PROFILING OF THE SORF-ENCODED PEPTIDES AND DISCOVERY OF THEIR POTENTIAL BIOLOGICAL FUNCTION USING MASS SPECTROMETRY

ZHANG Yuanliang

PhD

THE HONG KONG POLYTECHNIC UNIVERSITY

2025

The Hong Kong Polytechnic University

Department of Applied Biology and Chemical Technology

Profiling of the sORF-encoded Peptides and Discovery of Their Potential Biological Function using Mass Spectrometry

ZHANG YUANLIANG

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

June 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces neither material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

ZHANG Yuanliang

Abstract

The discovery of non-canonical proteins, such as small open reading frame-encoded peptides (sORF-encoded peptides, SEPs), alternative proteins (AltProts), and cryptic immunopeptides, has challenged the traditional boundaries of proteomics, revealing a hidden layer of the proteome. These biomolecules, often overlooked due to their small size, low abundance, and origins in genomic regions previously considered non-coding, challenge traditional proteomic paradigms. To overcome detection barriers, we developed advanced mass spectrometry-based approaches, integrating data-independent acquisition (DIA) proteomics, immunopeptidomics, and CRISPR-based functional screening. These methods, applied across three distinct studies corresponding to the core chapters of this thesis, provide novel insights into the biological roles of non-canonical proteins and establish robust tools for their characterization.

First in Chapter 2, we optimized a DIA-based proteomics workflow to enhance AltProt detection in mouse cardiac development. Recognizing the limitations of conventional data-dependent acquisition, we employed computationally predicted spectral libraries to improve sensitivity, achieving a twofold increase in AltProt identifications with reduced missing values. Applied to embryonic and adult mouse heart tissues, this approach identified over 50 differentially expressed AltProts enriched in pathways linked to cellular differentiation. Through targeted validation using Western blotting and parallel reaction monitoring, we confirmed ASDURF, an upstream open reading

frame (uORF)-derived protein, as a regulator of cardiomyocyte maturation, highlighting the developmental importance of AltProts and establishing a reliable pipeline for their study in other biological contexts.

Next in Chapter 3, we addressed the challenges of identifying MHC-presented peptides from non-canonical sources by developing a Pseudo-DIA Library Search Strategy for immunopeptidomics. This method, combining unrestricted DIA searches with in silico predicted libraries, increased immunopeptide detection by up to 3.8-fold compared to standard approaches. In human cell lines, we identified cryptic peptides and neoantigens encoded by sORFs, assessing their immunogenic potential for personalized cancer immunotherapy. To facilitate broader adoption, we created a user-friendly graphical interface for generating spectral libraries from MSFragger outputs, validated across independent datasets. This work expands the scope of immunopeptidomics and supports the development of targeted cancer therapies.

Finally in Chapter 4, we conducted a large-scale analysis of over 36,000 microproteins across 86 public mass spectrometry datasets, revealing their predominant origins in long non-coding RNAs and frequent proximity to oncogenes. Conservation analyses across 100 species indicated diverse evolutionary pressures, with mammalian-specific patterns suggesting adaptive roles. Using a CRISPR knockout library in three cancer cell lines, we demonstrated that several microproteins independently regulate proliferation, distinct from their associated canonical proteins, as evidenced by

comparisons with public DepMap data. These findings position microproteins as potential therapeutic targets in oncology.

Collectively, this thesis advances our understanding of the proteome's complexity by uncovering the functional roles of non-canonical proteins in development and disease. The methodologies developed provide practical tools for future studies, while the biological insights open new avenues for precision medicine and developmental biology research.

List of publications

The originality of the works presented here is based on my representative academic publications as follows:

1. **Zhang Y**, Yang Y, Li K, Chen L, Yang Y, Yang C, Xie Z, Wang H, Zhao Q. Enhanced Discovery of Alternative Proteins (AltProts) in Mouse Cardiac Development Using Data-Independent Acquisition (DIA) Proteomics. Anal Chem. 2025 Jan 28;97(3):1517-1527.
2. Yang Y, Chen C, Li K, **Zhang Y**, Chen L, Shi J, Mu Q, Xu Y, Zhao Q. Proteogenomic Profiling Reveals Small ORFs and Functional Microproteins in Activated T Cells. Mol Cell Proteomics. 2025 Feb 4:100914.
3. Yang Y, Wang H, **Zhang Y**, Chen L, Chen G, Bao Z, Yang Y, Xie Z, Zhao Q. An Optimized Proteomics Approach Reveals Novel Alternative Proteins in Mouse Liver Development. Mol Cell Proteomics. 2023 Jan;22(1):100480.
4. Chen, L.; **Zhang, Y.**; Yang, Y.; Yang, Y.; Li, H.; Dong, X.; Wang, H.; Xie, Z.; Zhao, Q.*, An Integrated Approach for Discovering Non-canonical MHC-I Peptides Encoded by Small Open Reading Frames. J Am Soc Mass Spectrom 2021, 32 (9), 2346-2357.
5. Chen L., Yang Y., **Zhang Y.**, Li K., Cai H., Wang H., Zhao Q., The Small Open Reading Frame-Encoded Peptides: Advances in Methodologies and Functional Studies. ChemBioChem. 2021, 23(8): e202100534.
6. Wang C, Liu Z, Zeng Y, Zhou L, Long Q, Hassan IU, **Zhang Y**, Qi X, Cai D, Mao B, Lu G, Sun J, Yao Y, Deng Y, Zhao Q, Feng B, Zhou Q, Chan WY, Zhao H.

ZSWIM4 regulates embryonic patterning and BMP signaling by promoting nuclear Smad1 degradation. EMBO Rep. 2024 Feb;25(2):646-671.

7. Yang H, Li L, Jiao Y, **Zhang Y**, Wang Y, Zhu K, Sun C. Thioredoxin-1 mediates neuroprotection of Schisanhenol against MPP^+-induced apoptosis via suppression of ASK1-P38-NF-κB pathway in SH-SY5Y cells. Sci Rep. 2021 Nov 3;11(1):21604.

Conference papers

1. **Zhang, Y.**, Zhao, Q., Large-scale identification of functional Microproteins (MPs) with immunopeptidomics and CRISPR screening. The 6th ABCT Research Postgraduate Symposium, 2025. (Oral Presentation)
2. **Zhang, Y.**, Chen, L., Zhao, Q., In-silico library and two-step search improve data-independent acquisition immunopeptidomics. The 72nd American Society for Mass Spectrometry Conference on Mass Spectrometry and Allied Topics, 2024. (Poster presentation)
3. **Zhang, Y.**, Chen, L., Zhao, Q., In-silico library and two-step search improve data-independent acquisition immunopeptidomics. The 5th ABCT Research Postgraduate Symposium, 2024. (Oral Presentation)
4. **Zhang Y.**, Yang Y., Li K., Chen L., Yang Y., Yang C., Xie Z., Wang H., Zhao Q., Enhanced Discovery of Alternative Proteins (AltProts) in Mouse Cardiac Development Using Data-Independent Acquisition (DIA) Proteomics. 2022. The 3rd ABCT Research Postgraduate Symposium in the Biology Discipline (Oral Presentation & Poster)
5. Chen L.; **Zhang Y.**; Yang Y.; Yang Y.; Li H.; Dong X.; Wang H.; Xie Z.; Zhao Q., An Integrated Approach for Discovering Noncanonical MHC-I Peptides Encoded by Small Open Reading Frames. The 1st Chinese American Society for Mass Spectrometry (CASMS), 2021.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Zhao Qian (赵倩教授), for her unwavering support, encouragement, and invaluable guidance throughout the past four years of my PhD journey. I am profoundly thankful for her trust and mentorship, as she chose me from a pool of applicants to join her research group as a PhD student. Under her supervision, I have benefited immensely, gaining not only a rigorous and meticulous approach to scientific research but also cultivating resilience and determination, qualities that will undoubtedly remain invaluable assets in my academic career and personal life. For this, I owe her a debt of gratitude that is beyond words.

I am thankful to the members of Prof. Zhao's lab, including Dr. Chen Lei Alyssa, Dr. Yang Yang, Dr. Liu Zhao, Dr. Zhang Qi, Miss Tse Yin-suen Chloe, Dr. Li Shuqi, Dr. Yang Ying, Dr. Shang Jin, Mr. Li Kecheng, Miss Yang Chenxi, Mr. Zhu Tianyang, Miss Liu Qi, and Mr. Qi Ao. Their camaraderie, support, and insightful discussions have greatly enriched my research experience and my personal growth. The past four years have been a fulfilling, meaningful, and joyful journey, and the lessons I have learned from working alongside them—both in science and in life—will always stay with me.

I would also like to extend my heartfelt thanks to the professors and researchers at The Hong Kong Polytechnic University (PolyU), who provided invaluable support

throughout my PhD. In particular, Prof. Mu Quanhua, Prof. Man-kin Wong, Dr. Wong Lai-king Iris, Dr. Lee Chun-kit Alan, Dr. So Pui-kin, and Dr. Sirius Tse have assisted me with data analysis, instrument training, and experimental troubleshooting, for which I am deeply grateful. My thanks also go to the Technical Staff at PolyU, including Mr. Chan Chi-ho Roy, Ms. Wan Hoi-ying Echo, Mr. Chung Yuk-fai Tony, Ms. Teo Tuang-ngo Ivy, and Mr. Cheng Lap-yung, whose meticulous assistance in laboratory operations has been crucial. Additionally, I am grateful to the Administrative Staff, namely Ms. Kwok Fung-yee Peggy, Ms. So Hiu Tung Esther, Ms. Chui Kei-yan Crystal, and Miss Yuen Chui-ying Ailsa, for their dedicated support in handling administrative matters, which has greatly facilitated my research progress.

I am also grateful to my collaborators from other institutions. I am especially thankful to Professor WONG Siu-Lun Alan's team at The University of Hong Kong (HKU), particularly Dr. Zhou Peng Nick and Dr. Wang Bei, who taught me essential skills in CRISPR and molecular cloning. Their generosity in sharing their expertise has been invaluable, and I have learned a great deal from them. I also wish to thank Prof. Chen Qiushi at HKU, Prof. Ren Yan at Shanghai University of Traditional Chinese Medicine, Dr. Yu Fengchao at the University of Michigan, and Mr. Guo Changcheng from Thermo Fisher for their guidance and support in various aspects of my research.

I am deeply thankful for the funding support that made this research possible. My work has been supported by the Research Grants Council-GRF (15305821, 15304819), CRF

Equipment (C5033-19E), and RGC-RIF (R5050-18), which enabled us to acquire the Thermo Fisher 480 mass spectrometer, which proved instrumental in my research. I also appreciate the funding provided by Prof. Zhao Qian through the RA position and the financial support from the Center for Eye and Vision Research (CEVR) under the Health@InnoHK Programme launched by ITC. Furthermore, I am grateful to the PolyU Research Facilities ULS and UCEA for providing access to essential resources.

I owe an immeasurable debt of gratitude to my parents, grandparents, and great-grandparents. Without their love, sacrifices, and unwavering belief in me, I would not have been able to embark on this journey from northern China to the Hong Kong Polytechnic University to complete my PhD. Their support is the foundation of my achievements.

Finally, I wish to express my heartfelt appreciation to my fiancée, Miss Hu Dandan, for her understanding, encouragement, and steadfast support during the countless days and nights. Her unwavering faith in me and companionship have been sources of strength and comfort. I am eternally grateful to have her by my side.

Grant support

This study was generously financed by,

1. RGC Collaborative Research Fund Equipment (C5033-19E),
2. RGC-GRF (15304819, 15305821),
3. RGC-RIF R5050-18.

I am deeply grateful for the generous support provided by the aforementioned grants.

List of Abbreviations

Artificial Intelligence	AI
Ammonium Bicarbonate	ABC
Alternative Open Reading Frames	AltORFs
Alternative Proteins	AltProts
Automatic Gain Control	AGC
Benjamini-Hochberg	BH
Biological Process	BP
Chromatographic Separation	CS
Co-Immunoprecipitation	Co-IP
Collision Energy	CE
CRISPR Knockout	CRISPR KO
Data-Dependent Acquisition	DDA
Data-Independent Acquisition	DIA
Differentially Expressed Proteins	DEPs
Dimethyl Pimelimidate	DMP
Dithiothreitol	DTT
Downstream Open Reading Frames	dORFs
Downstream overlapping ORFs	doORFs
Dulbecco's modified Eagle's medium	DMEM
Electrospray Ionization	ESI
False Discovery Rate	FDR

Fetal Bovine Serum	FBS
Fold Change	FC
Gas Phase Fractionation	GPF
Gene Ontology	GO
Higher Collisional Dissociation	HCD
Human Leukocyte Antigen	HLA
Immunopeptidomics	IP
Immunopeptides	IPs
Internal Open Reading Frames	intORFs
Iodoacetamide	IAM
Liquid Chromatography	LC
Long non-coding RNAs	lncRNAs
Major Histocompatibility Complex	MHC
Mass Spectrometry	MS
Maximum Injection Time	MIT
Microproteins	MPs
Non-Coding RNAs	ncRNAs
Oxidation	Ox
Open Reading Frames	ORFs
Peptide-Spectrum Matches	PSMs
Post-Translational Modifications	PTMs
Retention Time	RT

Principal Component Analysis	PCA
Ribosome Profiling	Ribo-seq
Retention Time	RT
Reference Proteins	RefProts
sORF-encoded Peptides	SEPs
Small Open Reading Frames	sORFs
Spectral Angle	SA
Tandem Mass Spectrometry	MS/MS
Target Decoy Strategy	TDS
Transcript Support Level	TSL
The Cancer Genome Atlas	TCGA
Untranslated regions	UTRs
Uniprot Protein Database	UniProt
Upstream Open Reading Frames	uORFs
Upstream overlapping ORFs	uoORFs

Table of Contents

Abstract ..... **I**

List of publications ..... **IV**

Acknowledgements ..... **VII**

Grant support ..... **X**

List of Abbreviations ..... **XI**

Table of Contents ..... **XIV**

List of Tables and Figures ..... **XVIII**

Chapter 1. Overview **1**

1.1 Overview of Microproteins Derived from Small Open Reading Frames 1
1.2 Microprotein Identification Based on DIA Mass Spectrometry 3
1.2.1 Introduction 3
1.2.2 Mass Spectrometry Scanning Methods 5
1.2.3 Library Construction for DIA search analysis 9
1.2.4 DIA Data search tools 10
1.3 Research goals and objectives 11
1.4 Outline of this thesis 15

Chapter 2. Enhanced Discovery of Alternative Proteins (AltProts) in Mouse Cardiac Development Using Data-Independent Acquisition (DIA) Proteomics **19**

2.1 Introduction 19
2.2 Materials and methods 21
2.2.1 Sample collection 21
2.2.2 Sample preparation 22
2.2.3 LC-MS analysis 22
2.2.4 MS Raw data searching 23

2.2.5 Differential expression analysis and spectra validation 24
2.2.6 Western blot analysis. 25
2.2.7 Identification of more microproteins using the PRM method 26
2.2.8 Public MS data reanalysis 26
2.2.9 AltProts database construction 27
2.2.10 Spectral angle calculation 28
2.3 Results and Discussion 28
2.3.1 The design of comparative evaluation of different MS-based workflows for AltProt analysis. 28
2.3.2 DIA outperformed DDA in identification and quantification of AltProts. 29
2.3.3 Different library construction methods influence AltProt identification. 33
2.3.4 Differentially expressed AltProts in the mouse heart development were detected by using an optimized DIA method 43
2.3.5 Validation of expression of ASDURF and other AltProts. 48
2.4 Conclusions 69
2.5 Limitations 70
Chapter 3. A Pseudo-DIA library search approach improves immunopeptidomics and neoantigen discovery 72
3.1 Introduction 72
3.2 Materials and methods 75
3.2.1 Cell Culture 75
3.2.2 MHC-I Immunopeptide Enrichment and Purification 75
3.2.3 Mass Spectrometry Data Acquisition 76

3.2.4 Mass Spectrometry Data Search 77

3.2.5 Prediction of Immunopeptide Binding and Immunogenicity 77

3.2.6 Two-Species Pseudo-Target FDR Estimation 78

3.2.7 RT and Spectrum Prediction Using DIA-NN 78

3.2.8 Public Data Sources 79

3.2.9 sORFs Genomic Localization 79

3.2.10 Mass Spectrometry Spectrum Annotation 79

3.3 Results 80

3.3.1 Pseudo-DIA Library Search Strategy for DIA Immunopeptide Identification 80

3.3.2 Mechanistic Evaluation of Identification Enhancement and Assessment of Identification Quality 84

3.3.3 Efficient Cryptic Immunopeptide Identification with a Pseudo-DIA Library Search Strategy. 87

3.3.4 Evaluation of the Pseudo-DIA Library Search Strategy Using Independent Datasets 96

3.4 Conclusions 114

Chapter 4. Large-scale identification of functional microproteins with immunopeptidomics and CRISPR screening 116

4.1 Introduction 116

4.2 Materials and methods 118

4.2.1 Immunopeptidomics Mass Spectrometry Database Search 118

4.2.2 Redundancy Removal of Identified Microproteins 119

4.2.3 Conservation Analysis of Microproteins 120

4.2.4 Selection of Microproteins for CRISPR Screening 121

4.2.5 sgRNA Spacer Sequence Design 122

4.2.6 CRISPR Library Construction 123

4.2.7 CRISPR Screening of Three Cancer Cell Lines 123

4.3 Results 124

4.3.1 Extensive Identification of Microproteins Through Immunopeptidomics 124

4.3.2 Quality of Immunopeptide Identification 127

4.3.3 Characterization of Microproteins 136

4.3.4 Microproteins candidate selection for CRISPR Library Construction 140

4.3.5 Quality Control of the CRISPR KO Library 142

4.3.6 Microproteins Are Functionally Significant 145

4.3.7 Microproteins May Function Independently of Their Associated Canonical Proteins 149

4.4 Conclusion 151

Chapter 5. Overall conclusions and future perspectives 154

References: 161

List of Tables and Figures

Figure 1-1 Comparison of Principles and MS1/MS2 Spectra Between DDA and DIA 5

Figure 1-2 The schematic DIA workflow of library construction and data searching. 11

Figure 2-1 Comparison of different mass spectrometry methods in term of AltProt analysis. 31

Figure 2-2 Comparison between DDA and DIA MS method using HCT116 sample.. 32

Figure 2-3 Diverse identifications by different strategies. 37

Figure 2-4 Peptide identification discrepancies in the mouse heart and HCT116 sample. 38

Figure 2-5 Reliability assessment for predicted library construction 39

Figure 2-6 Evaluation of identifications from different strategies. 40

Figure 2-7 Evaluation of identifications from different strategies using HCT116 sample. 41

Figure 2-8 Validation of AltProt peptide identification from different library construction methods.. 42

Figure 2-9 Application of DIA in mouse heart development. 45

Figure 2-10 Comparative profiling of peptide identifications in mouse heart development using different strategies 46

Figure 2-11 Different expression analyses. 47

Figure 2-12 Validation of ASDURF and other AltProts.. 50

Figure 2-13 Gene Ontology (GO) analysis and database sources in mouse heart

development. 51
Figure 2-14 Experimental spectrum validation using PRM and corresponding Prosit predicted spectrum 65
Figure 3-1 Data searching design for Pseudo-DIA library search strategy. 82
Figure 3-1 Data searching design for Pseudo-DIA library search strategy. 82
Figure 3-2 Comparison of the Pseudo-DIA and Traditional Library Search Workflows. 83
Figure 3-3 Assessing the reliability of Pseudo-DIA library search strategy. 86
Figure 3-4 Pseudo-DIA strategy enhancing non-canonical immunopeptides identification. 91
Figure 3-5 Visualization of Immunopeptide Spectrum Quality and TCGA Expression Differences for MORF4L2 uORF Protein (IP_302145). 93
Figure 3-6 Visualization of Immunopeptide Spectrum Quality and TCGA Expression Differences for ZNF146 intORF Protein (IP_273983) 94
Figure 3-7 Length Distribution of Canonical and Non-canonical Immunopeptides Identified by Various Methods. 95
Figure 3-8 Application of Pseudo-DIA library strategy with independent dataset to identify tumor-specific immunopeptides. 98
Figure 3-9 Graphical user interface (GUI) for the pseudo-DIA library search strategy. 99
Figure 3-10 Spectral Visualization of Immunopeptides with Single Amino Acid Mutations. 109

Figure 4-1 Non-canonical microproteins identified by immunopeptidomics. 126

Figure 4-2 Identification of microproteins using immunopeptides, with the PTEN uORF as an example. 130

Figure 4-3 Characteristics of immunopeptides.. 138

Figure 4-4 Heatmap of the conservation levels of microproteins from different ORF types across 100 species. 139

Figure 4-5 Examples of conservation levels for three microproteins from three different ORF types. 139

Figure 4-6 Microprotein Selection for the Next CRISPR Screening.. 141

Figure 4-7. Flowchart of the experimental design for CRISPR experiments. 143

Figure 4-8 Quality control of the CRISPR KO library. 144

Figure 4-9 NGS results of CRISPR data from three cell lines. 147

Figure 4-10 Comparison of CRISPR data across three cell lines 148

Figure 4-11 Comparison of our CRISPR results targeting microprotein KOs with public data from DepMap. 150

Figure 5-1 Diagrammatic Summary and Interconnections of the Three Research Projects. 158

Table 1-1 The list of mainstream DIA methods. 8

Table 2-1 Protein lists of different expressed AltProts in mouse heart development. .66

Table 3-1 List of Immunopeptides Containing Single Amino Acid Substitutions Identified in This Experiment via Pseudo-DIA Library Approach. 110

Table 4-1 The 36K Microprotein List Provides Immunopeptidomics Evidence for Previously Reported Microproteins 131

Chapter 1. Overview

1.1 Overview of Microproteins Derived from Small Open Reading Frames

Microproteins are a class of proteins encoded by small open reading frames (sORFs) within regions of the genome previously considered noncoding.[1] Historically, approximately 98% of the genome was thought to be noncoding and largely ignored in proteomic studies. However, emerging evidence has challenged this view, revealing that these regions can give rise to functional proteins with distinct biological roles in the cell. AltProts represent an exciting frontier in biology, as they expand our understanding of the proteome and its complexities.

Microproteins are implicated in a wide range of biological processes. To date, their functions have been increasingly uncovered, including their roles in RNA and protein processing, signaling, cancer cell survival, cellular trafficking, metabolism, cell communication, and apoptosis. For instance, Lima et al.[2] identified a novel microprotein that interacts with mRNA decapping proteins, facilitating the removal of the 5' cap and promoting 5'-3' decay, which directly influences RNA processing and expression in the cytoplasm. Similarly, Ray et al.[3] discovered a specific micropeptide in *Drosophila* that contributes to a molecular complex and plays a critical role in embryonic segmentation in the latter. These studies highlight the potential roles of microproteins in RNA and protein processing.

Niu et al.[4] identified a 17-amino-acid microprotein that cooperates with heat shock

cognate proteins to facilitate the transportation and presentation of major histocompatibility complex antigens. Exogenous injection of this microprotein significantly enhanced the immunological response in a mouse model, providing evidence of its involvement in cellular trafficking in vivo. Matsumoto et al.[5] identified a conserved microprotein localized in lysosomes that interacts with lysosomal v-ATPase in response to amino acid stimulation, thereby negatively regulating mTORC1 activation. This finding underscores the important role of microproteins in the process of signal transduction.

Stein et al.[6] and Zhang et al.[7] independently discovered microproteins localized in mitochondria, where they participate in the respiratory chain, enhance respiratory efficiency, and maintain mitochondrial homeostasis. Knockout experiments of genes encoding these microproteins in animal models have resulted in metabolic disorders, including TCA cycle perturbations, lactic acidosis, and early death, demonstrating their critical role in metabolism.

Pauli et al.[8] identified a conserved micropeptide in zebrafish embryogenesis that interacts with G protein-coupled APJ/apelin receptors as an activator, driving gastrulation movements. This discovery provides strong evidence for the involvement of microproteins in signal transduction and cellular communication. Finally, Guo et al.[9] identified a specific micropeptide involved in BAX protein-induced apoptosis that prevents apoptosis by interacting with and blocking BAX. Additionally, this

micropeptide inhibits the translocation of BAX from the cytosol to the mitochondria, further reducing apoptosis.

In contrast to the extensively studied large proteins, the biology of small proteins remains largely unknown. Although an increasing number of small proteins with significant biological functions have been identified, most annotated small proteins still lack direct evidence of their existence at the protein level, limiting the ability to study their functions in depth. This challenge primarily arises from the technical difficulties associated with the identification of small proteins in proteomics. However, as our understanding of the nature of small proteins continues to grow, advancements in sample preparation techniques and mass spectrometry are paving the way for more efficient and innovative identification strategies to be developed. These developments are expected to provide deeper coverage of small proteins, establishing a robust technical foundation for comprehensive investigations of their roles in biological processes.

1.2 Microprotein Identification Based on DIA Mass Spectrometry

1.2.1 Introduction

With advancements in technology and the development of novel instruments, mass spectrometry has become a widely used tool in proteomics[10], metabolomics[11], and lipidomics[12]. This powerful technique enables the identification of hundreds of biomolecules across a wide range of abundances, providing critical qualitative data for

life science research[13].

In omics research, the shotgun technique is one of the most classical and commonly employed analytical strategies[14]. In proteomics, this approach is referred to as Data-Dependent Acquisition (DDA). In DDA analysis, biological samples are first ionized via electrospray ionization (ESI), after which the ions pass through the quadrupole of a mass spectrometer and are analyzed in the orbitrap. This process generates a primary spectrum known as MS1. Subsequently, the quadrupole selects the top N most abundant ions, which are fragmented in the collision cell during the next scan, producing secondary spectra (MS2 spectra). The mass spectrometer then repeatedly cycled through this process until the sample analysis was complete. Since DDA selectively generates MS2 spectra based on ion abundance, an alternative non-selective strategy, known as Data-Independent Acquisition (DIA)[15], has also been developed.

In DIA, the MS1 scan is similar to that of DDA; however, during the MS2 scan, DIA fragments all MS1 ions within multiple pre-defined isolation windows. This generates co-eluting MS2 spectra for a comprehensive analysis. By capturing all MS2 ion information[16], DIA enables the collection of significantly more data than DDA. As a result, DIA offers several advantages, including higher identification rates, fewer missing values, and a broader quantitative dynamic range[17], making it increasingly popular in proteomics research. However, the DIA strategy imposes higher demands on the speed, accuracy, and stability[18]. Moreover, the massive datasets generated by DIA

present significant challenges for bioinformatics analysis[19].

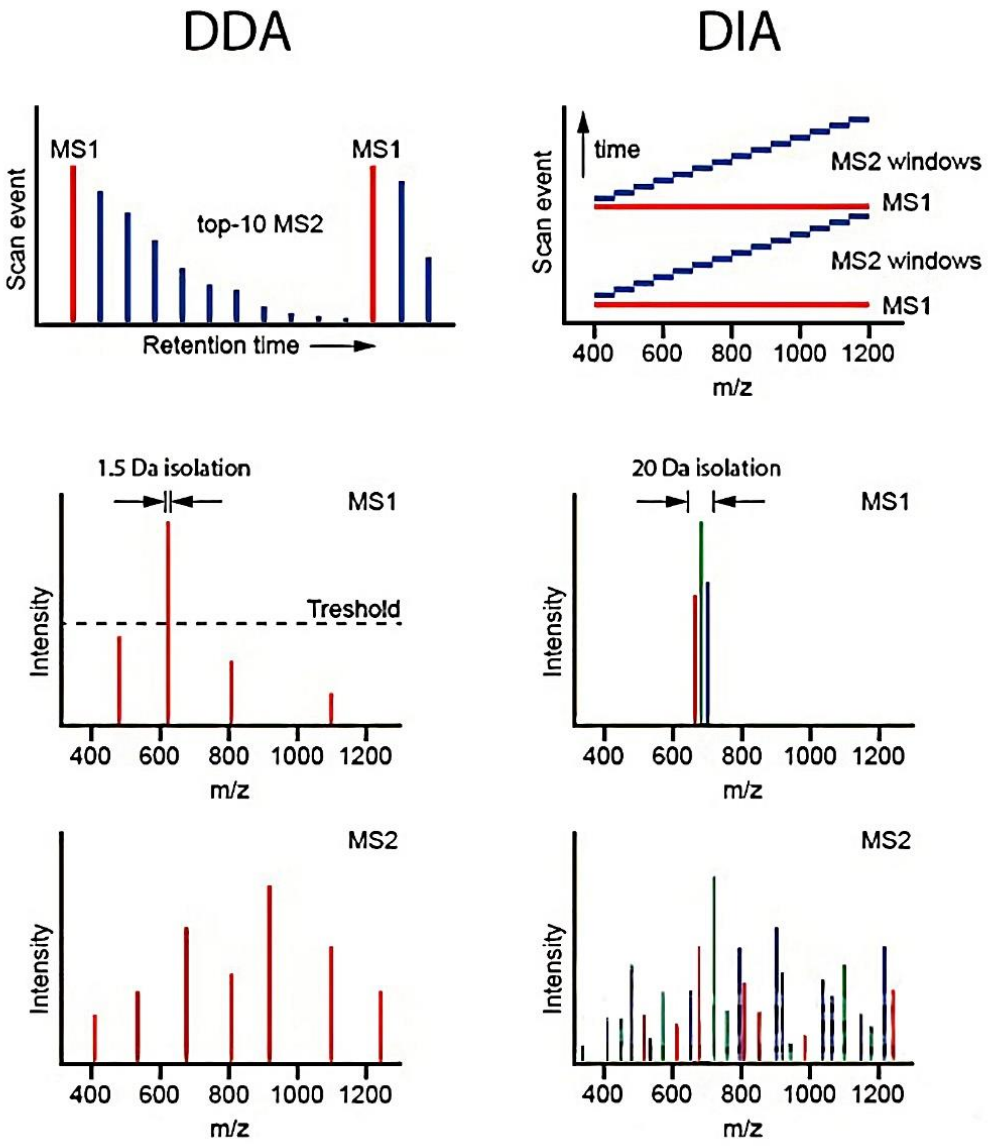


Figure 1-1 Comparison of Principles and MS1/MS2 Spectra Between DDA and DIA

1.2.2 Mass Spectrometry Scanning Methods

Numerous DIA analytical methods have been developed based on the DIA strategy by

various mass spectrometry manufacturers. These include, but are not limited to, SWATH[16], Scanning SWATH[20], DIA[21], IDA[22], PulseDIA[23], MSX[24], diaPASEF[25], and plexDIA[26]. Although all these methods adhere to the core principles of the DIA strategy, they are tailored to accommodate the specific technical features of different manufacturers' instruments and analytical requirements.

The Ruedi group pioneered the development of SWATH[16], a classic DIA method. Building on this, Guo et al. [27] were the first to combine pressure cycling technology (PCT) with DIA, enabling the detection of over 2,000 proteins in kidney samples with excellent reproducibility. Their approach successfully differentiated cancer subtypes in kidney tissues, demonstrating the clinical significance of the method.

Christoph et al. [20] enhanced the SWATH method by optimizing MS1 ion transport. Their improved Scanning SWATH allows simultaneous ion transport via the quadrupole and ion detection using time-of-flight (TOF). This innovation achieves exceptionally fast scanning speeds, detecting 40,000 peptides within a 5-minute gradient and nearly 20,000 peptides within 1 min in human cell line samples. These advancements hold great promise for clinical mass spectrometry applications.

The MacCoss group refined the MS2 window design by introducing the Multiplexed DIA (MSX) method[24]. This approach divides the MS2 window into five smaller 4 m/z windows, which are sequentially selected, accumulated, and scanned by an orbitrap.

MSX improves data quality and accuracy by reducing co-elution interference and contaminants in the MS2 spectrum.

Guo et al.[23] developed Pulse-DIA, a method that extends DIA scan times by splitting a complete DIA run into five separate runs. This strategy increased peptide identification by 50%, thereby enhancing the depth and coverage of proteomic analysis.

The Mann group, in collaboration with Bruker, introduced diaPASEF[25], a novel method that leverages the high-speed scanning ability of TOF and integrates it with an ion mobility device. This innovation adds an additional dimension of ion mobility to mass spectrometry data, resulting in the concept of "4D proteomics" diaPASEF delivers exceptional sensitivity, deep proteome coverage, and high reproducibility, even with as little as 10 μg of sample[25].

Slavov's group developed the plexDIA[26] technique by integrating the labeled quantification method mTRAQ with DIA analysis. This approach quantifies mTRAQ-labeled MS1 parent ions, reducing missing values by more than two-fold and exponentially decreasing the detection time for mixed samples. These advancements render plexDIA a powerful tool for high-throughput proteomics.

Table 1-1 The list of mainstream DIA methods.

Method	MS2 window	MS instrument	Author	Year	Highlights
SWATH	25	AB Sciex TripleTOF Series	Aebersold Group	2012	Classic
Scanning SWATH	5	AB Sciex TripleTOF Series	Ralser Group	2019	Super-fast
DIA	10-25	Thermo Orbitrap Series	Thermo Fisher Company	2004	Classic
PulseDIA	4	Thermo Orbitrap Series	Guo Group	2020	Improved sensitivity and reproducibility.
MSX	4	Thermo Orbitrap Series	MacCoss Groups	2013	Less interruptions
diaPASEF	25	Bruker timTOF Series	Mann Group Bruker Company	2019	high sensitivity and reproducibility
plexDIA	12.5,25,62.5	Thermo Orbitrap Series	Slavov Group	2021	Low cost and multiplex

1.2.3 Library Construction for DIA search analysis

The mainstream solution for analyzing the vast and complex datasets generated by DIA still relies on the use of spectral libraries. Retention time, m/z, and ion intensity information from a pre-established library can be utilized to annotate and resolve mixed MS2 spectra in DIA. Following filtering and scoring[15], the identified peptides were used to construct a proteome map. Several methods are available for library construction.

One common approach is DDA library construction, in which fractionation experiments are employed to reduce sample complexity. Multi-fractionated samples are analyzed using DDA to generate a deep proteomic dataset, and proteins are identified using traditional search engines such as Sequest[28], Mascot, Maxquant[29], and MSFragger[30]. The resulting identification data were converted into formats compatible with DIA library search software using tools such as SpectraST[31], easypqp, and TPP[32].

Another approach is DIA library construction, which focuses on directly interpreting DIA MS2 spectra, despite their complexity. Tools such as PECAN[33], which is integrated into the Walnut software[34], enable direct DIA data searching. DIAmeter[35], developed by Nobel's group, is a spectrum-centric search engine that analyzes DIA MS2 spectra using the Tides search engine, selects the top five Peptide-Spectrum Matches (PSM), and refines the results through multidimensional filtering and scoring. DIA-Umpire[36] adopts a different approach by converting DIA data into pseudo-tandem spectra based

on MS1 and MS2 features, enabling analysis using traditional DDA workflows. Once identification data are obtained, reformatting tools such as SpectraST can export them into a library for DIA workflows.

The third strategy is predicted library construction, which leverages machine learning to predict libraries with details comparable to experimental libraries. Using algorithms such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), tools such as Prosit[37], DIA-NN[38], deepDIA[39], and pDeep[40] can predict key ion properties, such as retention time, intensity, and ion mobility, contributing significantly to library construction. These advancements underscore the growing utility of computational methods in proteomics research.

1.2.4 DIA Data search tools

The mainstream strategy for DIA data searching is the target-decoy approach, which involves creating a decoy library by inverting PSM spectra. To obtain the final identification information, the MS2 spectra in DIA were compared with the target-decoy library to generate matching scores. These scores were then used for spectrum filtering, typically applying a 1% false discovery rate (FDR) threshold. Tools such as OpenSWATH[41], Skyline[42], Maxquant[29], DIA-NN[38], Spectronaut, EncyclopeDIA[34], and PEAKS[43] are among the most widely used search engines for DIA library analyses.

The interpretation of co-eluting MS2 spectra is crucial for the accurate analysis of DIA data. Comprehensive high-quality spectrum libraries combined with robust search

engines play a vital role in enabling detailed and reliable protein profiling in proteomics research.

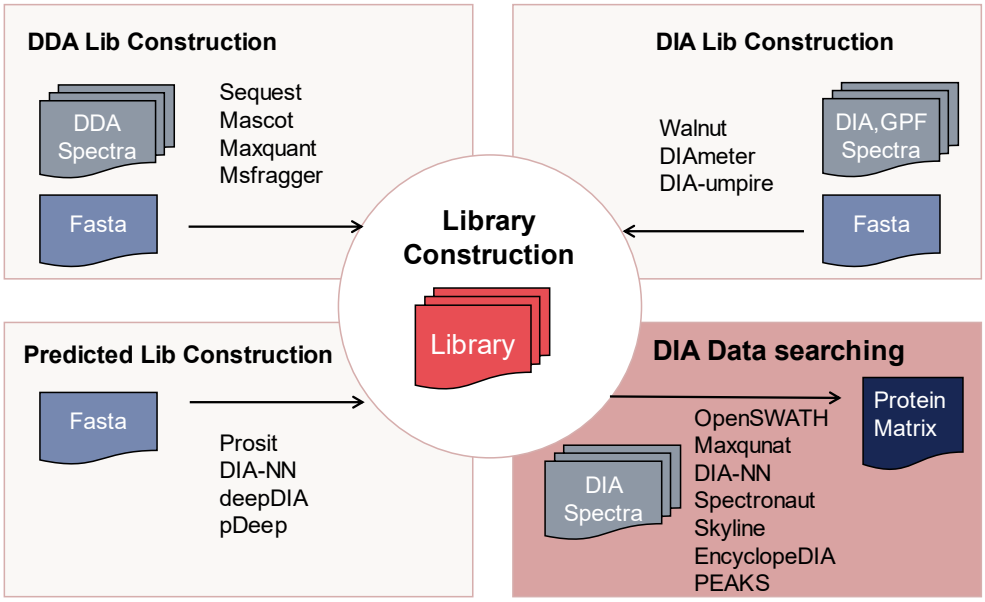


Figure 1-2 The schematic DIA workflow of library construction and data searching

1.3 Research goals and objectives

The primary goal of this research is to deepen our understanding of non-canonical proteins and peptides, such as microproteins, alternative proteins (AltProts), immunopeptides, and neoantigens, which represent a largely unexplored layer of the proteome with significant functional and therapeutic potential. These biomolecules, often overlooked in conventional proteomics workflows due to their small size, low abundance, and non-canonical origins, play critical roles in processes such as immune regulation, cancer progression, and cardiac development. By integrating advanced proteomics techniques, including data-independent acquisition (DIA) mass

spectrometry, immunopeptidomics, and CRISPR screening, we aim to establish robust methodologies for the identification and characterization of these peptides while uncovering their biological and therapeutic relevance.

A key objective of this study was to expand the known repertoire of microproteins using immunopeptidomics. By analyzing large-scale datasets comprising over 36,000 microproteins, we sought to determine their genomic origins, translation mechanisms, and functional roles in cellular processes, particularly in cancer biology. We aimed to investigate how microproteins regulate oncogenic signaling pathways, modulate immune responses, and contribute to cellular homeostasis. Furthermore, we will validate their existence and activity using experimental evidence, such as peptide-spectrum matches (PSMs) and co-immunoprecipitation, to ensure the reliability of our findings.

We also aim to optimize DIA-based workflows for the discovery of AltProts by addressing technical challenges, such as library construction, database redundancy, and false discovery rates. By systematically comparing library-building strategies, ranging from predicted spectral libraries to experimental fraction-based libraries, we sought to identify the most efficient and accurate approach for AltProt detection. Our research aims to demonstrate the superiority of DIA over traditional data-dependent acquisition (DDA) methods, with a particular focus on improving AltProt identification rates, reducing the number of missing values, and enhancing quantitative accuracy. These

optimized workflows will be applied to study AltProts in biological systems, such as their roles in mouse cardiac development, where we aim to uncover their contributions to embryonic and adult heart-tissue differentiation.

Another critical objective is to identify cryptic immunopeptides and neoantigens, which are essential for advancing cancer immunotherapy. Using our Pseudo-DIA Library Search Strategy, we aimed to detect cryptic peptides derived from small open reading frames (sORFs) and assess their potential as biomarkers for personalized cancer vaccines. By integrating public datasets and performing in silico mutation predictions, we aimed to identify neoantigens with high immunogenic potential and validate their existence using mass spectrometry and motif analysis. This study contributes to the development of novel immunotherapeutic strategies that harness the immune system to target tumor cells more effectively.

The functional characterization of non-canonical proteins is another major focus of this study. By employing CRISPR screening, we aimed to systematically evaluate the roles of microproteins and AltProts in cancer cell proliferation and survival. We will investigate whether these non-canonical proteins function independently of their canonical counterparts or exhibit co-regulatory mechanisms. For example, we aim to identify microproteins located near oncogenes or involved in cancer-related pathways and determine whether they serve as independent regulators of tumor progression, metastasis, and therapy resistance. This study provides insights into the functional

independence and biological significance of non-canonical proteins in diverse cellular contexts.

To validate our findings and ensure their translational relevance, we cross-referenced the identified proteins and peptides with public human datasets and performed conservation analyses across species. By examining the evolutionary conservation of non-canonical proteins, we sought to clarify their functional importance and potential roles in higher-order biological processes. Experimental validation, including PRM, Western blotting, and overexpression studies, will further confirm the existence and biological activity of key AltProts and microproteins, such as ASDURF, which has been implicated in cardiac development and protein synthesis.

Finally, we aim to apply our findings to address broader questions in biology and medicine, such as the dynamic regulation of the proteome during development and in disease. For instance, we will investigate the role of AltProts in mouse heart development, focusing on their differential expression in embryonic and adult tissues. By integrating proteomic data with transcriptomic analyses, we aimed to uncover novel regulatory mechanisms involving AltProts and their host genes. Additionally, we will explore the functional roles of cryptic peptides and neoantigens in cancer immunotherapy and identify potential therapeutic targets that could pave the way for personalized medicine.

In conclusion, our research aims to establish a comprehensive framework for studying

non-canonical proteins and peptides, from their discovery and validation to their functional characterization and therapeutic applications. By innovating proteomic methodologies and applying them to biological systems, we aim to contribute to fundamental scientific knowledge and address pressing challenges in cancer research, immunotherapy, and developmental biology. Through these efforts, we hope to unlock the full potential of non-canonical proteins as key players in cellular processes and promising targets for future therapeutic interventions.

1.4 Outline of this thesis

This thesis is organized into four chapters, each addressing a critical aspect of the discovery, characterization, and functional analysis of small open reading frame-encoded microproteins (SEPs) and related non-canonical proteins. The following is an outline of the thesis, summarizing the focus and contributions of each chapter.

Chapter 1: Introduction and Overview

Chapter 1 introduces the essential concepts of small open reading frame-encoded microproteins (SEPs) and the broader category of non-canonical proteins, such as alternative proteins (AltProts) and immunopeptides. This highlights the growing recognition of these biomolecules as functional entities encoded in genomic regions that were previously considered non-coding. This chapter provides an overview of their roles in cellular processes, including immune regulation, cancer progression, and metabolic homeostasis. Furthermore, this chapter reviews the current methods for

identifying non-canonical proteins, particularly mass spectrometry-based approaches, such as data-independent acquisition (DIA) proteomics and immunopeptidomics. The advantages and challenges of these technologies are discussed, including the critical need for optimized workflows, robust spectral libraries, and advanced computational tools for data analysis. The chapter concludes by stating the research goals and objectives, which aim to bridge the existing gaps in the systematic discovery and functional characterization of SEPs and related biomolecules.

Chapter 2: Enhanced Discovery of Alternative Proteins (AltProts) in Mouse Cardiac Development Using DIA Proteomics

Chapter 2 focuses on the development and application of an optimized DIA workflow for the discovery of AltProts. This workflow incorporates predicted spectral libraries, enabling a twofold increase in AltProt identification compared to traditional data-dependent acquisition (DDA) methods while reducing missing values by 50%. This chapter systematically evaluates the impact of different spectral library construction strategies, including experimental and predicted libraries, on AltProt detection. Using mouse cardiac tissue as a model, this study identified over 50 differentially expressed AltProts, highlighting their roles in embryonic and adult heart development. Among these, ASDURF, a uORF-encoded AltProt, has been validated as a novel regulator of cardiac development. This chapter establishes a comprehensive and reproducible DIA-based framework for AltProt analysis and highlights the biological significance of AltProts in organ development.

