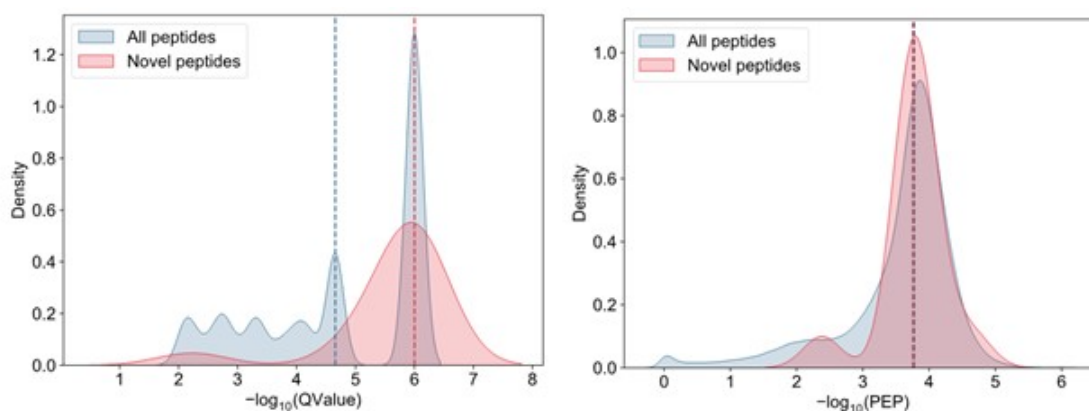


Figure 9.17: Comparison of novel and canonical peptide distributions for (left) q -values and (right) PEP values. The median value is represented by a dashed line.

- 1) high mass accuracy (5 ppm precursor ion and 20 ppm product ion mass error),
- 2) clearly annotated spectrum that was scrutinized for missed and extra peaks (we used a rough threshold of 25% maximum of unassigned fragment ions above 10% relative abundance) and 3) peptide length (minimum of 9 amino acids).

Additional manual validation criteria were applied to each novel peptide identification. Mass errors of novel peptide hits were compared against the mass error of other confident spectral assignments from the same raw file to check for consistency. Extra weight was given for the presence of sequence-specific characteristics. For example, the spectra assignment was considered more confident if a highly charged peptide was longer, had a higher number of basic residues; or if y -ions ending in proline were higher intensity than other fragment ions. Precursor co-isolation can complicate spectral annotation by resulting in fragment ions from multiple peptide origins. This complexity was taken into account by noting the percentage of MS2 total ion count (TIC) accounted for by annotated fragment ions. This percentage is represented by the decimal digits of the MetaMorpheus score. TIC coverage of 20% or more was used as a rough threshold but not a hard criterion.

We also investigated whether any peptide sequences from contaminant proteins,

GENCODE reference isoforms, single amino acid variant containing proteins or post-translationally modified proteins could provide better peptide assignments for the novel peptide spectra. Identifications were evaluated based on MetaMorpheus score, q -value and PEP. A contaminant protein database, included in MetaMorpheus, was searched alongside the PacBio-Hybrid database. No contaminant peptides were a match to the novel peptide spectra. In the search of the GENCODE reference database, no peptide identifications for the spectra in question were a better match than the novel peptides (Additional File 6: Table S4). To search for post-translationally modified peptides, Global Post-Translational Modification Discovery (GPTMD), with default settings, was performed with the GENCODE reference database.⁶ The subsequent search found no modified peptides were a match for the novel peptide spectra. To investigate if variant containing peptides could better account for the spectra supporting the novel peptide identifications, a proteogenomic database generated by Spritz using Jurkat short-read RNA-seq data published by Cesnik et. al.⁷ was searched. The results showed that no variant containing peptides were identified for the same spectra as the novel peptide assignments. Based on this stringent evaluation of novel peptide candidates, we are quite confident in the 14 novel peptide assignments that passed all criteria.

9.6 Supplementary Note 5: Rescue & Resolve algorithm abundance threshold optimization

The R&R algorithm requires selection of a transcript abundance threshold that is the basis for recovering a formerly eliminated protein identification. If the transcript abundance threshold is set too low, there is a higher probability of recovering protein isoforms that are not expressed (false positives). If the threshold is too high, there is

a higher probability of failing to rescue protein isoforms present in the sample (false negatives).

We evaluated 10 different abundance threshold values (5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 CPM) for the R&R algorithm. As expected, the lower the CPM abundance threshold, the more protein groups are rescued (**Figure 9.18**), however these larger values are not necessarily indicative of a higher true positive rate.

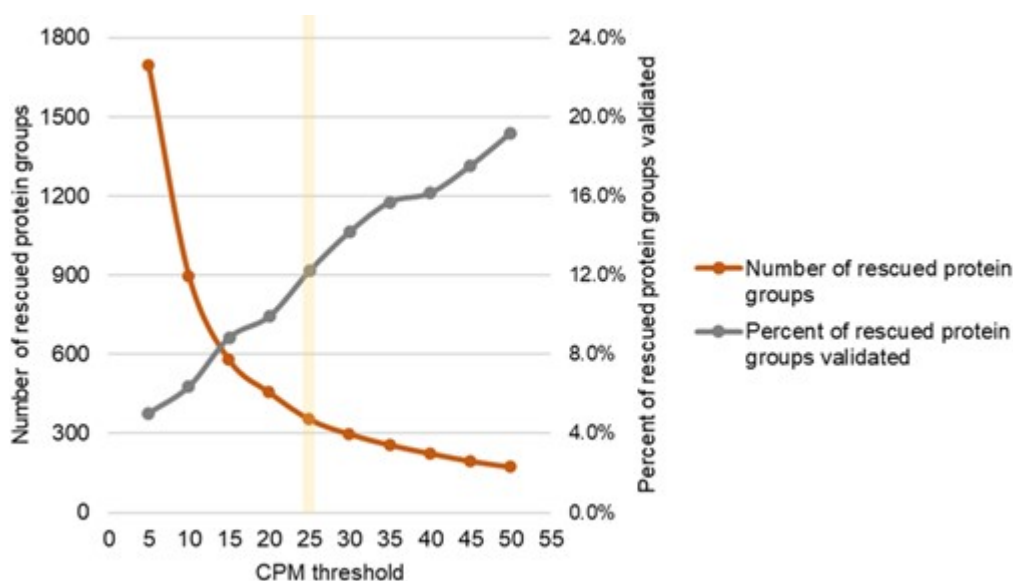


Figure 9.18: Abundance threshold evaluation for the Rescue & Resolve algorithm. The orange curve represents the number of protein groups rescued at 1% FDR for each CPM abundance evaluated. The grey curve represents the percent of rescued protein groups whose identity was validated in an independent multi-protease MS dataset.

Protein inference results obtained from searching a higher coverage MS dataset (i.e., a multi-protease proteomics dataset) can be utilized as a validation group. We can compare the rescued protein groups from the 28-fraction trypsin-only set to the protein inference results derived from a high coverage multi-protease digest data set, which serves as a “ground” truth set, to obtain a percent of rescued protein groups validated (**Section 9.7**). We found that as the abundance threshold increases, the

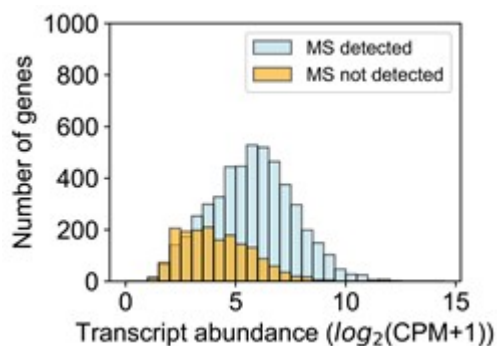


Figure 9.19: Relationship between transcriptional abundance and MS detectability. Distribution of genes either detected or not detected by MS as a function of cumulative transcriptional abundance.

percent of rescued groups validated in the multi-protease protein inference results increase (**Figure 9.18**). Based on this analysis we decided to set a conservative threshold to rescue transcripts with an abundance of 25 CPM or higher. Additionally, based on observed relationships between transcript abundances for genes with and without peptide evidence (**Figure 9.19**), with a CPM cutoff of 25 ($\log_2(\text{CPM}+1) = 4.6$) only 9.7% of genes in the high confidence space have no peptide support.

9.7 Supplementary Note 6: Multi-protease validation for Rescue & Resolve results

We hypothesize the R&R algorithm enables a more precise representation of sample isoform diversity than can be achieved using a traditional protein inference approach. However, the experimental validation of protein isoforms inferred from MS data is an ongoing challenge.^{8,9} Existing analytical standards fail to appropriately model the complexity of endogenous isoforms^{10,11}, and the lack of knowledge regarding protein isoforms present in a sample (i.e., ground truth) is a problem in and of itself.¹²⁻¹⁴ In place of experimental validation, studies have established

heuristic guidelines to compare protein inference results between different inference algorithms.^{8,15} Here we have developed a computational strategy that leverages the improved accuracy and precision of multi-protease protein inference to validate both the “rescued” and “resolved” protein isoforms.