Chapter 3: A Pseudo-DIA Library Search Approach for Improving Immunopeptidomics and Neoantigen Discovery

Chapter 3 introduces a novel Pseudo-DIA Library Search Strategy designed to address the challenges of immunopeptidomics and neoantigen discovery. Immunopeptides presented on major histocompatibility complex (MHC) molecules play critical roles in immune surveillance and cancer immunotherapy. The Pseudo-DIA strategy combines unrestricted DIA database searches with predicted spectral libraries to achieve up to 3.8 times more immunopeptide identifications than conventional approaches. This chapter demonstrates the method's ability to identify cryptic immunopeptides and neoantigens derived from small open reading frames (sORFs), which are crucial for advancing personalized cancer-immunotherapy. Additionally, this chapter describes the development of a user-friendly executable tool to streamline spectral library generation from MSFragger results, facilitating the broader adoption of this methodology in immunopeptidomics research. The robustness and versatility of the Pseudo-DIA strategy are validated using multiple independent datasets, emphasizing its potential to expand the landscape of immunopeptides and neoantigens.

Chapter 4: Large-Scale Identification of Functional Microproteins with Immunopeptidomics and CRISPR Screening

Chapter 4 presents a large-scale investigation of functional microproteins using a combination of immunopeptidomics, CRISPR screening, and bioinformatic analyses. By analyzing 86 datasets containing over 5,000 raw mass spectrometry files, the study

identified 36,494 microproteins, representing one of the most comprehensive datasets to date. These microproteins are categorized based on their genomic origins, start codon usage, and conservation across species. Functional screening using CRISPR knockout (KO) libraries in three cancer cell lines revealed that many microproteins play critical roles in cell proliferation, with some exhibiting functional independence from their canonical protein counterparts. This chapter highlights specific examples of microproteins that regulate oncogenic pathways, offering insights into their potential as therapeutic targets. It also discusses the evolutionary conservation of microproteins, particularly in mammals, and their preferential enrichment near oncogenes, highlighting their biological and clinical relevance.

Chapter 5: Overall conclusions and future perspectives

The thesis concludes by summarizing the key findings and contributions of each chapter, emphasizing the power of advanced proteomics techniques, such as data-independent acquisition (DIA) and immunopeptidomics, in uncovering the hidden proteome. This highlights the functional and therapeutic potential of SEPs, AltProts, and immunopeptides, particularly in cancer and developmental biology. The conclusion also outlines future research directions, including the need for further validation of the identified microproteins, exploration of their underlying mechanisms, and development of innovative therapeutic strategies targeting non-canonical proteins. The methodologies and findings presented in this thesis provide a solid foundation for future studies aimed at unlocking the full potential of non-canonical proteomes.

Chapter 2. Enhanced Discovery of Alternative Proteins (AltProts) in Mouse Cardiac Development Using Data-Independent Acquisition (DIA) Proteomics

2.1 Introduction

Over the past decades, 98% of the human genome has been presumed to be noncoding regions and remained largely unexplored[44-46,47]. However, emerging evidence suggests that many of these noncoding regions are translated, giving rise to what is known as alternative open reading frames (AltORFs). The proteins derived from these AltORFs, which have been termed alternative proteins (AltProts), are widely translated and perform unique functions in events such as respiratory chain reaction,[7] embryogenesis,[8] apoptosis[9], and signal transduction.[5] The novelty and importance of AltProts underscores the urgent need for their further investigation with advanced methodologies.[48, 49]

Although ribosome sequencing (Ribo-seq) is powerful in predicting the potential translation of AltProts, it does not provide direct evidence of translation products. In contrast, mass spectrometry (MS) can sequence and quantify AltProts with high confidence. Traditionally, the data-dependent acquisition (DDA) MS method has made great contribution to the constructing of the human proteome map,[50-55] the human protein Atlas[56], etc and has remained the mainstream method for detecting canonical proteins[5, 7, 8, 57-59]. In contrast, data-independent acquisition (DIA) has been demonstrated in recent years to outperform DDA in canonical protein analysis,[60-62].

Although there has been a report on the application of DIA in AltProt analysis, in which Martinez et al. used GPF and DDA to generate an experiment-specific spectral library, systematic comparisons and method optimization for DIA analysis of AltProts remain limited, likely due to several technical challenges[63]. First, DIA identification largely depends on the quality of the mass spectral library, particularly the accuracy of the peptide fragmentation spectrum and retention time (RT). Given the potentially high sample/temporal/spatial specificity of AltProts[64, 65], there is significant uncertainty regarding the quantity and abundance of AltProts. Therefore, a sample-specific spectral library would be ideal, and the demand for instrument time and effort to construct this sample-specific spectral library of AltProts is presumably much higher than that of the DDA method. Second, the use of DIA analysis to identify proteins from mixed spectra is more challenging, and the workflow is more complicated than that of DDA. Third, there are currently fewer search engines and less software available for DIA than for DDA. Theoretically, DIA could offer higher identification numbers and quantification accuracy of AltProts because the DIA mode fragments all precursor ions instead of just the top abundant ions, generating mixed secondary spectra and thus providing more ion information for peptide identification and quantification.[16, 66] However, its performance needs to be systematically evaluated. Therefore, an optimal and effective DIA-based workflow customized for AltProts analysis is urgently needed.

In response to these needs, we systematically compared four widely used AltProt databases, four DIA-library building strategies, and three software programs to provide

an optimal workflow. Our results indicated that DIA outperformed DDA in terms of AltProt identification number and reproducibility. It is also worth noting that different library construction methods significantly influenced AltProts identification and yielded complementary discoveries of AltProts. Next, we applied the DIA method to study functional AltProts in mouse heart development and identified over 50 differentially expressed AltProts. Among them, a representative ASDURF was validated to express a stable AltProt protein. The process involving ASDURF is likely to play a significant role in the development of the mouse heart[67]. Our work provides a reference for selecting an appropriate DIA data processing method for AltProt studies, as well as novel AltProts that potentially regulate heart development and have important functions.

2.2 Materials and methods

2.2.1 Sample collection

The C57BL/6 mice were purchased from Guangdong Medical Experimental Animal Center (Guangdong, China; License No: SCXK (YUE) 2018–0002). The embryonic and adult heart samples were collected from embryonic (E15.5) and adult (P42) C57BL/6 mice respectively, and immediately frozen and preserved in liquid nitrogen. All animal experiments were authorized by the Hong Kong Polytechnic University Animal Subjects Ethics Subcommittee and conducted in accordance with the Institutional Guidelines and Animal Ordinance of the Department of Health guidelines.

2.2.2 Sample preparation

Mouse heart samples were cut and homogenized with RIPA lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 2 mM EDTA, 1% NP40, 1% SDC) containing 1×cocktail protein inhibitor. Liquid nitrogen was added to the homogenizer (Bertin Technologies, France) according to the manufacturer's manual to avoid protein degradation. The cell sample was ultrasonicated with the same lysis buffer and cocktail inhibitor for 3min on ice. After extraction, the supernatant was collected after centrifugation at 16,000g, 4°C for 25min. BCA protein assay (Thermo Scientific) was applied to measure protein concentration. Dithiothreitol (DTT) at a final concentration of 10mM was added first and placed at 37 °C for 45min, followed by iodoacetamide (IAM) at a final concentration of 30mM for 30min in a dark room. The sp3 digestion experiment was performed using Sera-Mag beads (Cytiva) based on a previous study. Briefly, washed beads were added to the lysed sample, then an equal volume of pure methanol was added to bind the proteins and beads. Next, the sample was placed on a magnetic rack and the waste solution was discarded and the beads were washed three times with 80% methanol. Then, 45 μl of 50 mM Ammonium bicarbonate was added with Lys-C (enzyme to protein ratio 1:100) at 37°C for 4 hours. Lastly, Trypsin (enzyme to protein ratio 1:40) was added at 37°C overnight. Beads were washed with ddH2O and the supernatant was collected for LC-MS analysis.

2.2.3 LC-MS analysis

LC-MS analyses were performed using Orbitrap Exploris™ 480 Mass Spectrometer

(Thermo Fisher Scientific) coupled with UltiMate 3000 HPLC system (Thermo Fisher Scientific). Peptides were injected into LC and separated by a 2-hour gradient using the 25cm Aurora column (IonOpticks, Australia) with mobile phases buffer A (99.9% H_2O, 0.1% FA) and buffer B (80%ACN, 0.1% FA). The 2-hour gradient was (min, %B):0,8;3,8;104,32;113,90;120,90 and utilized by three different MS modes. For data-dependent acquisition (DDA) mode, MS1 resolution was set at 120,000 with 50ms max injection time (MIT) while MS2 was set at 30,000 with 50ms MIT. The isolation window was set as 1.6m/z and the cycle time of data-dependent mode was set as 3 seconds. For the data-independent acquisition (DIA) mode, MS1 was set at 60,000 with 50ms MIT and the precursor range starting from 400 to 1000. This range was evenly divided by 10m/z into 60 MS2 windows, each with an overlap of 1m/z. MS2 resolution was set as 30,000 at 32% HCD collision energy while desired minimum points across the peak were set as 8. For gas-phase fractionation (GPF) mode, the recommended 6× GPF-DIA acquisition settings were applied based on previous research.[68] For (PRM) mode, the m/z of target peptides were set in the MS2 transition table with a 1.2 m/z isolation window and 80 ms MIT.

2.2.4 MS Raw data searching

Regarding the spectral library construction method for DIA analysis, four methods were designed: The first method involved collecting data from eight fractions and constructing a spectral library using the software MSFragger and FragPipe (v17.1)[69] (DDA-frac-library). In FragPipe, the tabs called “MSFragger”, “Validation”, and “Spec Lib” were utilized to generate the spectral library with default parameters. The second

method involved collecting data from six GPF datasets, employing DIA-Umpire[70] and MSFragger for data searching and library construction (GPF-msf-library). In contrast to the first library construction strategy, this method utilized the “DIA-UMPIRE” tab. The third method also involved six GPF datasets, but it focused on searching against a fully predicted library using DIA-NN[71] to generate a sub-library (GPF-diann-library). The *.speclib file was generated using default parameters of DIA-NN, which included a 1% false discovery rate (FDR), fixed modification of carbamidomethylation on cysteine residues, one missed cleavage and trypsin protease mode. The fourth one directly used a fully predicted library (predicted-library). After obtaining the library, DIA-NN was used to analyze DIA data with the Match-between-run(MBR) search mode to reduce the library's redundancy.[72] For DDA, raw data were searched against different databases using Fragpipe(v17.1).[69] After MSFragger target-decoy searching with ±10ppm MS1 tolerance, ±0.02 Da MS2 tolerance, Percolator[73] and ProteinProphet were used for PSM validation and Protein Inference. EasyPQP was used for experimental spectral library generation. For PRM, raw data were analyzed by skyline (v21.2.0.369)[74]. After data searching, the identified peptides were classified as canonical peptides or AltProt peptides according to sequential mapping(Leucine and Isoleucine were treated as same) and BLAST[75]. All relevant raw data, fasta files, and DIA-NN search results have been uploaded and stored in the PRIDE database, with Accessing ID PXD045956.

2.2.5 Differential expression analysis and spectra validation

After log2 transformation, R package impute.knn (v1.66) was used to fill in the missing

values and those with more than 50% missing values were filtered out. However, if all missing values were present in the same biological group such as adult or embryonic, these proteins were considered differential proteins. In cases where all values were missing in either embryonic or adult samples, 13 differentially expressed proteins (DEPs) were identified in the analysis. We employed parallel reaction monitoring (PRM) to validate the absence of expression. Differentially expressed proteins were then filtered using a 2-fold change, 0.05 p-value, and the Benjamini-Hochberg algorithm to correct the p-value. The PCA plot was performed with R package pcaMethods(v1.84) and GO analysis was performed with R package clusterProfiler(v4.0.5).[76]

Normalized spectral angle(SA) was used to validate spectra quality according to previous research.[77, 78] Prosit[79] was used for spectra prediction to compare experimental spectra with predicted spectra and the high and low SA values represented the spectrum's relative quality.

2.2.6 Western blot analysis.

For The over-expression experiment was performed with transfection of plasmids containing AltProt cDNA sequences in the CHO cell line. The plasmids were constructed based on PB510-P2A-EGFP-control-3 vector and ordered from Gene Universal company (Delaware, USA)

Three micrograms plasmid of each candidate was diluted in 6μl Promega ViaFect™ Transfection Reagent and added to a 6-well plate. After 48 hours of transfection,

samples were collected and lysed with RIPA lysis buffer. 20ug proteins of cell lysed were separated by Tricine SDS-PAGE gel, followed by binding of anti-flag antibodies and anti-rabbit IgG antibodies.

2.2.7 Identification of more microproteins using the PRM method

Twenty-seven microproteins were selected from the Ribo-seq-based microprotein database for targeted PRM analysis (tier 3 level) to identify additional microproteins. Briefly, a fragmentation inclusion list of theoretically predicted tryptic peptides in the selected microprotein was generated to identify novel microproteins using high-resolution data-dependent scanning. A total of 51 unique peptide targets (corresponding to 27 microproteins) were selected in the inclusion list based on the following stringent screening criteria: peptides uncommon to RefProts, sequence length greater than 7 amino acids, and the absence of methionine oxidation.

2.2.8 Public MS data reanalysis

MS raw data of the three species mixed sample and HeLa sample were downloaded from ProteomeXchange with accession number PXD028735.[80] For pseudo-targets FDR estimation, the rice(4124 entries, Swiss-Prot) and human(20394 entries, Swiss-Prot) databases were combined to build a predicted library using DIA-NN with the default parameters which was then used in the subsequent search of HeLa DIA data. The estimated FDR was calculated by evaluating the frequency of occurrence of the rice peptide at various FDR conditions. For LFQbench analysis, human, yeast (6050 entries, Swiss-Prot), and E. coli. (4519 entries, Swiss-Prot)

were merged and comparable mass spectrometry analysis was performed as stated previously. R package LFQbench(v1.0.0)[66] was employed to generate scatter plots of proteins from three species.

Human MS raw data were downloaded from ProteomeXchange with accession number PXD016999.[81] The raw data was searched by MSFragger with N-terminal TMT modification, 229.16293. The PSMs were also validated by Percolator and ProteinProphet as mentioned above.

2.2.9 AltProts database construction

For the riboseq database, the previous Ribo-seq data[82] was reanalyzed by ten different softwares. After the adaptor removing using Cutadapt (v1.8), rRNA and tRNA removing using Bowtie2, as well as low-quality trimming using Sickle (v1.33), all remaining reads are mapped to GENECODE (version 28) using Tophat2. Ten softwares including RiboTISH (v0.2.1), ORFquant (v0.99.0), ORFRATER (v1.3.1), RiboCode (v1.2.11), riboHMM, Ribotricer (v1.3.1), RiboWave (v1.0), Rp-Bp (v2.0.0), RibORF (v1.0), and PRICE (v1.0.3b) were utilized for ORF and sORF detection using the longest strategy with the default threshold value. These ORF fasta sequences detected by different software were compiled into a complete riboseq database. For three databases(3db), the three public databases including Openprot[83], sORFs.org[84], and SmProt[85] were merged directly while identical ORFs were deleted.

2.2.10 Spectral angle calculation

For the riboseq Using the Prosit tool, the MS2 spectrum of a peptide's precursor ion is predicted. Concurrently, the experimental spectrum is extracted from Thermo *.raw files based on the scan number. After obtaining both spectra, the spectral angle score is calculated following established methods from prior research[77].

$$\mathrm{SA} = 1 - \frac{2\cos^{-1}(S_1 \cdot S_2)}{\pi}$$

2.3 Results and Discussion

2.3.1 The design of comparative evaluation of different MS-based workflows for AltProt analysis.

First, we designed a systematic scheme to evaluate commonly used DDA and DIA methods in terms of AltProts detection. As illustrated in **Figure 2-1A**, two distinct sample types, the HCT116 cell line and the adult mouse heart tissue, were chosen to ensure unbiased evaluation of methods. Subsequently, we compared three mass spectrometry scanning modes, DDA, DIA, and gas phase fractionation(GPF) [86], with four technical replicates. Additionally, we collected eight fraction data and six GPF data from two samples to assist DIA data analyses with various library construction methods. Three complementary databases were then selected to ensure a comprehensive coverage of AltProts and minimize false-negative discovery commonly caused by incomplete reference database. The first database was a reviewed RefProt database from UniProt[87] and the second database was a combined AltProt database that included

OpenProt,[83] sORFs.org,[84] and SmProt,[85] while the third one was our in-house Ribo-seq-based database[82]. Lastly, four different library construction methods (see methods) were designed to evaluate the impact of several DIA library construction methods on protein identification. Briefly, the four library construction methods were as follows: the first approach is to build a predicted spectral library based on machine learning; the second and third approaches involve using DIA-NN and MSFragger, respectively, to search gas-phase fractionation data to generate spectral libraries, while the last approach is to construct a sample-specific spectral library from experimental fractionated samples. In summary, we conducted a comprehensive evaluation of the impact of various data processing parameters on the identification and quantification of canonical proteins and AltProts. The assessment involved the use of two biological samples (human HCT116 cell line and mouse heart), two mass spectrometry data acquisition modes (data-independent acquisition and data-dependent acquisition), three protein sequence databases (RefProt DB, AltProt DB, Ribo-seq DB), and four spectral library construction methods. By employing these multiple factors, it provided a comprehensive assessment of the impact of different data processing parameters on the identification and quantification of canonical proteins and AltProts.

2.3.2 DIA outperformed DDA in identification and quantification of AltProts.

To evaluate the effectiveness of DIA proteomics with DDA proteomics in terms of AltProt analysis, we selected the traditional DDA-assisted spectral library construction method, named DDA-frac-library, for DIA data analysis and compared it with

traditional TopN DDA proteomics. Our results showed that DIA outperformed DDA both qualitatively and quantitatively in AltProt analysis **(Fig. 2-1C)**. DIA identified 1.48 times more peptides than DDA in mouse heart samples, with an average of 41,890 and 28,340 peptides, respectively. In addition, the missing values of canonical peptides identified by DIA (5.8%) were 2.64 times lower than those of DDA (15.3%). Meanwhile, for the AltProt peptides, DIA and DDA identified 22 peptides and 59 peptides on average **(Fig. 2-1B, C)**, respectively, and the number of AltProt peptides in DIA was 2.67 times more than DDA. As shown in **Figure 2-1C**, the missing values of DIA and DDA were 26.4% and 12.3% respectively, the former being 2.14 times greater than the latter.

Apart from identification, we also observed that DIA had a broader quantitative range and better quantitative correlation than DDA **(Fig. 2-1D, E)**. The broader distribution area of low abundance proteins detected by DIA implied its superior quantitative performance for these proteins in comparison with DDA. A similar trend was observed in HCT116 data for both canonical and AltProt peptides **(Fig. 2-2)**. Collectively, the mass spectrometry data from the mouse heart and HCT116 provided direct evidence for the superiority of the DIA method in the identification of canonical and AltProt peptides.

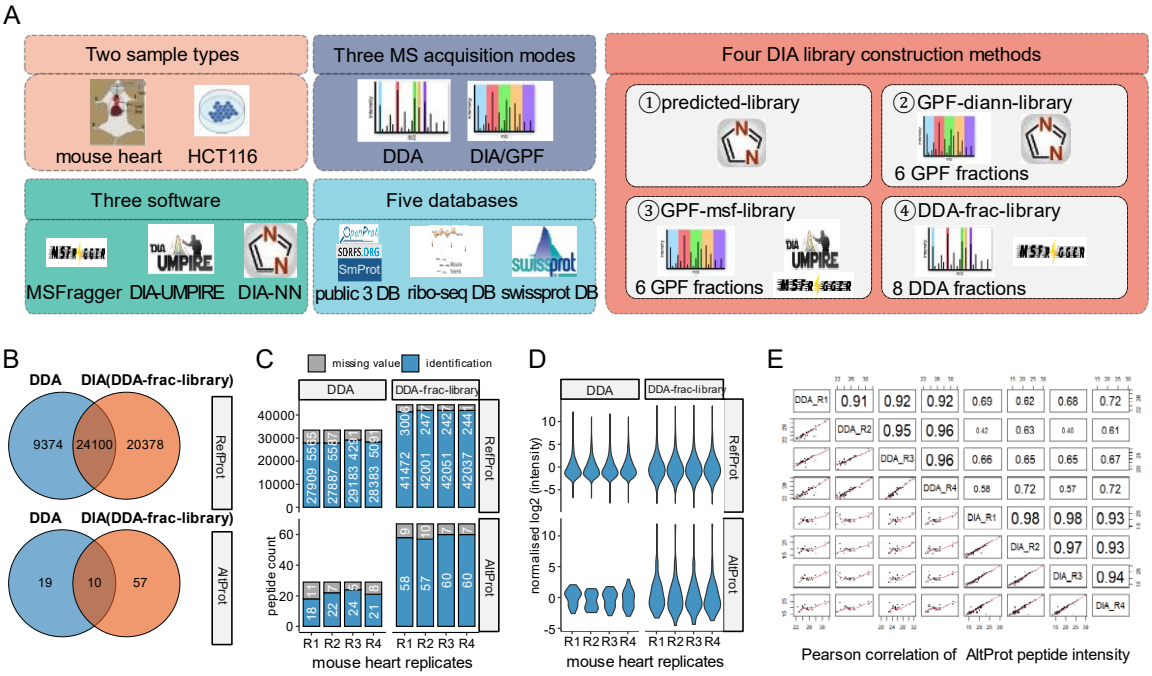


Figure 2-1 Comparison of different mass spectrometry methods in term of AltProt analysis. (A) Schematic workflow of DDA and DIA with four different library construction methods. (B-D) Venn diagram, box plot and violin diagram of identified canonical and AltProt peptides of mouse heart from DDA and DIA, respectively. (E) Pearson correlation of AltProt peptide intensity from four DDA and four DIA mouse heart replicates.

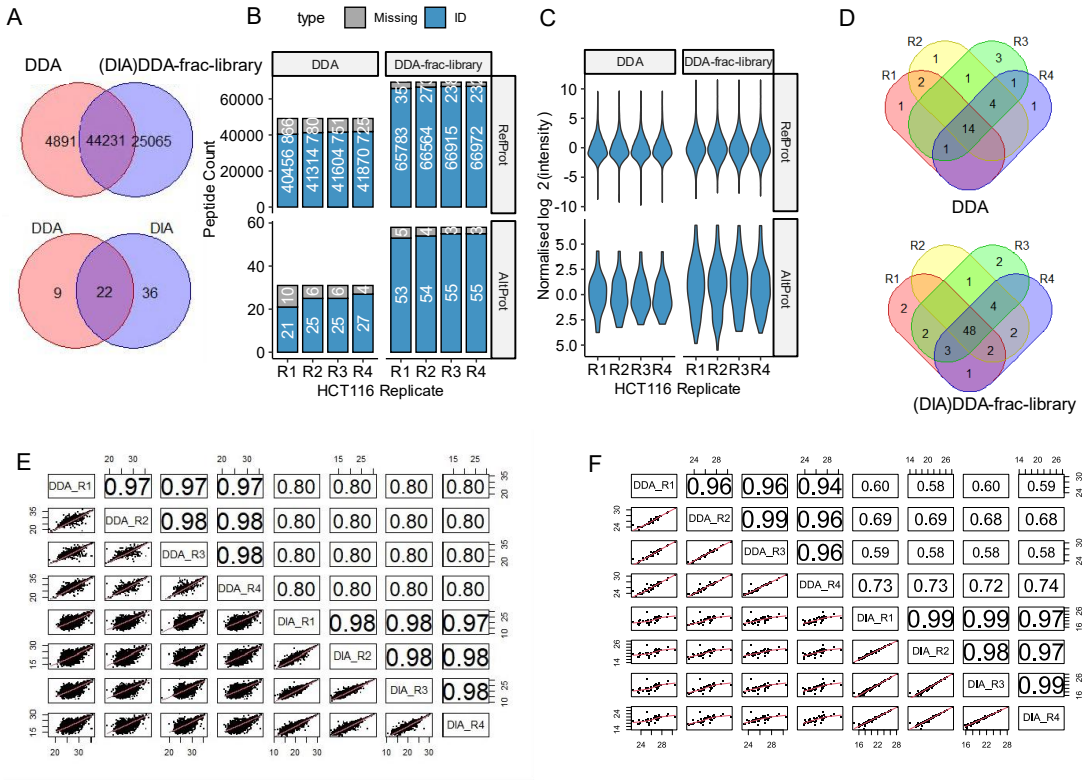


Figure 2-2 Comparison between DDA and DIA MS method using HCT116 sample.

(A-C) Venn diagram, box plot and violin diagram of identified canonical and novel peptides of HCT116 cell line sample from DDA and DIA, respectively. (D)Venn plot of identification repeatability of novel peptide in four technical replicates of DDA and DIA-frac-library method. Pearson correlation of canonical(E) and Novel(F) peptide intensity from 4 DDA and DIA HCT116 cell line replicates.

2.3.3 Different library construction methods influence AltProt identification.

There is growing evidence that DIA is more effective in identifying AltProt peptides, however, the identification of these peptides is influenced by the library construction method used, and the influence of various library construction methods on AltProt peptide identification should be examined further. Currently, the usual fractionation-based DIA procedure relies on an experimental library to aid the analysis of mixed spectra. However, with the development of deep learning, predicted peptide MS2 spectra and retention time (RT) can be directly used for DIA analysis, which can reduce the cost of experiments and increase identification number.[79, 88, 89] To evaluate the DIA data analysis method using predicted library, we calculated the false discovery rate (FDR) with a pseudo-targets FDR estimation strategy and LFQbench analysis were performed. Our observation was consistent with previous report that the experimental FDR was smaller than the set FDR, which demonstrated the accountability of the spectral library.**(Fig. 2-5A)**.[90] The LFQbench[66] plot illustrated that the qualitative and relative quantitative accuracy of the prediction library. **(Fig. 2-5B)**. Both pseudo-targets FDR and LFQbench analysis implied that the FDR of the prediction library construction method was controllable and the identification results were specific and sensitive.

Next, in order to examine the effect of different library construction methods on AltProt identification (see methods for library construction), the adult mouse heart data were searched with four different libraries against three different databases. The data

suggested the fully predicted library (predicted-library) and sub-library generated by GPF (GPF-diann-library) led to relatively high identification of 50K canonical peptides and 100 AltProt peptides, while the experimental fraction library (DDA-frac-library) only identified 41K canonical peptides and 67 AltProt peptides **(Fig. 2-3A and Fig. 2-4A)**. In terms of AltProt peptide identification, the predicted library method also yielded the highest number of AltProt identification. And the limited overlap in the identification of AltProt peptides across different library construction methods **(Fig. 2-4B)** suggested that the use of different library construction methods was more likely to result in the discovery of distinct AltProt peptides. A similar phenomenon was also observed in the HCT116 dataset **(Fig. 2-4B, D)**. And the DIA spectral library construction method influenced on the identification of AltProt peptides more dramatically than on canonical peptides. Several explanations for this phenomenon include the redundancy and inaccuracy of database entries. The number of entries in the three public databases ranges from 240,086 to 503,779, which may result in an increased number of false negatives. Furthermore, regarding the identification of canonical proteins, the four different library construction strategies were able to identify more than one-third of the proteins, whereas this proportion was less than 5% for identifying AltProt, as demonstrated in **Figure 2-3B** and **Figure 2-4D**. This discrepancy suggested that the AltProt databases utilized in this study may not be as comprehensive as the UniProt database, implying a potential limitation in the inaccuracy of the AltProt library compared to the UniProt database.

Besides quantity, we also assessed the quality of each identification by comparing each spectrum and the corresponding retention time (RT) with theoretical values predicted with Prosit. We chose all AltProt peptides and randomly selected similar amounts of canonical peptides to predict MS2 spectra and RT. By calculating the spectral angle (SA), the SA and RT correlation was plotted **(Fig. 2-6A)**. The chromatic representation of the data points indicated the quality of the spectrum and the spatial distribution of the points represented the correlation of RT. As illustrated in **Fig. 2-6B, C** for instance, the AltProt peptide GLGGGGGGGTAPGAHR, which was identified by the DDA-frac-library, exhibited a spectral similarity of only 0.04 to the predicted spectrum, whereas the AltProt peptide EPSDKASEENEAPNLHSR, which was derived from the predicted-library, had an SA value of 0.89**(Fig. 2-6B-C).** Additionally, we observed that the peptide GLGGGGGGGTAPGAHR, which had a lower SA value, deviated from the RT correlation. Based on RT and SA, we had reason to believe that the identification of EPSDKASEENEAPNLHSR was more reliable than that of GLGGGGGGGTAPGAHR.

In general, the results illustrated the SA of the two libraries from MSFragger (GPF-msf-library and DDA-frac-library) were relatively low and the SA value of the AltProt peptide on average was approximately 0.5 and the msbooster was not applied in the data searching process step of FragPipe. In contrast, the SA values obtained by the other two methods were in the range of 0.70 to 0.75, which was close to the SA value derived from the canonical peptides **(Fig. 2-6E)**. Moreover, the RT correlation of AltProt peptides derived from two MSFragger libraries was also inferior. The poor spectral

similarity and RT correlation implied a higher false discovery rate of AltProt peptides **(Fig. 2-7)**. Intriguingly, in the identification of canonical peptides, the SA and RT obtained by different library-building methods matched comparatively to the predicted spectra, indicating that the FDR of canonical peptides was relatively controlled. Obviously, the usage of experimental DDA libraries had the potential to result in more false AltProt peptide identifications. Potential false discoveries of AltProt were discovered during the DDA database searching process and could be propagated into the DIA searching results. Such false discoveries could also lead to perturbations in the FDR calculation during DIA database searching, and eventually reduce the quantity of identified AltProts. On contrary, the predicted strategy was constructed spectral library purely based on protein sequences from the database in an unbiased manner, which avoided steps that could introduce false discoveries in the spectral library.

Therefore, two mitigations are recommended to reduce the impact of false discoveries. One may utilize algorithms such as deeprescore[78] and msbooster[91] to optimize library construction procedures. Another option is to manually verify the spectra of specific AltProts following the data search processes with DIA or DDA, particularly for AltProt studies, as a cutoff of 1% FDR cannot ensure accurate identifications.

In addition to the quality of the spectra, the intensities of the AltProt peptides identified by the fully predicted library were significantly lower than those of the canonical peptides, with an average intensity that was just half of the canonical **(Fig. 2-6D)**. The

average peptide count number per AltProt was typically one **(Fig. 2-6F)**. These findings suggested that the lower abundance and smaller number of peptides from one AltProt could contribute to poor spectrum quality, resulting in different analysis methods leading to different AltProt identifications, thus explaining why AltProt detection was challenging.

Figure 2-3 Diverse identifications by different strategies. (A)Boxplot of identified AltProt peptide count by four different library construction methods and three different databases within four mouse heart replicate samples. (B)Venn diagram of identified AltProt peptides by four different library construction methods from different databases.

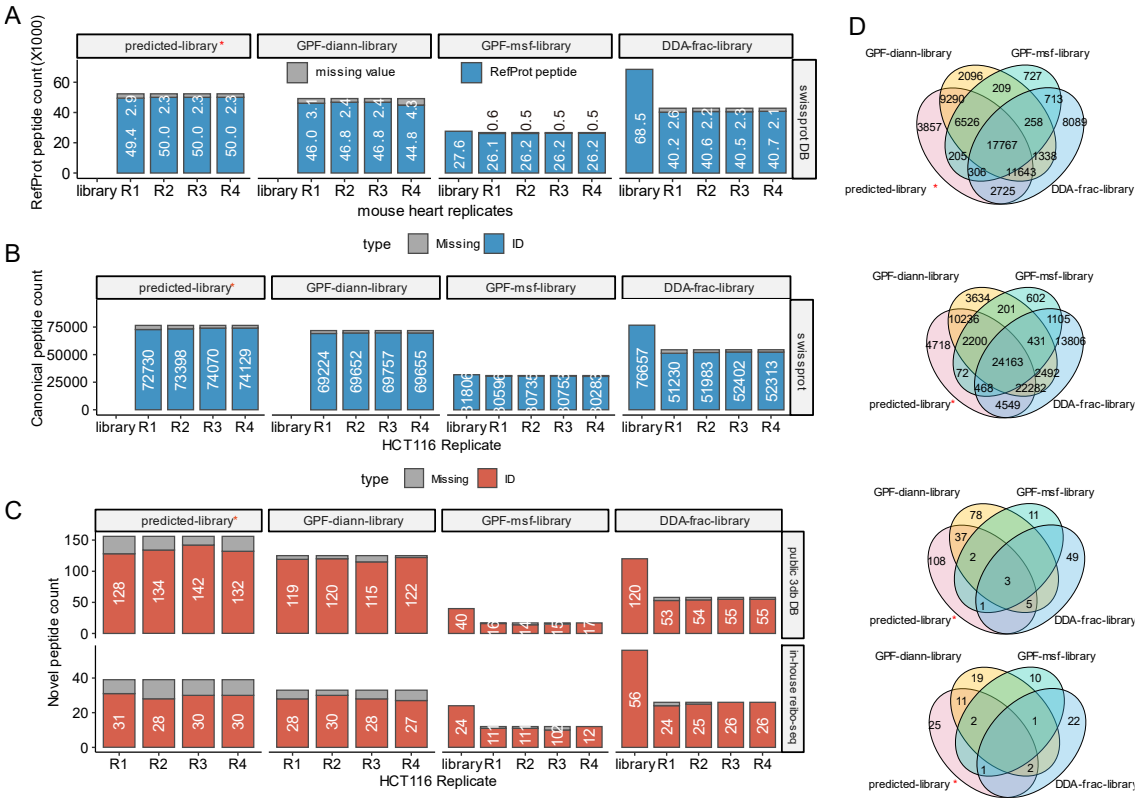


Figure 2-4 Peptide identification discrepancies in the mouse heart and HCT116 sample.

(A)Boxplot of identified canonical peptide count by four different libraries within four mouse heart sample replicates. Boxplot of identified canonical(B) and novel(C) peptide count by four different library construction methods and three different databases within four HCT116 sample replicates. (C)Venn diagram of identified peptides by four different library construction methods from different databases.

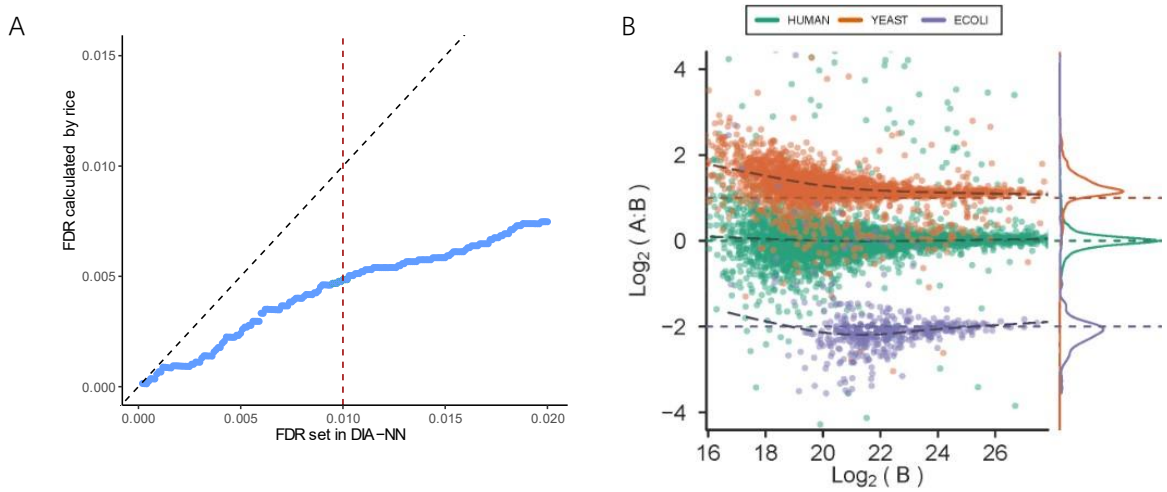


Figure 2-5 Reliability assessment for predicted library construction. (A) FDR estimation using Pseudo-targets searching approach. Rice proteome and human proteomes were mixed to construct predicted DIA library using DIA-NN. Based on the frequency of rice peptides observed in the final results, the experimental false discovery rate (FDR) was estimated. (B) Protein LFQbench result of predicted library-based DIA method. Human, yeast and *E. coli.* proteomes were used to build a predicted library using DIA-NN. Raw data from mixed samples of different species were searched by DIA-NN and dot plots were plotted by LFQbench.

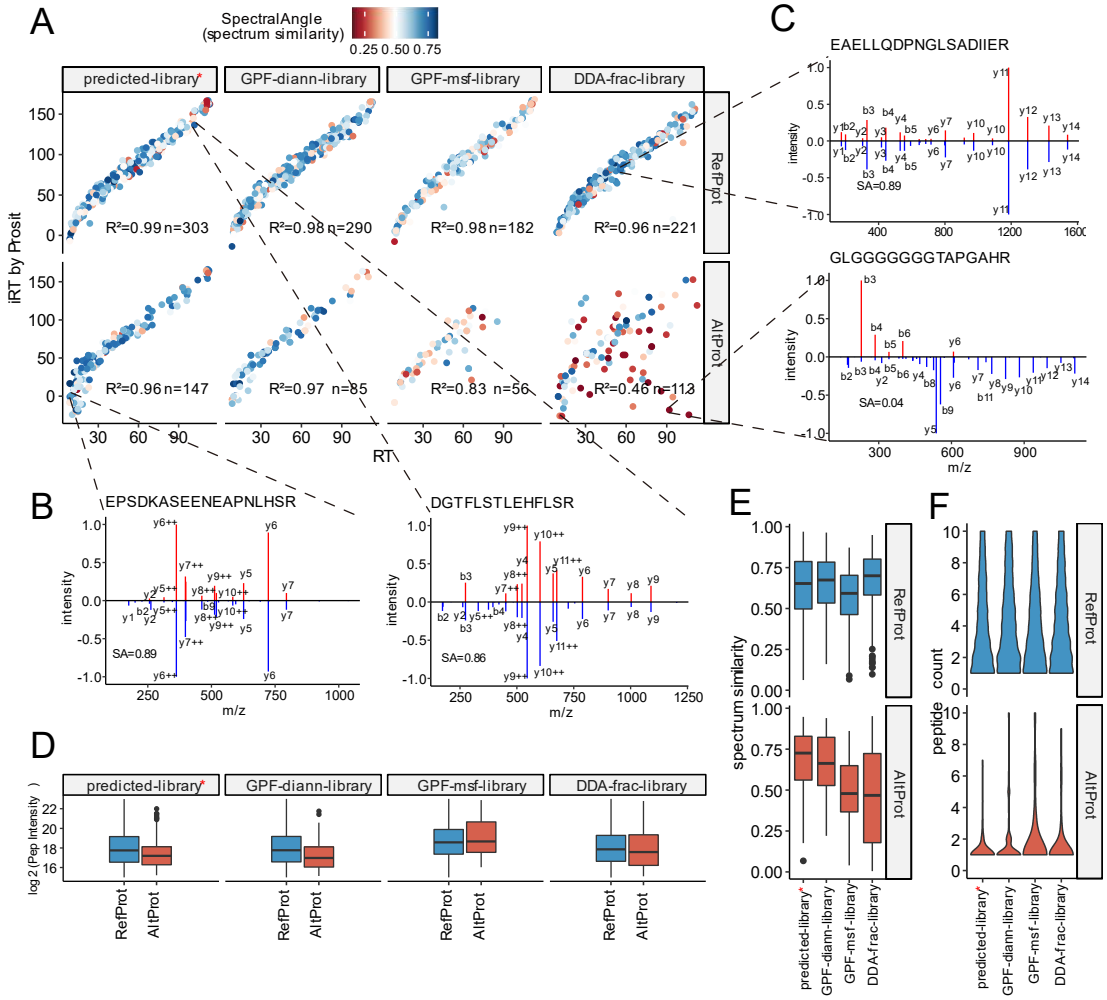


Figure 2-6 Evaluation of identifications from different strategies. (A)Correlation plots of the endogenous and predicted retention time of the canonical and AltProt and peptides, R square was calculated by Pearson method and SA value was indicated by colored dots. (B, C) Four examples of comparing spectra between endogenous and predicted peptides. (D)Distribution plot of peptide intensity identified by four strategies. (E) SA distribution plot of three types of peptides identified by four strategies. (F)Distribution of the number of peptide identifications per protein by four strategies.

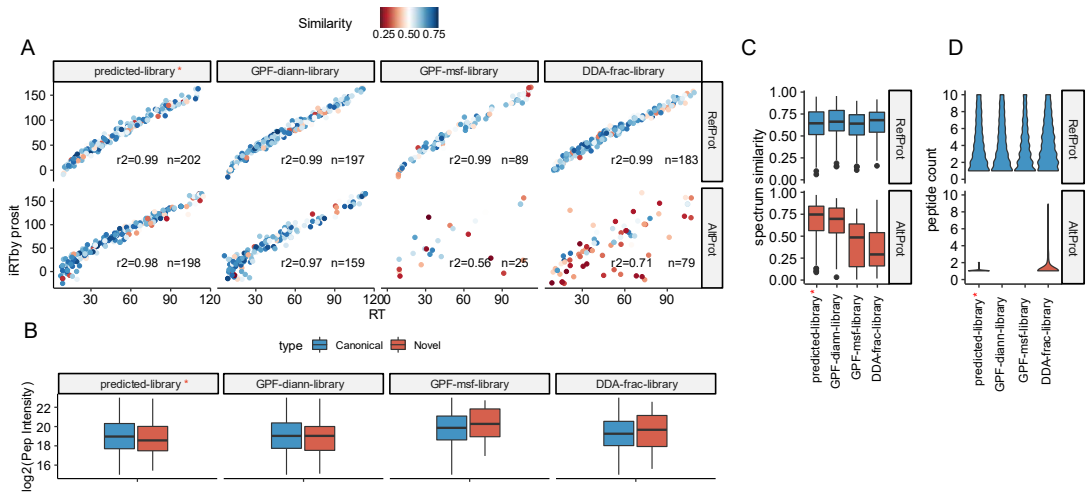


Figure 2-7 Evaluation of identifications from different strategies using HCT116 sample. (A)Correlation plots of the endogenous and predicted retention time of the canonical and novel peptides, R square was calculated by Pearson method and SA value was indicated by colored dots. (B) Distribution plot of peptide intensity identified by four strategies. (C) SA distribution plot of three types of peptides identified by four strategies. (D)Distribution of the number of peptide identifications per protein by four strategies.