It is well established that the use of multiple proteases improves proteomic results relative to those of a single protease (typically trypsin).^{16,17} Large portions of the proteome are inaccessible to any single protease, and the combination of peptide identifications from orthogonal proteolytic digests not only improves the number of protein identifications but also the percent protein sequence coverage obtained.^{16,17} The MetaMorpheus MS search software contains a multi-protease protein inference algorithm which enables all peptide identifications from several proteolytic digests to be considered in combination, and provides more accurate protein inference results than what is achieved by other protein inference algorithms, or by analysis of any single proteases data alone.¹⁶

For validation of “rescued” and “resolved” protein isoforms, an independently generated multi-protease dataset was used. Spectra from the analysis of six fractionated proteolytic digests (see **Section 4.6**) were searched against the PacBio-Hybrid database using MetaMorpheus, and the multi-protease protein inference algorithm was employed.¹⁶ The multi-protease protein inference results are considered to be more comprehensive, and reflective of the sample’s proteome compared to what can be achieved with trypsin alone. The use of orthogonal proteases provides higher, and more distinctive coverage of protein isoforms. One product of this is the identification of more unique peptides, which can confidently identify protein isoforms. Although the multi-protease protein inference results are not a perfect model of the isoforms expressed in the sample, for the purpose of our validation strategy, we will consider the results as a “ground truth” dataset.

For the purposes of validation, we determined if the protein isoforms that were “rescued” or “resolved” by “Rescue & Resolve” algorithm were isoforms that were identified in the multi-protease protein inference analysis. If a “rescued” or “resolved” protein isoform was identified—in the multi-protease analysis—as a single protein isoform, not as part of a multi-isoform protein group, the identification of the protein isoform in question was considered to be confirmed, or “validated”. These validated “rescued” or “resolved” isoforms, if identified in the multi-protease protein inference results, had sufficient peptide level evidence, such as an isoform-specific peptide, due to the identification of additional non-tryptic peptides derived from orthogonal proteases to support their confident identification. The percent of “rescued” and “resolved” protein isoforms whose presence were confirmed, or “validated”, in the multi-protease protein inference results can be calculated. This percent validation rate, when compared to expected rates, indicates how well the “rescue” and “resolve” portions of the “R&R” algorithm do at increasing the number of true positive protein isoform identifications.

Since the multi-protease protein inference results are still incomplete and subject to error due to incomplete peptide coverage of the proteome, the percent validated means very little on its own. To assess the significance of the percent of “rescued” or “resolved” isoforms validated, the experimentally determined value can be compared against the validation rates expected at random. Such values can be computed by calculating the percent of protein isoforms validated for a pool of randomly “rescued” or “resolved” protein isoforms.

For the evaluation of the “rescue” portion of the “Rescue & Resolve” algorithm we compared the rate of validation between the “rescued” isoforms and randomly selected isoforms (background null). We “rescued” 355 isoforms based on additional transcriptional evidence, from a pool of 15,700 isoforms that represent all the protein

isoforms that could possibly be “rescued”. The same number of protein isoforms that were rescued in the experimental results (N=355), were randomly selected, agnostic of transcriptional abundance, from the pool of 15,700 protein isoforms that were discarded in the protein inference process. Once the randomly “rescued” isoforms have been selected, we determined if such protein isoforms were identified in the multi-protease protein inference results, and the percent of isoforms validated was calculated just as was done for the experimental results. This process of randomly selecting 355 protein isoforms to “rescue” and determining the validation rate was repeated for a total of 1,000 permutations to generate a null distribution of validation rates against which the experimentally obtained results were compared, and statistical significance with a p -value <0.0001 was determined.

For the evaluation of the “rescue” portion of the “Rescue & Resolve” algorithm we compared the rate of validation between the “rescued” isoforms and randomly selected isoforms (background null). We “rescued” 355 isoforms based on additional transcriptional evidence, from a pool of 15,700 isoforms that represent all the protein isoforms that could possibly be “rescued”. The same number of protein isoforms that were rescued in the experimental results (N=355), were randomly selected, agnostic of transcriptional abundance, from the pool of 15,700 protein isoforms that were discarded in the protein inference process. Once the randomly “rescued” isoforms have been selected, we determined if such protein isoforms were identified in the multi-protease protein inference results, and the percent of isoforms validated was calculated just as was done for the experimental results. This process of randomly selecting 355 protein isoforms to “rescue” and determining the validation rate was repeated for a total of 1,000 permutations to generate a null distribution of validation rates against which the experimentally obtained results were compared, and statistical significance with a p -value <0.0001 was determined.

9.8 Supplementary Tables

Table 9.1: Protein Classifications Based on SQANTI Protein

SQANTI Protein class	N-terminus	Splicing	C-terminus	Note	Number of protein isoforms
pFSM	Known	Known	Known	-	15,549
pFSM	Known	-	Known	-	782
pNIC	Known	Known	Known	Novel combo of N/C-term	1,603
pNIC	Known	Combo	Known	-	6,039
pNNC	Known	Combo	Novel	-	1,910
pNNC	Known	Known	Novel	-	7,732
pNNC	Known	Novel	Known	-	3,317
pNNC	Known	Novel	Novel	-	4,759
pNNC	Novel	Combo	Known	-	150
pNNC	Novel	Combo	Novel	-	51
pNNC	Novel	Known	Known	-	2,078
pNNC	Novel	Known	Novel	-	368
pNNC	Novel	Novel	Known	-	537
pNNC	Novel	Novel	Novel	-	193

Table 9.2: Number of Isoforms for Each Transcript and Protein Isoform Classifications Between SQANTI and SQANTI Protein

SQANTI3 transcript isoform class	SQANTI Protein isoform class	Number of isoforms
FSM	pFSM	11,874
FSM	pNIC	390
FSM	pNNC	1,220
ISM	pFSM	371
ISM	pNIC	428
ISM	pNNC	4,752
NIC	pFSM	2,353
NIC	pNIC	6,398
NIC	pNNC	5,416
NNC	pFSM	1,733
NNC	pNIC	426
NNC	pNNC	9,707

Table 9.3: Summary of MetaMorpheus Search Results

Protein database	Peptides	Protein Groups	Genes	Gene space
GENCODE	76,255	7,717	7,666	All protein-coding genes
UniProt	76,718	7,623	7,524	All protein-coding genes
PacBio Hybrid	75,750	7,702	7,641	All protein-coding genes
GENCODE	52,341	5,126	5,036	High confidence space (HC space)
UniProt	52,494	5,055	4,960	High confidence space (HC space)
PacBio Hybrid	51,754	5,100	5,000	High confidence space (HC space)

Table 9.4: Search Parameters for MetaMorpheus

Search Task Parameters	
<i>Search Parameters</i>	
Protease: trypsin	Max Missed Cleavages: 2
Max Mods Per Peptide: 2	Min Peptide Length: 7
Max Peptide Length: None	Precursor Mass Tolerance: 5 ppm
Product Mass Tolerance: 20 ppm	Dissociation Type: HCD
Initiator Methionine : Variable	Separation Type: HPLC
<i>Modifications</i>	
Common Fixed: Carbamidomethyl on C & U	Common Variable: Oxidation on M
<i>Protein Parsimony</i>	
Apply protein parsimony and construct protein groups	
<i>Quantification</i>	
LFQ: Quantify peptides/proteins	Peakfinding tolerance: 5ppm
<i>Output Options</i>	
Write .mzID	Write Decoys
Write Contaminants	Write Individual File Results
Minimum score allowed: 5	
<i>Advanced Options: File Loading Parameters</i>	
Use Provided Precursor	Deconvolute Precursor
Deconvolution Max Assumed Charge State: 12	Trim MS2 Peaks
Top N Peaks per m/z window:200	Minimum intensity ratio: 0.01
<i>Advanced Options: Search Parameters</i>	
Search Mode: Classic Search	Number of Database Partitions: 1
Generate Target Proteins	Generate Decoy proteins
Generate Reversed Decoys	Max Modification Isoforms: 1024
Min Read Depth for Variants: 1	Max Heterozygous Variants for Combinations: 4
N-Terminal Ions	C-Terminal Ions
Max Fragment Mass (Da): 30000	Max Threads: 39
Mass Difference Acceptor Criterion: 1	Remove Contaminant
Missed Monoisotopic Peak	
Report PSM ambiguity	

9.9 References

- (1) Rodriguez, J. M.; Maietta, P.; Ezkurdia, I.; Pietrelli, A.; Wesselink, J. J.; Lopez, G.; Valencia, A.; Tress, M. L. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res* **2013**, *41*, Type: Journal Article, D110–7.
- (2) Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **2013**, *8*, Type: Journal Article, 1494–512.
- (3) Tang, S.; Lomsadze, A.; Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **2015**, *43*, Type: Journal Article, e78.
- (4) Kozak, M. Initiation of translation in prokaryotes and eukaryotes. *Gene* **1999**, *234*, Type: Journal Article, 187–208.
- (5) Deutsch, E. W.; Lane, L.; Overall, C. M.; Bandeira, N.; Baker, M. S.; Pineau, C.; Moritz, R. L.; Corrales, F.; Orchard, S.; Van Eyk, J. E.; Paik, Y. K.; Weintraub, S. T.; Vandembrouck, Y.; Omenn, G. S. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J Proteome Res* **2019**, *18*, Type: Journal Article, 4108–4116.
- (6) 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*, Type: Journal Article, 1844–1851.
- (7) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *J Proteome Res* **2021**, *20*, Type: Journal Article, 1826–1834.
- (8) Audain, E.; Uszkoreit, J.; Sachsenberg, T.; Pfeuffer, J.; Liang, X.; Hermjakob, H.; Sanchez, A.; Eisenacher, M.; Reinert, K.; Tabb, D. L.; Kohlbacher, O.; Perez-

- Riverol, Y. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J Proteomics* **2017**, *150*, Type: Journal Article, 170–182.
- (9) Claassen, M. Inference and validation of protein identifications. *Mol Cell Proteomics* **2012**, *11*, Type: Journal Article, 1097–104.
- (10) The, M.; Edfors, F.; Perez-Riverol, Y.; Payne, S. H.; Hoopmann, M. R.; Palmblad, M.; Forsstrom, B.; Kall, L. A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms. *J Proteome Res* **2018**, *17*, Type: Journal Article, 1879–1886.
- (11) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J Proteome Res* **2008**, *7*, Type: Journal Article, 96–103.
- (12) Ahrne, E.; Molzahn, L.; Glatter, T.; Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **2013**, *13*, Type: Journal Article, 2567–78.
- (13) Choi, M.; Eren-Dogru, Z. F.; Colangelo, C.; Cottrell, J.; Hoopmann, M. R.; Kapp, E. A.; Kim, S.; Lam, H.; Neubert, T. A.; Palmblad, M.; Phinney, B. S.; Weintraub, S. T.; MacLean, B.; Vitek, O. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. *J Proteome Res* **2017**, *16*, Type: Journal Article, 945–957.
- (14) Edfors, F.; Forsstrom, B.; Vunk, H.; Kotol, D.; Fredolini, C.; Maddalo, G.; Svensson, A. S.; Bostrom, T.; Tegel, H.; Nilsson, P.; Schwenk, J. M.; Uhlen, M. Screen-

- ing a Resource of Recombinant Protein Fragments for Targeted Proteomics. *J Proteome Res* **2019**, *18*, Type: Journal Article, 2706–2718.
- (15) Claassen, M.; Reiter, L.; Hengartner, M. O.; Buhmann, J. M.; Aebersold, R. Generic comparison of protein inference engines. *Mol Cell Proteomics* **2012**, *11*, Type: Journal Article, O110 007088.
- (16) 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*, Type: Journal Article, 3429–3438.
- (17) 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*, Type: Journal Article, 1323–9.

10 APPENDIX IV: SUPPORTING INFORMATION FOR "DISCOVERY OF
DEHYDROAMINO ACID RESIDUES IN THE CAPSID AND MATRIX
STRUCTURAL PROTEINS OF HIV-1"

This chapter has been published and is reproduced with permission from:

Miller, R.M.; Knoener, R.A; Benner, B.E.; Frey, B.L.; Shortreed, M.R.; Sherer, N.M.; Smith, L. M. Discovery of Dehydroamino Acid Residues in the Capsid and Matrix Structural Proteins of HIV-1. *Journal of Proteome Research* **2022**, *21*(4), 993–1001. <https://doi.org/10.1021/acs.jproteome.1c00867>.

Copyright © 2022 American Chemical Society.

10.1 Supplementary Note 1: Determination of the Reproducibility of Peptide Identifications

To determine the threshold at which a peptide would be considered reproducible for this experiment, all peptide identifications were compared across the four biological replicates. In the unlabeled data, 43% of peptides were only identified in a single biological replicate, and 57% were identified in 2+ biological replicates. A similar trend was observed for the glutathione-labeled data, with 47% of all peptides being identified in a single biological replicate, and 53% were identified in 2+ biological replicates. The exact breakdown of percent peptide overlap across biological replicates can be found in **Table 10.1**. Based on this information, identification of a peptide in 2+ biological replicates is sufficient to be considered reproducible.

Table 10.1: Percent of Peptides Identified in Multiple Biological Replicates

Number of Replicates a Peptide is identified In	Percent of Peptides at 1% FDR in Unlabeled Sample	Percent of Peptides at 1% FDR in Labeled Sample
1	43%	47%
2	18%	15%
3	15%	13%
4	24%	25%

10.2 Supplementary Note 2: Evaluation of Virion Sample Preparation

The efficacy of the virion isolation protocol at both removing contaminant proteins present in the culture media, and enriching HIV proteins relative to human proteins. The contaminant protein database, included with MetaMorpheus, contains proteins

known to be present in fetal bovine serum (FBS). Inclusion of the contaminant database enables the identification of PSMs which map to these FBS contaminant proteins. Only 4% of all non-decoy PSMs at 1% FDR were identified to map to these serum proteins in both the labeled and unlabeled searches (5,570 PSMs in unlabeled, 7,210 PSMs in labeled). These values indicate the sample preparation protocol was effective at removing media contamination, allowing human and HIV PSMs to represent the vast majority of PSMs identified at 1% FDR. The enrichment of viral proteins over human proteins in the sample was determined by comparing estimated abundance values. The number of PSMs identified for a given protein (PSM count) was used as an approximation of abundance, and the average PSM count was determined for all identified viral and human proteins. In the unlabeled sample, the average PSM count for human proteins is 46 while the average PSM count for HIV protein is 3,276. The same trend is observed in the labeled sample with the average PSM count of human proteins being 36 and the average PSM count for HIV proteins is 3,172. These results demonstrate the enrichment of HIV viral proteins over human proteins in the isolated HIV virions which is expected. All the results reported demonstrate the virion isolation protocol effectively isolates the HIV virions from the transfected host cells, as well as the culture media.

10.3 Supplementary Figures

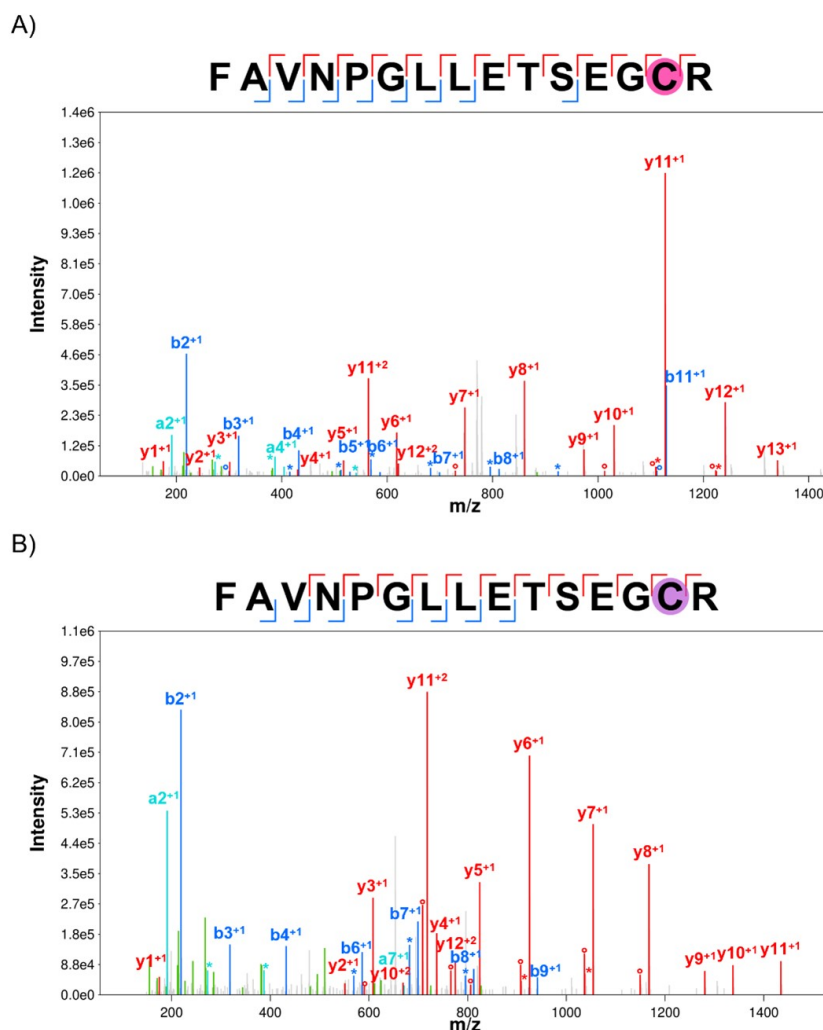


Figure 10.1: Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 56 in the HIV matrix protein. A) DHA and B) glutathione-labeled DHA, which is normally a cysteine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHA modification is highlighted by the pink circle, and the site of the glutathione-labeled DHA modification is highlighted by the purple circle. In panel A, y-ions 2-13 all confirm the DHA modification, and are shifted by -33.987 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, y-ions 2-12 all confirm the glutathione-labeled DHA modification and are shifted by $+273.096$ Da relative to the theoretical m/z of the peaks of an unmodified peptide, and by $+307.32$ Da relative to those of the DHA modified peptide.