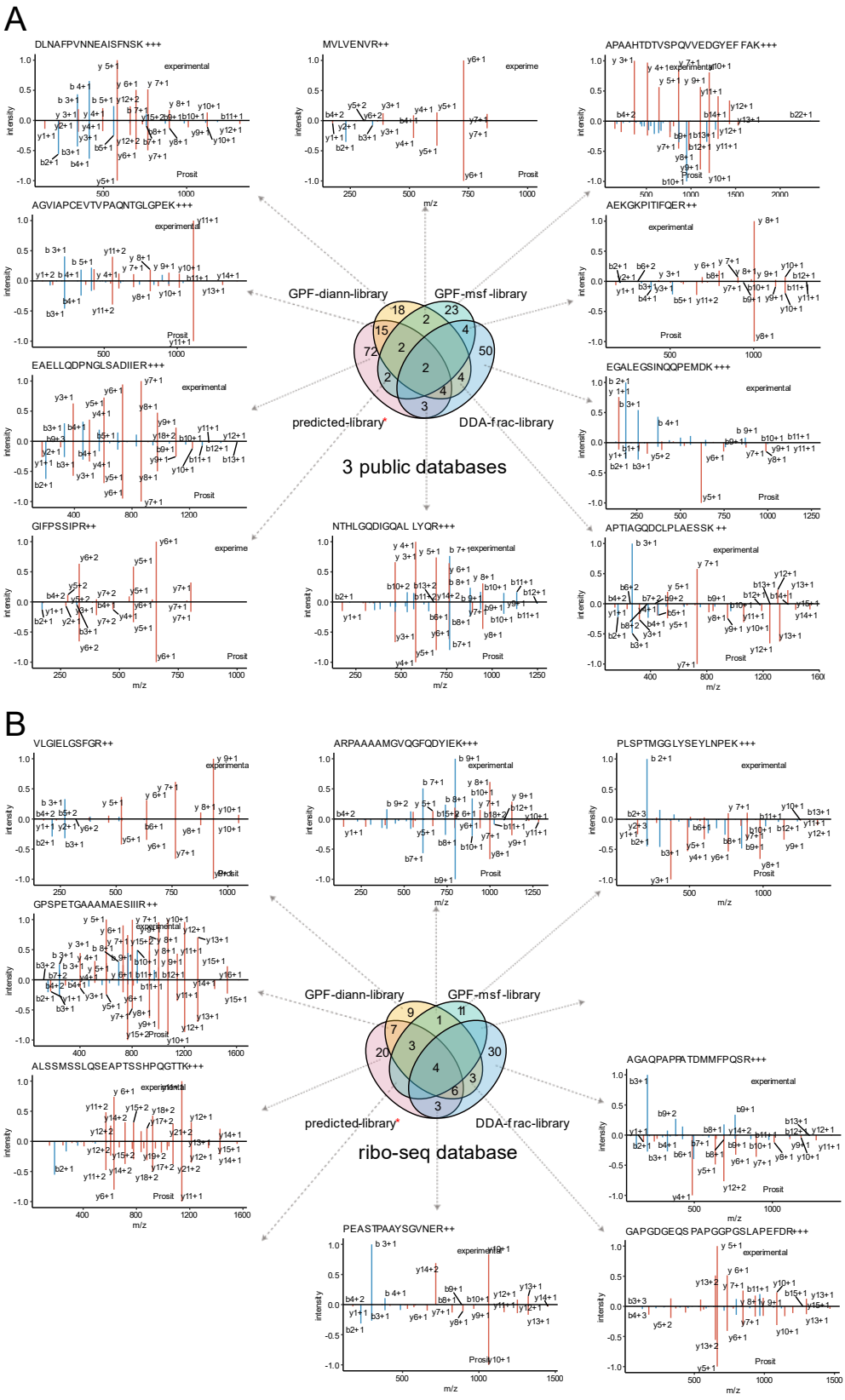


Figure 2-8 Validation of AltProt peptide identification from different library construction methods. Comparative spectra of experimental and Prosit-predicted

spectra driven from public 3DB (A) and in-house Riboseq database (B). This figure is connected to Fig 2-3 B. Each spectrum figure includes both experimental and predicted mass spectrometry data. The upper panel displays the experimental MS2 spectrum, with unannotated ions removed, while the lower panel shows the Prosit-predicted spectrum.

2.3.4 Differentially expressed AltProts in the mouse heart development were detected by using an optimized DIA method

To further delve into the practical biological applications of the DIA method in AltProt exploration studies, we utilized the same ten processing software in tandem with the previous Ribo-seq data to create a specific database to aid mass spectrometry analysis.[92] To circumvent the potential shortfall in AltProt identification due to the diversity of library search methods, we persistently employed the four library-building methods and three different databases for concurrent analysis**(Fig. 2-9A)** for identifying functional AltProts in mouse heart development.

Taking the Ribo-seq database as an example, we identified approximately 50K-80K peptides in both adult and embryonic mouse heart samples, and subsequently discovered nearly 300 AltProt peptides after mapping and blast filtering **(Fig. 2-9A, C).** We found that the N_extension held the top spot in terms of sORF type, a finding that aligned with *Na et al.* 's observation of the N_extension having the highest count.[93] These mass spectra results underscored the dynamic nature of genome translation and the incompleteness of the conventional protein database. Additionally, as shown in **Fig. 2-9D**, AltProts were identified across various chromosomes, and these AltProts were

often less than 200 amino acids in length. This revealed the widespread presence of AltProts in the human genome, warranting further investigation.

As illustrated in **Fig. 2-11A**, we performed a PCA analysis, that revealed a substantial divergence between the proteomes of adult mouse hearts and those of embryonic mouse hearts. Additionally, we assessed the correlation between the fold change in the mass spectrometry-based proteome and the fold change in the Ribo-seq-based transcriptome and got the result of a p-value of less than 2.2e-16**(Fig. 2-11B)**. These analyses lend credence to the relative reliability of our proteomic study. Upon applying a 2-fold change and a 0.05 adjusted p-value as filtering criteria, 14 and 41 differentially expressed AltProts in the Ribo-seq database and 3db were identified respectively **(Fig. 2-11C)** in addition to thousands of canonical differential proteins. GO pathways **(Fig. 2-11D，Fig. 2-13A)** revealed that in the biological process, embryonic mouse hearts were predominantly involved in RNA-related pathways, while proteins of adult mouse hearts were predominantly involved in metabolism and energy production, a finding that aligned perfectly with the biological phenotype. Within the GO pathways, we identified not only the uORF of Cd2bp2, the lncRNA of Ash1l, and the dORF of Smyd2, whose reference proteins were involved in pathways such as mRNA processing and covalent chromatin modification, but also detected the uORFs of Noct and Drd4. The RefProt of these two AltProts were in circadian rhythm pathway of GO. This suggested potential interactions and underlying roles between AltProts and their corresponding RefProts.

.

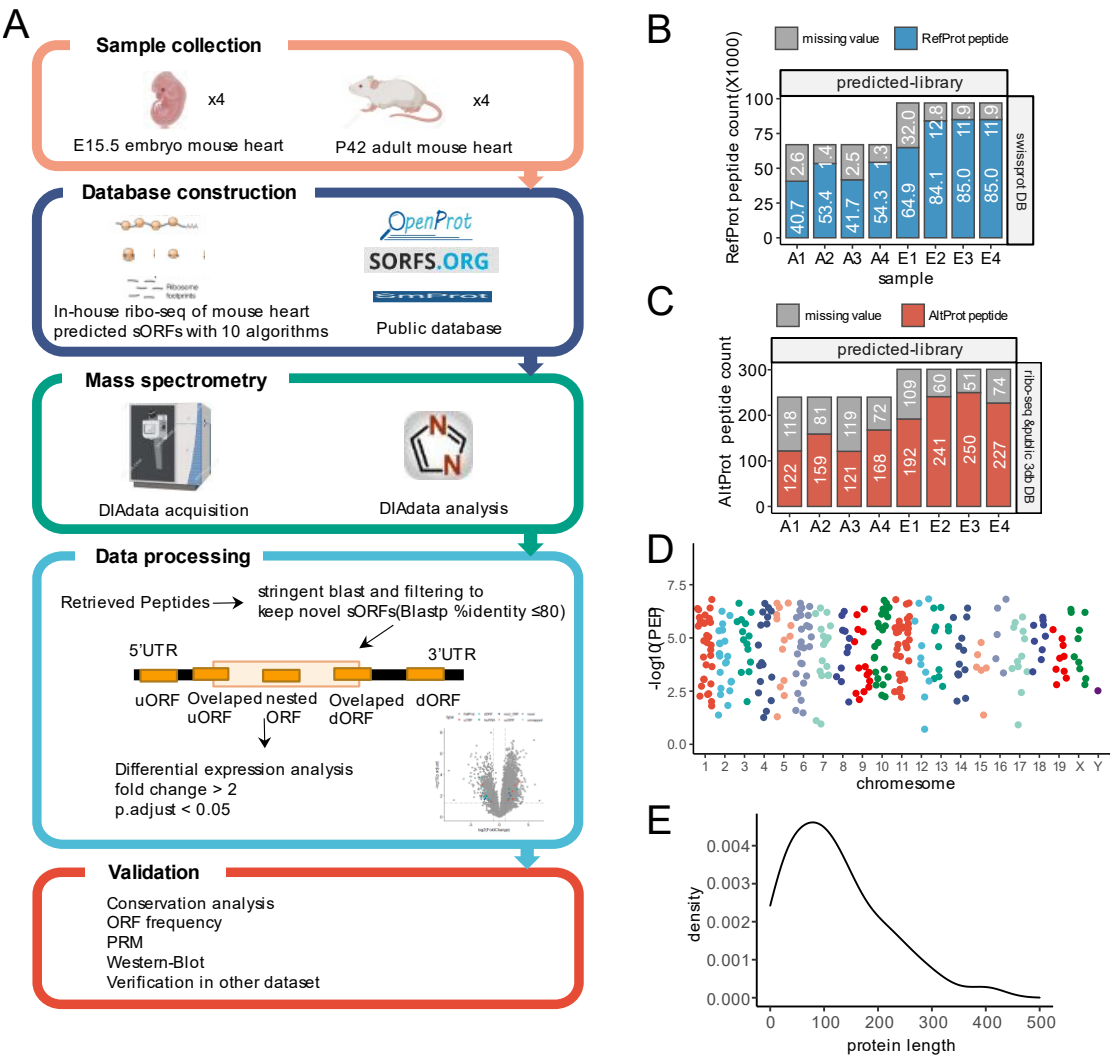


Figure 2-9 Application of DIA in mouse heart development. (A)Workflow of AltProt exploration in embryonic and adult mouse heart samples. (B-C) Count of canonical and AltProt peptides identified in eight samples by predicted-library. (D)Manhattan plot of gene chromosome location of AltProt peptides. (E)Protein length distribution of AltProt including uORF, overlapped uORF, AltProt ORF, overlapped dORF and dORF.

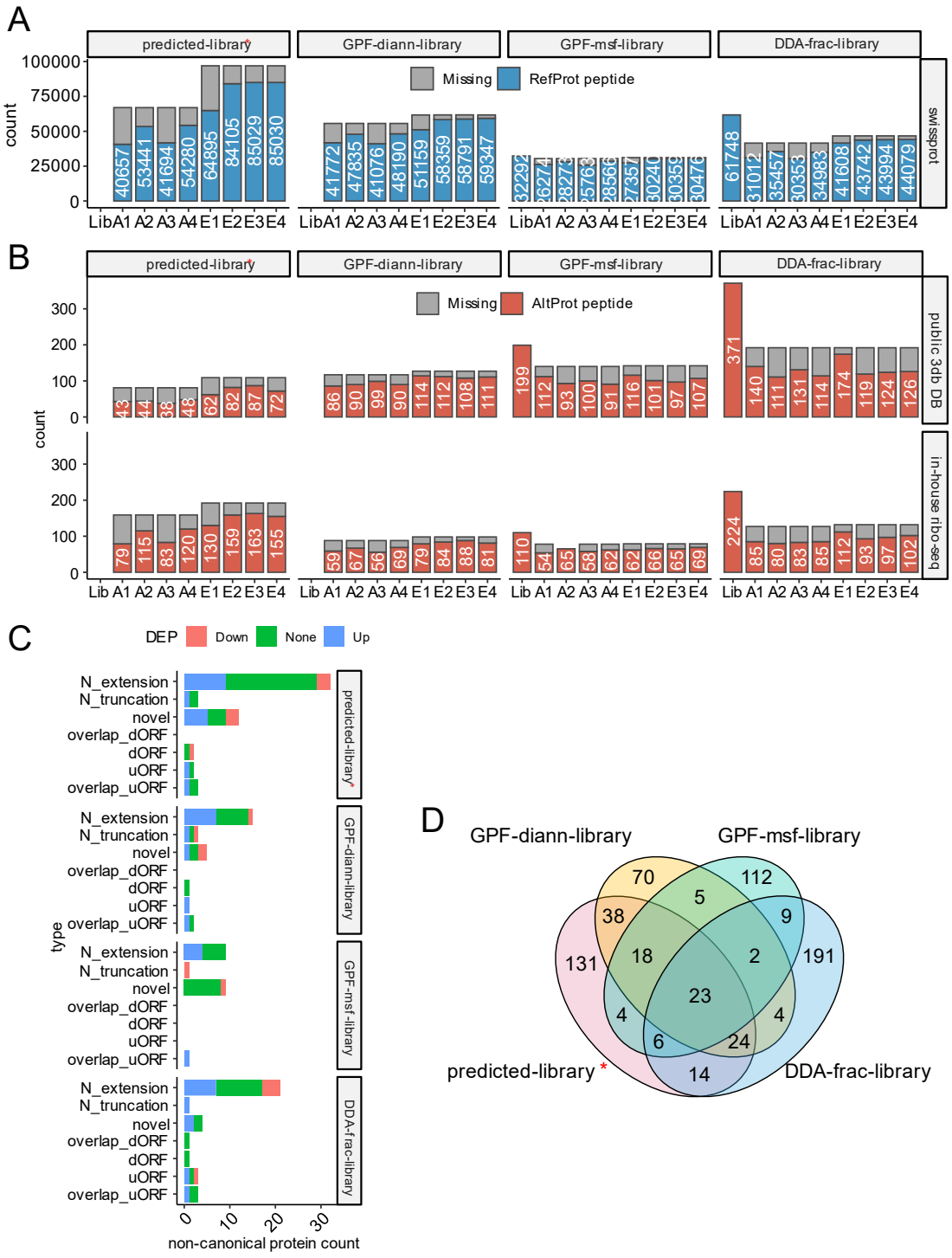


Figure 2-10 Comparative profiling of peptide identifications in mouse heart development using different strategies. Count of canonical (A) and novel (B) peptides identified in eight samples by four library construction methods in mouse heart development. (C) Bar plot of different expressed non-canonical protein count

numbers in four library construction methods. (D) Venn diagram of identified peptides by four different library construction methods.

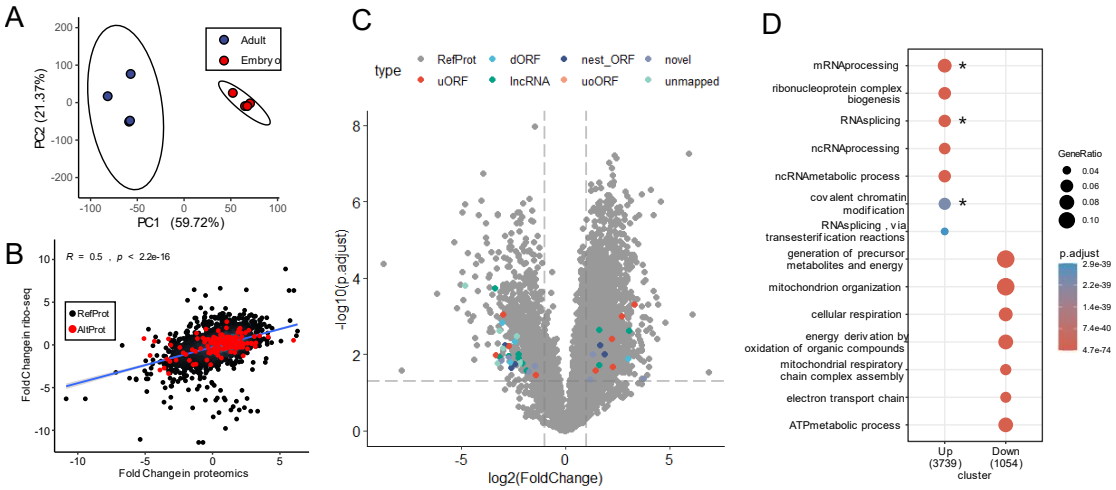


Figure 2-11 Different expression analyses. (A) PCA plot of eight samples where A represents adult mouse heart and E represents embryonic mouse heart. (B) Protein fold change correlation between Ribo-seq and mass spectrometry technique. (C) Volcano plot and protein type distribution of identified differentially expressed proteins. (D) GO terms (biological process) enriched in up- and down-regulated genes in embryonic development. The asterisks represent that certain proteins under this GO pathway have corresponding alternative proteins present.

2.3.5 Validation of expression of ASDURF and other AltProts.

We compiled a comprehensive list of differentially expressed AltProts derived from two databases and four library construction methodologies (please refer to Table 2-1). Further validation of these AltProts was conducted through PRM, western blot, conservation analysis, and cross-referencing with other datasets.

The PRM results demonstrated that 23 AltProts from the 3db database and 6 AltProts from the Ribo-seq database successfully passed manual verification **(Fig. 2-13 B, Fig. 2-14)**. We selected four AltProts for further translational confirmation via over-expression Western-Blot experiments **(Fig. 2-12C)**. Four distinct bands of AltProts were observed in the Western-Blot, indicating their translationality. Among the differentially expressed AltProts validated by PRM, a few previously reported AltProts including ASDURF (encoded by uORF)[67], ASH1L (encoded by lncRNA)[94], and RPS4L (encoded by lncRNA)[63, 95] were detected in our study, demonstrated the sensitivity of our method. Besides, we identified three uORFs that corresponded to the canonical protein of their respective annotated ORF. These three pairs of AltProts and canonical proteins exhibited identical expression trends in mouse hearts development. This may be due to the transcription and translation mechanisms, or there may be some unidentified regulatory mechanism between uORF and main ORF. For instance, Lipoprotein lipase (LPL) was highly expressed in adult mouse hearts, while ER degradation-enhancing alpha-mannosidase-like protein (Edem) and Mesoderm-specific transcript homolog protein (mest) were highly expressed in embryonic hearts **(Fig. 2-12B).** Their co-expression patterns with uORFs hint at potential unidentified

regulatory mechanisms, although further evidence is required to elucidate their interactions.

Lastly, we discovered the AltProt ASNSD1 Upstream Open Reading Frame (ASDURF) in our dataset. Previous reports have suggested that this AltProt is involved in the construction of the PAQosome,[96, 97] which plays a crucial role in various fundamental cellular activities, including protein synthesis, ribosome biogenesis, transcription, and splicing[98]. Furthermore, it has been reported that phenotypically it can enhance the survival of medulloblastoma cells.[99] Our PRM spectra confirmed the presence and elevated expression of ASDURF in mouse embryonic hearts **(Fig. 2-12D)**, ASDURF has been identified as a previously uncharacterized component of the PAQosome complex[96], which is involved in the biogenesis of various protein complexes, transcription, splicing, and other cellular processes[98]. Moreover, in addition to ASDURF, we observed that other subunits of the PAQosome, including RPAP3, PIH1D1, PFDN2, and RUVBL1, are similarly highly expressed in mouse embryonic hearts. This observation suggests that the PAQosome, comprising components such as ASDURF, plays a significant role in functions including protein synthesis and transcriptional regulation during cardiac development. To further validate this, we cross-referenced ASDURF with public human MS data[81] and found ASDURF was identified in several different organs, especially in the thyroid **(Fig. 2-12E)**. Moreover, conservation analysis revealed that this AltProt was highly conserved across several species **(Fig. 2-12F)**. Although ASDURF is currently classified as a reviewed human

protein in the UniProt database, it is not yet classified as such in mice. The mass spectrometry findings presented herein provide the first compelling evidence of the presence of ASDURF in the mouse proteome.

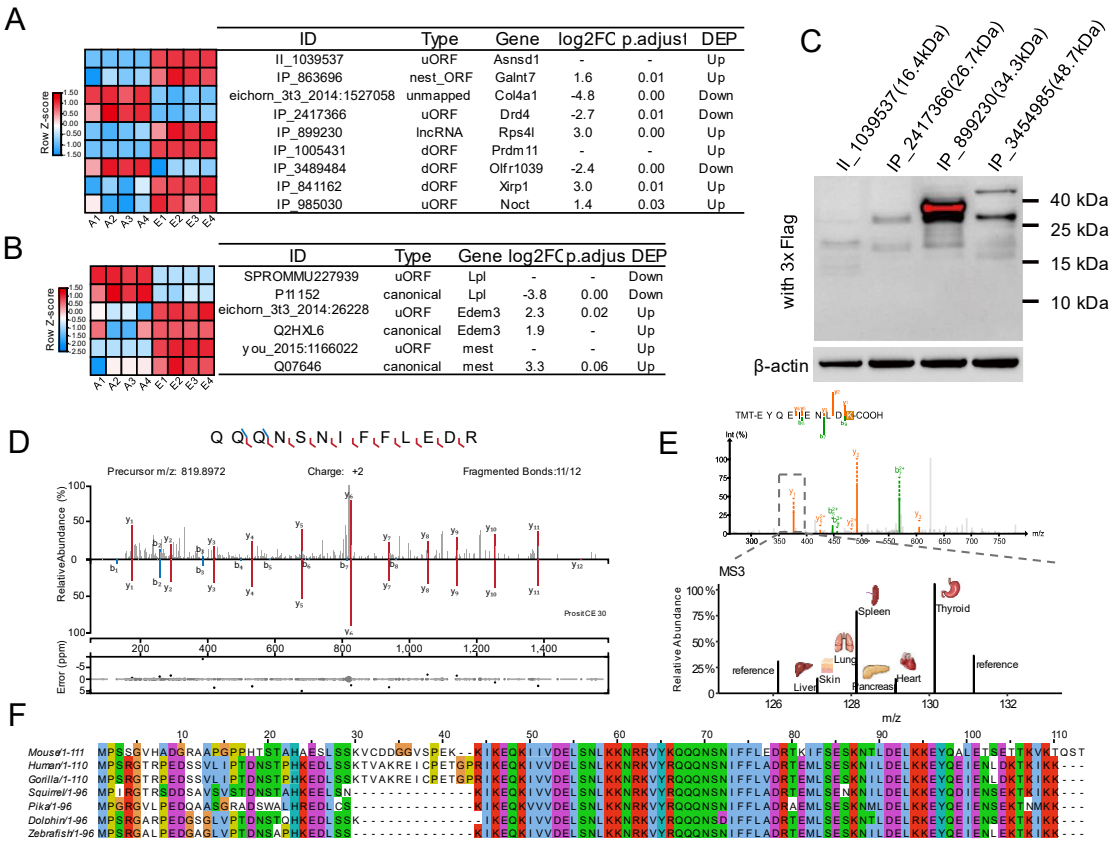


Figure 2-12 Validation of ASDURF and other AltProts. (A) Heatmap of nine differently expressed AltProt examples. (B)Heatmap of three uORFs with corresponding main ORF. (C) Over-expression validation of four selected AltProts by western-blot experiment. The labeled molecular weight had been added with the 3.3kDa of flag tag size. (D)Comparison between predicted and endogenous spectra of ASDURF. (E) MS2 and MS3 spectra of AltProt peptide EYQEIENLDK from ASDURF. The y1 ion was triggered for the MS3 event and each TMT peak represented the protein abundance existed in different human organs. (F) Protein sequence alignment plot of ASDURF in seven species.

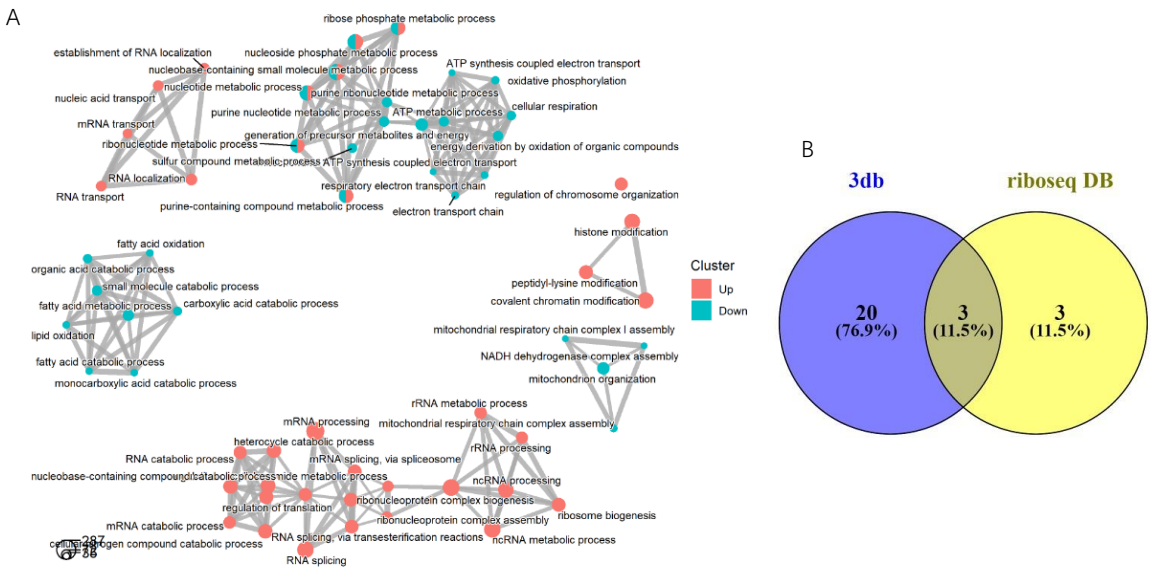
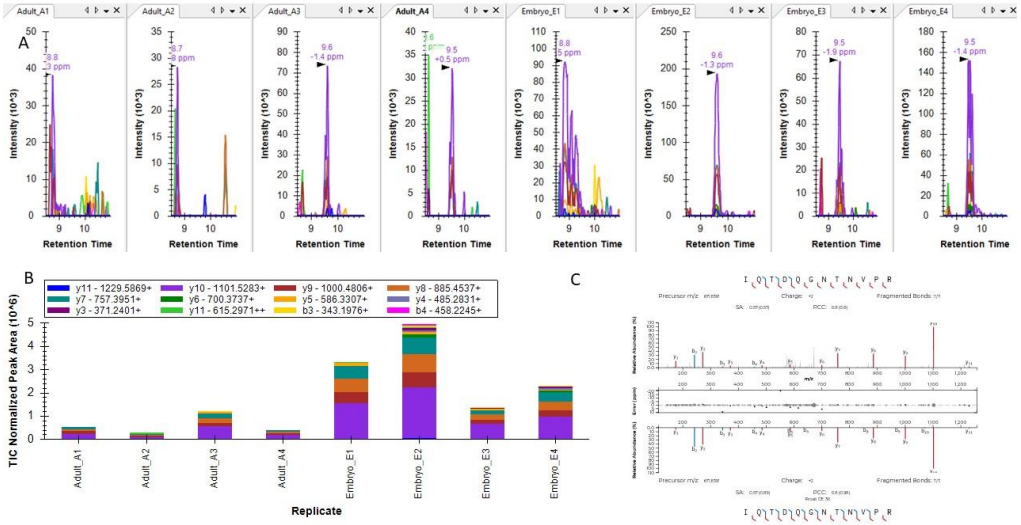
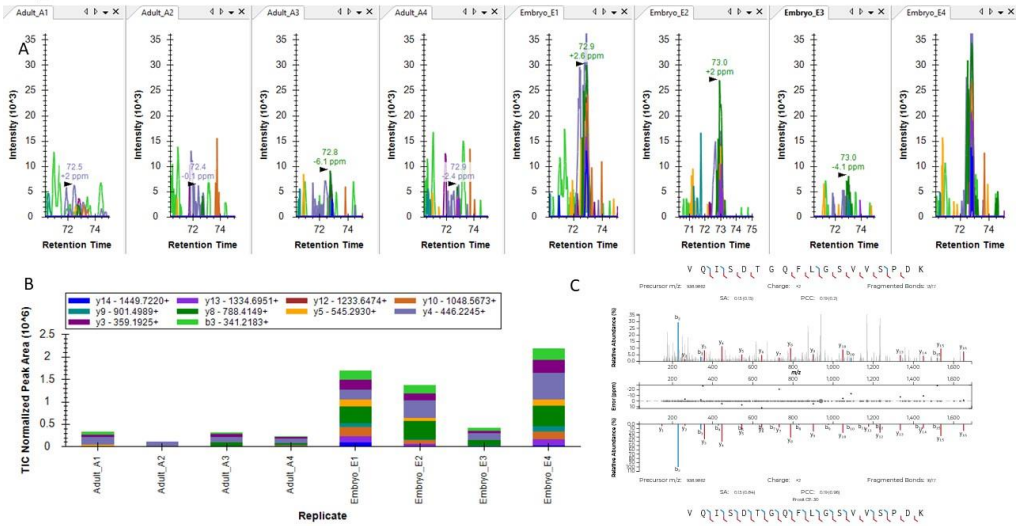


Figure 2-13 Gene Ontology (GO) analysis and database sources in mouse heart development. (A) Emapplot of different expressed proteins in mouse heart development. Red dots represent up regulated biological progress in embryo mouse heart and blue dots represent down regulated biological progress in mouse adult heart. (B) Venn diagram of different expressed AltProts in ribo-seq database and 3db after PRM validation.

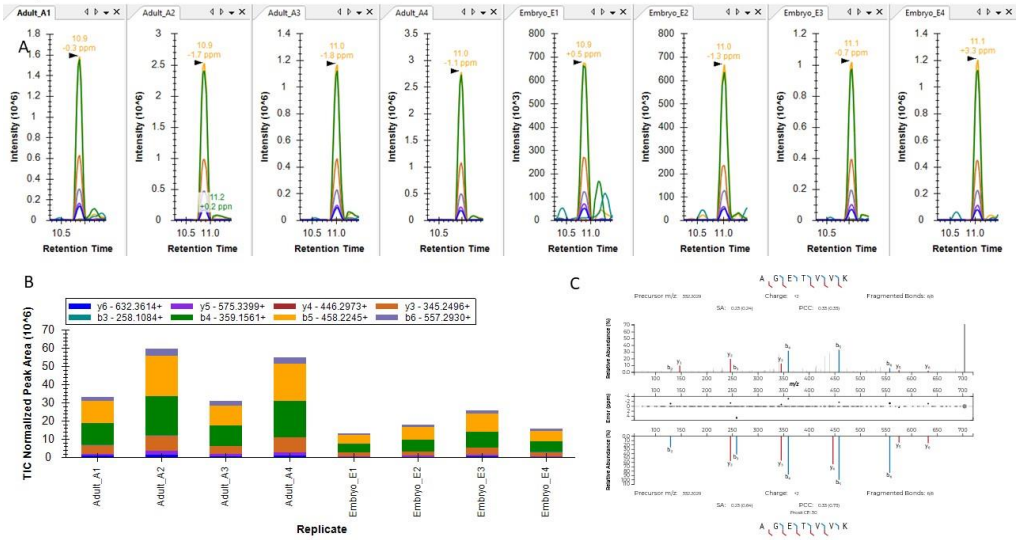
IP_985030, Noct, uORF, IQTDQGNTNVPR



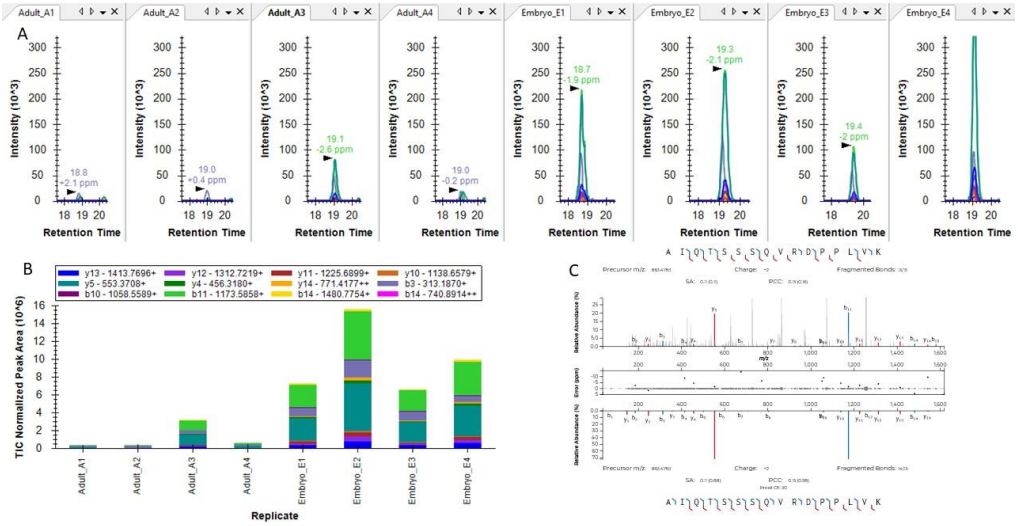
IP_985030, Noct, uORF, VQISDTGQFLGSVVSPDK



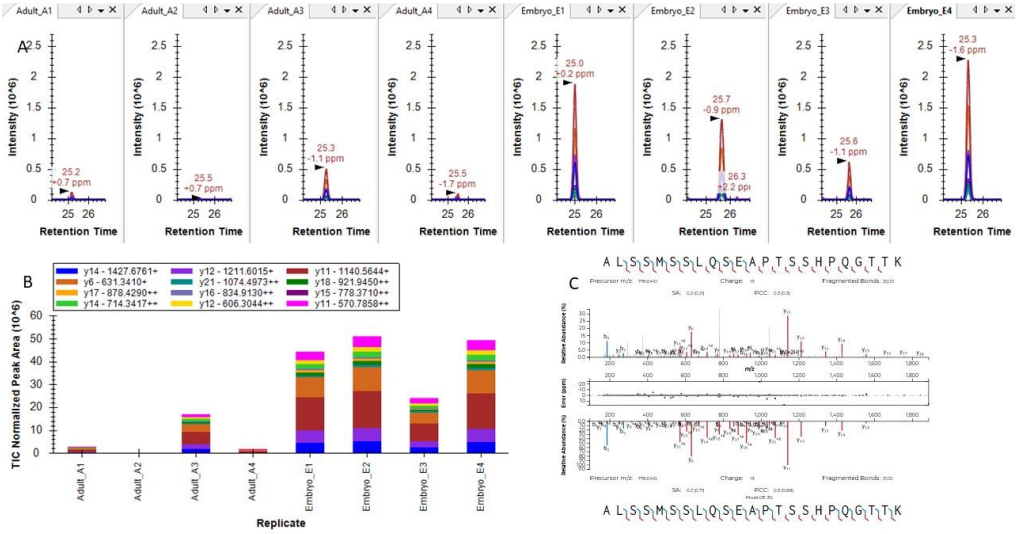
IP_985759, Ash1l, lncRNA, AGETVVK



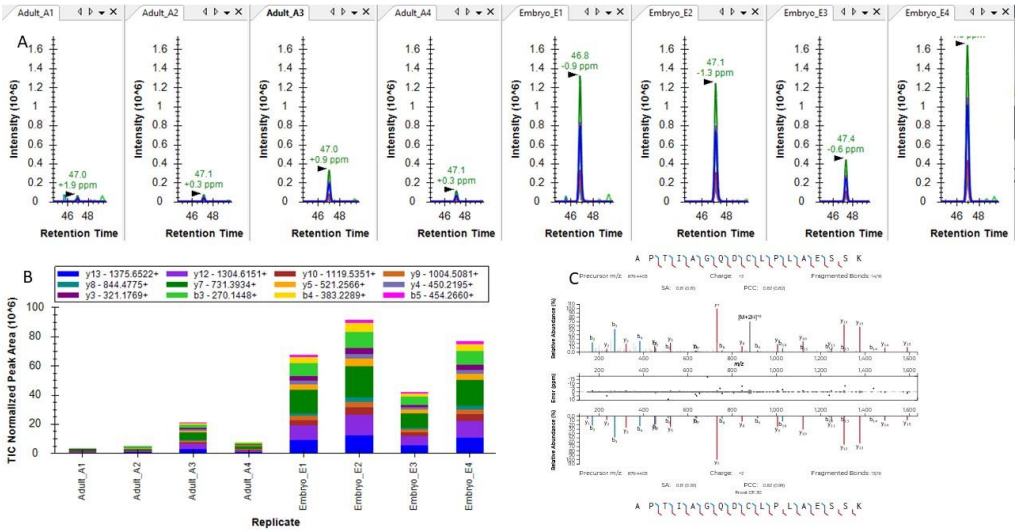
IP_841162, Xirp1, dORF, AIQTSSSQVRDPPLVK



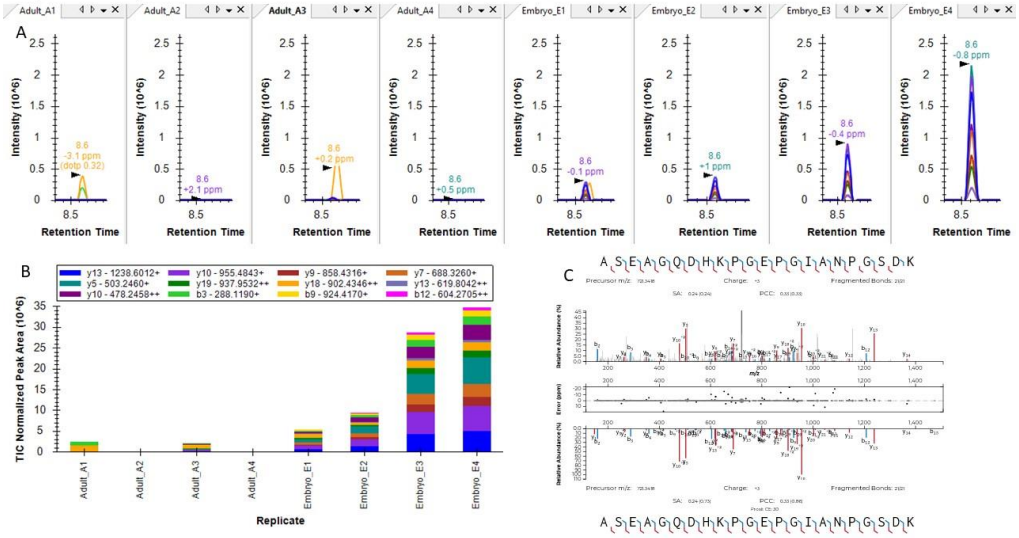
IP_841162, Xirp1, dORF, ALSSMSSLQSEAPTSSHPQGTTK



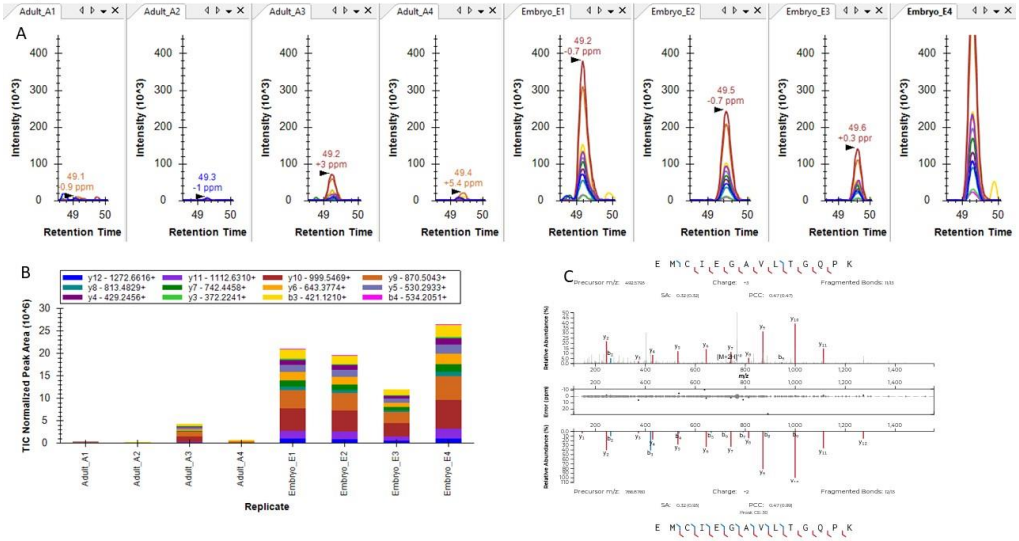
IP_841162, Xirp1, dORF, APTIAGQDCLPLAESSK



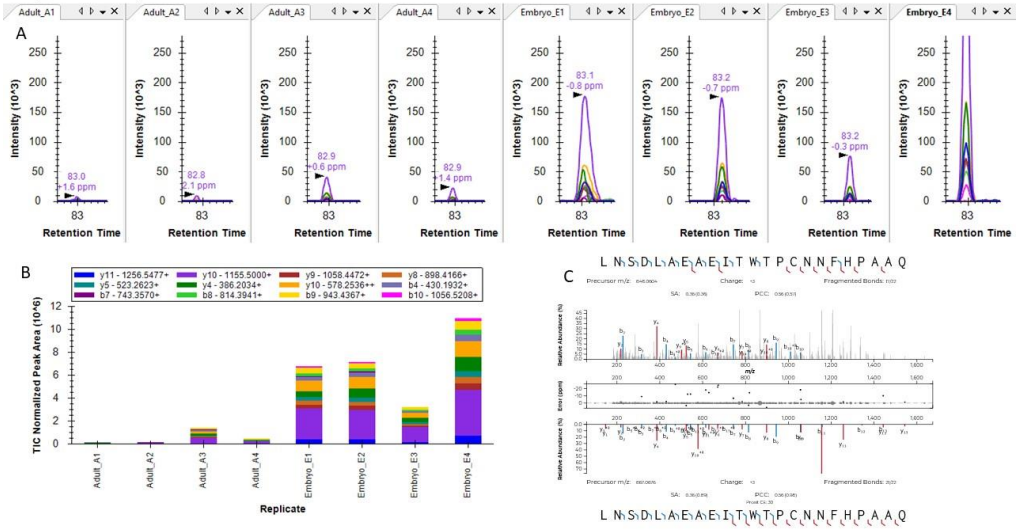
IP_841162, Xirp1, dORF, ASEAGQDHKPGEPGIANPGSDK



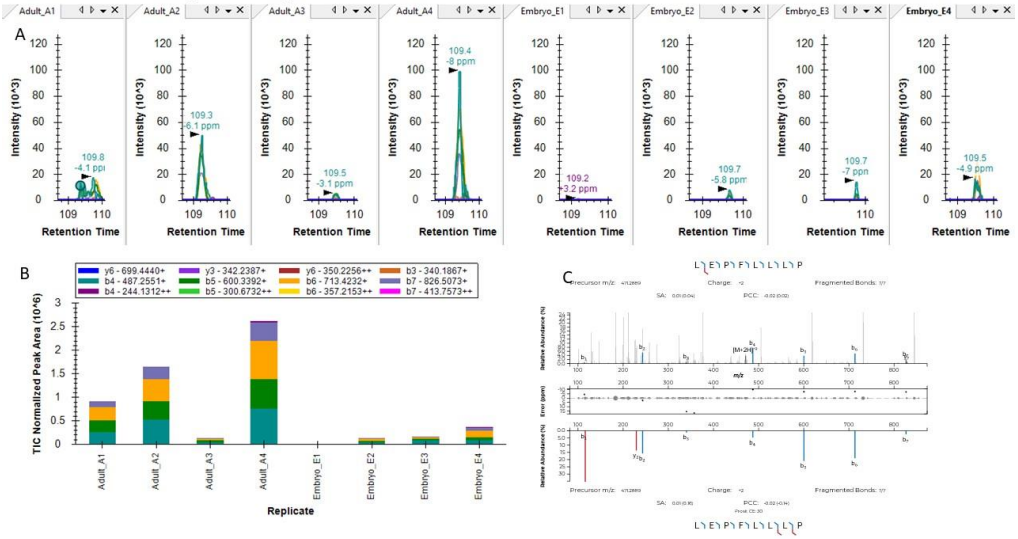
IP_841162, Xirp1, dORF, EMCIEGAVLTGQPK



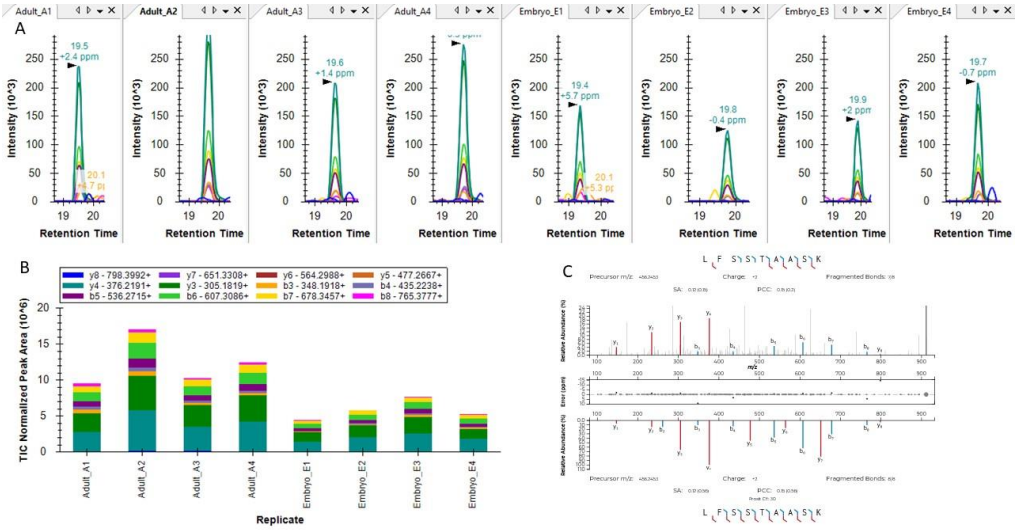
IP_841162, Xirp1, dORF, LNSDLAEAEITWTPCNNFHPAAQ