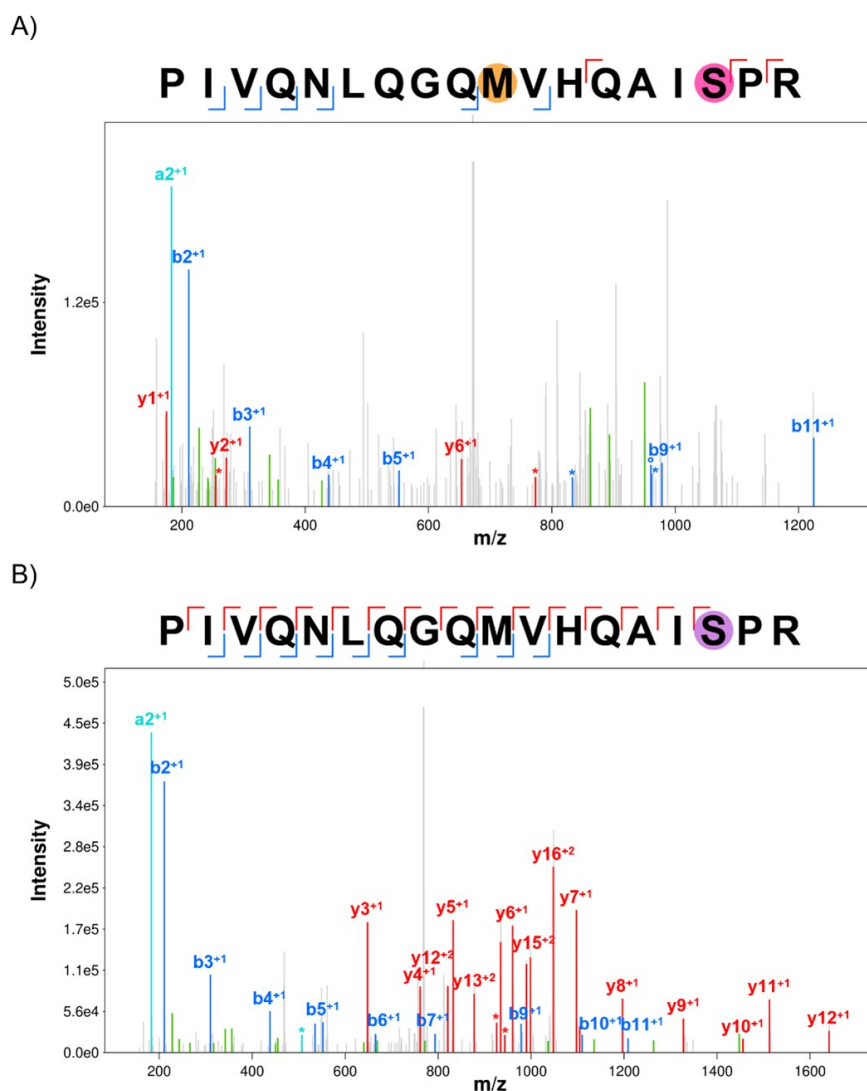


Figure 10.2: Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 16 in the HIV capsid protein. A) DHA and B) glutathione-labeled DHA, which is normally a serine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHA modification is highlighted by the pink circle, and the site of the glutathione labeled DHA modification is highlighted by the purple circle. The orange circle indicates oxidation of methionine. In panel A, the y6 ion confirms the DHA modification, and is shifted by -18.01 Da relative to the theoretical m/z peak of an unmodified peptide. In panel B, y-ions 3-15 all confirm the glutathione-labeled DHA modification and are shifted by +289.07 Da relative to the theoretical m/z of the peaks of an unmodified peptide, and by 307.32 Da relative to those of the DHA modified peptide.

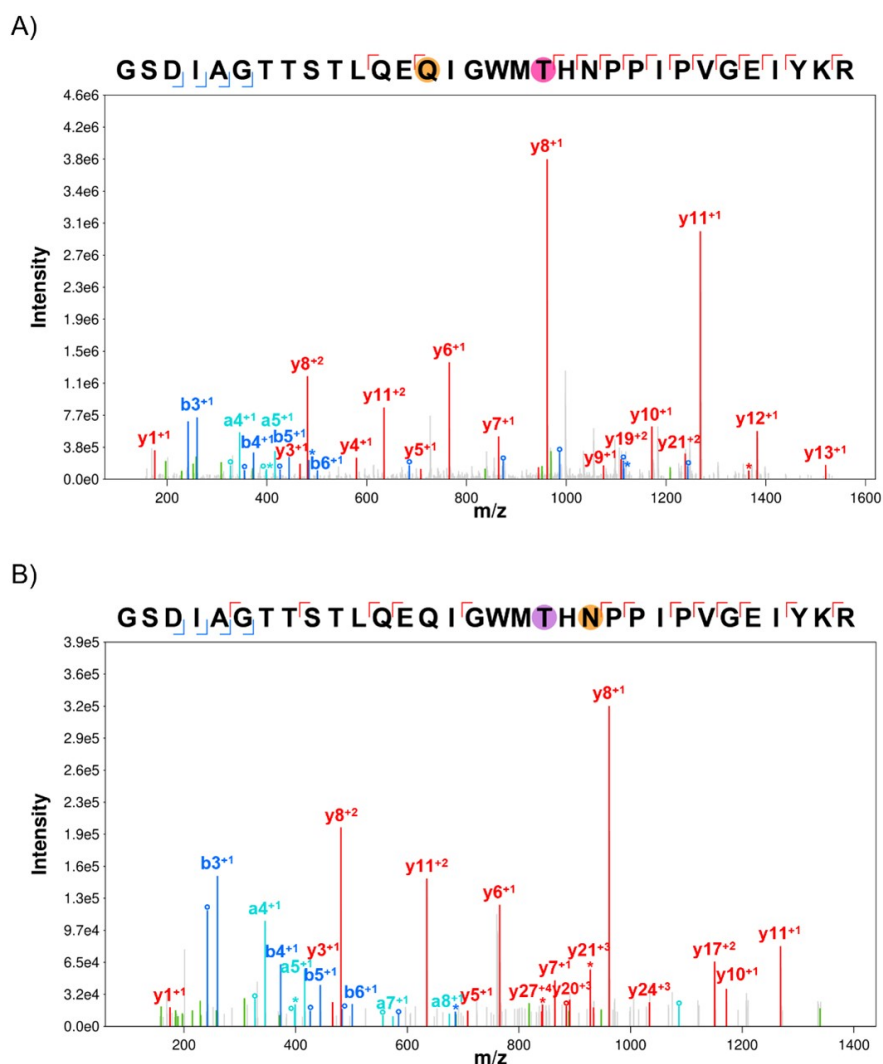


Figure 10.3: Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 119 in the HIV capsid protein. A) DHB and B) glutathione-labeled DHB, which is normally a threonine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHB modification is highlighted by the pink circle, and the site of the glutathione-labeled DHB modification is highlighted by the purple circle. The orange circle indicates deamidation in panel A and hydroxylation in panel B. In panel A, the y-ions 19 and 21 confirm the DHB modification, and are shifted by -18.01 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, y-ions 17, 20, 21, 24, and 27 all confirm the glutathione-labeled DHB modification and are shifted by $+289.07$ Da relative to the theoretical m/z of the peaks of an unmodified peptide, and by 307.32 Da relative to those of the DHB modified peptide.

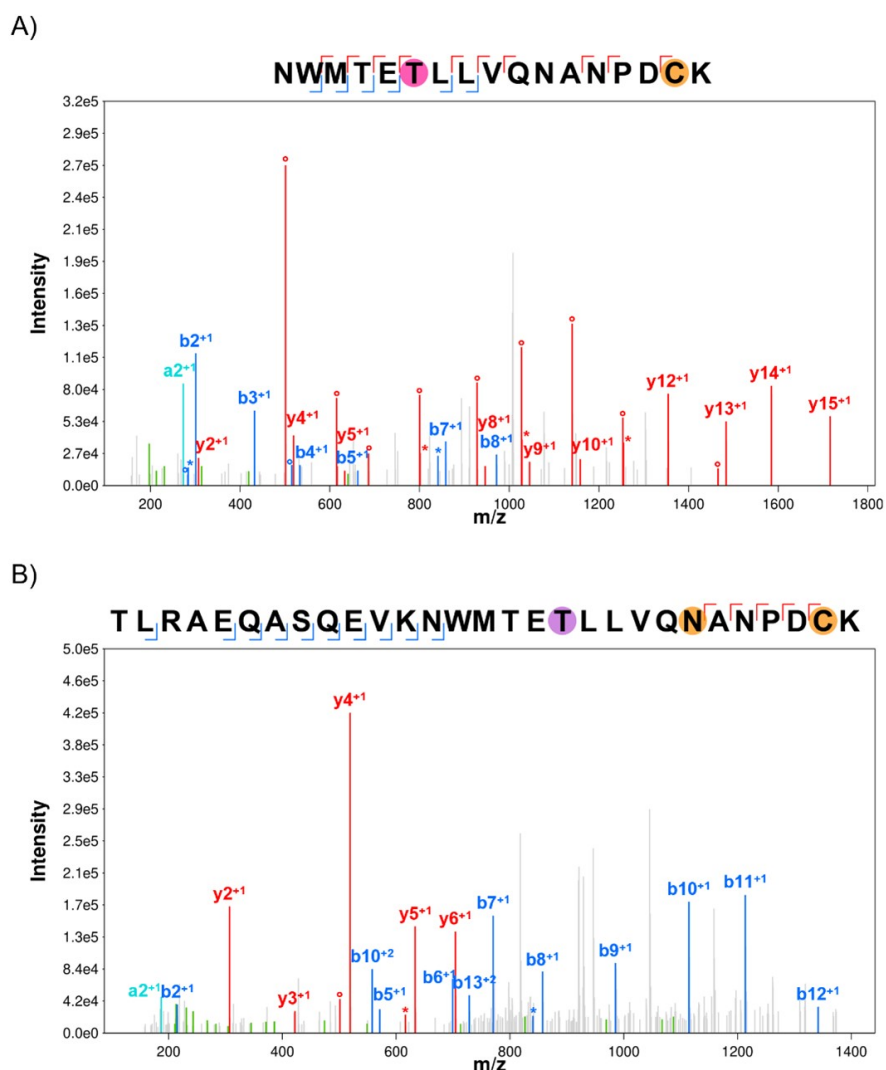


Figure 10.4: Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 188 in the HIV capsid protein. A) DHB and B) glutathione labeled DHB, which is normally a threonine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHB modification is highlighted by the pink circle, and the site of the glutathione-labeled DHB modification is highlighted by the purple circle. The orange circle covering cysteines in panel A and B indicate carbamidomethylation, the other orange circle in panel B indicates hydroxylation. In panel A, the y-ions 12-15 confirm the DHB modification, and are shifted by -18.01 Da relative to the theoretical m/z of the peaks of an unmodified peptide.