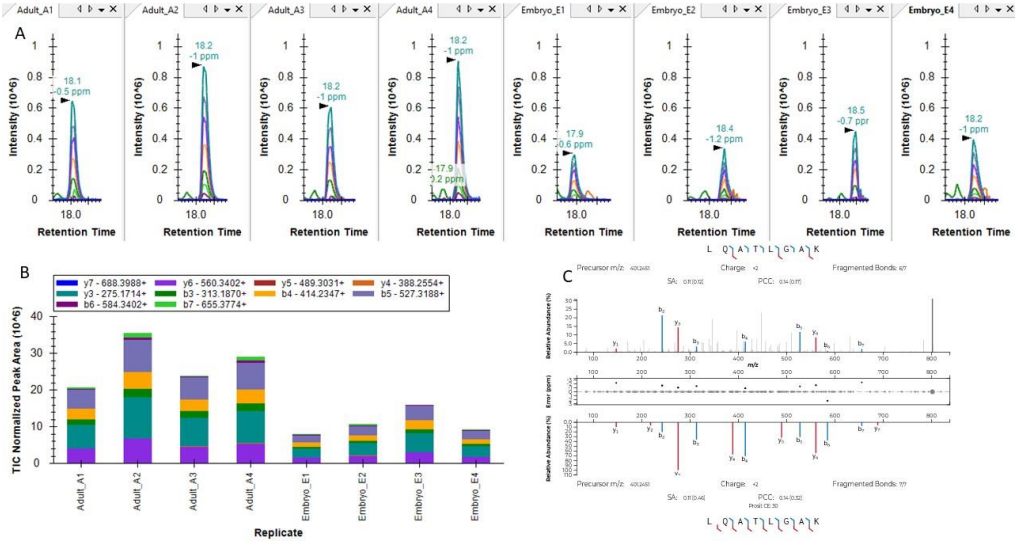
SPROMMU221938, Zfp518b, nest_ORF, LEPFLLLP



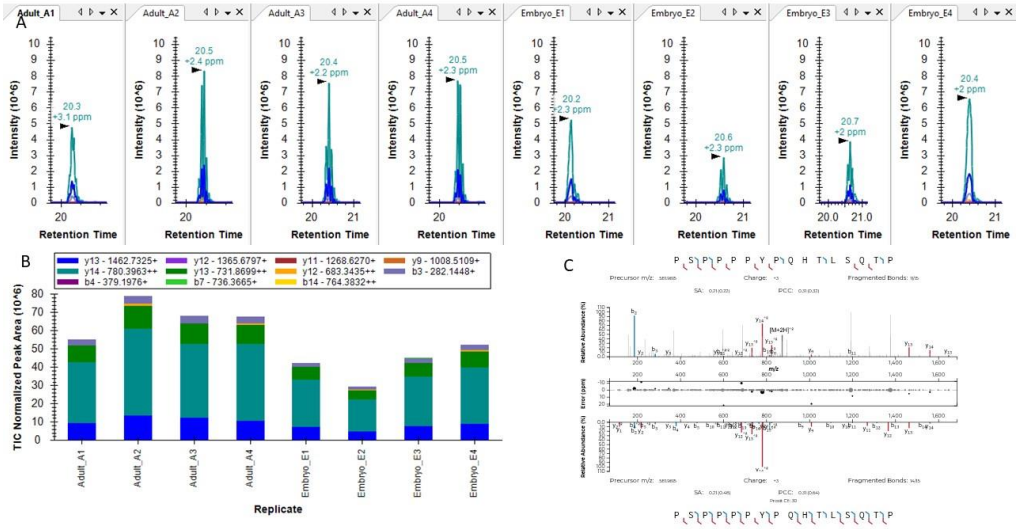
IP_872056, Gm44916, ncRNA, LFSSTAASK



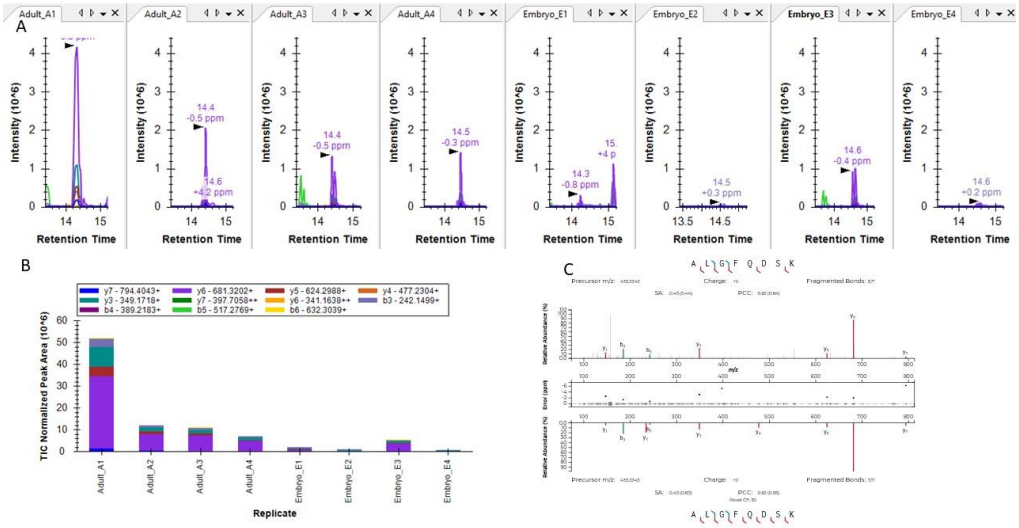
IP_2502222, Gm42260, ncRNA, LQATLGAK



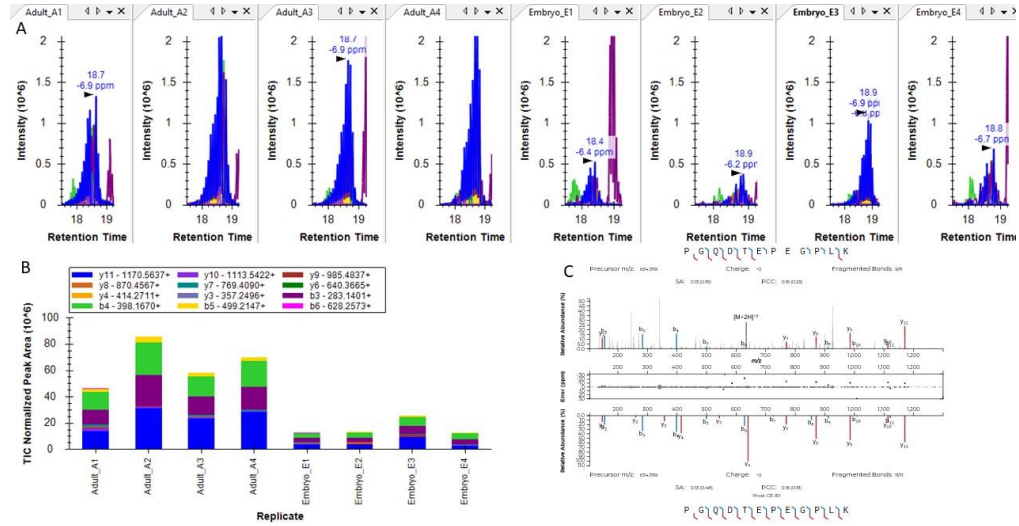
IP_3442236, Gm10277, ncRNA, PSPPPPYPQHTLSQTP



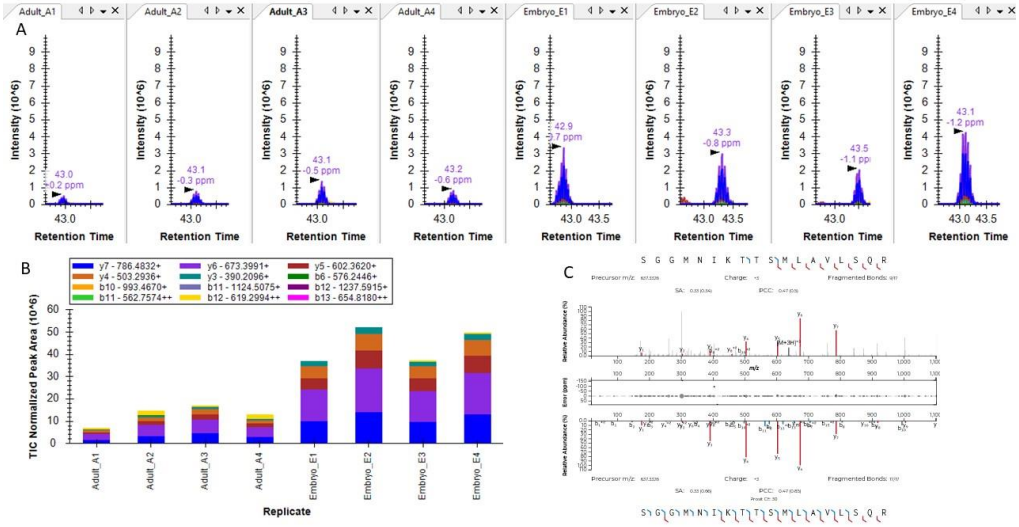
eichorn_3t3_2014:1527058, Col4a1, unmapped, ALGFQDSK



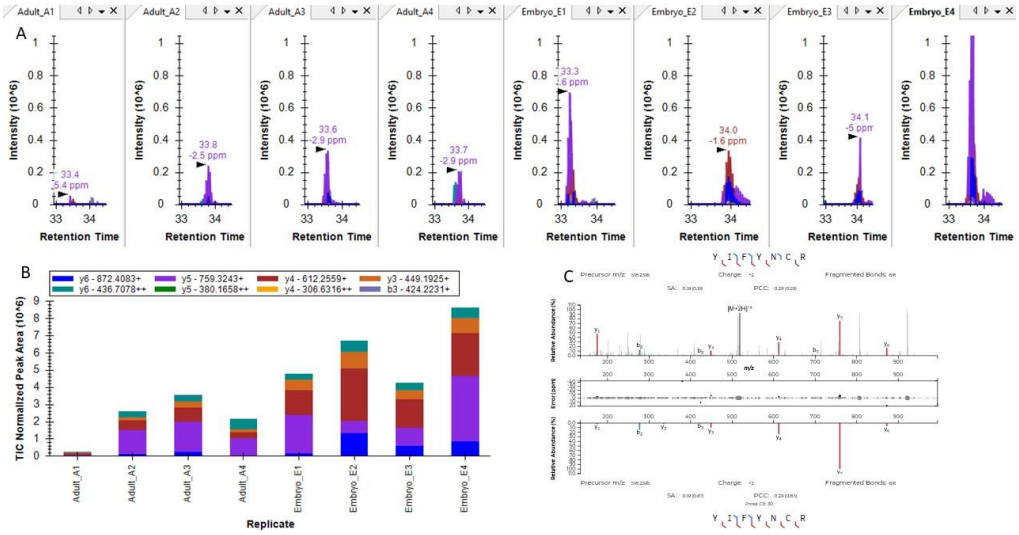
IP_2417366, Drd4, uORF, PGQDTEPEGPLK



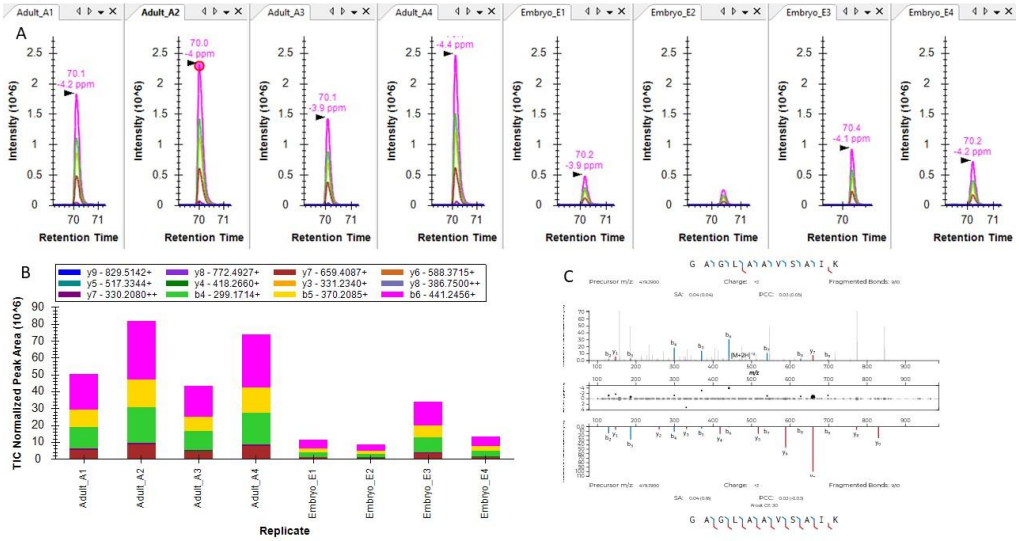
IP_863696, Galnt7, nest_ORF, SGGMNIKTTSMLAVLSQR



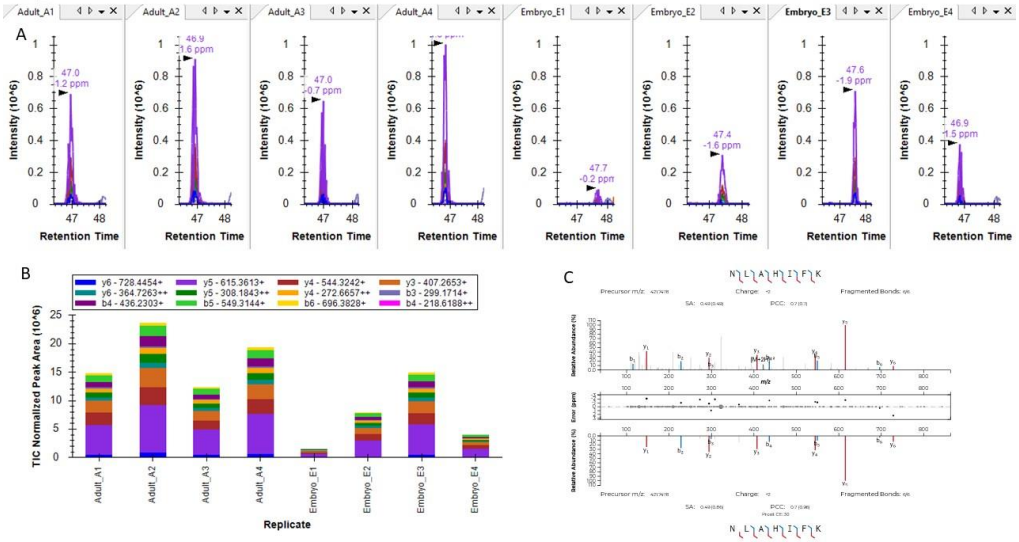
IP_935570, Gbp6, lncRNA, YIFYNCR



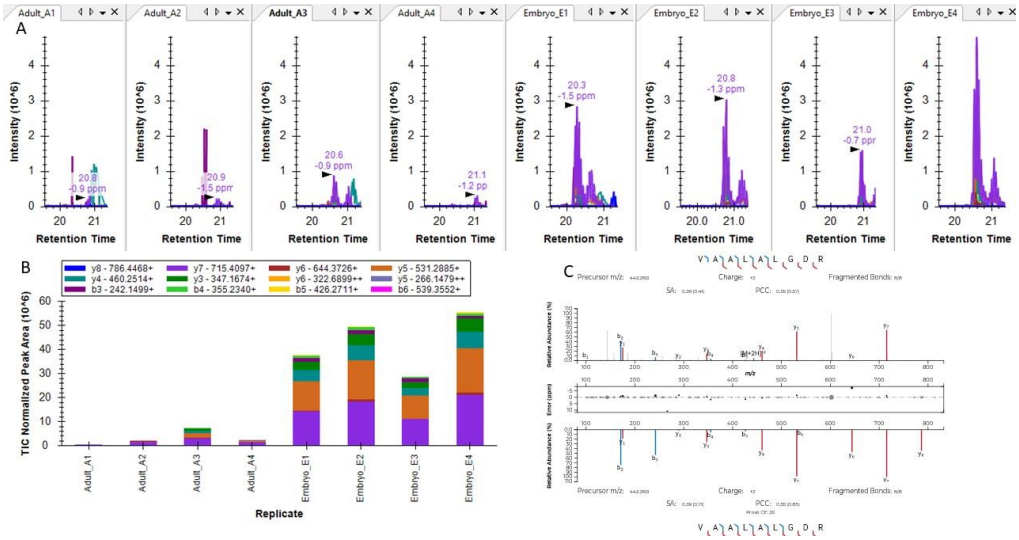
IP_859576, Car7, dORF, GAGLAAVSAIK



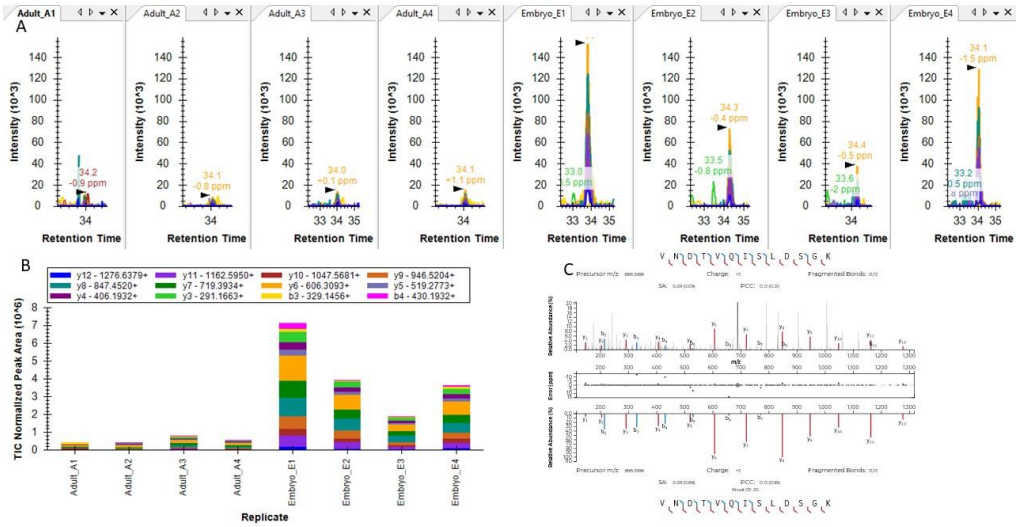
IP_3489484, Olfr1039, dORF, NLAHIFK



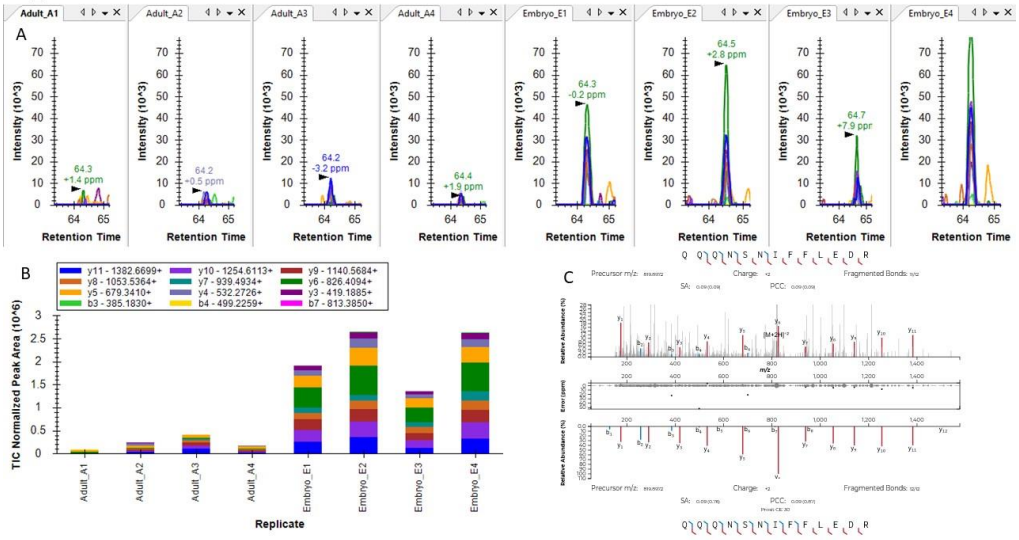
eichorn_3t3_2014:26228, Edem3, uORF, VAALALGDR



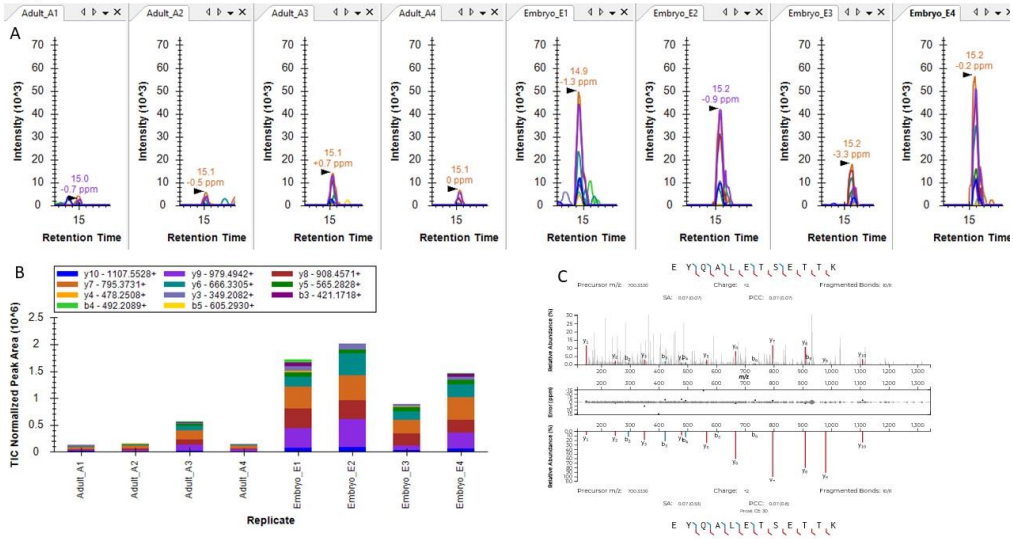
IP_899230, Rps4l, lncRNA, VNDTVQISLDSGK



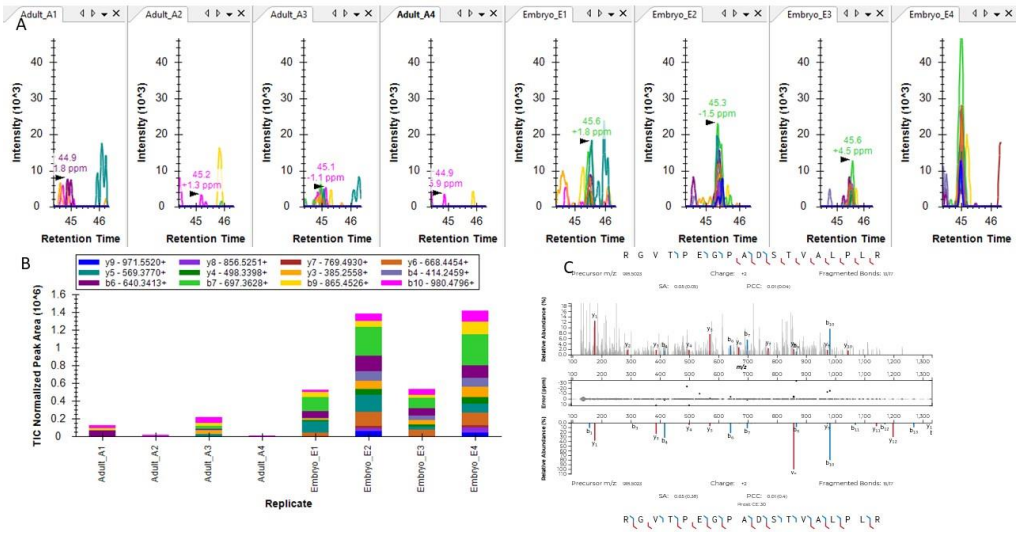
II_1039537, Asnsd1, uORF, QQQNSNIFFLEDR



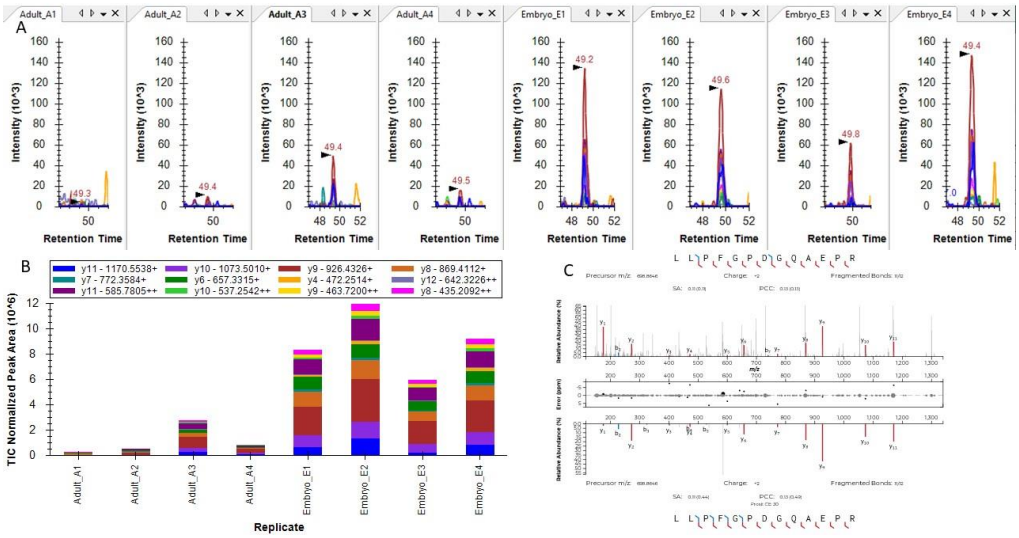
II_1039537, Asnsd1, uORF, EYQALETSETTK



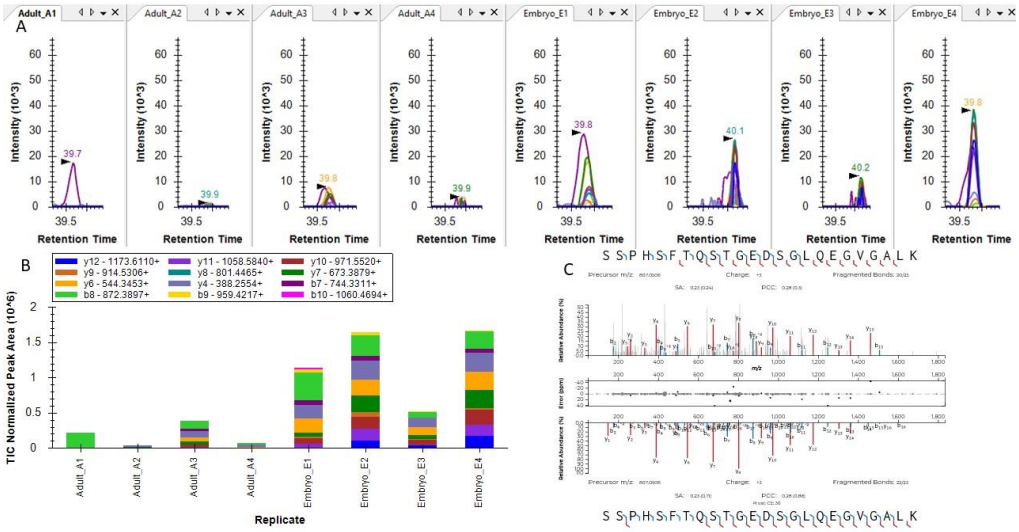
IP_3454985, Gm42047, lncRNA, RGVTPEGPADSTVALPLR



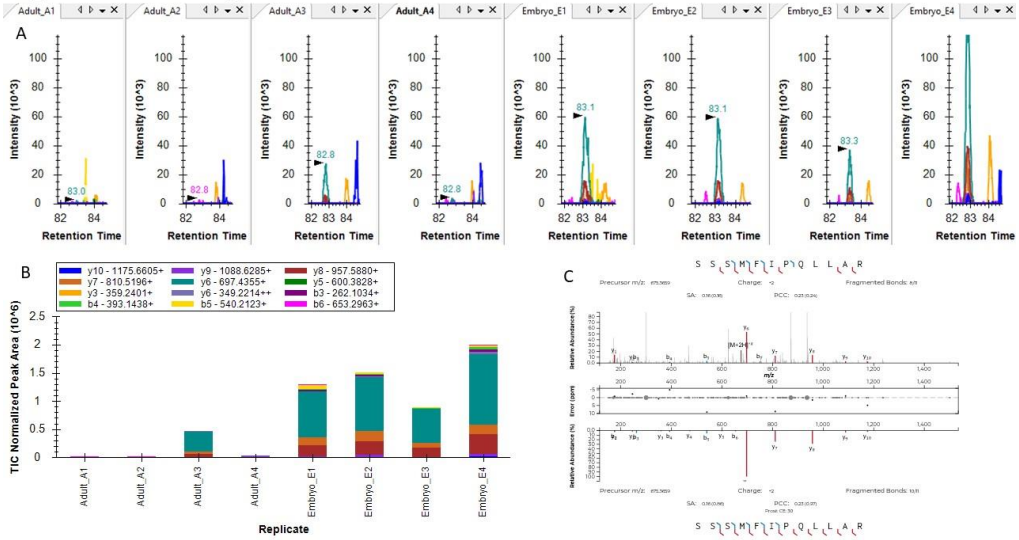
II_3451328, Nedd4l, unmapped, LLPFGPDGQAEPR



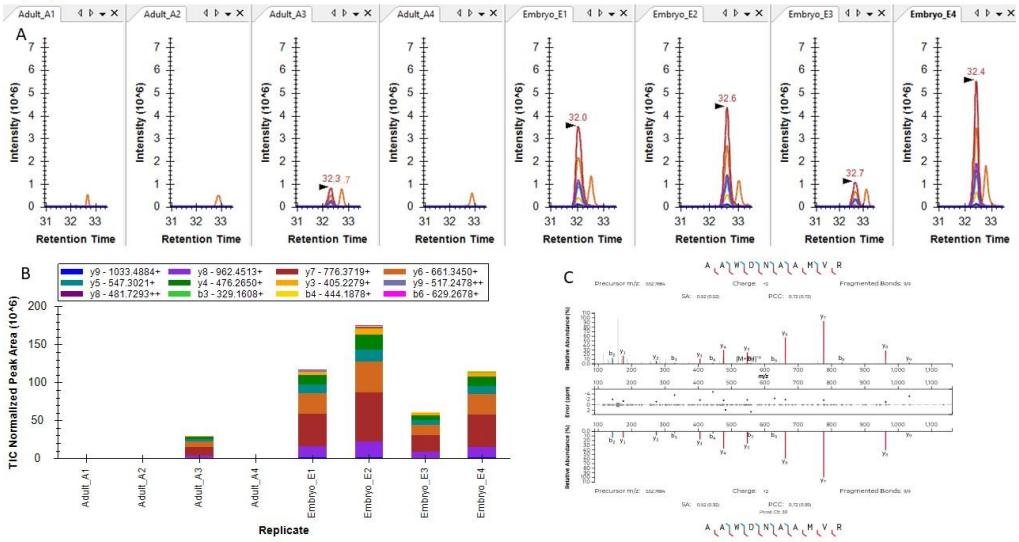
II_3451328, Nedd4l, unmapped, SSPHSFTQSTGEDSGLQEGVGALK



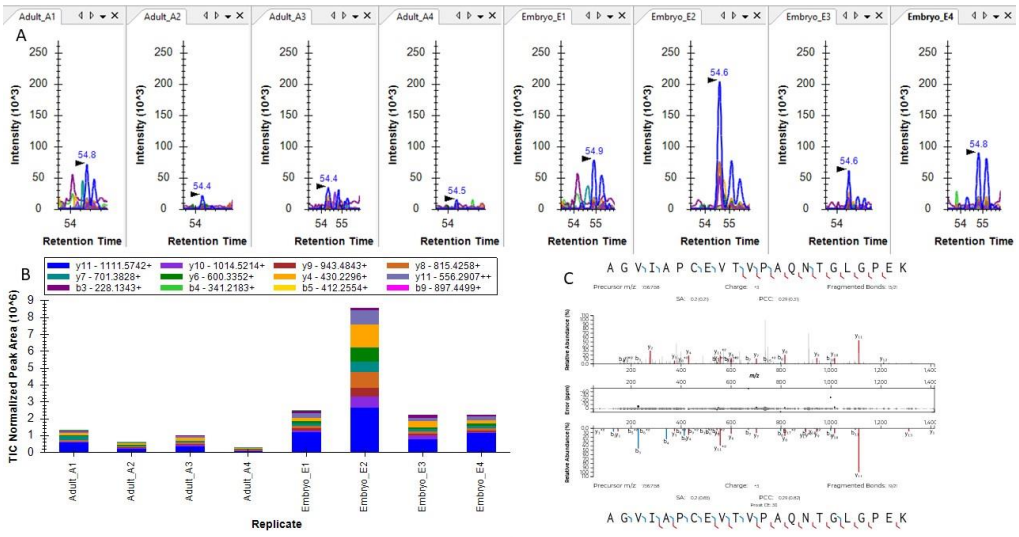
II_3451328, Nedd4l, unmapped, SSSMFIPQLLAR



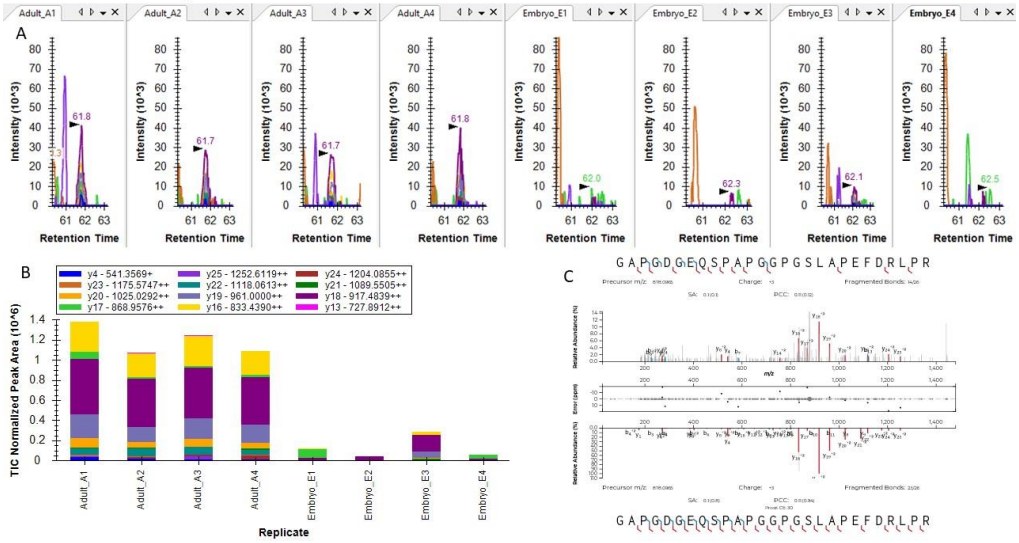
you_2015:1166022, mest, uORF, AAWDNAAMVR



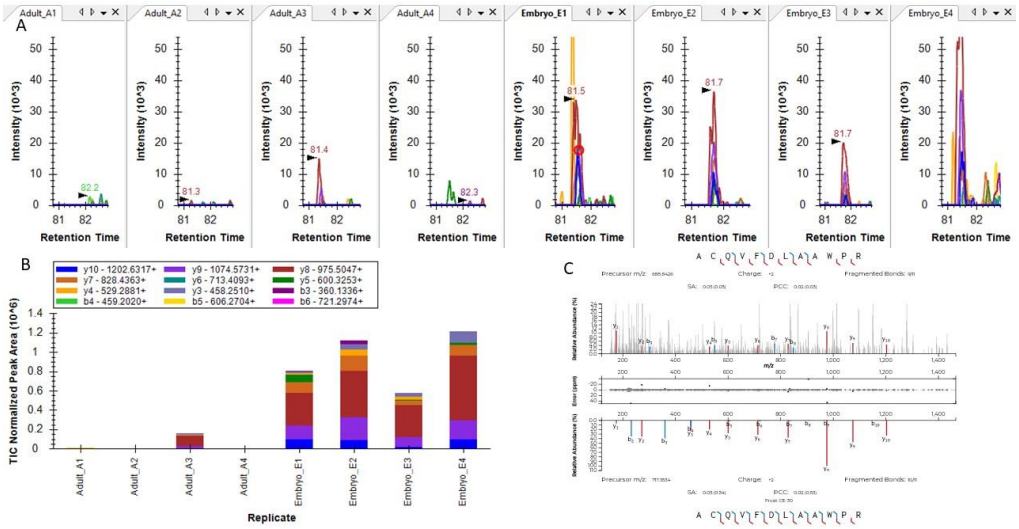
SPROMMU147363, Gm5948, lncRNA, AGVIAPCEVTVPAQNTGLGPEK



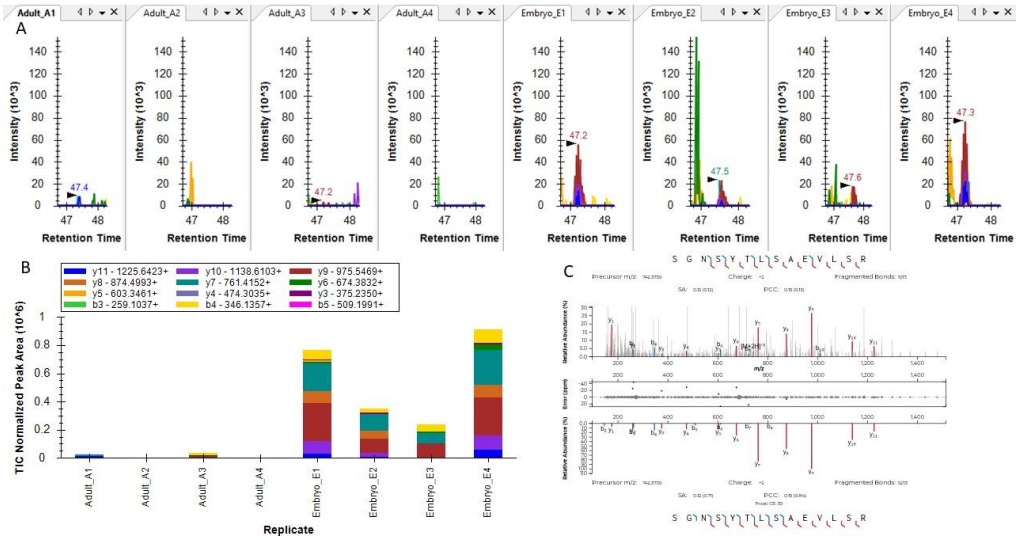
SPROMMU227939, Lpl, uoORF, GAPGDGEQSPAPGGPGSLAPEFDRLPR



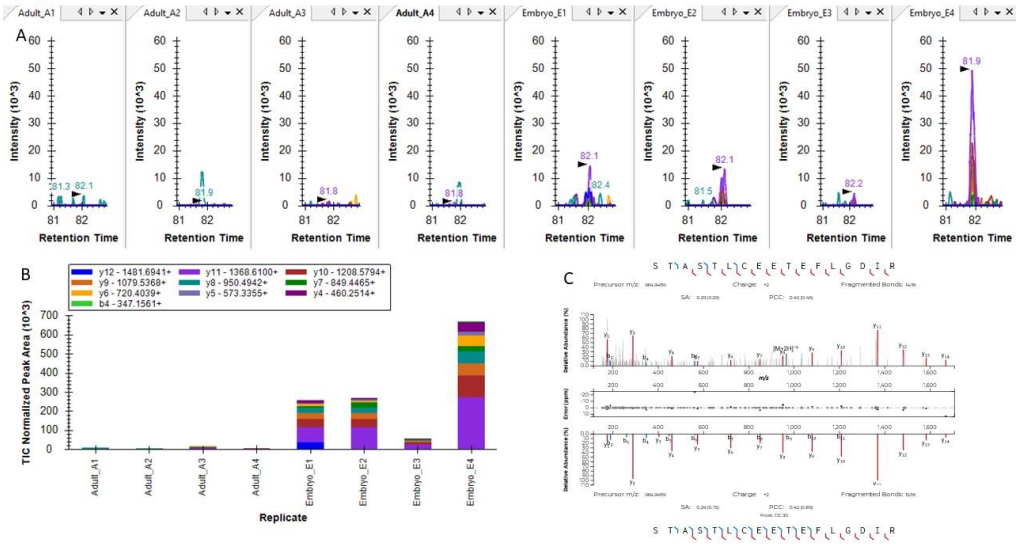
IP_1005431, Prdm11, dORF, ACQVFDLAAWPR



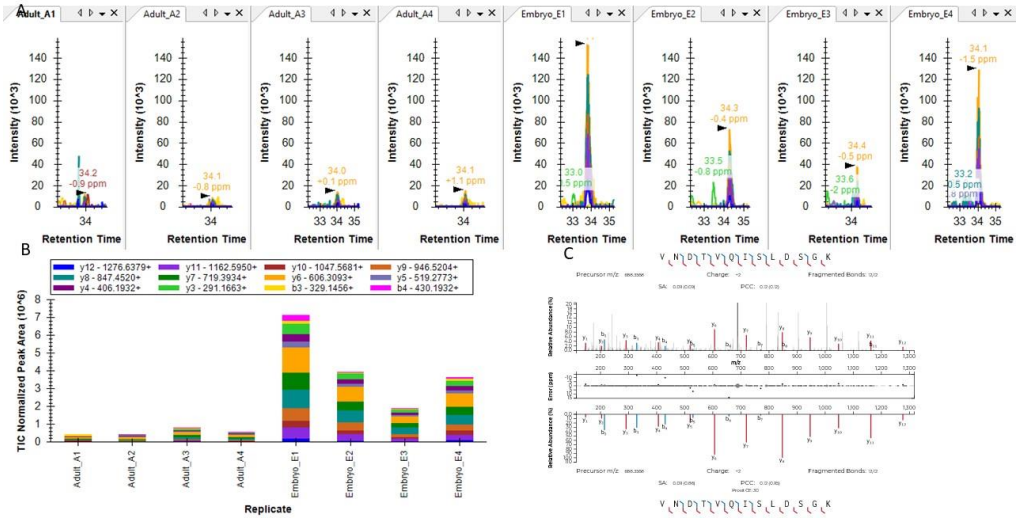
IP_1005431, Prdm11, dORF, SGNSYTLSAEVLSR



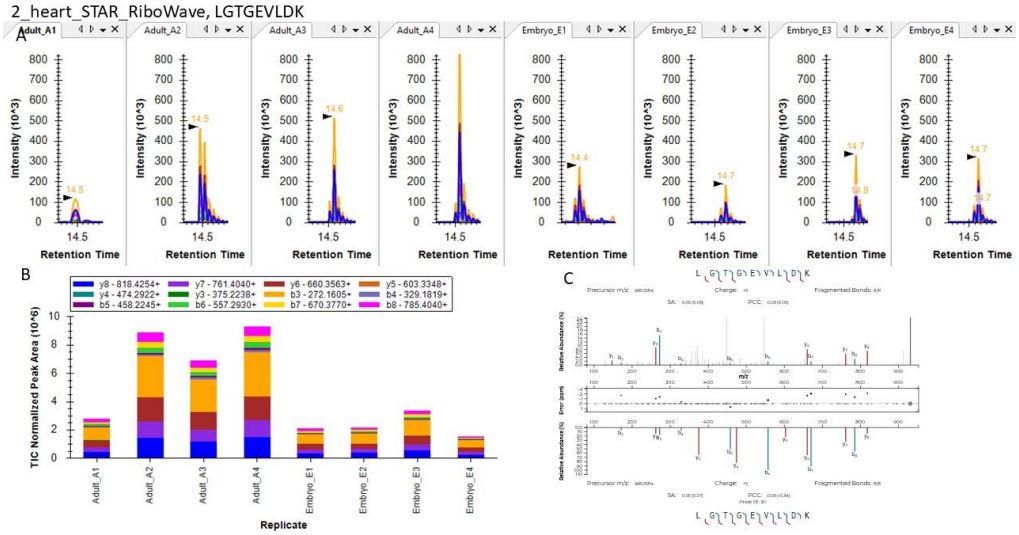
IP_1005431, Prdm11, dORF, STASTLCEETEFLGDIR



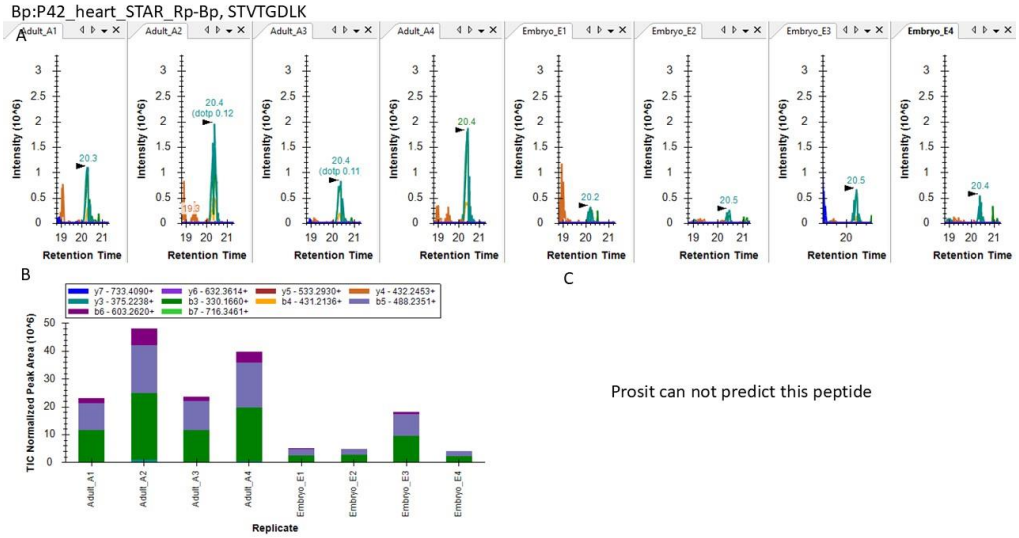
chr6,+,ENSMUSG00000063171.4_148354665_148355498_277,novel,ENSMUST00000071745.3,943,E15_heart_STAR_Ribo-TISH-longest:E15_heart_STAR_RiboWave:P42_heart_STAR_Ribo-TISH-, VNDTVQISLDSGK



chr8,+,ENSMUSG00000109708.1_61051522_61051578_18,novel,ENSMUST00000211661.1,3894,P42_heart_STAR_riboHMM:P42_heart_STAR_RiboWave, LGTGEVLDK

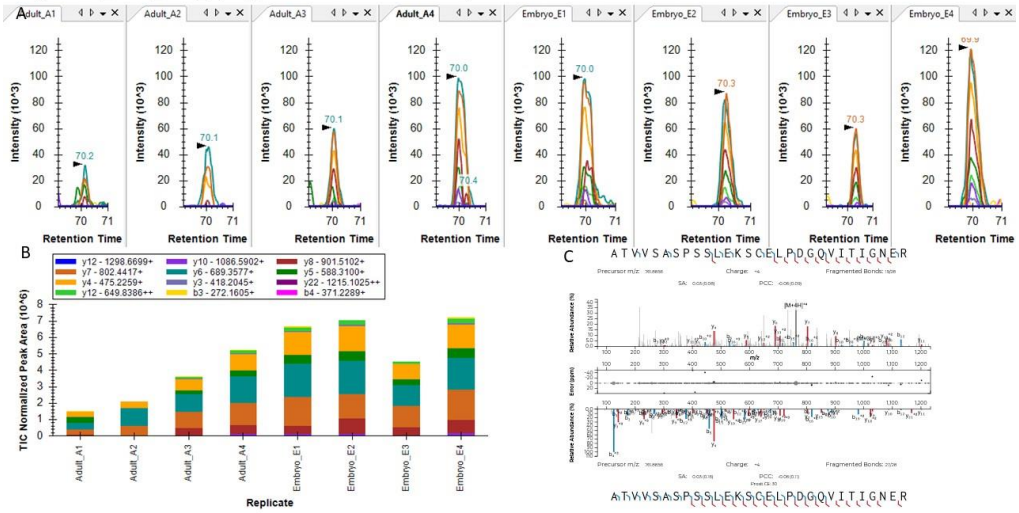


chr2,+,ENSMUSG00000068686.12_104069854_104069889_11,uORF,ENSMUST00000111131.8,555,E15_heart_STAR_Rp-Bp:P42_heart_STAR_Rp-Bp, STVTGDLK

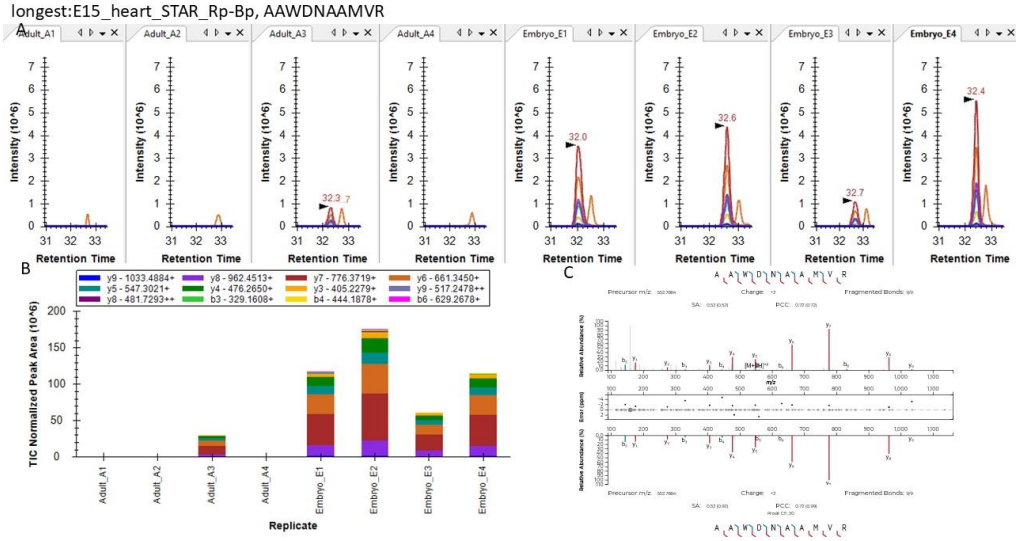


Prosit can not predict this peptide

chr11,-,ENSMUSG00000083859.1_12937289_12936834_151,novel,ENSMUST00000118518.1,1105,E15_heart_STAR_Ribo-TISH-longest:E15_heart_STAR_Rp-Bp, ATVVSASPSSLEKSCELPDGQVITIGNER



chr11,-,ENSMUSG00000083859.1_12937289_12936834_151,novel,ENSMUST00000118518.1,1105,E15_heart_STAR_Ribo-TISH-longest:E15_heart_STAR_Rp-Bp, AAWDNAAMVR



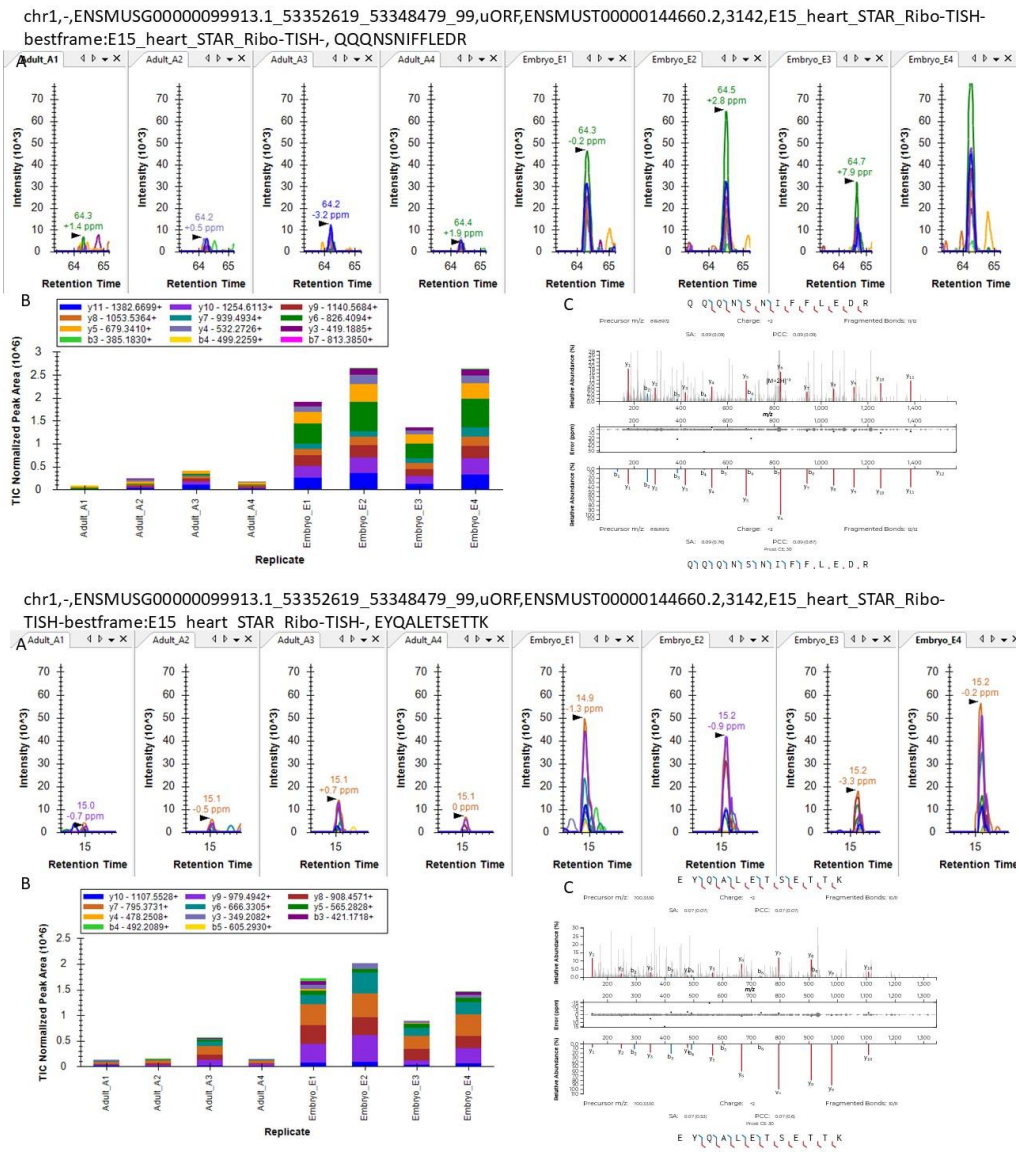


Figure 2-14 Experimental spectrum validation using PRM and corresponding Prosit predicted spectrum. (A) Chromatographic peak profile of peptide from AltProt protein detected by PRM. (B) Intensity plot of fragmented ions of AltProt peptides after TIC normalization in mass spectrometry. (C) Comparative plot of experimental spectrum by PRM and predicted spectrum by Prosit.

Table 2-1 Protein lists of different expressed AltProts in mouse heart development.

Protein ID	sORF type	DEP	Gene name	FC	Pvalue	type	PRM
IP_985030	uORF	Up	Noct	1.4089	0.00863	Openprot	Pass
IP_985759	lncRNA	Down	Ash1l	-2.756	0.00234	Openprot	Pass
IP_841162	dORF	Up	Xirp1	2.98	0.00297	Openprot	Pass
SPROMMU221938	nest_ORF	Down	Zfp518b	-2.917	0.0008	SmProt	Pass
IP_872056	ncRNA	Down	Gm44916	-2.275	0.00281	Openprot	Pass
IP_2502222	ncRNA	Down	Gm42260	-2.266	0.00193	Openprot	Pass
IP_3442236	ncRNA	Down	Gm10277	-2.076	0.00452	Openprot	Pass
eichorn_3t3_2014:1527058	unmapped	Down	Col4a1	-4.801	1.10E-06	sORF.org	Pass

IP_2417366	uORF	Down	Drd4	-2.744	0.00088	Openprot	Pass
IP_863696	nest_ORF	Up	Galnt7	1.6355	0.00079	Openprot	Pass
IP_935570	lncRNA	Up	Gbp6	1.6165	0.00512	Openprot	Pass
IP_3441380	lncRNA	Down	Gm6209	-1.871	0.00846	Openprot	Pass
IP_859576	dORF	Down	Car7	-3.036	8.60E-05	Openprot	Pass
IP_3489484	dORF	Down	Olfr1039	-2.416	0.0006	Openprot	Pass
eichorn_3t3_2014:26228	uORF	Up	Edem3-203	2.2612	0.00613	sORF.org	Pass
IP_899230	lncRNA	Up	Rps4l	3.0316	0.00019	Openprot	Pass
II_1039537	uORF	Up	Asnsd1	NA	NA	Openprot	Pass
IP_3454985	lncRNA	Up	Gm42047	NA	NA	Openprot	Pass

II_3451328	unmapped	Up	Nedd4l	NA	NA	Openprot	Pass
you_2015:1166022	uORF	Up	mest	NA	NA	sORF.org	Pass
SPROMMU147363	lncRNA	Up	Gm5948	NA	NA	SmProt	Pass
SPROMMU227939	uoORF	Down	Lpl	NA	NA	SmProt	Pass
IP_1005431	dORF	Up	Prdm11	NA	NA	Openprot	Pass

2.4 Conclusions

We have innovatively and systematically assessed the performance of Data-Dependent Acquisition (DDA) and Data-Independent Acquisition (DIA) in the identification of both canonical and AltProt peptides. This was achieved through comprehensive comparisons across diverse samples, various mass spectrometry data type databases, library construction methodologies, and mass spectrometry analysis software. Our findings underscored the obvious superiority of DIA over DDA in both qualitative and quantitative dimensions. Particularly in the identification of AltProt peptides, the DIA method outperformed DDA by a factor of two, while also halving the number of missing values. This suggests that the DIA method holds significant potential for enhancing the identification of both canonical and AltProt peptides.

However, our study also has revealed that the choice of library construction method exerts a substantial influence on the final identification results. Different library construction methods can lead to variations in peptide identification, especially in the case of AltProt peptides. Notably, AltProt peptides derived from the experimental fraction library construction method were found to have a higher false discovery rate compared to those obtained through the fully predicted library method.

In our application of DIA to a mouse heart development model, we identified nearly 50 differentially expressed AltProts through rigorous filtering. Following multiple validation techniques, we discovered ASDURF and three uORFs, three of which have

the potential to regulate the host gene. Furthermore, we provided the first evidence of ASDURF in the mouse proteome, thereby substantiating the existence of AltProts and hinting at their potentially significant role in the proteome.

We therefore posit that our comparative analysis of previous methods and practical biological applications can serve as a valuable guide for selecting DIA analysis methods and refining spectral library, paving the way for future exploratory studies on AltProts.

2.5 Limitations

In this article, we conducted a systematic evaluation of detecting AltProts using mass spectrometry techniques, including DDA, DIA, library construction methods, and databases. In this study, we did not adopt AltProt enrichment strategies. We believe that current enrichment strategies, such as SDS-PAGE in-gel digestion63, multiple-protease64, and multidimensional separation65, should have some enrichment effects. However, our focus in this article was primarily on mass spectrometry DDA/DIA mode and library construction methods. Different AltProt enrichment methods should be evaluated in the future for seamless integration with DIA. Regarding databases, we collected as many potential AltProt entries as possible from public data. While the number of potential AltProts has increased, so has the prevalence of false positives in database. Currently, we do not have more effective methods to reduce library redundancy. Although DIA-NN can perform two-step searches to reduce spectral library redundancy, a high-quality AltProt library can theoretically better assist in

AltProt identification, which is an area we should explore further. Furthermore, regarding various library construction methodologies, our data indicated that the experimental library method based on DDA fractions may exhibit a higher rate of false positives, a conclusion derived from comparisons with the Prosit-predicted spectral library. Although the synthesis of peptides for PRM validation of AltProt spectra could more effectively confirm the existence of AltProts, we did not implement this approach in the present study. We acknowledge that conducting PRM experiments with several hundred synthesized peptides presents significant challenges.

Lastly, regarding the potential functional AltProts identified in mouse heart development, many AltProts require further detailed validation. For example, confocal microscopy experiments are needed to determine the localization of AltProts. For the uORF of LPL, Edem3, and mest, subsequent co-immunoprecipitation experiments should be conducted to assess whether there is indeed a physical interaction between the uORF and the main ORF.

Chapter 3. A Pseudo-DIA library search approach improves immunopeptidomics and neoantigen discovery

3.1 Introduction

Immunopeptides, which are small peptide fragments derived from protein degradation through proteasomal processing, play a critical role in the immune response.[100-102] These peptides, particularly those presented by Major Histocompatibility Complex (MHC) molecules, serve as essential signals for T-cell recognition, thereby influencing adaptive immunity[103-105]. The vast diversity of immunopeptides, estimated to reach 167 million for sequences of 8-15 amino acids, poses significant challenges for their identification and characterization in immunopeptidomics studies. Unlike conventional tryptic peptides, which typically yield identification numbers in the range of millions, immunopeptides exhibit a complexity nearly two orders of magnitude greater.[106] This complexity arises not only from the sheer number of potential peptides but also from the intricacies of post-translational modifications (PTMs), which can further diversify peptide sequences.

Current methodologies for immunopeptide identification predominantly rely on data-dependent acquisition (DDA) mass spectrometry coupled with traditional database search engines such as MaxQuant[107], PD, Mascot, Comet[108], and MSFragger.[109] Although these techniques have served as the backbone of proteomic analysis, they often fail to effectively capture the full spectrum of immunopeptides. DDA methods can introduce biases, primarily focusing on the most abundant ions while neglecting

low-abundance peptides. [110, 111] As a result, the full diversity of the immunopeptide landscape may remain insufficiently characterized in these studies.

Data-independent acquisition (DIA)[112] is a promising alternative method. DIA strategies enable the simultaneous acquisition of all ions within a specific mass range, thereby enhancing the likelihood of detecting low-abundance peptides. This method not only improves identification rates but also broadens the quantification range, reduces missing values, and enhances data stability.[110, 111] Despite these advantages, the application of DIA for immunopeptide identification remains fraught with challenges, particularly because of the expansive landscape of potential immunopeptides. Constructing a comprehensive spectral library to accommodate this vast array is a daunting task because it may result in significant redundancy and further complicate the analysis.[113]

To address these challenges, we propose a novel Pseudo-DIA Library Search Strategy tailored specifically for the identification of immunopeptides. Our approach began with an unrestricted DIA database search, enabling the generation of a comprehensive list of potential immunopeptides without predefined false discovery rate (FDR) thresholds. This initial step significantly reduced the peptide space to a more manageable scale, allowing subsequent analysis. By incorporating the predicted spectra and retention times, we constructed a spectral library that facilitates the efficient and accurate identification of immunopeptides during DIA analysis.

We validated our Pseudo-DIA Library Search Strategy using publicly available datasets, demonstrating its effectiveness in identifying immunopeptides across different cell types, including JY, 0D5P, and RA957 cells. Our results indicate that our approach achieved up to 3.8 times more immunopeptide identification than traditional methods, underscoring its potential to enhance detection rates while maintaining high specificity. Importantly, the identification of cryptic immunopeptides and neoantigens, peptides generated from small open reading frames (sORFs) within non-canonical reading frames, further highlights the significance of our method.[114] These neoantigens are crucial for personalized immunotherapy and offer new avenues for targeted cancer treatments.[115]

Moreover, we developed an executable program for our Pseudo-DIA Library Search Strategy, designed to operate in a Windows environment. This tool simplifies the process of generating spectral libraries directly from MSFragger[30] results, thus streamlining workflows and making the method more accessible to researchers. Through this integration, we aimed to enhance the overall efficiency of immunopeptide identification and facilitate the broader application of our strategy in proteomic research.

In summary, the Pseudo-DIA Library Search Strategy represents a significant advancement in the identification of immunopeptides and addresses the complexity and challenges inherent to their detection. Our approach contributes to ongoing efforts in immunology and cancer therapy by improving identification rates and expanding the

scope of detectable peptides, paving the way for more effective therapeutic strategies in the future.

3.2 Materials and methods

3.2.1 Cell Culture

HCT116, SW1573, and H358 cell lines were cultured in RPMI-1640 medium (Gibco, Thermo Fisher Scientific, USA) supplemented with 10% fetal bovine serum (FBS) (Gibco) and 1% penicillin-streptomycin (Gibco). The cells were maintained at 37°C in a humidified incubator with 5% CO_2. The culture medium was refreshed every 2–3 days, and the cells were subcultured upon reaching 70–80% confluency using 0.25% trypsin-EDTA solution (Gibco) for detachment.

3.2.2 MHC-I Immunopeptide Enrichment and Purification

Immunopeptides bound to MHC-I complexes were enriched and purified following a previously described protocol with minor modifications.[107] Briefly, W6/32 MHC-I-specific antibodies were cross-linked to Protein A-sepharose beads using a 40 mM dimethyl pimelimidate (DMP) solution in 200 mM triethanolamine (pH 8.3–9.0). The reaction was quenched with 100 mM ethanolamine (pH 8.3–9.0), and the antibody-conjugated beads were stored in 1× PBS containing 0.02% NaN_3 at 4 °C until use. For peptide isolation, 1–5 × 10^8 cells were lysed with lysis buffer containing 0.5% sodium deoxycholate, 2.0% octyl-β-D-glucopyranoside, and protease inhibitors. The lysates were clarified by centrifugation at 17,000 × g for 50 min at 4 °C. The supernatant was incubated with antibody-conjugated beads for co-immunoprecipitation of MHC-I

complexes. After sequential washes with buffers containing increasing concentrations of NaCl (150 mM, 400 mM) and Tris-HCl (20 mM, pH 8.0), the MHC-I complexes were eluted with 1% trifluoroacetic acid (TFA). The eluted peptides were further purified using a Sep-Pak C18 cartridge, sequentially washed with 0.1% trifluoroacetic acid (TFA), and eluted with 28% acetonitrile (ACN) containing 0.1% TFA. The purified peptides were dried by vacuum centrifugation at 30 °C and resuspended in 10 μL of 2% ACN with 0.1% TFA for downstream mass spectrometry analysis.

3.2.3 Mass Spectrometry Data Acquisition

Peptide samples were analyzed using a Thermo Scientific UltiMate 3000 nano-LC system coupled to a Thermo Scientific Orbitrap Exploris 480 Mass Spectrometer. Chromatographic separation was performed on an IonOpticks Aurora Ultimate 25 cm C18 column at a column temperature of 50°C. Both data-dependent acquisition (DDA) and data-independent acquisition (DIA) modes were performed with a 2-hour LC gradient using 0.1% formic acid (FA) in water (solvent A) and 80% acetonitrile (ACN) with 0.1% FA (solvent B). The flow rate was set to 300 nL/min with the following gradient: 0 min, 5% B; 3 min, 8% B; 80 min, 20% B; 110 min, 90% B; 115 min, 90% B; 116 min, 5% B; 120 min, 5% B.

For DDA, MS1 spectra were acquired at a resolution of 120,000, with an automatic gain control (AGC) target of 300%, and a maximum injection time (MIT) set to auto. MS2 spectra were acquired at a resolution of 30,000 using an isolation window of 1.6 m/z, a maximum injection time of 64 ms, and an AGC target of 100%. Fragmentation

was performed using a higher-energy collisional dissociation (HCD) energy of 30%. For DIA, MS1 spectra were acquired at a resolution of 60,000 with an AGC target of 300%. MS2 spectra were acquired at a resolution of 30,000 using 50 isolation windows across the mass range, with an AGC target of 1000% and a maximum injection time set to auto. Fragmentation was performed at an HCD energy of 33%.

3.2.4 Mass Spectrometry Data Search

Raw mass spectrometry data were analyzed using FragPipe (version 20) and DIA-NN (versions 1.8–2.0). For FragPipe, the MS1 mass tolerance was set to ±20 ppm, and the MS2 mass tolerance was set to 0.02 Da. The peptide length was restricted to 8–15 amino acids. Following the initial search, MSBooster was employed to enhance the scores by incorporating the predicted spectral and retention time information. Percolator was then used to rescore peptide-spectrum matches (PSMs) and control the false discovery rate (FDR). For DIA-NN, the analysis was performed using a spectral library generated by the software. The peptide length was restricted to 8–15 amino acids. FDR thresholds were set at 1% for both protein and precursor levels. All identified peptides were subjected to sequential mapping. Peptides that could be matched to canonical proteins were excluded from classification as non-canonical immunopeptides. Additionally, isoleucine (I) and leucine (L) were treated the same, as mass spectrometers could not distinguish these amino acids.

3.2.5 Prediction of Immunopeptide Binding and Immunogenicity

Multiple computational tools were employed to evaluate the properties of the identified

immunopeptides. NetMHCpan[116] (version 4.1) was used to predict the binding affinity of peptides to specific MHC-I alleles. MixMHCp[117] (version 2.1) was used to perform motif clustering and identify sequence patterns characteristic of MHC-I-restricted peptides. PRIME[118] (version 2.0) and BigMHC[119] were used to assess the potential immunogenicity of the identified immunopeptides to elicit immune responses.

3.2.6 Two-Species Pseudo-Target FDR Estimation

The maize and human proteomes were downloaded from the UniProt database (April 2022). A mixed FASTA database of human and maize was constructed and used for an initial Presearch of the DIA data. Identified peptides present in both species were removed to ensure species specificity of the results. Retention times (RT) and spectral predictions were generated to build the spectral library, and DIA data were subsequently searched using DIA-NN without applying an FDR threshold. False discovery rate (FDR) estimation was performed using the diann-rpackage (https://github.com/vdemichev/diann-rpackage), with the FDR calculated based on the proportion of maize-specific peptides in the dataset. The differing protein quantities in the maize and human proteomes were taken into account during the FDR calculation.

3.2.7 RT and Spectrum Prediction Using DIA-NN

The retention time (RT) and spectral prediction of immunopeptides were performed using DIA-NN (version 1.8.2).[38] Additionally, the Prosit[37] tool was used to predict the spectra for further verification of the MS search results.

3.2.8 Public Data Sources

T OpenProt 2.0 Database: Data were downloaded from the OpenProt website (https://www.openprot.org/downloads).[120]

TCGA Transcript Expression Data: Transcript expression data were obtained from the Xenabrowser platform (https://xenabrowser.net/datapages/?dataset=TcgaTargetGtex_rsem_isopct&host=https%3A%2F%2Ftoil.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443).[121]

HCT116 Mutation Data: Mutation information for the HCT116 cell line was downloaded from the COSMIC database (https://cancer.sanger.ac.uk/cosmic/sample/overview?id=2301978).[122]

3.2.9 sORFs Genomic Localization

The ORF mapping process utilized the gencode-riboseqORFs tool (https://github.com/jorruior/gencode-riboseqORFs).[1] Custom modifications were applied to the code to incorporate support for the NCBI annotation data, enhancing the functionality of the original tool.

3.2.10 Mass Spectrometry Spectrum Annotation

The annotation of secondary mass spectrometry spectra was conducted using the Universal Spectrum Explorer (USE) tool (https://www.proteomicsdb.org/use/),[123] which provides a detailed visualization and annotation of spectral data.

3.3 Results

3.3.1 Pseudo-DIA Library Search Strategy for DIA Immunopeptide Identification

Immunopeptides, products of protein degradation through proteasomal processing, are generated in much greater quantities than tryptic peptides. As shown in **Figure 3-1A**, considering 8-15 amino acid sequences[109], the potential number of immunopeptides can reach up to 167 million, while traditional proteomics, focusing on 7-35 amino acid lengths, only yields approximately 1 million peptides or 2 million when considering single enzyme cleavage sites. This indicates that the complexity of the immunopeptide landscape is nearly two orders of magnitude higher than that of tryptic peptides, even without considering post-translational modifications (PTMs). This inherent complexity of immunopeptide generation poses a significant challenge to their identification.

There are two main search strategies for DIA data analysis: spectrum-centric and peptide-centric.[15, 124] An effective approach is to establish a spectral library to assist in DIA data analysis. Two methods can be used to construct such a library: 1) fractionating the sample, performing DDA experiments, and building a sample-specific spectral library. This approach is more sample-specific but requires more biological resources, instrument time and data analysis. 2) Machine learning was used to predict the spectra and retention times by directly constructing a library.[39, 110] Previous studies have shown that this library prediction method can achieve relatively high identification numbers, even for low-abundance proteins and peptides, while maintaining a controllable false-

positive rate. [110] However, this search strategy cannot be directly applied to immunopeptide data because of the vast immunopeptide landscape, which encompasses approximately 167 million potential peptides. Constructing a comprehensive library for such a large peptide space would result in significant redundancy and pose a significant challenge.[106]

To address this, we proposed A Pseudo-DIA Library Search Approach for immunopeptide identification. First, we performed an unrestrained DIA database search to obtain a full list of potential immunopeptides without setting any false discovery rate (FDR) threshold. This reduces the potential immunopeptide space to a few million. We then used the predicted spectra and retention times to construct a library and complete the DIA data analysis **(Figure 3-1C, Figure 3-2).**

We tested and validated our strategy using a publicly available DIA dataset from Pak et al.[113], which included three cell types: JY, 0D5P, and RA957. As shown in Figure 1D, our method identified three times (11,651 vs. 3,785), 2.5 times (16,508 vs. 6,482), and 1.6 times (21,108 vs. 13,322) more immunopeptides than those identified in the original study. Importantly, as illustrated in Figure 1E, our approach captured a significant portion of the identifications made in the previous study. Notably, the original authors acquired DDA data and built sample-specific libraries for each cell type, whereas our method achieved increased identifications without the need for precious sample resources or additional mass spectrometry instrument time, demonstrating the

efficiency of our approach.

A

Peptide counts by two in-silico digestion methods

Omics	Immunopeptdomics	Proteomics	Proteomics
Length	8-15	7-35	7-35
Digestion	non-specific	trypsin	trypsin
Missed cleavage	-	1	0
Peptide count	167,253,581	2,602,016	1,020,972

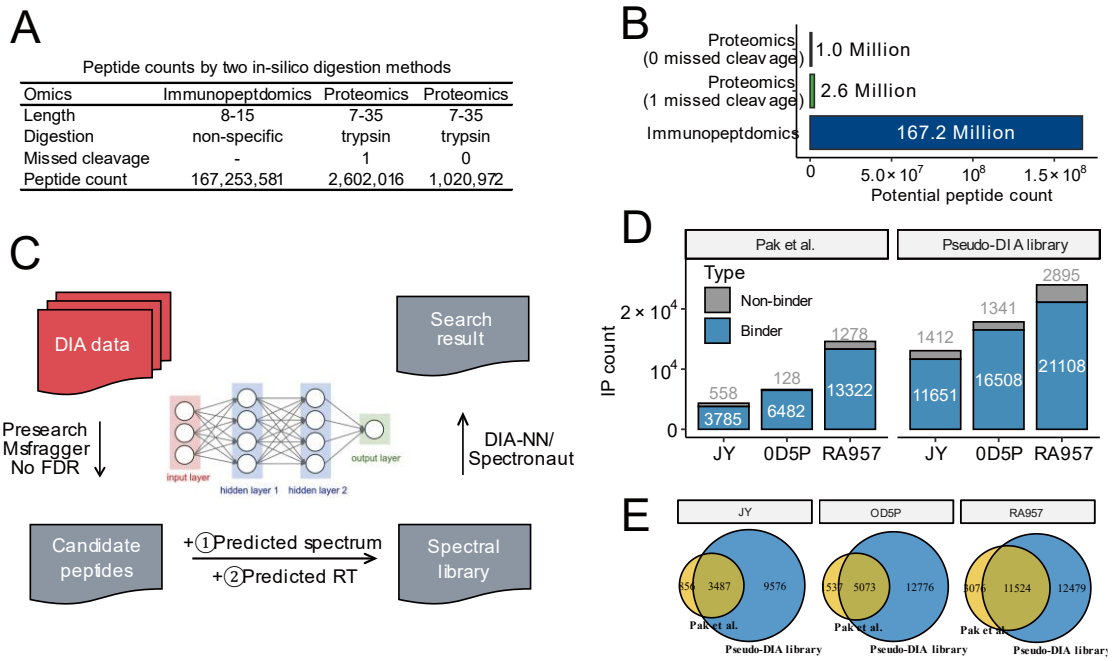


Figure 3-1 Data searching design for Pseudo-DIA library search strategy. (A-B) The bar plot of potential peptide counts from proteomics and immunopeptidomics. (C) The schematic design of Pseudo-DIA library search strategy. (D)The immunopeptide ID number of Pseudo-DIA library search strategy. (E)The immunopeptide overlap between Pak et al. and Pseudo-DIA library search strategy.

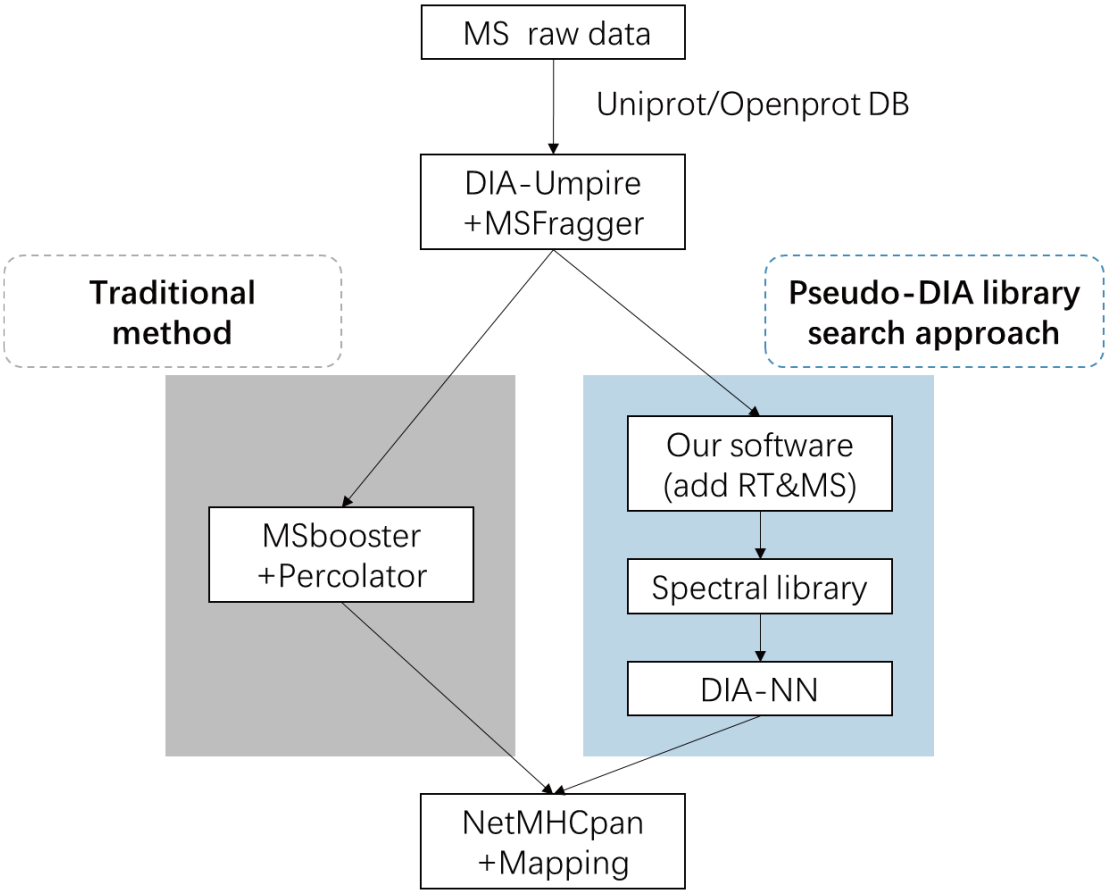


Figure 3-2 Comparison of the Pseudo-DIA and Traditional Library Search Workflows. The commonly used approach for processing DIA immunopeptides involves converting DIA data into pseudo-DDA data using DIA-Umpire, followed by library searching using MSFragger. Subsequently, tools such as MSBooster and Percolator are used to score the spectra. Our Pseudo-DIA Library Search Approach also utilizes DIA-Umpire and MSFragger; however, instead of applying FDR filtering at this stage, we predict the retention times (RT) and spectra for all peptide candidates and construct a library. The library was then passed to DIA-NN for analysis, which output the final FDR-filtered results.

3.3.2 Mechanistic Evaluation of Identification Enhancement and Assessment of Identification Quality

In this section, we investigate the reasons for the increase in immunopeptide identification using the Pseudo-DIA library search strategy. We compared the Q value of the same spectrum analyzed by the first-round search, MSFragger, and second-round search, DIA-NN. As shown in **Figure 3-3A**, for immunopeptide data from three different cell types, one-third of the spectra analyzed by MSFragger had false discovery rates (FDRs) greater than 0.01. However, after the second round of analysis with DIA-NN[38], all spectra with FDRs below 0.01 were identified successfully. Although we do not assert that DIA-NN outperforms MSFragger, given that it is designed for different data acquisition modes (DDA and DIA), we believe that the incorporation of machine-learning-derived spectral information and retention time data significantly improves identification efficiency.

Furthermore, as illustrated in **Figure 3-3B**, we observed a correlation between the scores assigned by MSFragger and DIA-NN for the same spectra. Specifically, lower q-values from MSFragger corresponded to lower q-values from DIA-NN. This suggests that both software tools effectively assess the reliability of spectra, reinforcing their functionality in this context.

We also employed a two-species Pseudo-target FDR estimation method[35] to evaluate the potential false positives in our approach. We analyzed the raw data using our

strategy of mixing maize and human protein data. We assessed false positives in immunopeptide identification based on the frequency of maize peptides. We evaluated the thresholds for both protein and peptide FDRs, as well as for peptide-level FDRs alone. As shown in **Figure 3-3C**, when both the protein and peptide FDRs were set at 1%, approximately 1.5% of the identified immunopeptides were false-positive. This indicates a slightly elevated false-positive rate for our strategy. When the FDR threshold was increased to 3%, the actual FDR fell below 3%, demonstrating that our approach maintained a relatively controllable false-positive rate. However, it is noteworthy that without setting a protein FDR and relying solely on peptide FDR, the false-positive rate for peptides could reach 5%. Although our primary focus is on peptides, controlling the protein FDR is essential when using DIA-NN for immunopeptide identification.

Finally, we compared the motifs of the immunopeptides. As depicted in **Figure 3-3D**, the motifs identified by Pak et al. based on the experimental DDA library were considered reliable. Notably, the motifs identified by Pak et al. and our strategy exhibited similarities, suggesting that reliable results can be obtained using the predicted library, even without using real samples for DDA library construction.

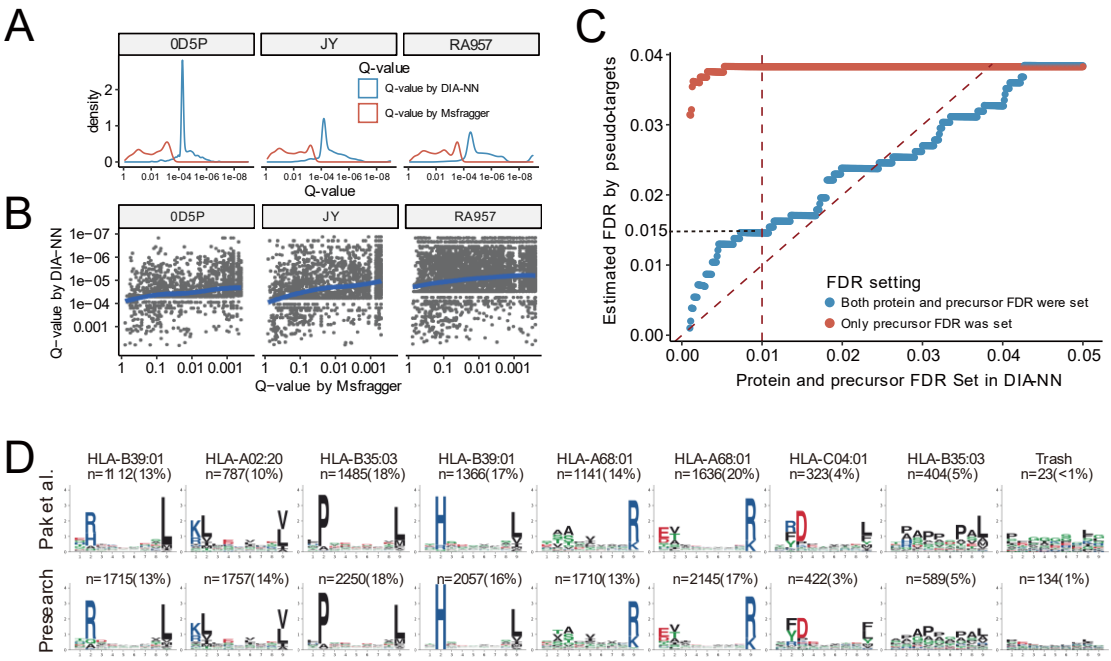


Figure 3-3 Assessing the reliability of Pseudo-DIA library search strategy. The q-value distribution(A) and correlation(B) between first round searching by MSFragger and second round searching with predicted MS information. (C) Applying pseudo-targets searching strategy to validate false discovery rate of Pseudo-DIA method. (D) The Motif comparison of RA957 sample between original searching result by Pak et al. and Pseudo-DIA library search strategy.

3.3.3 Efficient Cryptic Immunopeptide Identification with a Pseudo-DIA Library Search Strategy.

Cryptic immunopeptides are non-canonical immunopeptides generated during the degradation of proteins encoded by small open reading frames (sORFs).[125] These peptides are derived from regions of the genome not typically recognized as coding sequences, which leads to the production of unique immunogenic epitopes.[114, 126] The production of cryptic immunopeptides involves the translation of sORFs under specific cellular conditions, such as stress or particular developmental stages.[127, 128] This process allows for the expression of peptides that may play critical roles in modulating the immune response.[129, 130]

Cryptic immunopeptides are particularly significant because they can function as neoantigens, which are crucial targets for immunotherapy, especially in the context of cancer treatment.[126, 131] Their unique ability to elicit T-cell responses makes them valuable for developing personalized cancer vaccines and other immune-based therapies.[126] Furthermore, cryptic immunopeptides are believed to enhance tumor antigenicity and influence how the immune system recognizes and responds to cancer cells.[114] Understanding and identifying cryptic immunopeptides is essential for advancing therapeutic strategies that harness the immune system to combat malignancies.[132] By exploring the mechanisms underlying their production and their role in immune recognition, researchers can unlock new opportunities for developing innovative cancer immunotherapies.

To identify cryptic immunopeptide identification, the mainstream approach uses data-dependent acquisition (DDA) for mass spectrometry data collection,[126, 132, 133] coupled with traditional search engines such as MaxQuant[29], PD, Mascot, Comet[108], and MSFragger.[30] Although these methods are classic, data-independent acquisition (DIA) is more effective, offering higher identification rates, broader quantification ranges, fewer missing values, and greater stability.[110, 113] However, identifying cryptic immunopeptides from DIA data presents a significant challenge. Given that OpenProt[120] includes approximately 666,123 sORFs, which is approximately 300 times larger than UniProt, the challenge becomes even greater. While UniProt has already generated 167 million immunopeptides, OpenProt could yield a staggering 593 million immunopeptides when considering sequences of 8-15 amino acids. Mapping this vast array of peptides to DIA chromatograms is indeed a substantial challenge.

Our method is well-suited for addressing this scenario, enabling the identification of cryptic non-canonical immunopeptides, as illustrated in **Figure 3-4A**. We tested our Pseudo-DIA library search strategy using the aforementioned public datasets. The database search workflow is shown in **Figure 3-2.** The results showed that for the three cell types, JY, 0D5P, and RA957, our method identified 3.8 times (545/145), 2.7 times (636/232), and 2.2 times (578/268) more peptides, respectively. As shown in **Figure 3-4B**, our method almost completely encompassed the peptides identified using conventional methods. This demonstrates the effectiveness of our approach in

identifying cryptic immunopeptides.