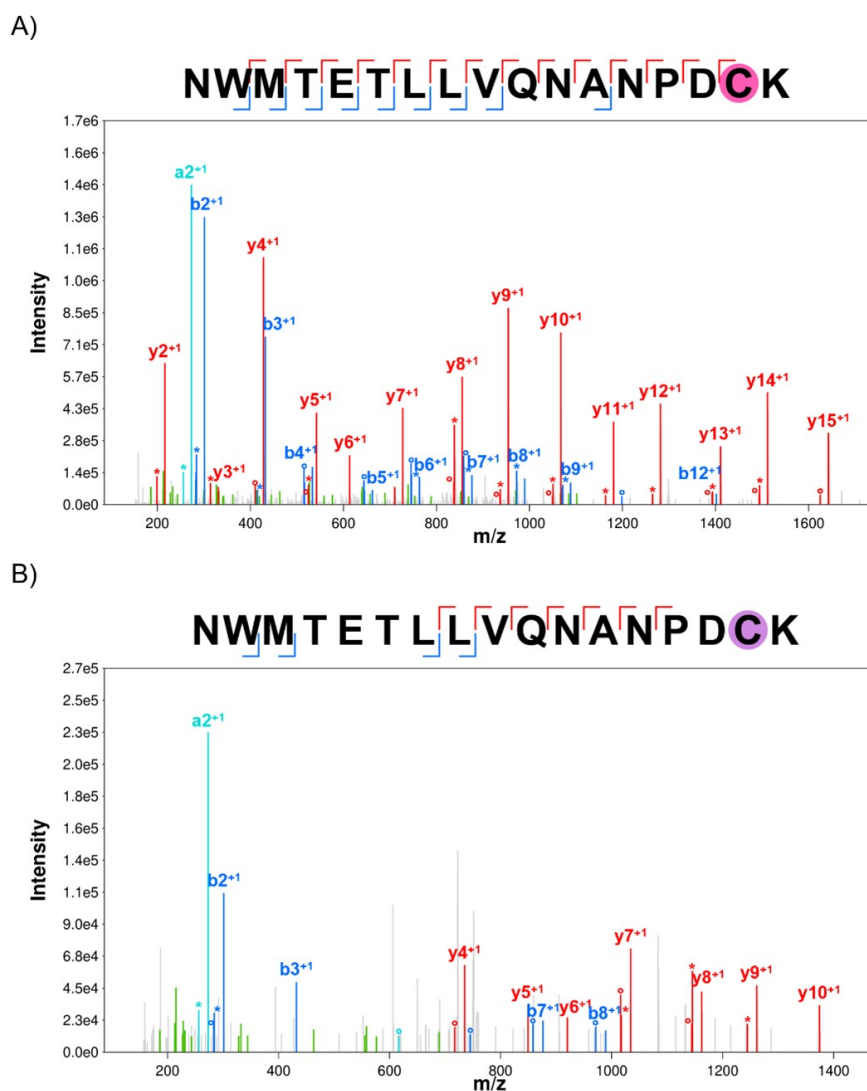


Figure 10.5: Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 198 in the HIV capsid protein. A) DHA and B) glutathione-labeled DHA, which is normally a cysteine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHA modification is highlighted by the pink circle, and the site of the glutathione-labeled DHA modification is highlighted by the purple circle. In panel A, y-ions 2-15 all confirm the DHA modification, and are shifted by -33.987 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, y-ions 4-10 all confirm the glutathione-labeled DHA modification and are shifted by $+273.096$ Da relative to the theoretical m/z of the peaks of an unmodified peptide, and by $+307.32$ Da relative to those of the DHA modified peptide.

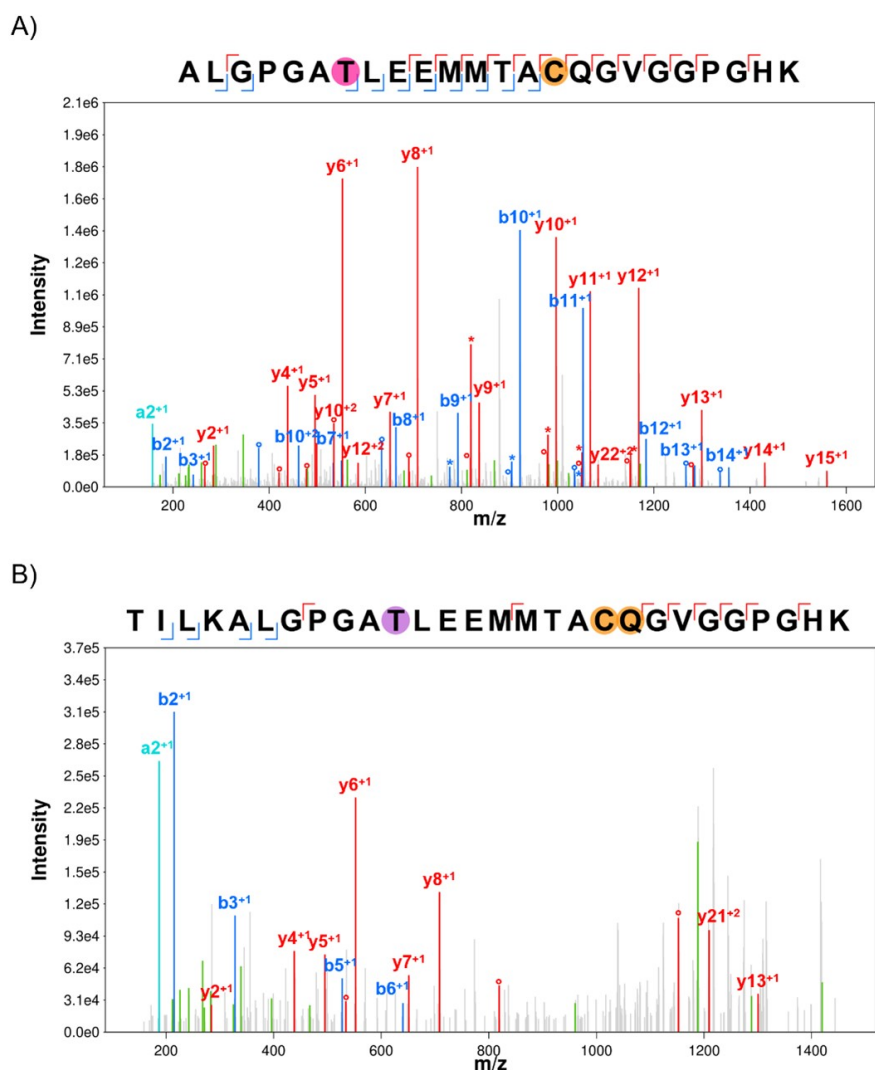


Figure 10.6: Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 210 in the HIV capsid protein. A) DHB and B) glutathione-labeled DHB, which is normally a threonine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHB modification is highlighted by the pink circle, and the site of the glutathione labeled DHB modification is highlighted by the purple circle. Orange circles in the annotated sequence for both panels, indicate carbamidomethylation. In panel b the second orange circle indicates deamidation. In panel A, the y22 ion as well as the b-ions 7-14 all confirm the DHB modification, and are shifted by -18.01 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, the y21 ion confirms the glutathione-labeled DHB modification and is shifted by +289.07 Da relative to the theoretical m/z peak of an unmodified peptide, and by +307.32 Da relative to that of the DHB modified peptide.