Using the Pseudo-DIA Library Search Strategy, we identified 1,223 non-canonical immunopeptides, which were assigned to 1,083 OpenProt IDs. Genomic localization analysis revealed that approximately one-third of these non-canonical immunopeptides were derived from lncRNA regions **(Fig. 3-4C)**. Among these sORFs, we identified several located around canonical proteins associated with tumorigenesis and cancer progression, such as the uORF of RBM10,[134] the uoORFs of ATF4[135] and SLC19A1, the intORF of ZNF146,[136] as well as sORFs derived from tumor-associated lncRNAs, including DTNB-AS1, CHASERR,[137] and LINC00649 (see Supplementary Table 2).

As shown in **Fig. 3-4D**, we identified an sORF driven by LINC00649, IP_290788, with two unique immunopeptides detected (spectra and chromatograms are shown in **Fig. 3-4F**). Previous studies have shown that LINC00649 promotes cancer progression,[138-141] and consistent with this, our analysis of TCGA data revealed a high expression of LINC00649 across multiple cancer types **(Fig. 3-4E)**. Based on spectral evidence and differential expression profiles, we propose that these non-canonical immunopeptides hold potential for clinical applications in cancer immunotherapy.

Additional examples are presented in **Fig. 3-5, S3-6,** where sORFs derived from tumor-associated genes also produce immunopeptides. These peptides are promising

targets for cancer immunotherapy and may enable the development of personalized cancer vaccines. By incorporating these non-canonical peptides, we can expand the repertoire of tumor antigens, offering new directions for immunotherapy and advancing precision oncology research.

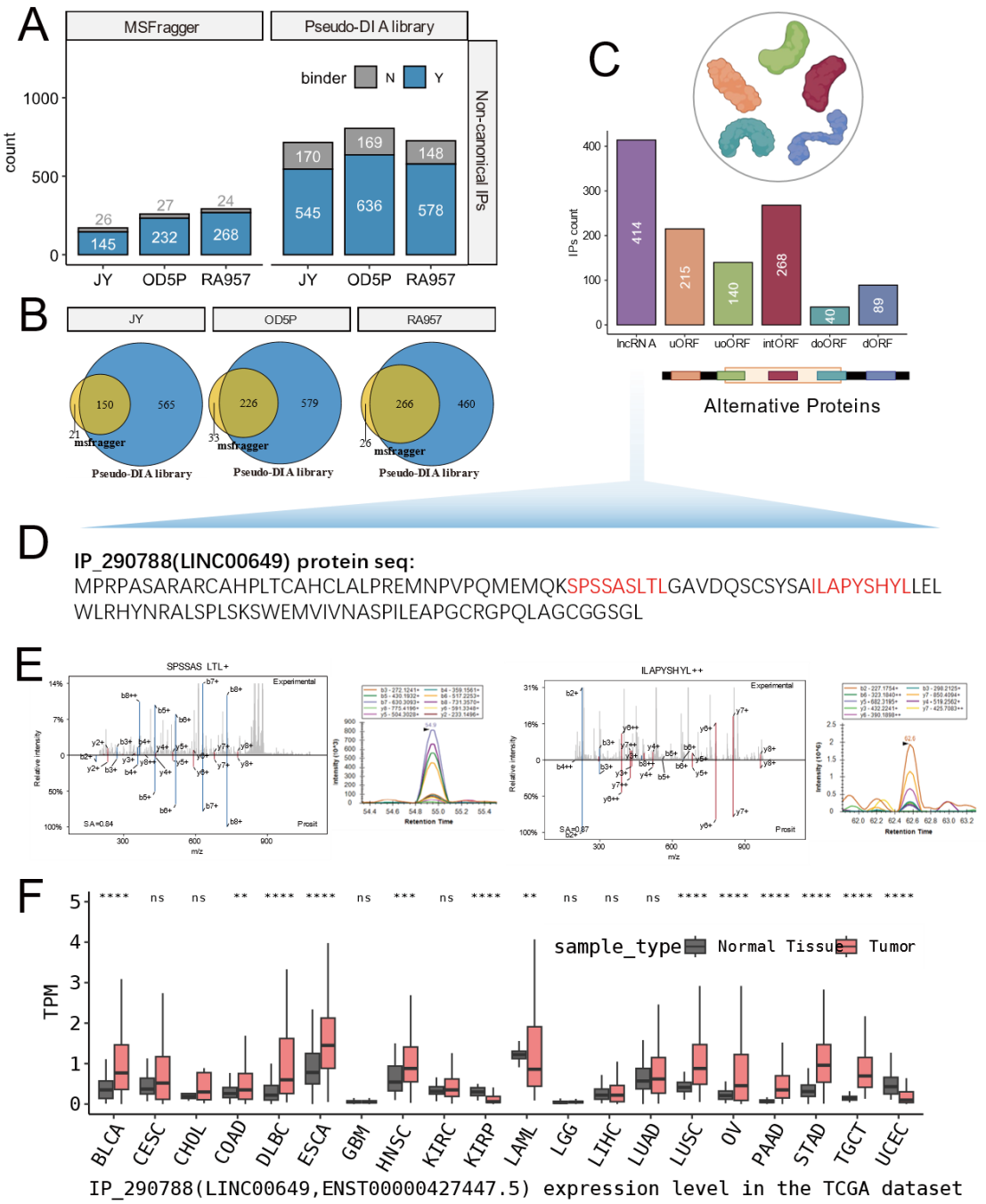


Figure 3-4 Pseudo-DIA strategy enhancing non-canonical immunopeptides identification. (A) The non-canonical immunopeptide id number of Pseudo-DIA library search strategy. (B) The overlap of non-canonical immunopeptides identified by Pak et al. and Our Method. (C) The genomic localization categories of sORFs corresponding to the identified immunopeptides. (D) The protein sequence of LINC00649 showing the positions of immunopeptides; detected immunopeptide

fragments are highlighted in red. (E) The mass spectra and chromatograms of the two identified immunopeptides. (F) The expression levels of the LINC00649 transcript ENST00000427447.5 across various cancer types and normal tissues based on TCGA data.

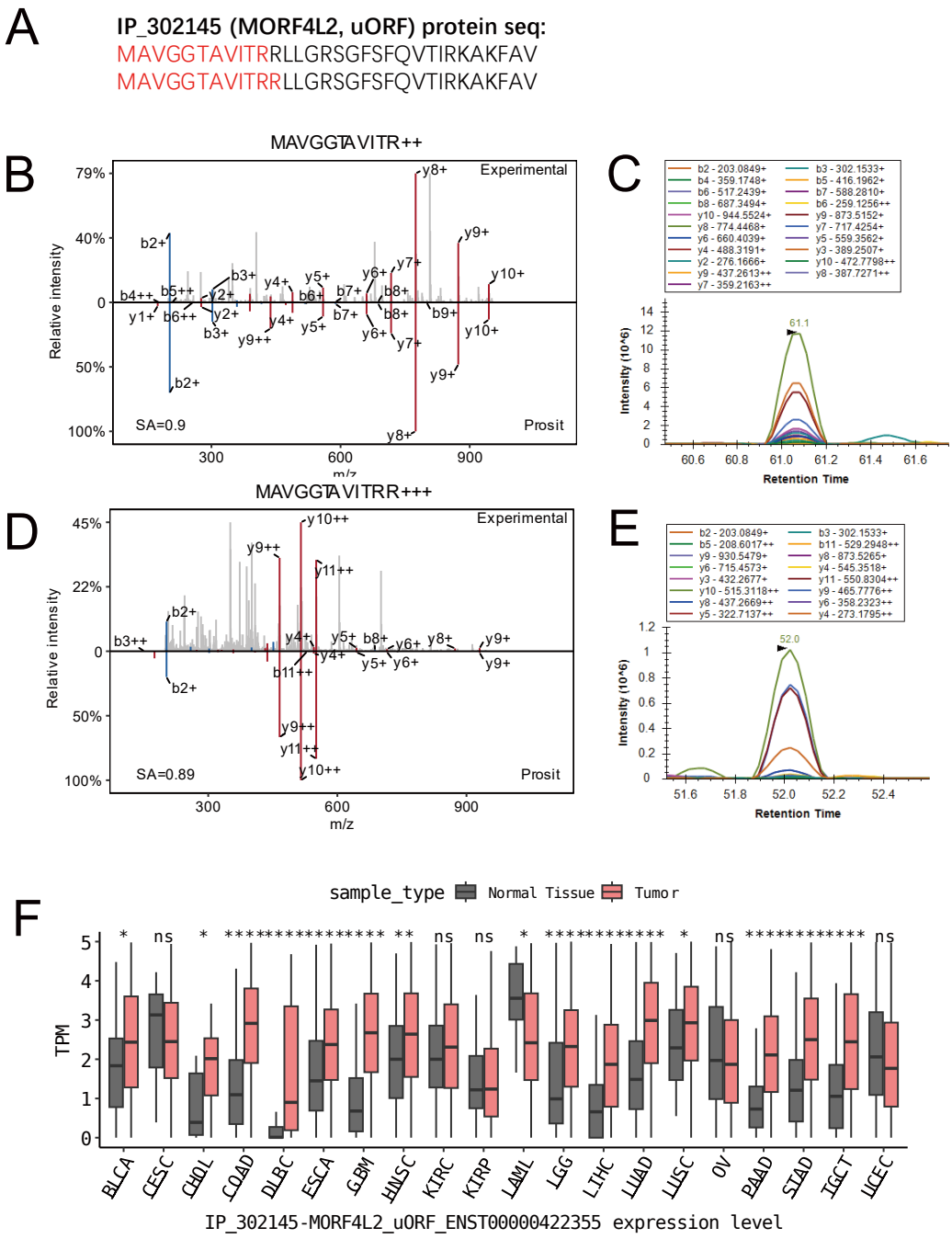
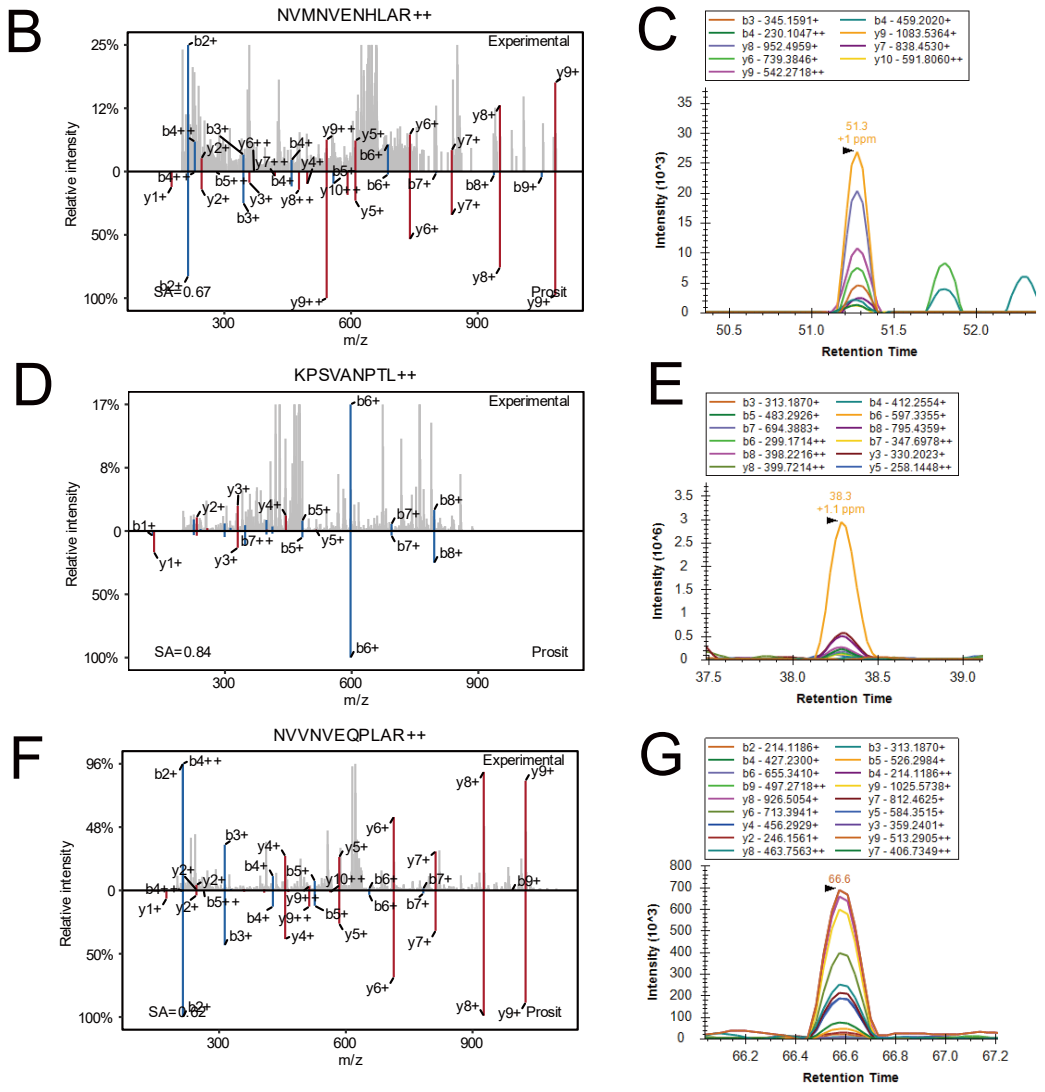


Figure 3-5 Visualization of Immunopeptide Spectrum Quality and TCGA Expression Differences for MORF4L2 uORF Protein (IP_302145). (A) Positions of the two immunopeptides of IP_302145 within the protein. (B-E) Spectral quality and chromatograms of the two immunopeptides, MAVGGTAVITR and MAVGGTAVITRR, from IP_302145. (F) Differential expression levels of the IP_302145 transcript in tumor and normal samples from TCGA RNA expression data.

A **IP_273983 (ZNF146, intORF) protein seq:**

MSIFTRERNLLNVTSVEKPLAKSSMSLNIRTPILARSFSNVMNVENHLARRKTSLRTRKFTLEKNLLSV
KIAGKLSFRSQTSSDTRELTQERSPLYVRSVEKPSVANPTLLSMRKSILERSLLNVVNVEQPLARRSTS



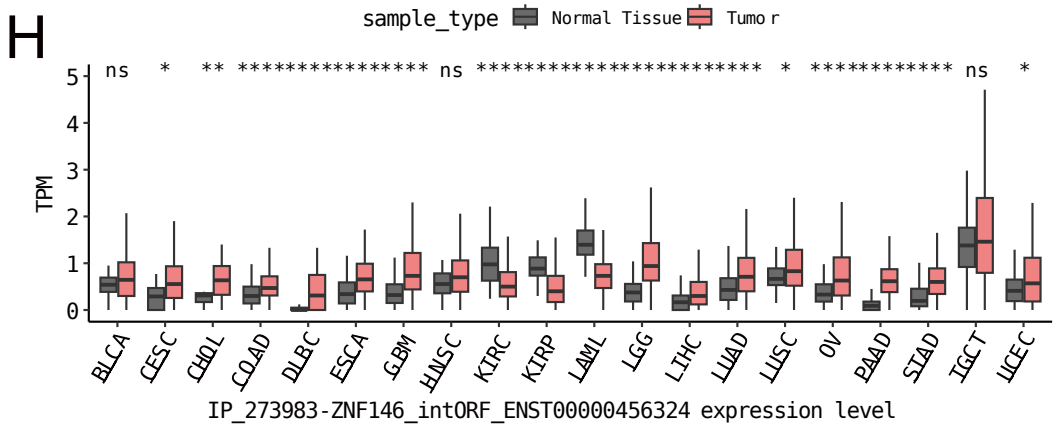


Figure 3-6 Visualization of Immunopeptide Spectrum Quality and TCGA

Expression Differences for ZNF146 intORF Protein (IP_273983). (A) Positions of the two immunopeptides of IP_273983 within the protein. (B-G) Spectral quality and chromatograms of three immunopeptides, NVMNVENHLAR, KPSVANPTL, and NVVNVEQPLAR, from IP_302145. (H) Differential expression levels of the IP_273983 transcript in tumor and normal samples from TCGA RNA expression data.

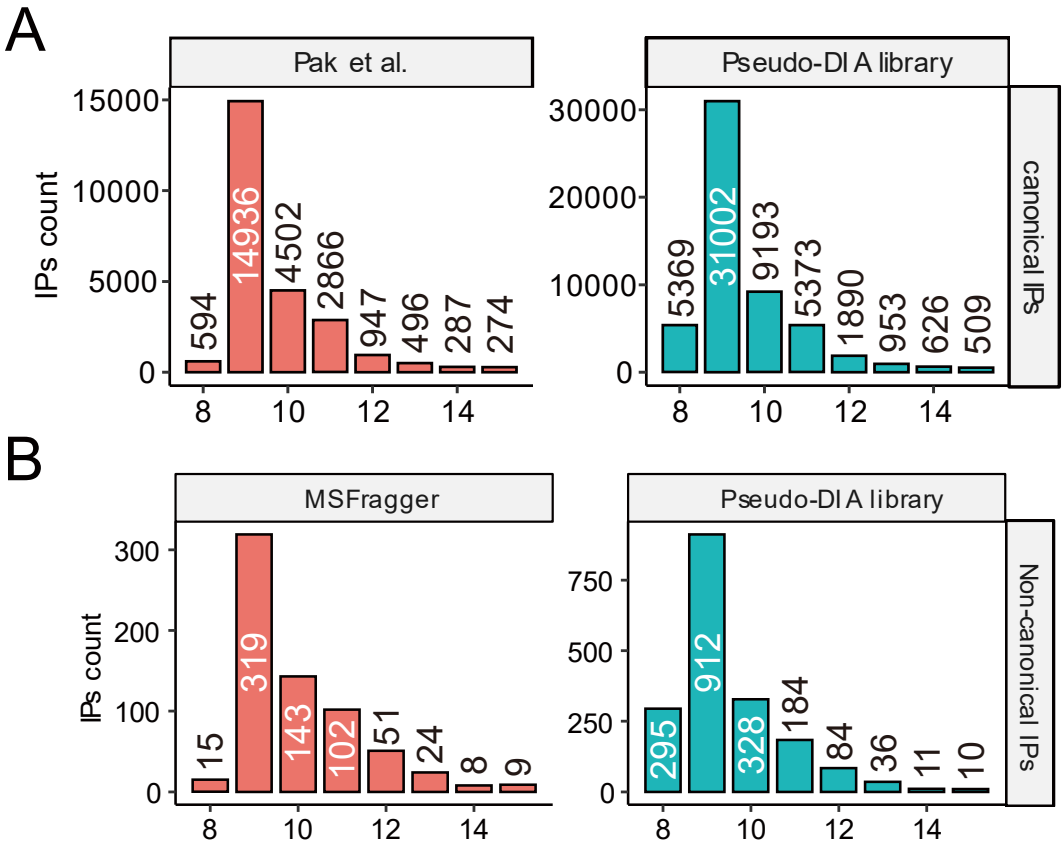


Figure 3-7 Length Distribution of Canonical and Non-canonical Immunopeptides Identified by Various Methods. (A) Comparison of Immunopeptide Length Distributions Identified by Pak et al. and Pseudo-DIA Library Search Approach. (B) Comparison of Non-canonical Immunopeptide Length Distributions Identified by Traditional MSFragger method and Pseudo-DIA Library Search Approach.

3.3.4 Evaluation of the Pseudo-DIA Library Search Strategy Using Independent Datasets

To further evaluate our Pseudo-DIA Library Search Strategy, we collected immunopeptide data from three cell lines: H358, HCT116, and SW1573. We acquired data in both data-dependent acquisition (DDA) and data-independent acquisition (DIA) formats for each cell line. As shown in **Figure 3-8A**, our method achieved the highest identification numbers in this independent dataset, with a significant overlap, as shown in **Figure 3-8B.**

In addition to general immunopeptide identification, we focused on sample-specific mutations known as neoantigens. Neoantigens are generated through mutations in tumor cells, resulting in unique peptide sequences, making them potential targets for immune responses. Using mutation information sourced from the COSMIC[122] database, we derived mutated protein sequences for each cell line. Subsequently, we conducted in silico predictions to identify all potential 8-15 amino acid immunopeptides that contained these mutations **(Fig. 3-8C).** The in-silico mutation sequences were extracted and combined with the peptides generated from our Pseudo-DIA library strategy to create a comprehensive library. Ultimately, we identified approximately 30 neoantigens **(Fig. 3-8D)**. Notably, the DDA approach identified only 13 neoantigens, even though we had already utilized MSbooster to enhance the identification efficiency of DDA searching,[91, 142] demonstrating that our DIA and predictive library strategies are significantly more effective in this context.

Among the identified peptides, we confirmed two derived from the CHMP7 p.A324T mutation, QTDQMVFNTY and QTDQM(ox)VFNTY, with one exhibiting an oxidation modification. Additionally, we identified the neoantigen KVIDIYEQV from the NAPA p.A181V mutation and AENGKLVTNGNPIT from the GAPDH p.I69T mutation, as shown in **Fig. 3-8E**. By comparing our findings with the Prosit-generated spectra and reviewing the chromatograms, we confirmed the reliability of the spectral data. Moreover, interestingly, previous studies have also identified the same neoantigen, further supporting the reliability of the spectra as well as the robustness of our method.[143]

These findings not only validate the sensitivity of our method but also provide valuable resources and insights for the field of immunology. The identification of neoantigens is crucial, as they can serve as biomarkers for personalized cancer immunotherapy, guiding the development of targeted treatments that enhance the immune system's ability to recognize and eliminate tumor cells. This underscores the importance of our Pseudo-DIA Library Search Strategy, which not only improves peptide identification but also contributes significantly to advancing immunotherapeutic strategies in oncology.

We developed an executable program for the Pseudo-DIA Library Search Strategy that operates within a Windows environment **(Figure 3-9)**. This program was designed to

directly read the results generated by MSFragger, facilitating the efficient and user-friendly construction of spectral libraries. By streamlining the process, the executable allows users to seamlessly generate a spectral library that can be directly fed into subsequent DIA search engines for comprehensive data analysis. This integration simplifies the workflow and enhances the overall efficiency of immunopeptide identification, making it accessible to researchers in the field.

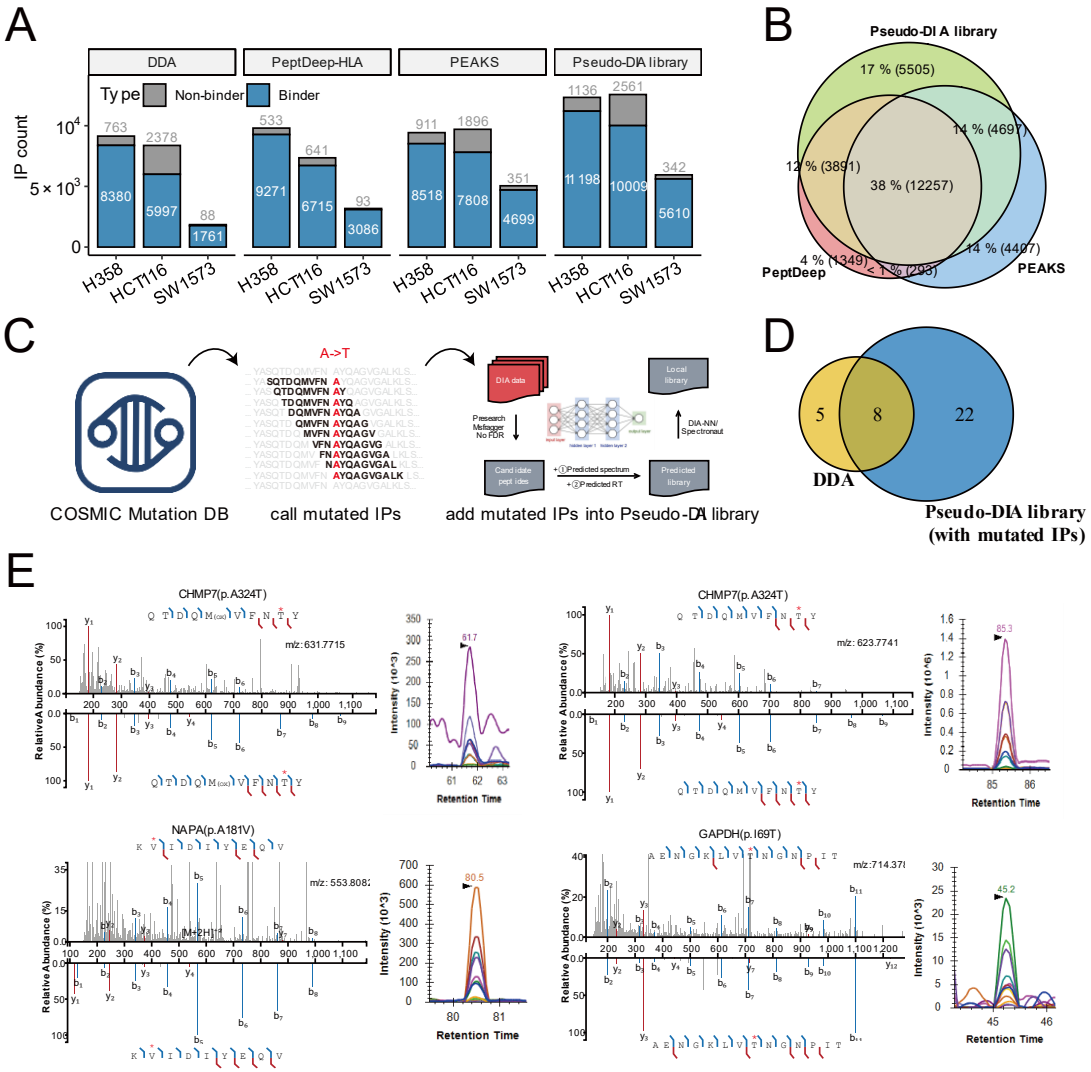


Figure 3-8 Application of Pseudo-DIA library strategy with independent dataset to identify tumor-specific immunopeptides. (A) The immunopeptide id number of

HCT116, H358, SW1573 cell line identified by DDA, PeptDeep-HLA, PEAKS and Pseudo-DIA library search strategy. (B) The Venn diagram of three data searching strategies. (C) The workflow for mutated IP identification. (D) The Venn diagram of mutated immunopeptide identified by DDA and Pseudo-DIA library search strategy. (E) Representative PSM of mutated immunopeptide identified from Pseudo-DIA library search strategy

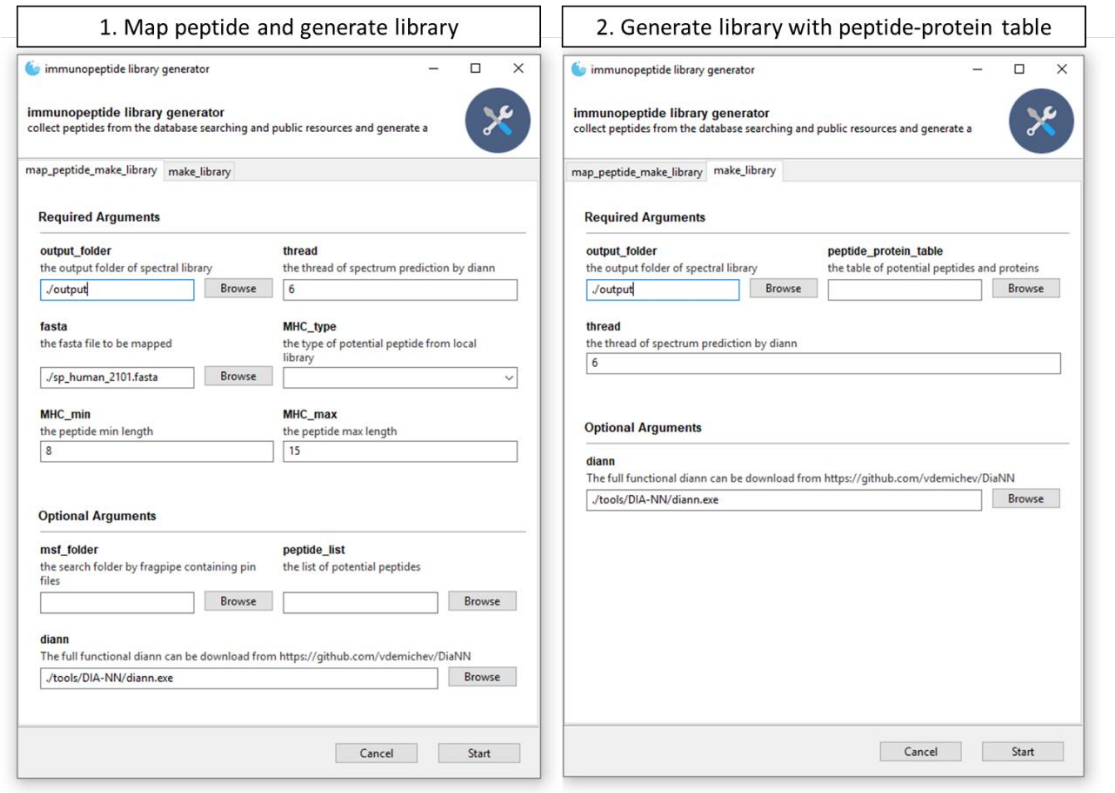
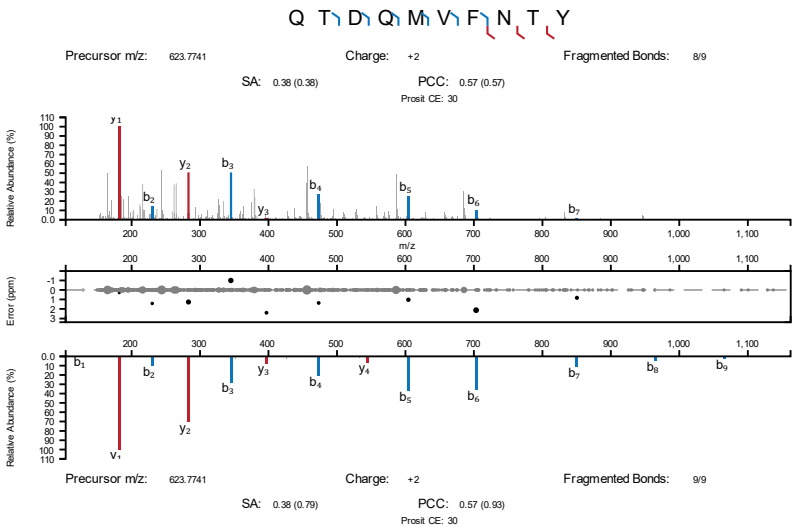
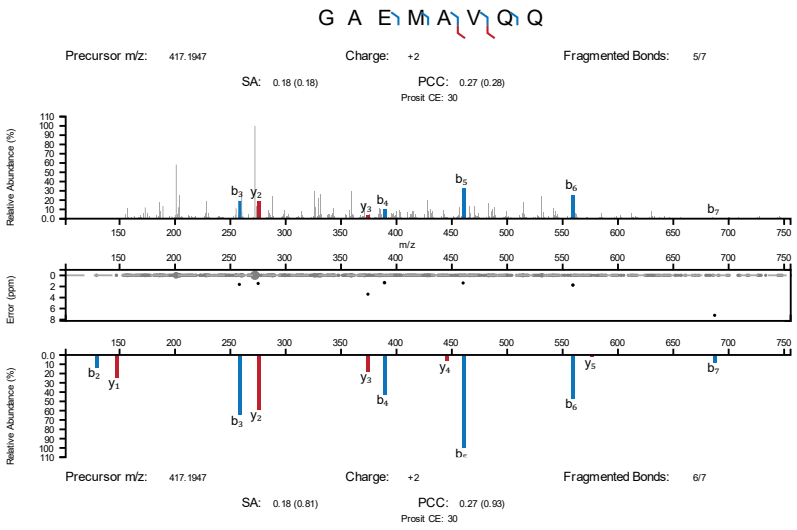
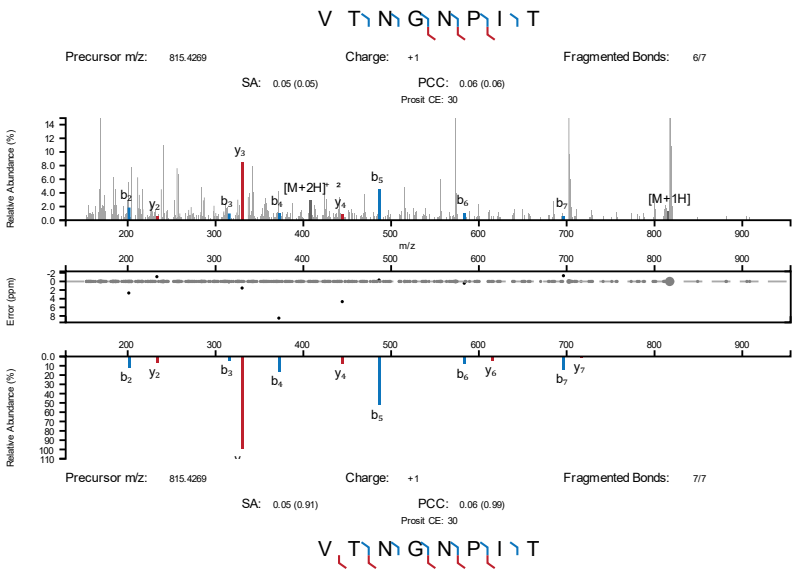
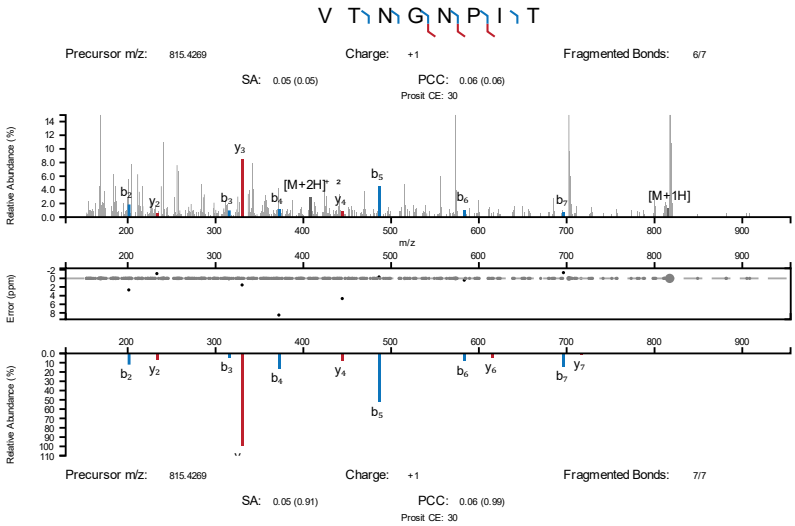


Figure 3-9 Graphical user interface (GUI) for the pseudo-DIA library search strategy.

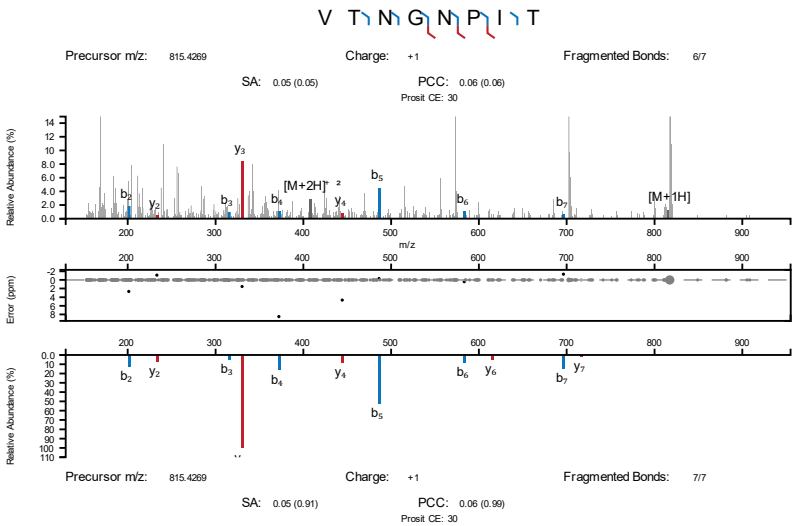

Q T D Q M V F N T Y
Precursor m/z: 623.7741
Charge: +2
Fragmented Bonds: 8/9
SA: 0.38 (0.38)
PCC: 0.57 (0.57)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 623.7741
Charge: +2
Fragmented Bonds: 9/9
SA: 0.38 (0.79)
PCC: 0.57 (0.93)
Prosit CE: 30
Q T D Q M V F N T Y


G A E M A V Q Q
Precursor m/z: 417.1947
Charge: +2
Fragmented Bonds: 5/7
SA: 0.18 (0.18)
PCC: 0.27 (0.28)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 417.1947
Charge: +2
Fragmented Bonds: 6/7
SA: 0.18 (0.81)
PCC: 0.27 (0.93)
Prosit CE: 30
G A E M A V Q Q

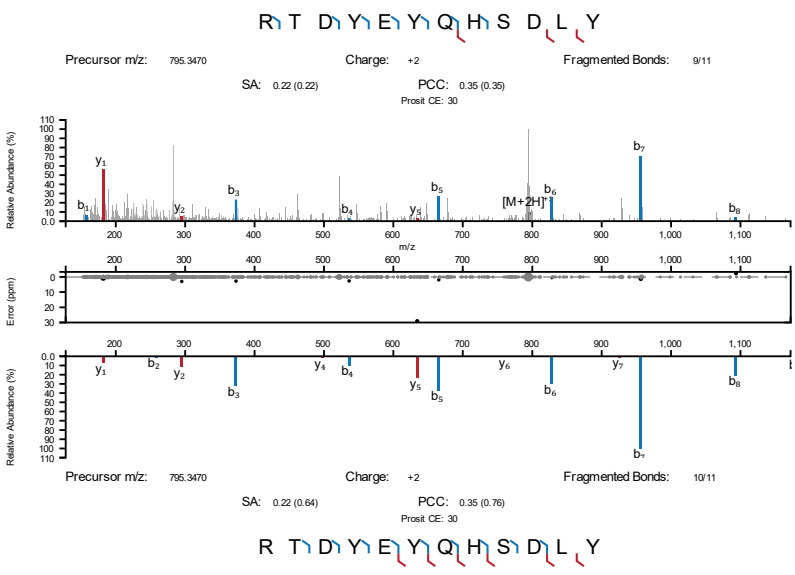

V T N G N P I T
Precursor m/z: 815.4269
Charge: +1
Fragmented Bonds: 6/7
SA: 0.05 (0.05)
PCC: 0.06 (0.06)
Prosit CE: 30
Relative Abundance (%)
[M+2H]+2
[M+1H]
m/z
Error (ppm)
Precursor m/z: 815.4269
Charge: +1
Fragmented Bonds: 7/7
SA: 0.05 (0.91)
PCC: 0.06 (0.99)
Prosit CE: 30
V T N G N P I T



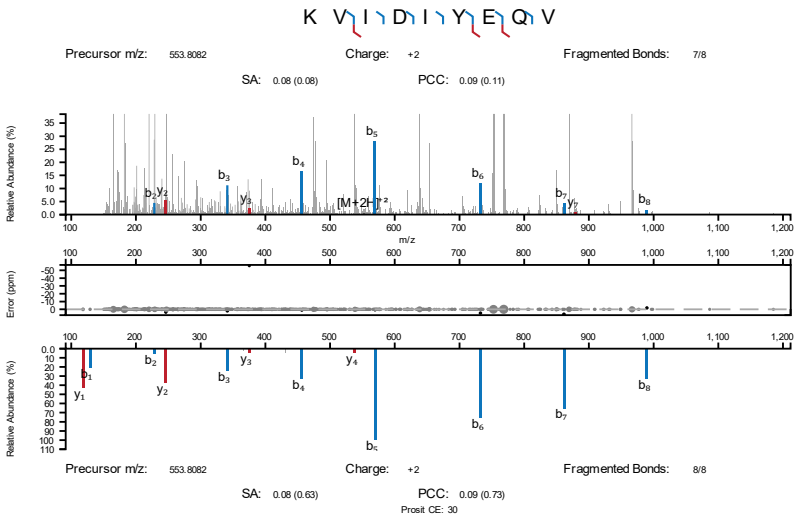
V T N G N P I T

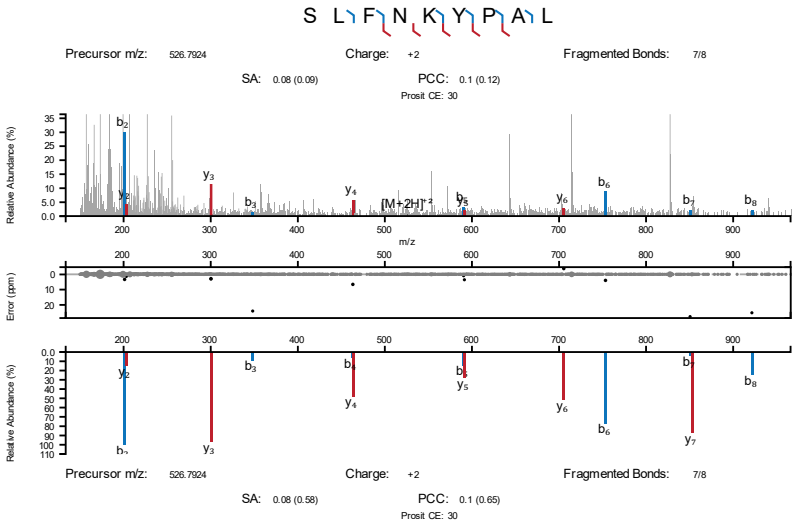


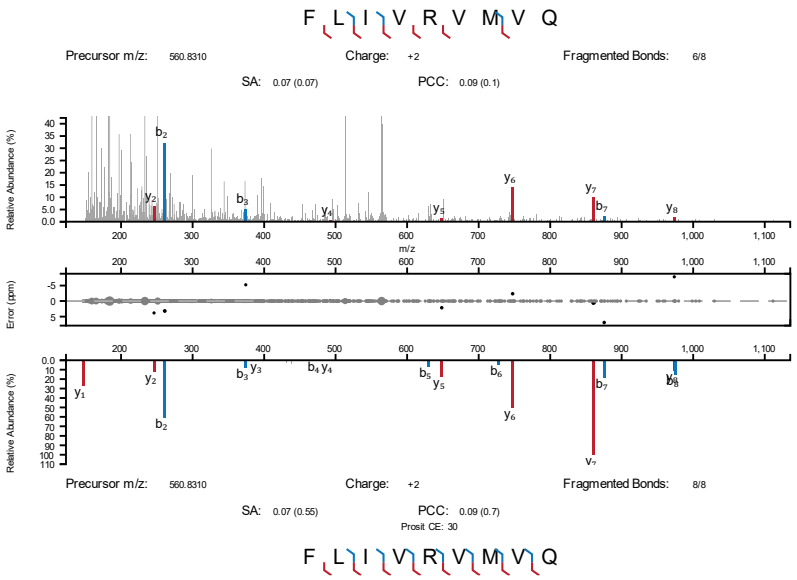
V T N G N P I T

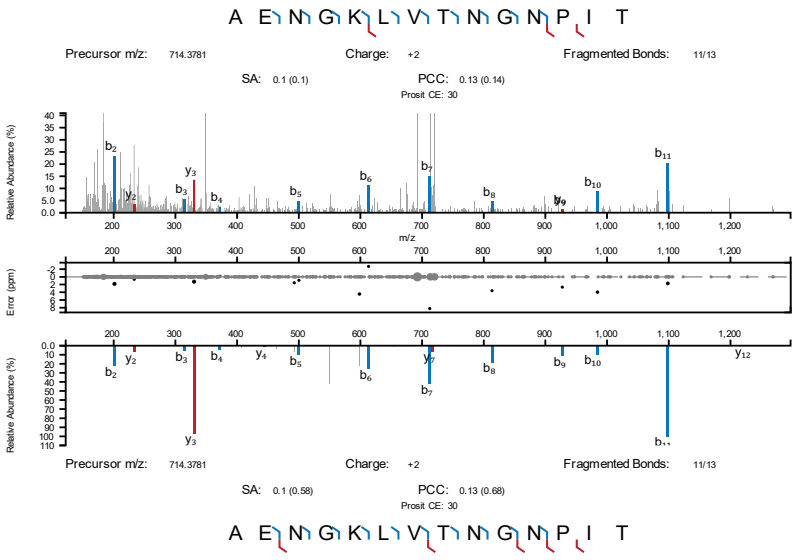


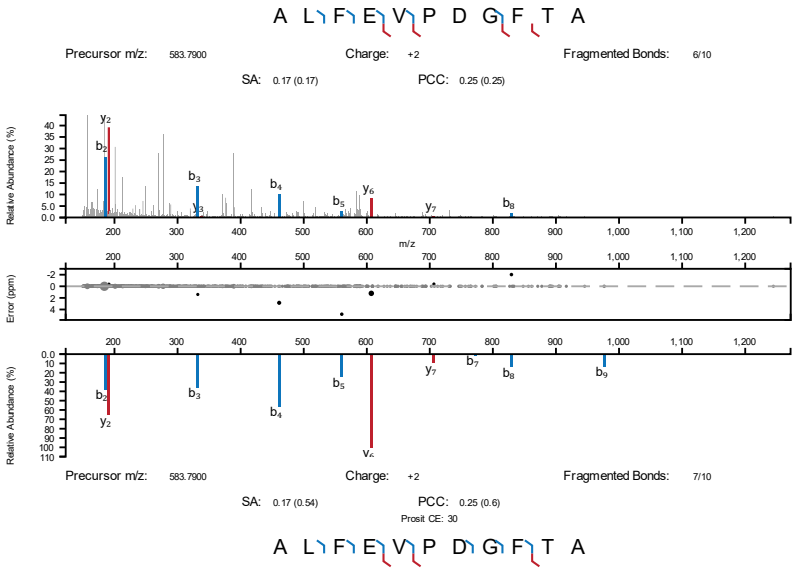
R T D Y E Y Q H S D L Y

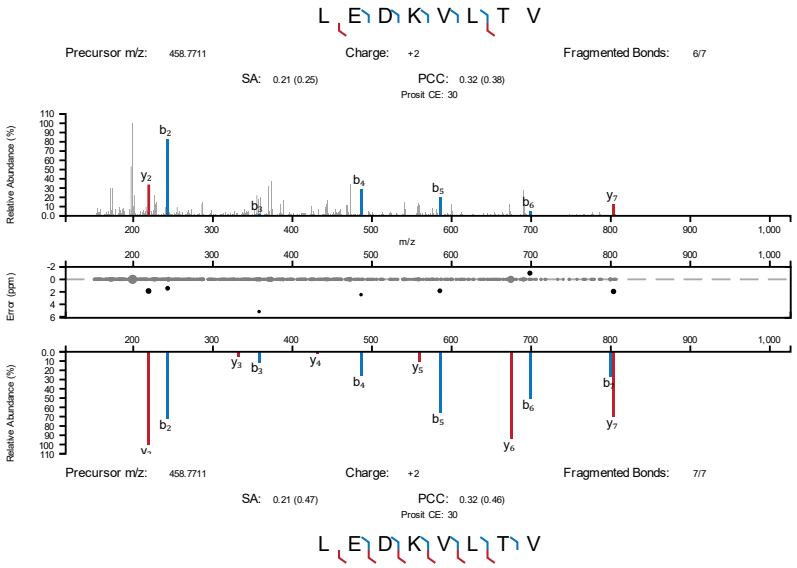

K V I D I Y E Q V
Precursor m/z: 553.8082
Charge: +2
Fragmented Bonds: 7/8
SA: 0.08 (0.08)
PCC: 0.09 (0.11)
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 553.8082
Charge: +2
Fragmented Bonds: 8/8
SA: 0.08 (0.63)
PCC: 0.09 (0.73)
Prosit CE: 30
K V I D I Y E Q V