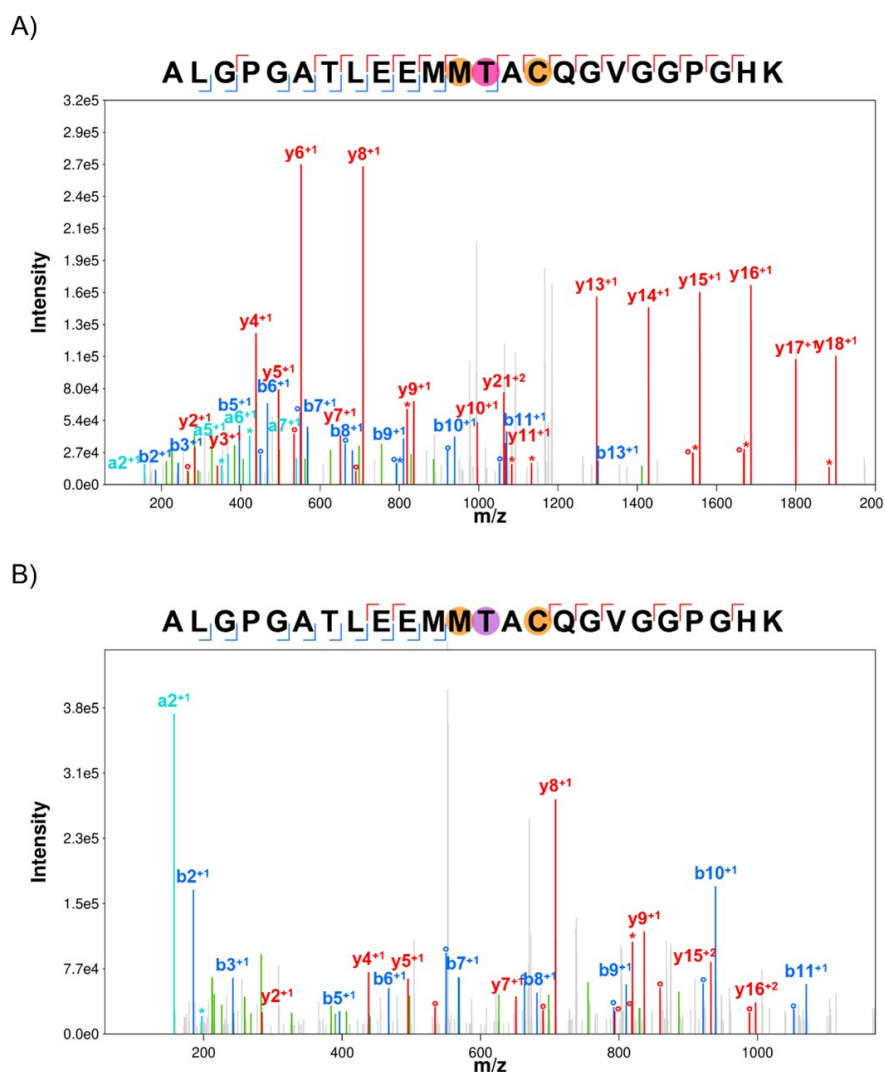


Figure 10.7: Annotated spectra for peptide identifications confirming the DHB and glutathione-labeled DHB at residue 216 in the HIV capsid protein. A) DHB and B) glutathione-labeled DHB, which is normally a threonine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHB modification is highlighted by the pink circle, and the site of the glutathione-labeled DHB modification is highlighted by the purple circle. Orange circles in the annotated sequence for both panels, indicate carbamidomethylation and oxidation. In panel A, y-ions 13-18 and 21 as well as the b13 ion all confirm the DHB modification, and are shifted by -18.01 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, y-ions 15 and 16 all confirm the glutathione-labeled DHB modification and are shifted by +289.07 Da relative to the theoretical m/z of the peaks of an unmodified peptide, and by +307.32 Da relative to those of the DHB modified peptide.

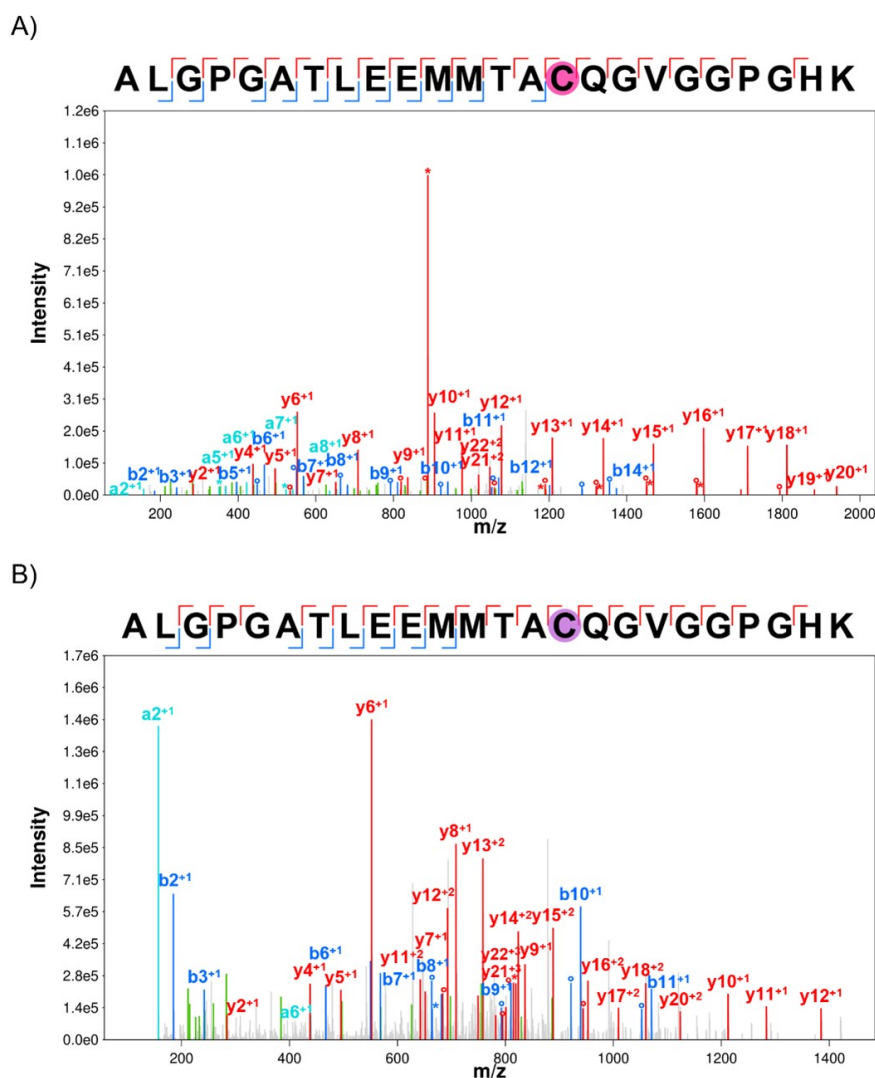


Figure 10.8: Annotated spectra for peptide identifications confirming the DHA and glutathione-labeled DHA at residue 218 in the HIV capsid protein. A) DHA and B) glutathione-labeled DHA, which is normally a cysteine in an unmodified peptide. Identified y-ions are in red, b-ions in blue, a-ions in cyan and internal ions in green. Water loss ions are annotated with a degree symbol, and ammonia loss ions are annotated with a star. Within the annotated sequence, the site of DHA modification is highlighted by the pink circle, and the site of the glutathione-labeled DHA modification is highlighted by the purple circle. In panel A, y-ions 10-22 all confirm the DHA modification, and are shifted by -33.987 Da relative to the theoretical m/z of the peaks of an unmodified peptide. In panel B, y-ions 10-18 and 20-22 all confirm the glutathione-labeled DHA modification and are shifted by $+273.096$ Da relative to the theoretical m/z of the peaks of an unmodified peptide, and by $+307.32$ Da relative to those of the DHB modified peptide.

10.4 Supplementary Tables

Table 10.2: Summary of Proteomic Search Results for Unlabeled and Glutathione-Labeled Samples

Unlabeled Samples						
Sample	Target PSMs	Contaminant PSMs	Target Peptides	Contaminant Peptides	Target Protein Groups	Contaminant Protein Groups
	1% FDR	1% FDR	1% FDR	1% FDR	1% FDR	1% FDR
All	169,860	8,229	48,541	1,324	3,742	63
a	41,630	2,133	24,969	879	3,101	59
b	41,509	1,729	26,397	817	3,070	59
c	43,280	2,418	27,007	1,012	3,044	57
d	43,411	1,949	28,851	1,035	2,996	58
Glutathione Labeled Samples						
Sample	Target PSMs	Contaminant PSMs	Target Peptides	Contaminant Peptides	Target Protein Groups	Contaminant Protein Groups
	1% FDR	1% FDR	1% FDR	1% FDR	1% FDR	1% FDR
All	120,210	6,568	34,249	1,064	3,273	57
G _a	26,205	1,544	16,510	689	2,436	55
G _b	27,969	1,420	17,075	647	2,374	52
G _c	25,503	1,610	16,232	721	2,388	54
G _d	40,533	1,985	24,601	891	2,866	56

Table 10.3: Relative Abundance of DHA/DHB containing PSMs in the 100 Most Abundant Human Proteins

Protein Group	Total PSMs	DHA/ DHB PSMs	% DHA/ DHB	# of DHA/ DHB PSMs	# of DHA/ DHB PSMs (2+ res.)	% of DHA/ DHB PSMs (2+ res.)	# of DHA/ DHB PSMs (2+ res.)	# of DHA/ DHB PSMs (2+ res.)	% of DHA/ DHB PSMs (2+ res.)	# of DHA/ DHB PSMs (2+ res.)	# of DHA/ DHB PSMs (2+ res.)
Q00610	4,742	62	0.013	10	10	0	0	0	0	0	0
P68104	2,999	100	0.033	14	9	5	0.0017	1	1	1	1
Q5VTE0	2,740	96	0.035	13	8	5	0.0018	1	1	1	1
P62937	2,451	65	0.027	8	3	14	0.0057	1	1	1	1
P0DMV8 P0DMV9	1,601	13	0.0081	2	2	0	0	0	0	0	0
P63261	1,456	65	0.045	5	5	0	0	0	0	0	0
P60709	1,369	66	0.048	6	6	0	0	0	0	0	0
P62805	1,358	16	0.012	4	3	0	0	0	0	0	0
Q05639	1,314	53	0.040	7	5	0	0	0	0	0	0
Q5QNW6	1,247	43	0.034	5	3	0	0	0	0	0	0

Continuation of Table 10.3

Protein Group	Total	DHA/ %		# of #		# of #		GSH %		# of #		# of #		
		PSMs	DHB	DHA/	DHA/	DHA/	DHA/	DHA/	DHB	PSMs	GSH	PSMs	GSH	PSMs
Q99880	1,246	43	0.035	9	5	3	7	0.0056	0	0	0	0	0	
O60814	1,237	43	0.035	9	5	3	0	0	0	0	0	0	0	
P33778	1,183	29	0.025	5	3	2	0	0	0	0	0	0	0	
Q8N257	1,145	29	0.025	5	3	2	6	0.0052	0	0	0	0	0	
P11142	1,087	2	0.0018	0	0	0	0	0	0	0	0	0	0	
P53675	872	7	0.0080	3	1	1	0	0	0	0	0	0	0	
Q8WUM4	870	27	0.031	4	4	4	0	0	0	0	0	0	0	
Q16777	782	3	0.0038	2	0	0	0	0	0	0	0	0	0	
P13639	776	5	0.0064	2	1	1	0	0	0	0	0	0	0	
P10809	744	5	0.0067	2	1	1	0	0	0	0	0	0	0	
Q09666	704	2	0.0028	2	0	0	0	0	0	0	0	0	0	
Q71DI3	675	15	0.022	3	1	1	2	0.0030	2	0	0	0	0	

Continuation of Table 10.3

Protein Group	Total	DHA/ %		# of #		# of #		GSH %		# of #		# of #		
		PSMs	DHB	DHA/	DHA/	DHA/	DHA/	DHA/	DHA/	PSMs	PSMs	PSMs	PSMs	PSMs
				DHB	DHB	DHB	DHB	DHB	DHB	DHB	DHB	DHB	DHB	DHB
				PSMs	PSMs	PSMs	PSMs	PSMs	PSMs	PSMs	PSMs	PSMs	PSMs	PSMs
				peps	peps	peps	peps	peps	peps	peps	peps	peps	peps	peps
				(2+	(2+	(2+	(2+	(2+	(2+	(2+	(2+	(2+	(2+	(2+
				reps)	reps)	reps)	reps)	reps)	reps)	reps)	reps)	reps)	reps)	reps)
Q9H4B7	111	1	0.0090	0	0	0	0	0	0	0	0	0	0	0
Q9Y277	104	2	0.019	1	1	1	0	0	0	0	0	0	0	0
P26640	100	1	0.010	1	0	0	0	0	0	0	0	0	0	0
P49755	92	2	0.022	1	0	0	0	0	0	0	0	0	0	0
P09104	91	4	0.044	1	1	1	0	0	0	0	0	0	0	0
P63092 Q5JWF2	88	1	0.011	0	0	0	0	0	0	0	0	0	0	0
End of Table 10.3														

Table 10.4: HIV-1 Protein Sequences for Search Database

Begin Table 10.4	Protein Sequence
Protein Accession P04591_WTJB474	MGARASVLSGGELDKWEKIRLRPGGKKQYKCLKHIVWASRELERFAVNPGL LETSEGCRQLGQLQPSLQTGSEELRSLYNTIAVLYCVHQRIDVKDTKEALD KIEEEQNKSKKKAQQAADTGNNSQVSQNYPIVQNLQGQMVHQAI SPRTL NAWVKVVEEKAFSPEVPMFSALSEGATPQDLNNTMLNTVGGHQAAAMQMLK ETINEEAAEWDRLHPVHAGPIAPGQMREPRGSDIAGTTSTLQEIQIGWMTH NPPIPVGEIYKRWIILGLNKIVRMYSPSILDIRQGPKEPFRDYVDRFYKTLRA EQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGVGGP GHKARVLAEAMSVTNPATIMIQGNFRNQRKTVKCFNCGKEGHI AKNCR APRKKGWCWKCQKEGHQMKDC'ETERQANFLGKIWPSHKGRPGNFLQSRPEPT APPEESFRFGEETTPSQKQEPIDKELYPLASLSLFGSDPSSQ

Continuation of Table 10.4

P04585_WTJB474

MGARASVLSGGELDKWEKIRLRPGGKKQYKXKHIVWASRELERFAVNPGL
 LETSEGRQLGQLQPSLQTSEELRSLYNTIAVLYCVHQRIDVKDTKEALD
 KIEEEQNKSKKKAQQAADTGNNSQVSNYPVQNLQGMVHQAI SPRTL
 NAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNMLNTVGGHQAAAMQMLK
 ETINEEA AEWDRLHPVHAGPIAPGMREPRGSDIAGTTSTLQEIGWMTH
 NPPPIVGEIYKRWILGLNKIVRMYPTSILDIRQGPKEPRDYVDRFYKTLRA
 EQASQEVKNWMTETLLVQNANPDKTILKALPGATLEEMMTACQGVGGP
 GHKARVLAEAMSQVTNPATIMIQQGNFRNQRKTVKCFNCCKEGHIAKNCR
 APRKKGCKWCKEGHQMKDCTERQANFFREDLAFQGKAREFSEQTRANSP
 TRRELQVWGRDNNSLSEAGADRQGTVSFSFPQITLWQRPLVTIKIGGQLKEALLD
 TGADDTVLEEMNLPGRWKPKMIGGIGGFIVRQYDQILIEICGHKAIGTVLVGP
 TPVNIIGRNLLTQIGCTLNFPIPIETVPVKKLPGMDGPKVKQWPLTEEKIK
 ALVEICTEMEKEGKISKIGPENYPNTPVFAIKKDKSTKWRKLYDFRELNKRT
 QDFWEVQLGIPHPAGLKQKKSVTVLDVGDAYFSVPLDKDFRKYTAFTIPSIN
 NETPGIRYQYNVLPQGWKGSPAIQCSMTKILEPFRKQNPDIVYQYMDDDL
 VGSDEIGQHRTKIEELRQHLLRWGFTTPDKKHQKEPPFLWMGYELHPDKWT
 VQPIVLEPEKDSWTVNDIQKLVGKLNWASQIYAGIKVRQLCKLLRGTKALTEV
 VPLTEEAELAEENREILKEPVHGYYDPSKDLIAEIQKQGGQWYTYQYQE
 PFKNLKTGKYARMKGAHTNDVKQLTEAVQKIATESIVIWGKTPKFKLPIQKE
 TWEAWWTEYWQATWIPEWEFVNTPLV'KLWYQLEKEPII GAETFFYVDGAANR
 ETKLGKAGYVTDGRQKVVPLTDTTNQKTELQAIHLALQDSGLEVNIVTDSQ
 YALGIIQAQPDKSESELYSQIEQLIKKEKVVYLAWVPAHKGIGNEQVDKLVS
 GIRKVLFLDGDIDKAQEEHEKEYHSNWRAMASDFNLPPVVAKEIVASCDCQLK
 GEAMHGQVDCSPGIWQLDCTHLEGVILVAVHVASGYIEAEVIPAETGQETA
 YFLKLAGRWPVKTVHTDNGSNFTSTTVKAAACWWAGIKQEFGIPYNIPQSQGV
 IESMNKELKIIIGQVRDQAEHLKTAVQMAVFIHNFKRKGGIGGYSAGERIVD
 IIAATDIQTKELQKITKIQNFRVYRDSRDPVWKGPAKLLWKGEGAVVIQDN
 SDIKVVPRRKAKIIRDYGKQ MAGDDCVASRQDED