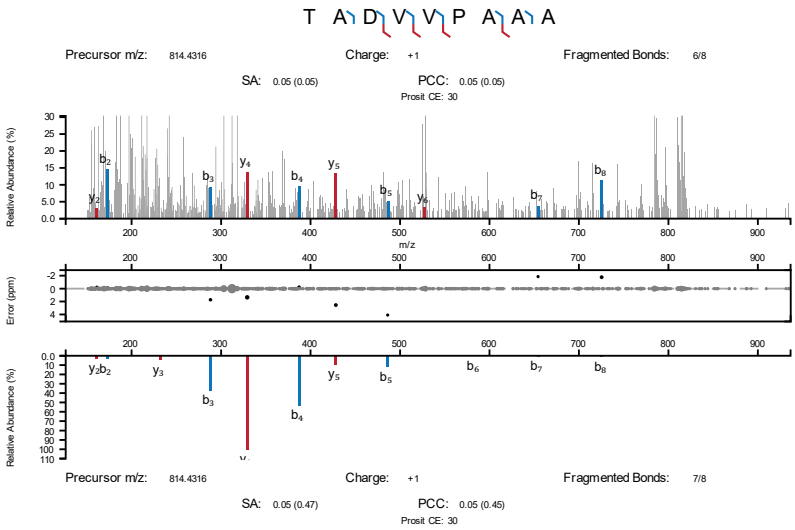

S L F N K Y P A L
Precursor m/z: 526.7924
Charge: +2
Fragmented Bonds: 7/8
SA: 0.08 (0.09)
PCC: 0.1 (0.12)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 526.7924
Charge: +2
Fragmented Bonds: 7/8
SA: 0.08 (0.58)
PCC: 0.1 (0.65)
Prosit CE: 30
S L F N K Y P A L

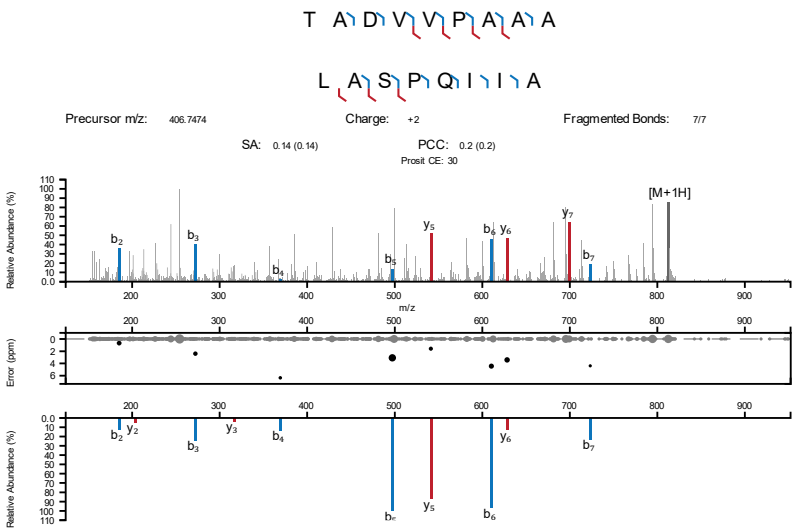

F L I V R V M V Q
Precursor m/z: 560.8310
Charge: +2
Fragmented Bonds: 6/8
SA: 0.07 (0.07)
PCC: 0.09 (0.1)
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 560.8310
Charge: +2
Fragmented Bonds: 8/8
SA: 0.07 (0.55)
PCC: 0.09 (0.7)
Prosit CE: 30
F L I V R V M V Q

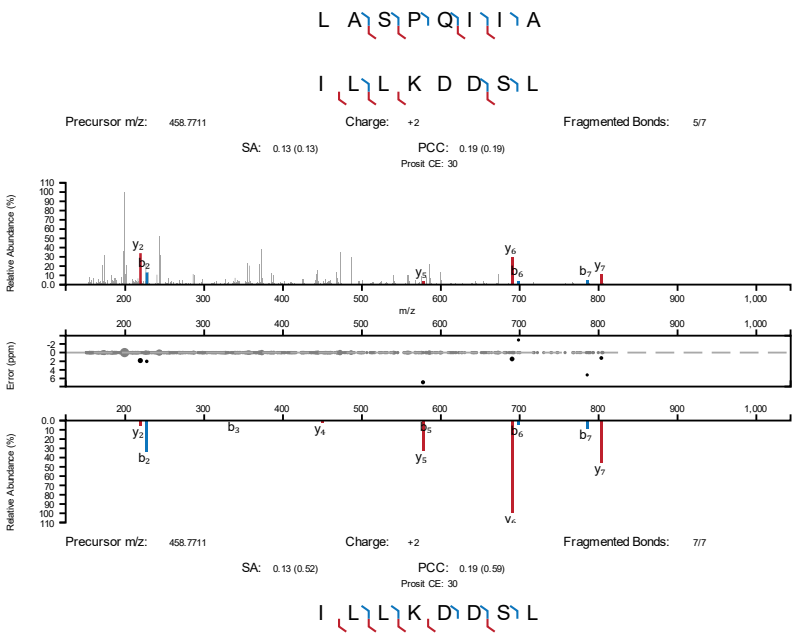

A E N G K L V T N G N P I T
Precursor m/z: 714.3781
Charge: +2
Fragmented Bonds: 11/13
SA: 0.1 (0.1)
PCC: 0.13 (0.14)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
SA: 0.1 (0.58)
PCC: 0.13 (0.68)
A E N G K L V T N G N P I T

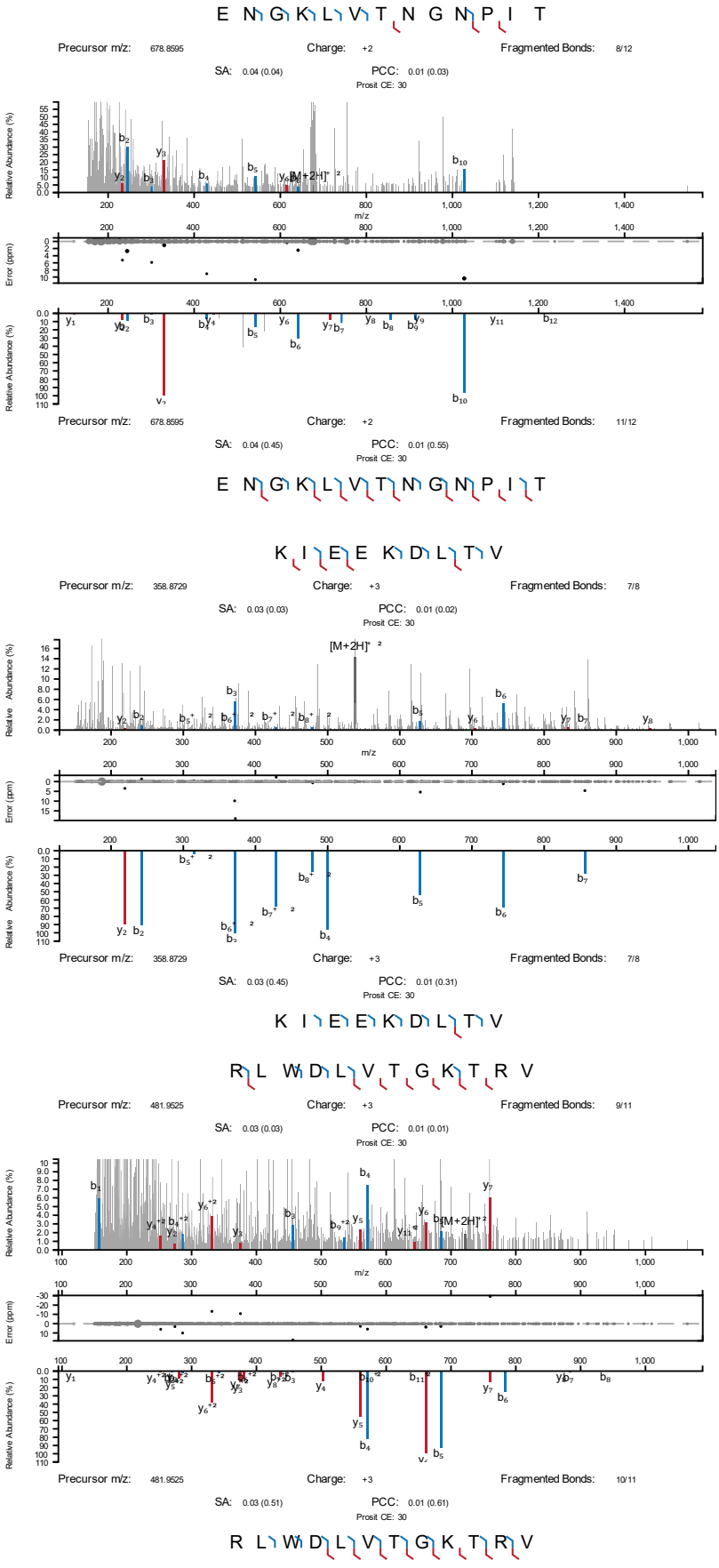

A L F E V P D G F T A
Precursor m/z: 583.7900
Charge: +2
Fragmented Bonds: 6/10
SA: 0.17 (0.17)
PCC: 0.25 (0.25)
Relative Abundance (%)
m/z
Error (ppm)
SA: 0.17 (0.54)
PCC: 0.25 (0.6)
Prosit CE: 30
A L F E V P D G F T A

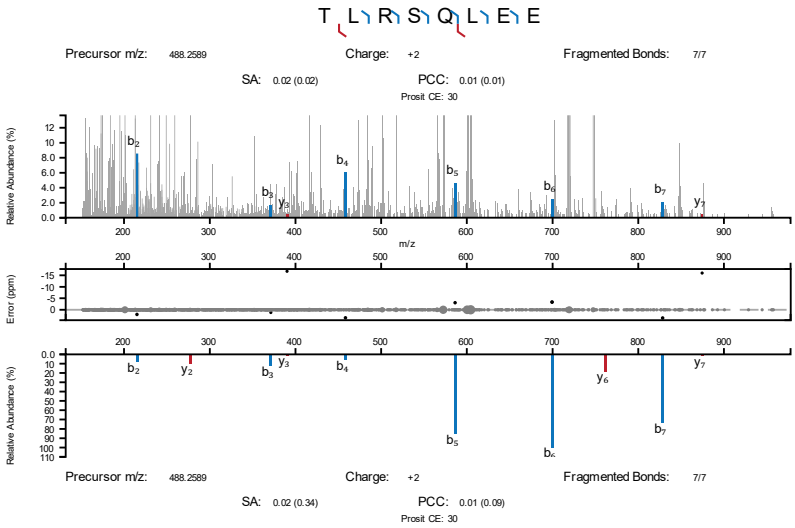

L E D K V L T V
Precursor m/z: 458.7711
Charge: +2
Fragmented Bonds: 6/7
SA: 0.21 (0.25)
PCC: 0.32 (0.38)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
SA: 0.21 (0.47)
PCC: 0.32 (0.46)
L E D K V L T V

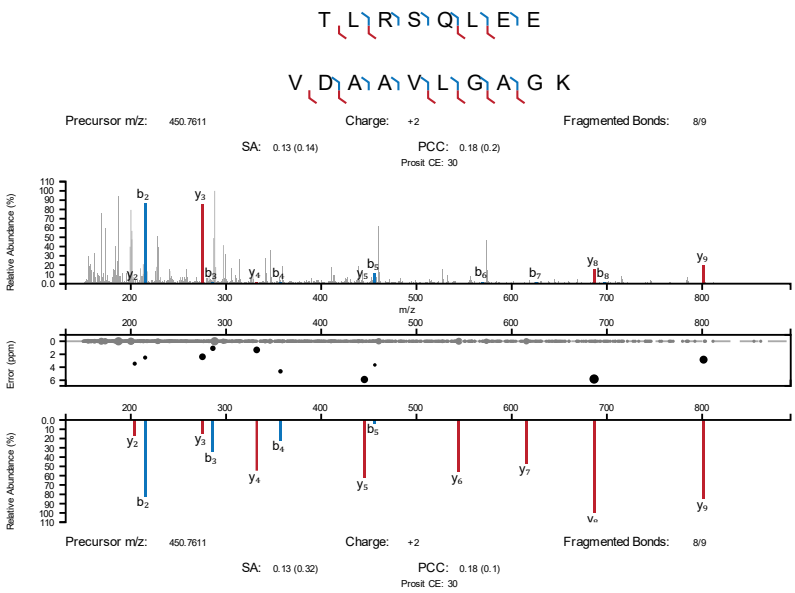

T A D V V P A A A
Precursor m/z: 814.4316
Charge: +1
Fragmented Bonds: 6/8
SA: 0.05 (0.05)
PCC: 0.05 (0.05)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 814.4316
Charge: +1
Fragmented Bonds: 7/8
SA: 0.05 (0.47)
PCC: 0.05 (0.45)
Prosit CE: 30

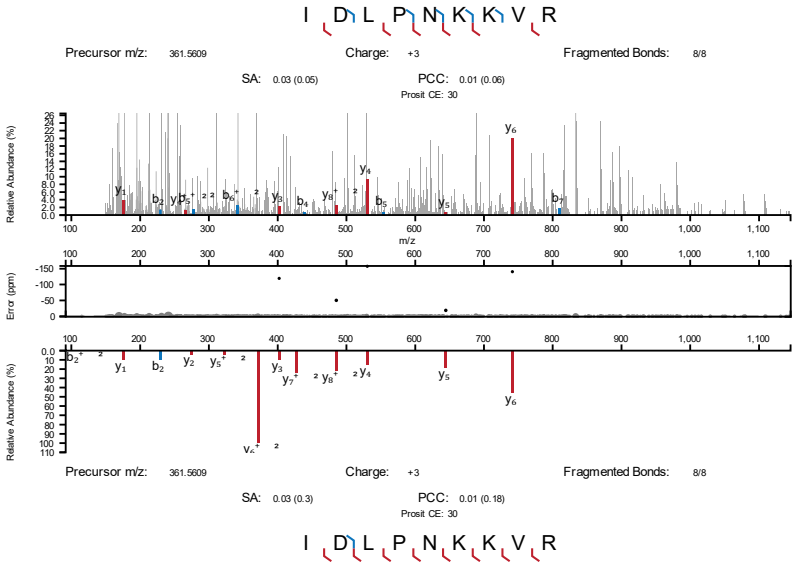

T A D V V P A A A
L A S P Q I I A
Precursor m/z: 406.7474
Charge: +2
Fragmented Bonds: 7/7
SA: 0.14 (0.14)
PCC: 0.2 (0.2)
Prosit CE: 30
[M+1H]
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 406.7474
Charge: +2
Fragmented Bonds: 6/7
SA: 0.14 (0.53)
PCC: 0.2 (0.42)
Prosit CE: 30

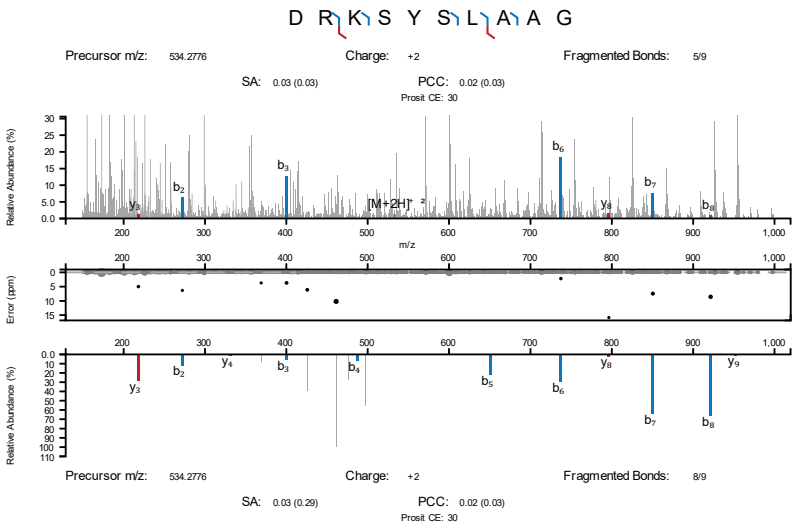

L A S P Q I I A
I L L K D D S L
Precursor m/z: 458.7711
Charge: +2
Fragmented Bonds: 5/7
SA: 0.13 (0.13)
PCC: 0.19 (0.19)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 458.7711
Charge: +2
Fragmented Bonds: 7/7
SA: 0.13 (0.52)
PCC: 0.19 (0.59)
Prosit CE: 30

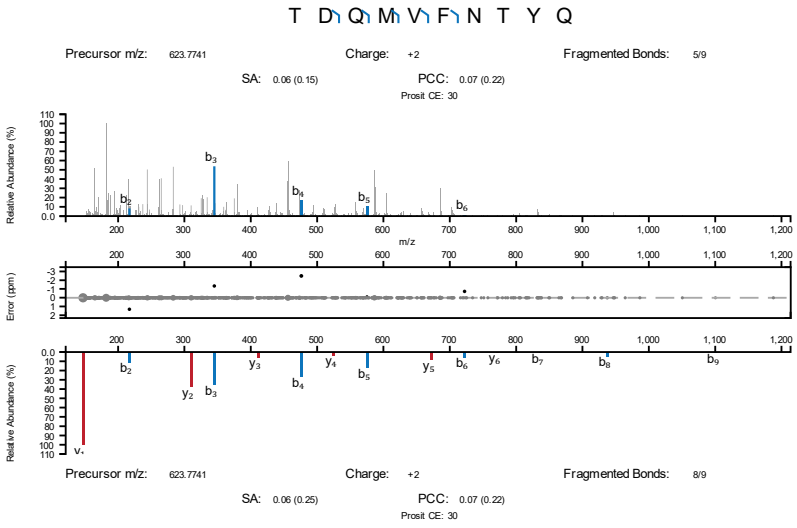

E N G K L V T N G N P I T
Precursor m/z: 678.8595
Charge: +2
Fragmented Bonds: 8/12
SA: 0.04 (0.04)
PCC: 0.01 (0.03)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 678.8595
Charge: +2
Fragmented Bonds: 11/12
SA: 0.04 (0.45)
PCC: 0.01 (0.55)
Prosit CE: 30
E N G K L V T N G N P I T
K I E E K D L T V
Precursor m/z: 358.8729
Charge: +3
Fragmented Bonds: 7/8
SA: 0.03 (0.03)
PCC: 0.01 (0.02)
Prosit CE: 30
Precursor m/z: 358.8729
Charge: +3
Fragmented Bonds: 7/8
SA: 0.03 (0.45)
PCC: 0.01 (0.31)
Prosit CE: 30
K I E E K D L T V
R L W D L V T G K T R V
Precursor m/z: 481.9525
Charge: +3
Fragmented Bonds: 9/11
SA: 0.03 (0.03)
PCC: 0.01 (0.01)
Prosit CE: 30
Precursor m/z: 481.9525
Charge: +3
Fragmented Bonds: 10/11
SA: 0.03 (0.51)
PCC: 0.01 (0.61)
Prosit CE: 30
R L W D L V T G K T R V

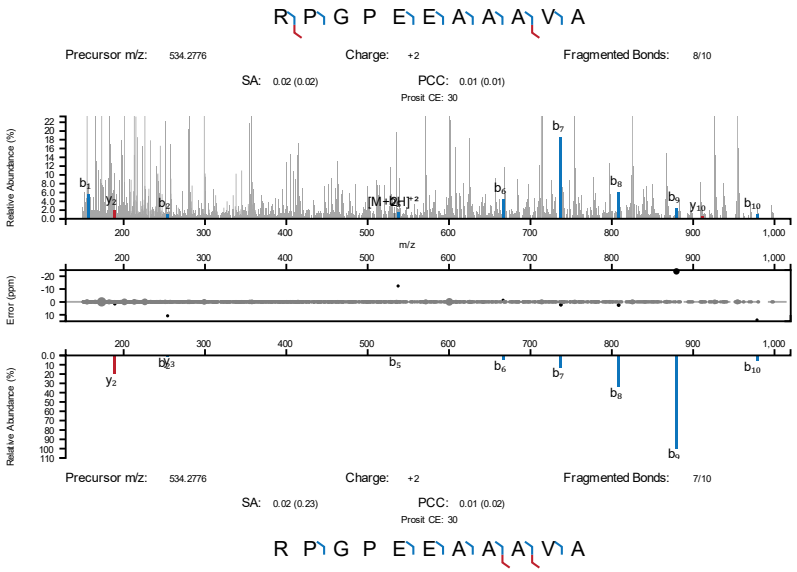

T L R S Q L E E
Precursor m/z: 488.2589
Charge: +2
Fragmented Bonds: 7/7
SA: 0.02 (0.02)
PCC: 0.01 (0.01)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
SA: 0.02 (0.34)
PCC: 0.01 (0.09)

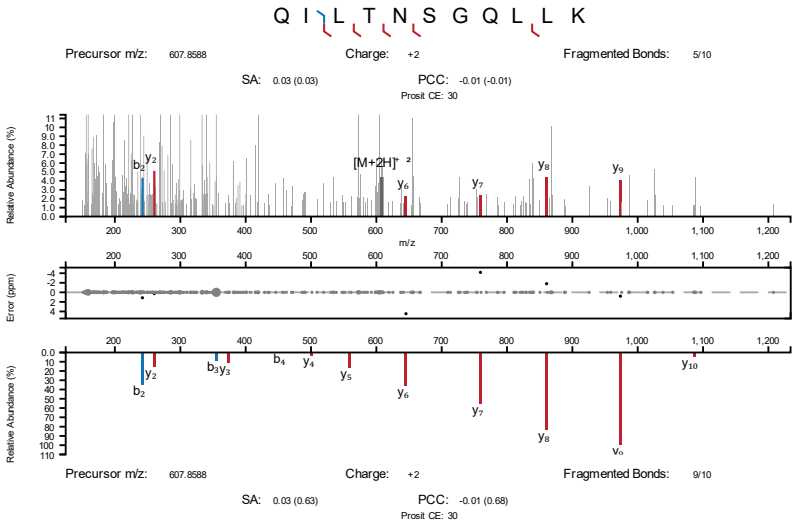

V D A A V L G A G K
Precursor m/z: 450.7611
Charge: +2
Fragmented Bonds: 8/9
SA: 0.13 (0.14)
PCC: 0.18 (0.2)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
SA: 0.13 (0.32)
PCC: 0.18 (0.1)

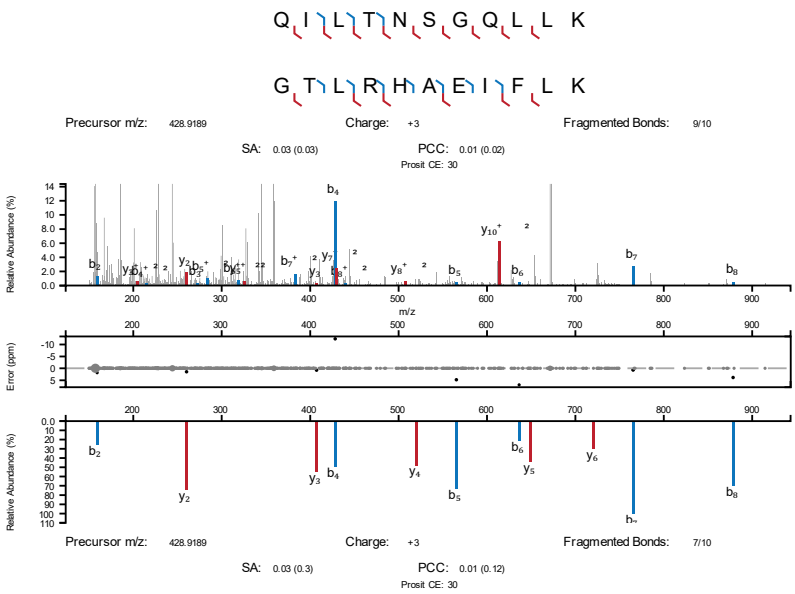

I D L P N K K V R
Precursor m/z: 361.5609
Charge: +3
Fragmented Bonds: 8/8
SA: 0.03 (0.05)
PCC: 0.01 (0.06)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
SA: 0.03 (0.3)
PCC: 0.01 (0.18)

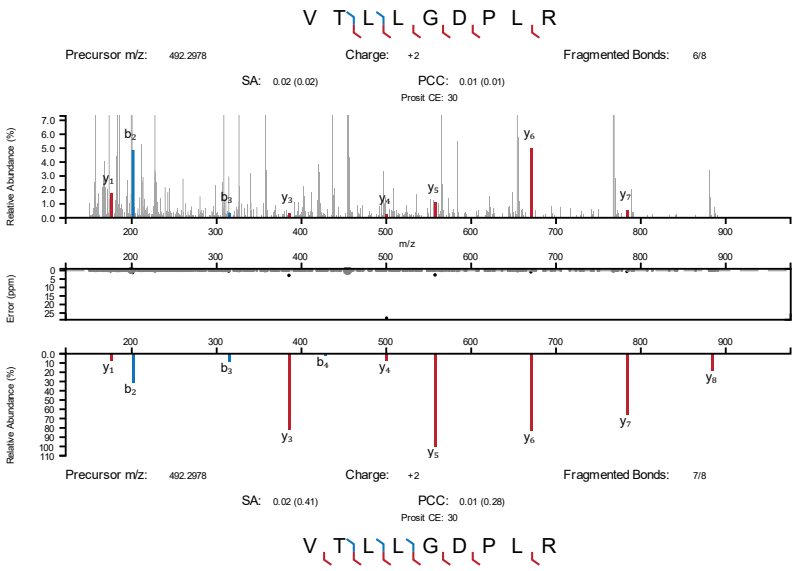

D R K S Y S L A A G
Precursor m/z: 534.2776
Charge: +2
Fragmented Bonds: 5/9
SA: 0.03 (0.03)
PCC: 0.02 (0.03)
Prosit CE: 30
Relative Abundance (%)
m/z
[M+2H]+2
Error (ppm)
Precursor m/z: 534.2776
Charge: +2
Fragmented Bonds: 8/9
SA: 0.03 (0.29)
PCC: 0.02 (0.03)
Prosit CE: 30
D R K S Y S L A A G


T D Q M V F N T Y Q
Precursor m/z: 623.7741
Charge: +2
Fragmented Bonds: 5/9
SA: 0.06 (0.15)
PCC: 0.07 (0.22)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 623.7741
Charge: +2
Fragmented Bonds: 8/9
SA: 0.06 (0.25)
PCC: 0.07 (0.22)
Prosit CE: 30
T D Q M V F N T Y Q


R P G P E E A A A V A
Precursor m/z: 534.2776
Charge: +2
Fragmented Bonds: 8/10
SA: 0.02 (0.02)
PCC: 0.01 (0.01)
Prosit CE: 30
Relative Abundance (%)
m/z
[M+2H]+2
Error (ppm)
Precursor m/z: 534.2776
Charge: +2
Fragmented Bonds: 7/10
SA: 0.02 (0.23)
PCC: 0.01 (0.02)
Prosit CE: 30
R P G P E E A A A V A


Q I L T N S G Q L L K
Precursor m/z: 607.8588
Charge: +2
Fragmented Bonds: 5/10
SA: 0.03 (0.03)
PCC: -0.01 (-0.01)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 607.8588
Charge: +2
Fragmented Bonds: 9/10
SA: 0.03 (0.63)
PCC: -0.01 (0.68)
Prosit CE: 30


G T L R H A E I F L K
Precursor m/z: 428.9189
Charge: +3
Fragmented Bonds: 9/10
SA: 0.03 (0.03)
PCC: 0.01 (0.02)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 428.9189
Charge: +3
Fragmented Bonds: 7/10
SA: 0.03 (0.3)
PCC: 0.01 (0.12)
Prosit CE: 30


V T L L G D P L R
Precursor m/z: 492.2978
Charge: +2
Fragmented Bonds: 6/8
SA: 0.02 (0.02)
PCC: 0.01 (0.01)
Prosit CE: 30
Relative Abundance (%)
m/z
Error (ppm)
Precursor m/z: 492.2978
Charge: +2
Fragmented Bonds: 7/8
SA: 0.02 (0.41)
PCC: 0.01 (0.28)
Prosit CE: 30

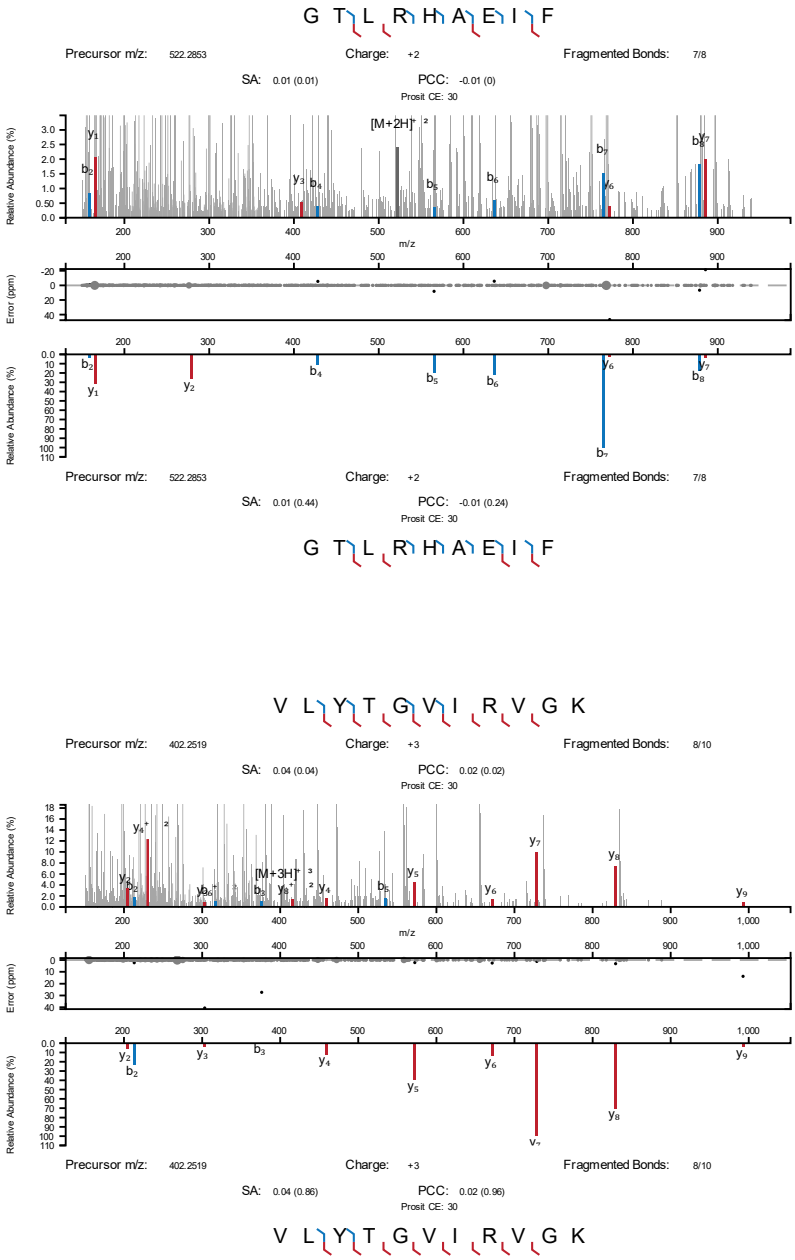


Figure 3-10 Spectral Visualization of Immunopeptides with Single Amino Acid Mutations.

Table 3-1 List of Immunopeptides Containing Single Amino Acid Substitutions Identified in This Experiment via Pseudo-DIA Library Approach.

sample	Gene	Mutation	Gene name	scan	m/z	intensity	RT
HCT116	CHMP7	A324T	QTDQM(UniMod:35)VFNTY2	50520	631.7723	495526	61.6318
HCT116	CHMP7	A324T	QTDQMVFNTY2	69848	623.7749	4448710	85.37
HCT116	CHMP7	A324T	TDQMVFNTYQ2	69899	623.7749	2470410	85.37
HCT116	GAPDH	I69T	AENGKLVTNGNPIT2	37013	714.3789	106496	45.2738
HCT116	GAPDH	I69T	ENGKLVTNGNPIT2	36295	678.8604	39674.3	44.4
HCT116	GAPDH	I69T	VTNGNPIT1	17847	815.4269	86274.8	21.9545
HCT116	NAPA	A181V	KVIDIYEQV2	65914	553.8091	2010450	80.4974
HCT116	RNPEP	I195F	ALFEVPDGFTA2	91213	583.7909	81928.4	111.424

HCT116	ANAPC1	E1292K	DRKSYSLAAG2	20573	534.2784	249184	25.2983
HCT116	KIF13B	T1254M	FLIVRVM(UniMod:35)VQ2	84530	560.8319	103655	103.238
HCT116	TTC33	A115G	GAEMAVQQ1	23306	833.3833	606103	28.6107
HCT116	TTC33	A115G	GAEMAVQQ2	23315	417.1956	108826	28.6227
HCT116	ATOX1	C41R	IDLPNKKVR3	18075	361.5616	37749.1	22.152
HCT116	TBC1D7	R36S	ILLKDDSL2	48054	458.7719	1792500	60.2304
HCT116	TPK1	K118T	KIEEKDLTV3	28427	358.8736	808420	34.8035
HCT116	COG7	V193I	LASPQIIA1	68898	812.4887	446030	84.1407
HCT116	IQGAP3	S1070T	LEDKVLTV2	48003	458.7719	2466700	60.2304
HCT116	PLRG1	A361T	RLWDLVTGKTRV3	75393	481.9533	165913	92.0159

HCT116	TMEM158	A70V	RPGPEEAAAVA2	20573	534.2784	347446	25.3599
HCT116	MVB12B	N247D	RTDYEYQHSDLY2	33859	795.3478	175843	41.4413
HCT116	PLS3	N372S	SLFNKYPAL2	78457	526.7932	228093	95.8134
HCT116	CCNI2	A91V	TADVVPAAA1	47376	814.4316	39950.3	57.8021
HCT116	BFSP1	A220T	TLRSQLEE2	27402	488.2597	161284	33.5844
HCT116	DEAF1	A167T	TTGLKGPT1	20852	774.4367	70460.3	26.1484
HCT116	PPP1R18	P499L	VDAAVLGAGK2	26838	450.7619	325927	32.9004
HCT116	GSTP1	G208V	VNLPINGNVKQ2	49547	598.3441	330710	60.5126
H358	CTH	Q323R	GTLRHAEIF2	48826	522.2861	334916	59.6172
H358	CTH	Q323R	GTLRHAEIFLK3	41268	428.9197	276813	50.4762

H358	GTF2A1	Q164K	QILTNSGQLLK2	59799	607.8596	60082.8	73.1255
H358	SPATA5	Q234H	SLHLSQLDL2	76518	513.2858	456018	93.3885
H358	SIRPG	Q234L	VTLLGDPLR2	75853	492.2987	462555	92.742
H358	ANAPC1	V504I	VLYTGVIRVGK3	59626	402.2527	340958	72.854

3.4 Conclusions

In this study, we introduced a novel Pseudo-DIA Library Search Strategy for the identification of immunopeptides to address the inherent complexities associated with their detection in complex biological samples. Our approach leverages the high potential diversity of immunopeptides, estimated at 167 million for sequences of 8-15 amino acids, effectively reducing this vast landscape to a manageable size for analysis. Our results demonstrate that this strategy significantly enhances identification rates compared to traditional methods, achieving up to 3.8 times more immunopeptide identification in various cell lines.

Moreover, our method can be used to identify cryptic immunopeptides and neoantigens that are crucial for developing personalized immunotherapy. By utilizing both DDA and DIA formats, we validated our strategy across multiple independent datasets, revealing its robustness and applicability in diverse contexts. Notably, the efficiency of our approach is underscored by our executable program, which simplifies the generation of spectral libraries directly from MSFragger results, allowing seamless integration into existing workflows.

The successful identification of neoantigens further emphasizes the significance of our findings, as these peptides represent promising targets for immunotherapy. Overall, our Pseudo-DIA Library Search Strategy not only advances the field of proteomics but also opens new avenues for cancer research and personalized medicine, facilitating the

development of effective therapeutic interventions.

Chapter 4. Large-scale identification of functional microproteins with immunopeptidomics and CRISPR screening

4.1 Introduction

The traditional view of the proteome has long been centered on canonical proteins encoded by annotated open reading frames (ORFs).[1] However, recent advances in high-throughput sequencing,[144] ribosome profiling,[64, 145] and mass spectrometry[146] have uncovered a previously hidden layer of the proteome: microproteins. These small proteins, often encoded by small ORFs (sORFs) within non-coding RNAs, untranslated regions (UTRs), and other non-canonical regions of the genome, challenge the conventional boundaries of transcriptome and proteome annotations. Microproteins, typically less than 100 amino acids in length,[147] represent an intriguing and underexplored class of biomolecules, with emerging evidence suggesting that they play critical roles in diverse biological processes.[148, 149]

Despite their small size, microproteins have been implicated in fundamental cellular functions, such as regulation of immune responses,[150] control of gene expression,[151] modulation of signaling pathways,[152] and maintenance of cellular homeostasis.[153] For example, several microproteins have been shown to regulate cancer-related pathways, including oncogenic RAS signaling and DNA damage repair mechanisms.[154] However, the majority of microproteins remain functionally uncharacterized, and their precise roles in health and disease are still poorly understood[155-157]. This knowledge gap is partly due to challenges associated with identifying and validating microproteins, as

their small size, low abundance, and lack of well-defined annotations often exclude them from traditional proteomics pipelines.[110, 147]

Immunopeptidomics,[158] a mass spectrometry-based approach that identifies peptides presented on major histocompatibility complex (MHC) molecules, has emerged as a powerful tool for uncovering novel microproteins.[102, 148] Unlike traditional trypsin-based proteomics, immunopeptidomics does not rely on specific enzymatic cleavage sites and can detect peptides derived from non-canonical sORFs.[107, 159] Furthermore, the enrichment of low-abundance peptides through co-immunoprecipitation circumvents many limitations of conventional proteomics, making it uniquely suited for the discovery of microproteins.[160] Recent studies have successfully used immunopeptidomics to reveal thousands of previously unannotated peptides, suggesting that the repertoire of microproteins is far larger and more complex than previously appreciated.[161]

In addition to their potential biological significance, microproteins located near oncogenes or other cancer-related genes are of particular interest in cancer research. Several studies have reported that microproteins derived from uORFs, lncRNAs, and other non-coding regions regulate oncogenic pathways, including tumor progression, metastasis, and resistance to therapy.[162, 163] These findings underscore the importance of systematically exploring the microproteome in cancer to uncover novel regulators and therapeutic targets.

To address these challenges and opportunities, we performed a comprehensive study to identify, characterize, and functionally validate microproteins using a combination of immunopeptidomics, CRISPR screening, and bioinformatic analyses. First, we used immunopeptidomics to identify over 36,000 microproteins across 86 datasets, representing one of the largest microprotein datasets reported to date. We then categorized these microproteins based on their genomic origins, start codon usage, and conservation among species. To evaluate their functional significance, we designed a CRISPR screening library targeting over 2,000 microproteins and performed a high-throughput functional analysis in three cancer cell lines. Finally, we integrated our CRISPR results with the DepMap data to investigate whether microproteins function independently of their associated canonical proteins.

This study provides novel insights into the microproteome, uncovering not only the sheer diversity of microproteins but also their critical roles in cellular processes, particularly in cancer. Our findings highlight the importance of microproteins as a distinct and functional class of biomolecules, setting the stage for future research to elucidate their mechanisms and therapeutic potential.

4.2 Materials and methods

4.2.1 Immunopeptidomics Mass Spectrometry Database Search

A total of 86 datasets, including 5,406 raw mass spectrometry data files, were downloaded and analyzed using FragPipe software (version 20.0) and MSFragger

(version 3.8) for database searching. The search parameters were set with an MS1 mass tolerance of 20 ppm and an MS2 mass tolerance of 0.02 Da. The peptide lengths were restricted to a range of 8–15 amino acids, while oxidation and carbamidomethylation were included as variable modifications.

After completing the database searches, pin files were generated using MSFragger. These files were further processed using MSBooster to predict the secondary spectra and retention times. Subsequently, Percolator software was employed to score the spectra, with a stringent false discovery rate (FDR) threshold of 1% applied to ensure high confidence in the results.

Three distinct small open reading frame (sORF) databases were utilized for the database search: OpenProt 2.0, smProt2, and sORFs.org. In addition, the UniProt database, which contains 20,000 canonical proteins, was downloaded. These four databases were merged into a single FASTA file to facilitate a comprehensive MSFragger-based search of immunopeptidomics data.