P04618_WTJB474

MAGRSGDSDEELIRTVRLIKLLYQSNPPNPEGTRQARRNRRRRWRERQRQIHSISE
 RILSTYLGRSAEPVPLQLPPLERLTLDCNEDCGTSGTQGVGSPQILVESPTVLESGTKE

Continuation of Table 10.4	
P05919_WTJB474	MQPIIWAIVALVVAIIIIVVWSVIEIYKILRQRKIDRLIDRLIERAEDSGNESE GEVSALVEMGVEMGHHPWDIDDL
P69723_WTJB474	MENRWQVMIVWQVDRMRINTWKRLVKHHMYSRKAADWFWYRHHYESTNPKIS SEVHIPLDGAKLVITTYWGLHTGERDWHLGQGVSEIWRKKRYSTQVDPDLAD QLIHLHYFDCFSES AIRNTILGRIVSPRCEYQAGHNKVGSLQLALAALIKPKQIKP PLPSVRKLTEDRWNKPKQTKGHRGSHTMNGH
P04608_WTJB474	MEPVDPRLEPWKHPGSQPKTACTNICYCKKCCFHCQVCFMTKALGISYGRKKRRQ RRRAHQNSQTHQASLSKQPTSQSRGDPITGPKE
P04591_MA_WTJB474	MGARASVLSGGELDKWEKIRLPGGKKQYKLVKHIWASRELERFAVNPGLLET SEGCRQILGQLQPSLQTSSEELRSLYNTIAVLYCVHORIDVKDTKEALDKIEEEQ NKSKKKAQQAADTGNNSQVVSQNY
P04591_CA_WTJB474	PIVQNLQGMVHQVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDL NTMLNTVGGHQAAMQMLKETINEEA AEWDRLHPVHAGPIAPGQMPREPRGSDI AGTTSTLQEQIGWMTHTNPPPIPVGEIYKRWILGLNKIVRMYSPTSILDIRQG PKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQANANPDCKTILKALGPG ATLEEMMTACQGVGGPGHKARVL
P04591_SP1_WTJB474	AEAMSQVTNPATIM
P04591_NC_WTJB474	IQKGNFRNQRTVKCFNCGKEGHIANKNCRAPRKKGCWKCGKEGHQMKDCTE RQAN
P04591_SP2_WTJB474	FLGKIWPSHKGRPGNF
P04591_P6_WTJB474	LQSRPEPTAPPEESFRFGEETTTTSPQKQEPIDKELYPLASLSLFGSDPSSQ
P04591_CA_SP1_WTJB474	PIVQNLQGMVHQVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDL NTMLNTVGGHQAAMQMLKETINEEA AEWDRLHPVHAGPIAPGQMPREPRGSDI AGTTSTLQEQIGWMTHTNPPPIPVGEIYKRWILGLNKIVRMYSPTSILDIRQG PKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQANANPDCKTILKALGPG ATLEEMMTACQGVGGPGHKARVLAEAMSQVTNPATIM
P04591_SP1_NC_WTJB474	AEAMSQVTNPATIMIQKGNFRNQRTVKCFNCGKEGHIANKNCRAPRKKGCWK GKEGHQMKDCTERQAN
P04591_SP1_NC_SP2_WTJB474	AEAMSQVTNPATIMIQKGNFRNQRTVKCFNCGKEGHIANKNCRAPRKKGCWK GKEGHQMKDCTERQANFLGKIWPSHKGRPGNF

Continuation of Table 10.4

P04591_SP1_NC_SP2_P6_WTJB474	AEAMSVTNPATIMIQKGNFRNRQKTVKFCNCGKEGHIAKNCRAPRKKGCWK KKEGHQMKDCTERQANFLGKIWPSHKGRPGNFLQSRPEPTAPPEESFRFGEETT TPSQKQEPIDKELYPLASLRSLFGSDPSSQ
P04591_NC_SP2_P6_WTJB474	IQKGNFRNRQKTVKFCNCGKEGHIAKNCRAPRKKGCWKCGKEGHQMKDCTE RQANFLGKIWPSHKGRPGNFLQSRPEPTAPPEESFRFGEETTTPSQKQEPIDKEL YPLASLRSLFGSDPSSQ
P04591_SP2_P6_WTJB474	FLGKIWPSHKGRPGNFLQSRPEPTAPPEESFRFGEETTTPSQKQEPIDKELYPLA SLRSLFGSDPSSQ
P04585_TF_WTJB474	FFREDLAFPQGGKAREFSEQTRANSPTRRELQVWGRDNNLSLSEAGADRQGTVSFSF
P04585_PR_WTJB474	PQITLWQRPLVTIKIGGQKAEALLDTGADDIVLEEMNPLGRWPKMIGGIGGF IKVRQYDQILIEICGKKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
P04585_RT51_WTJB474	PISPIETVPVKKLPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENP YNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGHPAGLKQKKSVTVLD VGDAYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPIAIFQCSMT KILEPFRKQNPDIVIYQYMDLTVGSDLEIGQHRTKIEELRQHLLRWGFTTPDKK HQKEPPFLWGYELHPDKWTVQPIVLPKDSWTVNDIQKLVGKLNWASQIYAGIK VRQCKLLRGTKALTEVPLTEEALELAENREILKEPVHGVYDPSKDLIAEIQ KQGGQGWTYQIYQEPFKNLKTGKYARMKGAHTNDVKQLTEAVQKIATESIVI WGKTPKFKLPIQKETWEAWWTEYWQATWIPEWEFVNTPLVWKLWYQLEKEPI IGAETF
P04585_RT65_WTJB474	PISPIETVPVKKLPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPN TPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGHPAGLKQKKSVTVLD VGDAYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPIAIFQCS MTKILEPFRKQNPDIVIYQYMDLTVGSDLEIGQHRTKIEELRQHLLRWGFTT PDKKHQKEPPFLWGYELHPDKWTVQPIVLPKDSWTVNDIQKLVGKLNWASQ IYAGIKVRQCKLLRGTKALTEVPLTEEALELAENREILKEPVHGVYDPS KDLIAEIQKQGGQGWTYQIYQEPFKNLKTGKYARMKGAHTNDVKQLTEAVQKI ATESIVWGTTPKFKLPIQKETWEAWWTEYWQATWIPEWEFVNTPLVWKLWYQ LEKEPIGAETFYVDGAANRETKLGKAGYVTRDRQKVVPLTDTTNQKTELQA IHLALQDSGLEVNIVTDSQYALGIQQAQPKSESELYSQIIEQLIKKEKVLAWV PAHKGIGGNEQVDKLVSAIRKVL

Continuation of Table 10.4	
P04585_InterRT_WTJB474	YVDGAANRETKLGKAGYVTDGRQKVVVPLTDTTNQKTELQAIHLALQDSGL EVNIVTDSQYALGIIQAQPDKSESELYSQIEQLIKKEKVVYLAWVPAHKGIGGNE QVDKLVSA GIRKVL
P04585_IN_WTJB474	FLDGIDKAQEEHEKEYHSNWRAMASDFNLPVVVAKEIVASCDKCKQLKGEAMH GQVDCSPGIWQLDCTHLEKGVILVAVHVASGYIEAEVIPAETGQETAYFLKLA GRWPVKTVHTDNGSNFTSTTVKAAACWWAGIKQEFGIPYNPQSQGVIESMNKELK KIIGQVRDQAEHLKTAVQMAVFIHNFKRKGGIGGYSAGERIVDIIATDIQTKEL QKQITKIQNFRVYYRDSRDPVWKGPAKLLWKGEGAVVIQDNSDIKVVPRRKAKI IRDYGGKQ MAGDDCVASRQDED
P04585_TF_PR_WTJB474	FFREDLAFPQGKAREFSSEQTRANSPTRRELQVWGRDNNLSLSEAGADRQGTVS FSFPQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWPKMIGGIG GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQJGCTLNF
End of Table 10.4	

Table 10.5: MetaMorpheus GPTMD Modifications

Common Biological	Common Biological Cont.	Less Common	Common Artifact	Metal	Custom
Acetylation on K	HexNAC on S	DHA on S	Ammonia loss on C	Calcium on D	GSH on S
Acetylation on X	HexNAC on T	DHA on C	Ammonia on N	Calcium on E	GSH on C
ADP-ribosylation on S	Hydroxybutyrylation on K	DHB on T	Carbamyl on C	Cu[I] on D	GSH on T
Butyrylation on K	Hydroxylation on K		Carbamyl on K	Cu[I] on E	
Carboxylation on D	Hydroxylation on N		Carbamyl on M	Fe[II] on D	
Carboxylation on E	Hydroxylation on P		Carbamyl on R	Fe[II] on E	
Carboxylation on K	Malonylation on K		Carbamyl on X	Fe[III] on D	
Citrullination on R	Methylation on K		Deamidation on N	Fe[III] on E	
Crotonylation on K	Methylation on R		Deamidation on Q	Magnesium on D	
Dimethylation on K	Nitrosylation on C		Water loss on E	Magnesium on E	
Dimethylation on R	Nitrosylation on Y			Potassium on D	
Formylation on K	Phosphorylation on S			Potassium on E	
Glu to PyroGlu on Q	Phosphorylation on T			Sodium on D	
Glutarylation on K	Pyridoxal phosphate on K			Sodium on E	
HexNAC on Nxs	Succinylation on K			Zinc on D	
HexNAC on Nxt	Sulfonation on Y			Zinc on E	
Trimethylation on K					

Table 10.6: MetaMorpheus Search Task Settings

Search Mode	In silico Digestion Params	Fragment Ion Params	Mass Difference Acceptors	Dif-Params	Ambiguity Params	Scoring Options	Post-Search Analysis
Classic Search	Protease: Trypsin	Dissoication Type: HCD	Missed Monoisotopic Peak	Report ambiguity	PSM ambiguity	Minimum Score allowed: 5	Apply protein parsimony and construct protein groups
Generate proteins	target	Max Threads: 39					
Generate proteins	decoy	Max Fragment Mass (Da): 3000					
Max Missed Cleavages: 2		N-Terminal Ions					
Initiator Ions: Variable	Methionine	C-Terminal Ions					
Max Modification Isoforms: 1024							
Min Length: 7	Peptide						
Max Length: None	Peptide						
Max Mods per Peptide: 2							

COLOPHON

This document was typeset with \LaTeX . It is based on the University of Wisconsin dissertation template created by William C. Benton (available at <https://github.com/willb/wi-thesis-template>).