4.2.2 Redundancy Removal of Identified Microproteins

Given that some immunopeptides could map to multiple highly similar genomic sORF regions, a prioritization strategy was implemented to select the most probable sORF for each peptide. The prioritization was based on the following criteria: (1) Transcript Support Level (TSL): Microproteins corresponding to transcripts with higher TSL scores were prioritized to ensure high-confidence annotations. (2) Start Codon

Preference: Start codons were ranked in the following order of preference: ATG > CTG > GTG > others, reflecting the likelihood of translation initiation from these codons. (3) Protein Length: Longer microproteins were given higher priority, as longer sequences typically provide more robust evidence for their existence. (4) Annotation Source: Among the various annotation sources, Ensembl annotations were prioritized due to their reliability and extensive curation. (5) Database Source: Among the three sORF databases, sORFs annotated in OpenProt were given the highest priority, followed by smProt and sORFs.org.

4.2.3 Conservation Analysis of Microproteins

The conservation of microproteins across species was evaluated following the methodology outlined by Sandmann, et al.[48] Briefly, genomic alignment data from 120 species were downloaded from the UCSC Multiz 100-Way Alignment. Homologous regions in other species were identified using the genomic coordinates of human microproteins as references.

The DNA sequences corresponding to these homologous regions were translated in silico into protein sequences, generating potential microprotein sequences for each species examined. To quantify the degree of conservation, BLAST was used BLAST to compare human microproteins with their homologs in other species. The similarity was scored using the e-value, and the final conservation score was calculated as -log10(e-value). A higher score indicates greater similarity and stronger conservation.

To visualize the alignment and conservation of microproteins across species, Jalview[164] software was employed, providing a detailed display of sequence similarities and variations.

4.2.4 Selection of Microproteins for CRISPR Screening

To identify suitable microprotein candidates for CRISPR screening, a two-tiered selection strategy was employed, focusing on both the existence and potential functionality of the microproteins.

For **existence-driven selection,** the following criteria were applied: (1) Microproteins derived from uORFs, dORFs, and lncRNAs were selected to avoid targeting canonical proteins. (2) Preference was given to microproteins with ATG, CTG, or GTG as start codons, as these are more likely to initiate translation. (3) The protein length was required to be at least 10 amino acids to ensure its functional potential. (4) Only microproteins with a peptide-spectrum match (PSM) count of ≥2 were included to ensure high-confidence identification. (5) The transcript support level (TSL) of the corresponding transcripts was set to ≤3, indicating strong transcript-level evidence.

For **functionality-driven selection**, the following criteria were applied: (1) The expression level of the transcript encoding the microprotein, measured in transcripts per million (TPM), needed to be ≥0.5 to ensure sufficient expression of the microprotein. (2) The transcript had to exhibit differential expression between normal and cancer samples in TCGA dataset. (3) Survival analysis required the transcript to show a statistically significant association with patient survival (p-value ≤0.05). To broaden

the scope of potentially functional microproteins, criteria (2) and (3) for functionality-driven selection were combined into a union set. This comprehensive approach ensured the selection of high-confidence microprotein candidates with potential biological and clinical relevance for CRISPR screening.

4.2.5 sgRNA Spacer Sequence Design

To design sgRNAs targeting microproteins, the cDNA sequences of the microproteins were extracted from the genome, with an additional 15 bp of flanking DNA regions included upstream and downstream. The following parameters were applied during sgRNA spacer design and selection:

(1) Two tools, SSC[165] and CRISPick[166], were used for sgRNA design, with an on-target score threshold of 0.2. The scores from both tools were balanced to achieve optimal design, as illustrated in Fig. 8A. (2) Off-target scores were filtered with a cutoff of <100 to minimize non-specific targeting. (3) Bowtie[167] software was used to identify the number of matches for each sgRNA spacer in the human genome. Only spacers with a unique match (match = 1) were retained. (4) The GC content of each spacer was required to fall within 20%–80% to ensure stability and efficiency.[168] (5) Spacers containing poly-T signals were excluded to avoid premature transcription termination. (6) Spacers and plasmid vectors were checked to ensure that they did not form Esp3I restriction enzyme sites. (7) A maximum of 10 sgRNAs were designed for each microprotein to ensure sufficient coverage.

For control sgRNAs, over 400 sgRNAs targeting 86 essential human genes (e.g., ribosomal proteins, EIF proteins, PSMA/PSMB proteins) were selected from the GECKO dataset.[169] Additionally, negative controls included non-targeting sgRNAs and sgRNAs targeting the safe harbor AAVS1 region. In total, 18,000 sgRNAs, including both microprotein-targeting and control sgRNAs, were designed.

4.2.6 CRISPR Library Construction

A custom vector was designed based on the FUGW vector. The vector contained the following key elements: -cPPT-CMV promoter-Puromycin-U6 promoter -ESP3I restriction sites-spacer sequence-ESP3I restriction sites-sgRNA scaffold-WPRE-cPPT-. The ssDNA library was synthesized by TWIST Bioscience and amplified through eight rounds of PCR and purification. The amplified DNA fragments were ligated into the vector using Esp3I enzymes. The resulting plasmid library was transformed into competent cells using the NEB electroporation kit, following the manufacturer's protocol. The amplified plasmid library was then sent to Novogene for next-generation sequencing (NGS).

4.2.7 CRISPR Screening of Three Cancer Cell Lines

Three different cancer cell lines were cultured in 15 cm dishes, with an initial cell number of approximately 2×10^7 cells per dish. The cells were then infected with viruses packaged using 293T cells, maintaining a multiplicity of infection (MOI) of approximately 30% to ensure that each cell was infected with only one virus. Two to three days post-infection, puromycin was added to the culture at a final concentration

of 1.5 μg/mL to select for successfully transduced cells. After 2–3 days of puromycin selection, the first batch of cells was collected and designated as T1, representing the starting time point. After five doubling times, the second batch of cells was collected and designated as T2. After ten doubling times, the final batch of cells was collected and designated as T3. For each cell line, samples were collected at three time points, with two replicates for each time point. The collected samples were then sent for next-generation sequencing.

4.3 Results

4.3.1 Extensive Identification of Microproteins Through Immunopeptidomics

Using immunopeptidomics, we analyzed 86 datasets containing over 5,000 raw data files. By employing MSFragger to search against three sORF databases,[120, 170, 171] we successfully identified a total of 36,494 microproteins, which represents one of the most comprehensive datasets of microproteins to date.[172] To further characterize these microproteins, we categorized them using established localization and classification methods.

As shown in Figure 4-1B, the majority of these microproteins were derived from lncRNAs, with over 12,000 identified, followed by intORFs (7230), uORFs (5,258), dORFs (4197), uoORFs (2,539), and doORFs (1215), which were the least frequent category, with only 1,347 identified. This distribution highlights the diverse origins of

microproteins and emphasizes the significant contribution of lncRNAs to the microproteome.

We also analyzed the start codons of these microproteins **(Figure 4-1D)**. The vast majority (87%) of the genes used the canonical ATG start codon. This finding aligns with the known translation initiation mechanisms for canonical proteins. Additionally, we observed that the second, third, and fourth most common start codons were CTG (6.8%), GTG (2.3%), and TTG (0.8%), respectively, a ranking that is consistent with previously reported studies.[148]

One particularly interesting observation was that approximately 3,000 microproteins were located in the vicinity of known oncogenes **(Figure 4-1B)**, including cancer-related genes with identified uORFs, dORFs, and intORFs. These findings suggest that microproteins may play important roles in cancer biology. The presence of microproteins near oncogenes raises intriguing questions about their potential regulatory functions and biological significance, which warrant further investigation.

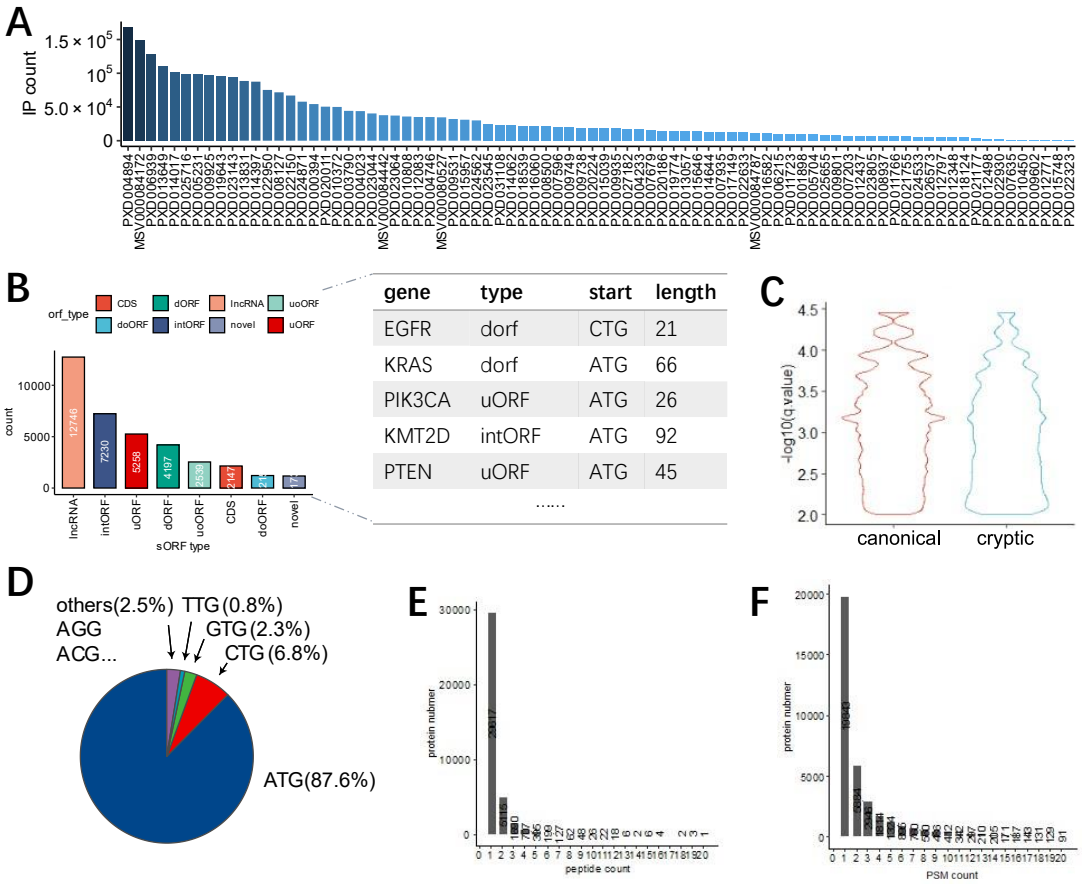


Figure 4-1 Non-canonical microproteins identified by immunopeptidomics.

(A) Bar chart showing the number of identified immunopeptides for each of the 86 datasets. (B) Distribution of different ORF types identified in the immunopeptide dataset, highlighting microproteins associated with oncogenes. (C) q-value score distribution of cryptic non-canonical and canonical immunopeptides. (D) Bar chart showing the distribution of different ORF start codons. (E) Statistical chart of the number of unique immunopeptides contained in each ORF. (F) Statistical chart of the number of PSMs contained in each ORF.

4.3.2 Quality of Immunopeptide Identification

The quality and reliability of immunopeptide identification are critical for ensuring the validity of these findings. This is particularly true for MS2 spectra, which provide the foundation for identifying peptides and their corresponding microproteins.

As shown in **Figure 4-1E**, more than 8,000 microproteins were supported by at least two unique immunopeptides, demonstrating a strong level of confidence in their identification. Furthermore, as illustrated in **Figure 4-1F**, over 18,000 microproteins were associated with at least two PSMs in the present study. These results indicate that a significant proportion of the identified microproteins is supported by robust experimental evidence.

To further evaluate the reliability of our data, we compared the q-values of the canonical proteins and microproteins **(Figure 4-1C)**. Both groups showed highly similar scoring trends, with minimal differences in the q-value distributions. This strongly suggests that the spectral quality of microproteins is comparable to that of canonical proteins, reinforcing the validity of our microprotein identification.

As a specific example, we highlight the uORF of PTEN,[63] a microprotein that comprises 45 amino acids. For this microprotein, we identified nine unique immunopeptides with a total of 62 PSMs, as shown in **Figure 4-2**. These peptides are unlikely to have arisen randomly or by chance. Instead, their presence implies transcription, translation, degradation, and subsequent presentation on the cell

membrane. These peptides were enriched by co-immunoprecipitation and detected via mass spectrometry, providing strong evidence that this microprotein is actively translated and functions.

We hypothesize that the ability of immunopeptidomics to identify a larger number of microproteins compared to traditional methods is due to several key factors:[102] (1) Higher potential peptide yield: Immunopeptidomics generates a significantly larger number of peptides than traditional tryptic digestion, often by orders of magnitude, and this higher yield increases the likelihood of identifying unique peptides corresponding to microproteins. (2) Independence from trypsin cleavage patterns: Trypsin cleavage relies on specific residues (K/R), which can result in no peptides >8 amino acids for some microproteins, whereas immunopeptides are not constrained by this limitation. (3) Enrichment of low-abundance peptides: Immunopeptides are enriched through co-immunoprecipitation, reducing interference from high-abundance proteins and allowing the detection of low-abundance microproteins. (4) HLA diversity across samples: Different HLA types in individuals act as a natural "fractionation system," contributing unique immunopeptides from different samples and significantly enhancing the coverage of microprotein identification.[102]

Additionally, many previously reported microproteins were also identified in our study, suggesting that our dataset is consistent with and expands upon prior work. As shown in Table 4-1, these include microproteins derived from lncRNAs (e.g.,

MIR155HG,[4] LINC00511,[154] and GAS5[173]) and uORFs (for example, MKKS,[174] PTEN,[63] GTF2H1[151]). These microproteins are known to exhibit diverse biological functions, including (1) regulating immune and inflammatory responses, such as suppressing autoimmune inflammation by modulating antigen presentation. (2) Participation in cancer-related pathways, such as controlling oncogenic RAS signaling, inhibiting triple-negative breast cancer progression, and enhancing translation initiation in hematopoietic malignancies. (3) Maintaining DNA damage repair mechanisms, which are crucial for cellular homeostasis.

These previously reported microproteins served as positive controls, strongly supporting the functional relevance of the microproteins in our dataset. Moreover, they suggested that many additional functional microproteins remain to be discovered within the 36 K dataset.

A

IP_185775, PTEN_uORF_45, protein sequence:

MRDGGGRGPEPLSAPVSSRGGSALGEPAGLRRRQRRRFSPPLRLF

1 MRDGGGRGPEPL

2 APVSSRGGSA

3 RRRFSPPLRLF

4 RRRFSPPLR

5 RRFSPPLR

6 RRFSPPLRLF

7 RRFSPPLRL

8 RFSPPLRLF

9 FSPPLRLF

B

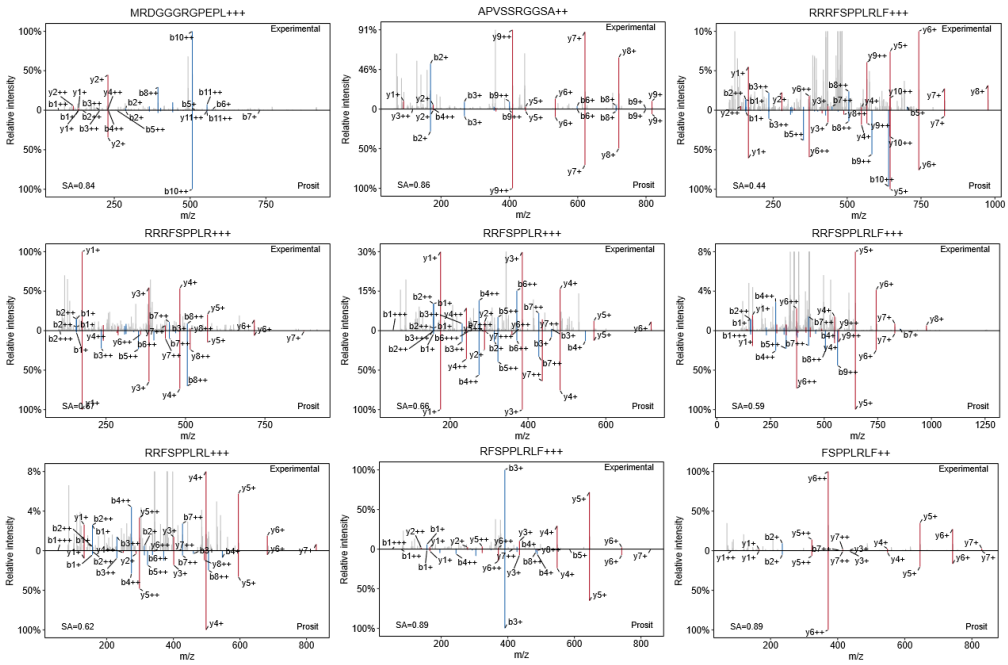


Figure 4-2 Identification of microproteins using immunopeptides, with the PTEN uORF as an example. (A) The protein amino acid sequence of the PTEN uORF, with the positions of 9 immunopeptides highlighted in red. (B) MS2 spectra of the 9 immunopeptides.

Table 4-1 The 36K Microprotein List Provides Immunopeptidomics Evidence for Previously Reported Microproteins

Year	Author	Pubmed ID	Journal	Gene name	Orf type	Length	Peptide count	PSM count
2020	Honglin Wang[4]	32671205	Sci Adv	MIR155HG	lncRNA	17	2	8
2022	Nu Zhang[154]	36241718	Cell Res	LINC00511	lncRNA	108	5	10
2022	Jun Li[175]	35849344	NAR	CTBP1-DT	lncRNA	186	2	2
2022	Jianhua Yang[176]	35609991	Genome Res	KLRK1-AS1	lncRNA	100	2	5
2013	Hitoshi Endo[174]	23671934	BBA	MKKS	uORF	50	11	22
2013	Hitoshi Endo[174]	23671934	BBA	MKKS	uORF	63	26	88
2024	Zhe Ji[151]	38431639	Nat Commun	PGRMC1	intORF	95	2	5
2024	Zhe Ji[151]	38431639	Nat Commun	CGGBP1	uORF	105	9	58

2024	Zhe Ji[151]	38431639	Nat Commun	TFAM	uoORF	99	12	39
2024	Zhe Ji[151]	38431639	Nat Commun	GTF2H1	uORF	91	4	34
2024	Zhe Ji[151]	38431639	Nat Commun	SLC5A6	uORF	77	3	13
2024	Zhe Ji[151]	38431639	Nat Commun	MAPKAPK5	uORF	44	5	12
2024	Zhe Ji[151]	38431639	Nat Commun	INPP5F	uORF	106	2	2
2020	Jonathan S. Weissman[148]	32139545	Science	HAUS6	uORF	15	1	2
2020	Jonathan S. Weissman[148]	32139545	Science	FBXO9	uORF	23	2	3
2020	Jonathan S. Weissman[148]	32139545	Science	TBPL1	uORF	42	2	3
2020	Jonathan S. Weissman[148]	32139545	Science	ABHD17A	uORF	43	9	37
2012	Alan Saghatelian[177]	23160002	Nat Chem Biol	DRAP1	uoORF	95	3	8

2012	Alan Saghatelian[177]	23160002	Nat Chem Biol	CACTIN	intORF	102	2	45
2024	Susan Carpenter[178]	38781216	PNAS	SLC39A13-AS1	lncRNA	96	1	9
2024	Susan Carpenter[178]	38781216	PNAS	SLC39A13-AS1	lncRNA	59	1	1
2013	Carla Scaroni[179]	23555276	PLoS Genet	CDKN1B	uORF	29	1	1
2019	Sebastiaan van Heesch[173]	31155234	Cell	GIHCG	lncRNA	64	4	6
2019	Sebastiaan van Heesch[173]	31155234	Cell	MOCS2-DT	lncRNA	85	8	12
2019	Sebastiaan van Heesch[173]	31155234	Cell	CDC37L1-DT	lncRNA	33	1	2
2019	Sebastiaan van Heesch[173]	31155234	Cell	ENSG00000275055	lncRNA	50	3	5

2019	Sebastiaan van Heesch[173]	31155234	Cell	GAS5	lncRNA	73	9	98
2019	Sebastiaan van Heesch[173]	31155234	Cell	IDI2-AS1	lncRNA	204	2	3
2019	Sebastiaan van Heesch[173]	31155234	Cell	KDM7A-DT	lncRNA	68	2	2
2019	Sebastiaan van Heesch[173]	31155234	Cell	KDM7A-DT	lncRNA	33	7	58
2019	Sebastiaan van Heesch[173]	31155234	Cell	LINC02941	lncRNA	59	2	6
2019	Sebastiaan van Heesch[173]	31155234	Cell	ILRUN-AS1	lncRNA	52	1	25
2019	Sebastiaan van Heesch[173]	31155234	Cell	CD63-AS1	lncRNA	69	3	11
2019	Sebastiaan van Heesch[173]	31155234	Cell	SNHG6	lncRNA	23	2	11
2019	Sebastiaan van Heesch[173]	31155234	Cell	SNHG8	lncRNA	36	7	30
2019	Sebastiaan van Heesch[173]	31155234	Cell	SNHG8	lncRNA	52	6	14

2019	Sebastiaan van Heesch[173]	31155234	Cell	SNHG16	lncRNA	59	12	17
2019	Sebastiaan van Heesch[173]	31155234	Cell	TBX5-AS1	lncRNA	79	1	2
2019	Yifeng Zhou[180]	31755573	EMBO J	LINC00665	lncRNA	52	1	5
2021	Yueqin Chen[94]	34555354	Mol Cell	ASH1L-AS1	lncRNA	145	3	5
2023	Alan Saghatelian[63]	36599300	Cell Metab	PTEN	uORF	45	9	62

4.3.3 Characterization of Microproteins

Microproteins are generally much shorter than canonical proteins. As shown in **Figure 4-3C**, the average length of microproteins in our dataset was 66 amino acids, compared to 500 amino acids for canonical proteins. This makes microproteins approximately eight times shorter than canonical proteins, consistent with previous studies.[148, 151, 181] In addition to length, prior studies[182] have suggested that non-canonical sORFs often possess a hydrophobic tail, which makes them more susceptible to degradation by the BAG6 system than canonical sORFs. We applied similar analytical methods to our dataset and observed the same trend **(Figure 4-3A)**. Microproteins, except for uoORFs, displayed significantly more hydrophobic C-terminal tails compared to canonical proteins. The exception for uoORFs is likely due to their overlap with canonical ORFs, as their "tails" correspond to the start regions of the canonical proteins. These findings align with prior studies, suggesting that microproteins are more prone to degradation by BAG6 or similar systems.[182]

Interestingly, we observed a distinct phenomenon in the length distribution of microproteins: a sharp boundary at 28–29 amino acids in the length distribution. Specifically, microproteins are more likely to be ≥29 amino acids in length. The biological significance of this observation is currently unclear; however, it may reflect an evolutionary or functional threshold for microprotein stability or activity.

To investigate the evolutionary conservation of microproteins, we analyzed alignment

data from 100 species using UCSC resources.[48, 183] As shown in **Figure 4-4**, conservation patterns revealed a clear distinction between 62 mammalian species and 38 non-mammalian species, suggesting that microproteins are more prevalent in mammals. Within mammals, microproteins showed higher conservation in primates, indicating that they may have evolved alongside more advanced and specialized functions.[48]

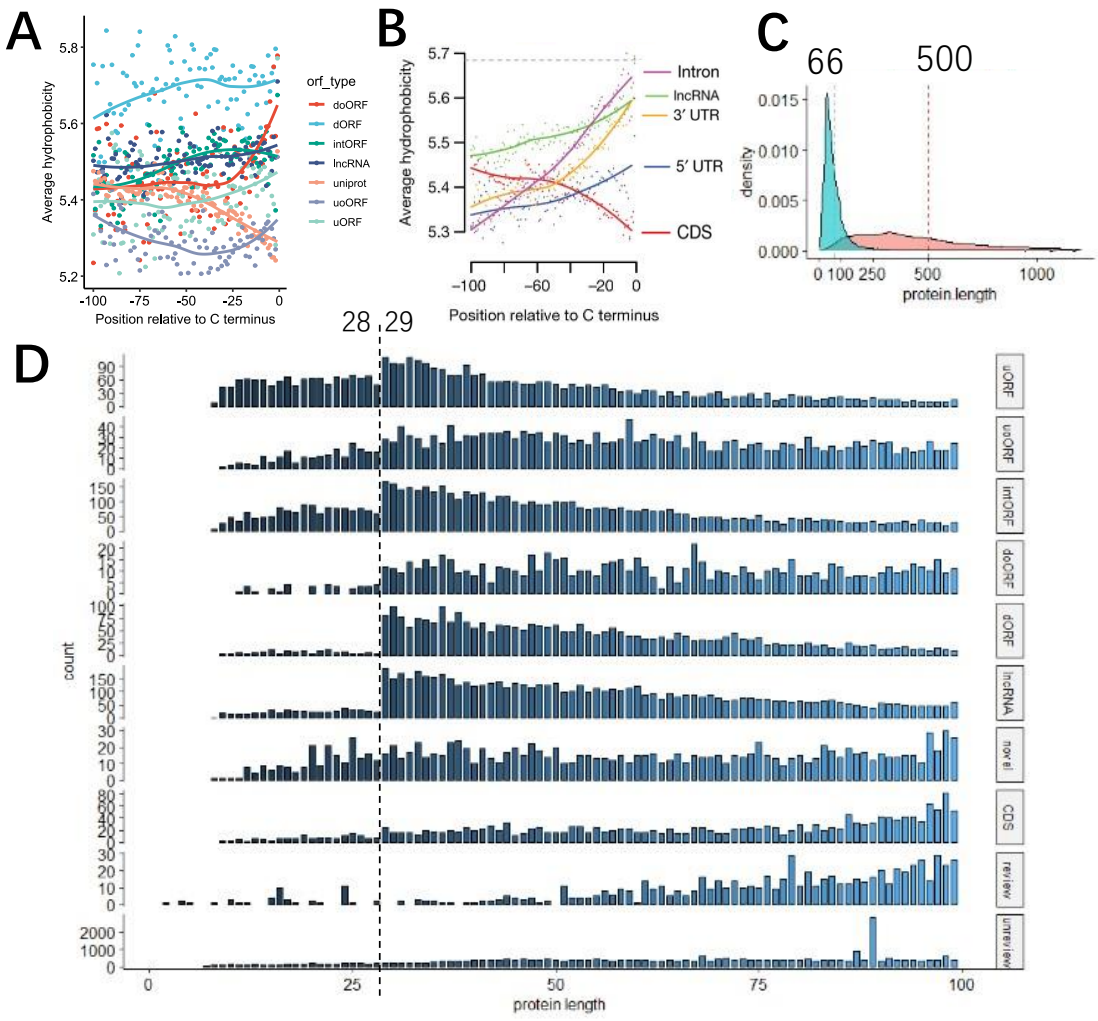


Figure 4-3 Characteristics of immunopeptides. (A) Average hydrophobicity of C-terminal amino acids in microproteins. (B) Average hydrophobicity of C-terminal amino acids in microproteins from the Wu, et al. study.[182] (C) Density plot comparing the protein lengths of microproteins and UniProt canonical proteins. (D) Distribution of protein lengths across different ORF types of microproteins.

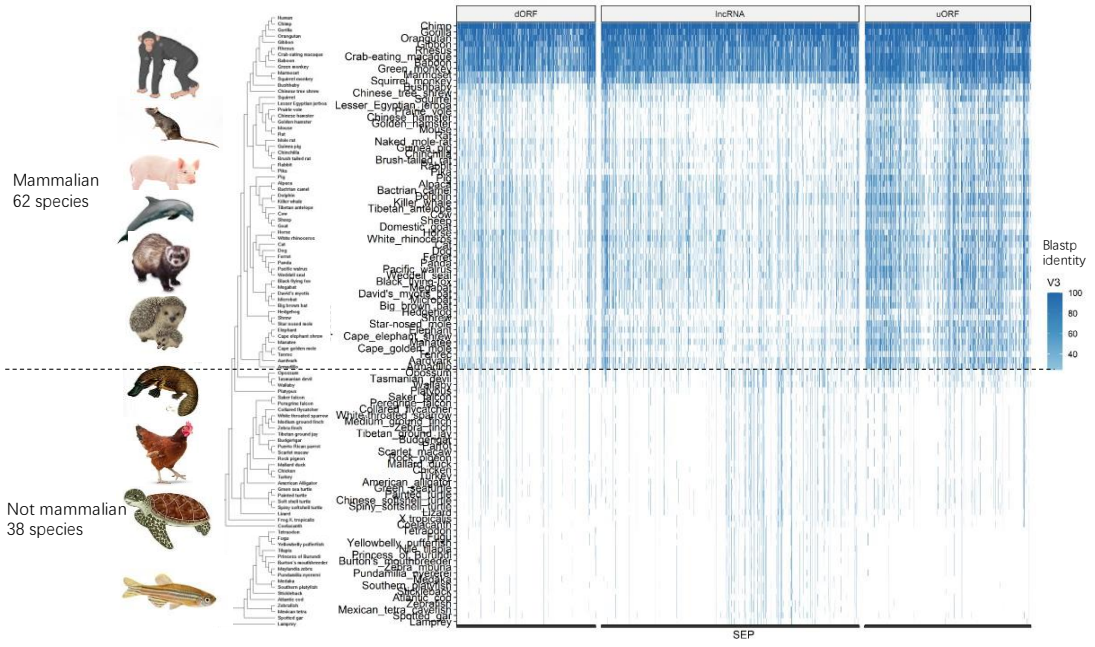


Figure 4-4 Heatmap of the conservation levels of microproteins from different ORF types across 100 species.

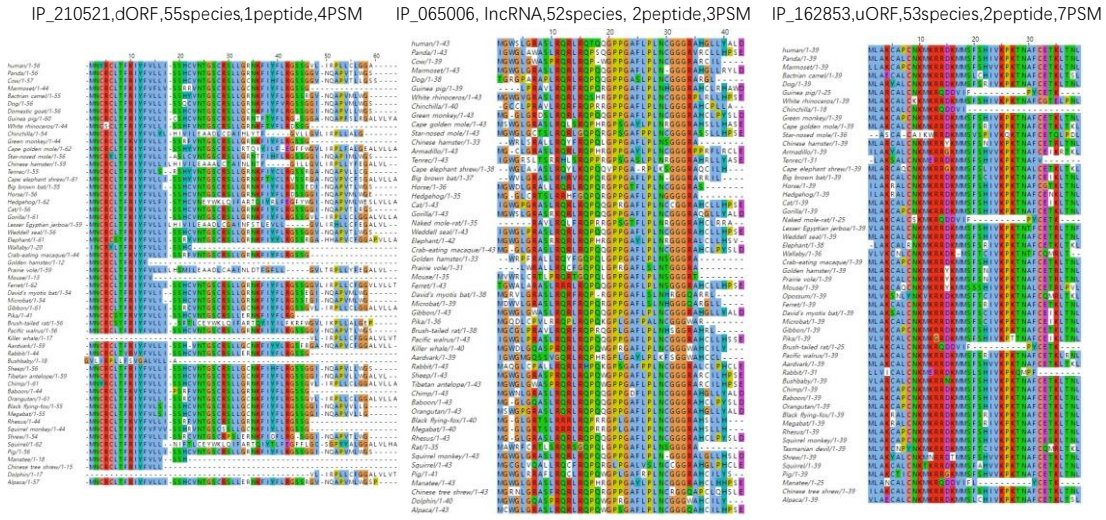


Figure 4-5 Examples of conservation levels for three microproteins from three different ORF types.

4.3.4 Microproteins candidate selection for CRISPR Library Construction

To explore the functions of microproteins, we designed a CRISPR screening experiment for high-throughput functional analysis. Candidate microproteins were selected based on transcript expression data from The Cancer Genome Atlas (TCGA)database.[184] Transcripts that were differentially expressed between cancer and normal samples were prioritized **(Figure 4-6A).** Additionally, transcripts with a p-value ≤ 0.05 in the survival analysis were included **(Figure 4-6B)**. After manual filtering, over 2,000 microproteins were selected for CRISPR library construction. For each microprotein, we designed up to 10 sgRNAs, along with positive controls targeting 86 essential genes (e.g., ribosomal proteins and EIF proteins). The negative controls included non-targeting sgRNAs and sgRNAs targeting the safe harbor AAVS1 locus, which has a minimal impact on cellular function. In total, we designed 18,000 sgRNAs, which were synthesized, amplified, and cloned into the lentiviral vectors.

The CRISPR KO library was used to infect three cancer cell lines: MDA-MB-231 (triple-negative breast cancer), LN229 (glioblastoma), and OVCAR8-ADR (drug-resistant ovarian cancer) cell lines. These cell lines were chosen to provide diverse biological contexts and robust data. We measured the sgRNA abundance at three time points: T1 (infection), T2 (5 doublings), and T3 (10 doublings). By comparing the sgRNA abundance between T1 and T3, we identified microproteins whose knockout

significantly impacted cell proliferation.

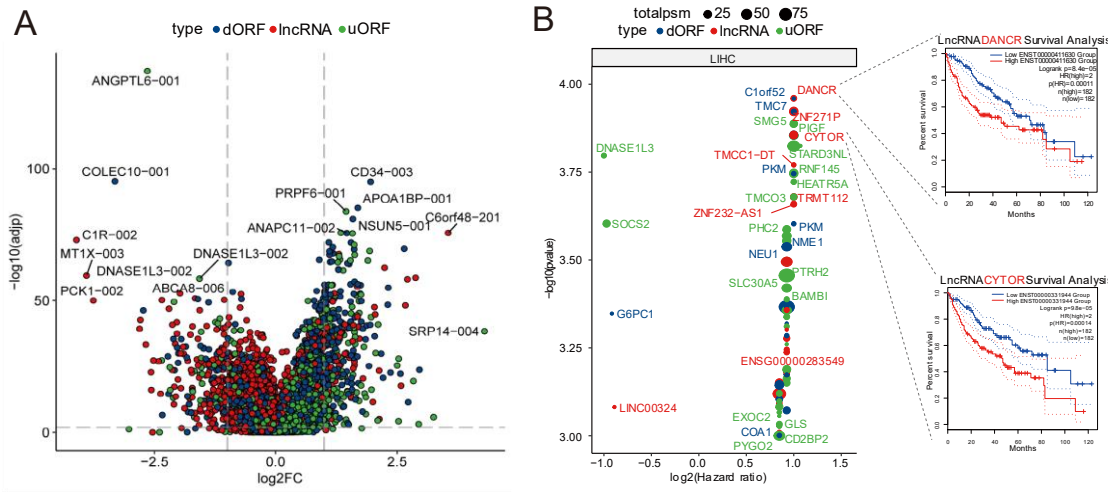


Figure 4-6 Microprotein Selection for the CRISPR Screening. (A) Volcano plot of source transcripts for microproteins in cancer versus normal samples. Differentially expressed transcripts corresponding to microproteins were selected to construct the CRISPR library. (B) Survival analysis of source transcripts for microproteins. Microproteins corresponding to transcripts with a p-value ≤ 0.05 were selected to construct the CRISPR library.

4.3.5 Quality Control of the CRISPR KO Library

To ensure the reliability and effectiveness of the CRISPR KO library, we performed a series of quality-control assessments. We evaluated the on-target efficiency of the sgRNAs using two widely recognized tools, SSC and CRISPick. As shown in **Figure 4-8A**, more than 73% of the sgRNAs achieved an on-target score greater than 0.4, indicating high targeting specificity. This suggests that the sgRNAs designed for microproteins are of sufficient quality to ensure reliable knockout experiments in mice. After constructing the CRISPR KO library, we performed targeted PCR on the sgRNA region, followed by NGS sequencing, to evaluate the uniformity of sgRNA abundance. As shown in **Figure 4-8B**, over 94% of sgRNAs exhibited read counts within a narrow range (log2 between 9.2 and 10.2). This result indicated that the sgRNAs were distributed evenly in the library, with less than two-fold variation in abundance. Such uniformity is critical for ensuring that all sgRNAs are adequately represented during the screening process, minimizing bias, and maximizing the reliability of the results.

To enhance the richness of our results and gain a more comprehensive understanding of microprotein phenotypes across different cancer types, we conducted CRISPR screenings in three distinct cancer cell lines (MDA-MB-231, LN229, and OVCAR8-ADR). Each condition was tested in two biological replicates to ensure reproducibility. As shown in **Figures 4-8C, E**, the correlation coefficients between replicates ranged from 0.79 to 0.90, demonstrating high consistency between biological replicates.

These results confirm the stability of the experimental system and validate the reliability of the data generated by CRISPR screening.

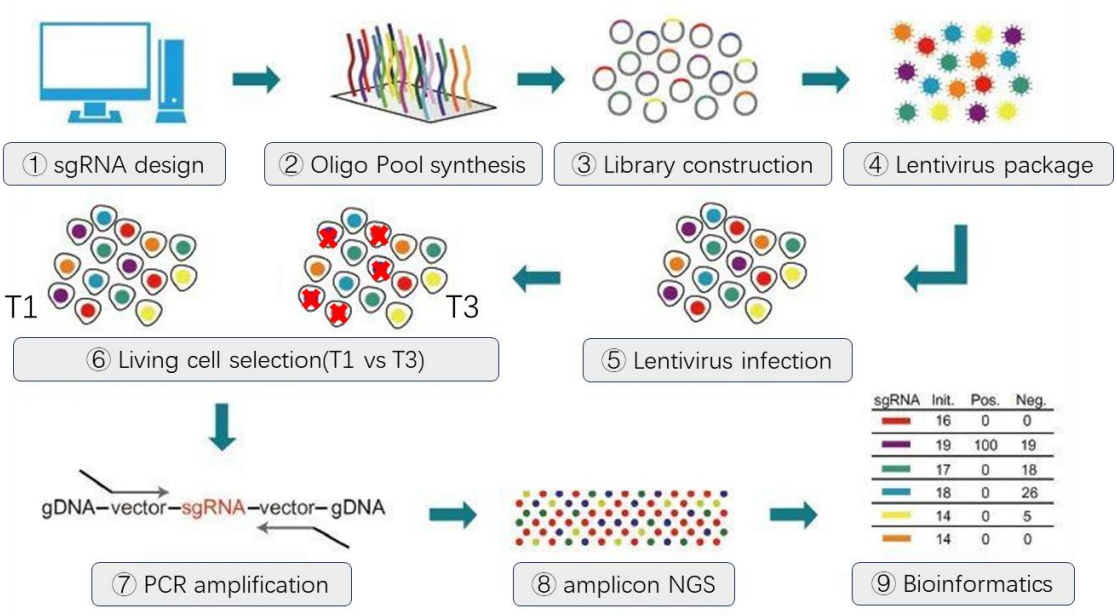


Figure 4-7. Flowchart of the experimental design for CRISPR experiments.

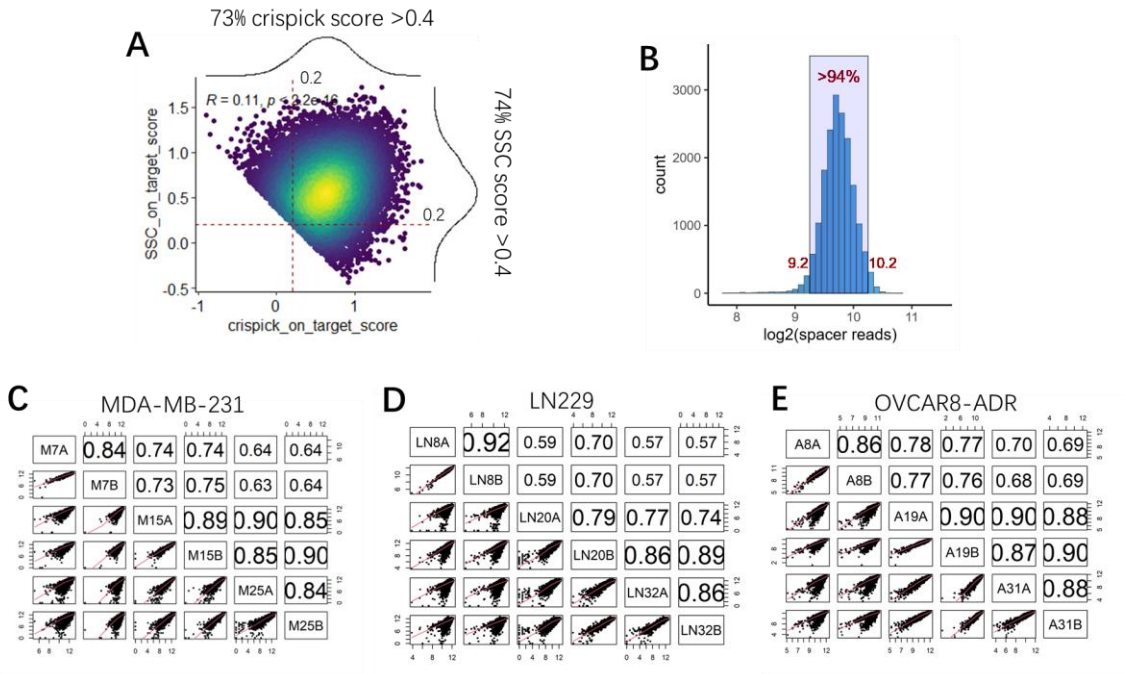


Figure 4-8 Quality control of the CRISPR KO library. (A) Score distribution of designed sgRNAs evaluated by two software tools, SSC and CRISPick. (B) Uniformity of sgRNAs in the CRISPR KO library, with over 94% of sgRNA abundances differing by less than twofold. (C-E) The technical reproducibility correlation of NGS sequencing at different time points across three different cell lines.

4.3.6 Microproteins Are Functionally Significant

We performed CRISPR screening in three cancer cell lines (MDA-MB-231, LN229, and OVCAR8-ADR) to systematically investigate the functional significance of microproteins **(Fig. 4-7)**. These cell lines were selected to provide a diverse representation of cancer types: MDA-MB-231 is a triple-negative breast cancer cell line, LN229 is derived from glioblastoma, and OVCAR8-ADR is a drug-resistant ovarian cancer cell line. This diversity was intended to ensure that the data were broadly applicable and captured functional microproteins relevant to various cancer types.

The results of the CRISPR screens are summarized in **Figure 4-9**, which presents a volcano plot of the changes in sgRNA abundance between T3 (10 doubling times) and T1 (initial infection). In the plot, the x-axis represents the log2 fold change (log2FC), where smaller values indicate greater depletion of sgRNAs targeting specific microproteins, while the y-axis shows the -log10(p-value), representing the statistical significance of the depletion. Positive controls (red points), which included sgRNAs targeting essential genes such as ribosomal proteins and EIF proteins, were significantly depleted, with log2FC values ranging from -2 to -10, indicating that knockout of these genes severely impaired cell viability, as expected. Negative controls (gray points), consisting of non-targeting sgRNAs or those targeting the safe harbor AAVS1 locus, clustered near the origin, demonstrating a minimal impact on cell proliferation. Microprotein-targeting sgRNAs (blue points) exhibited

intermediate depletion levels, which were between those of the positive and negative controls. Many microproteins showed moderate to significant effects on cell proliferation, highlighting their important functional role.

Interestingly, a subset of microproteins exhibited depletion levels comparable to the positive controls, with log2FC values approaching or even exceeding those of the essential genes. This suggests that these microproteins are critical for cell proliferation and may play roles as important as those of canonical essential genes.

To identify conserved functional microproteins across different cancer types, we analyzed the overlap among the 150 most significant microproteins in each of the three cell lines. As shown in **Figure 4-10A**, approximately 50% of the microproteins were shared across all three cell lines, despite their diverse origins. This overlap suggests that a core set of microproteins may play a universal role in cancer biology. We highlight three examples of shared microproteins: ENSG00000290937_lncRNA_35, WBP1_uORF_36, and GPRC5A_dORF_50. As shown in **Figure 4-10B**, multiple sgRNAs targeting these microproteins consistently depleted all three cell lines over time (T1, T2, T3). Additionally, these microproteins were supported by solid immunopeptide evidence, further validating their existence and their functional relevance. These findings strongly suggest that these microproteins are essential for cell survival and may play a critical role in cancer progression.

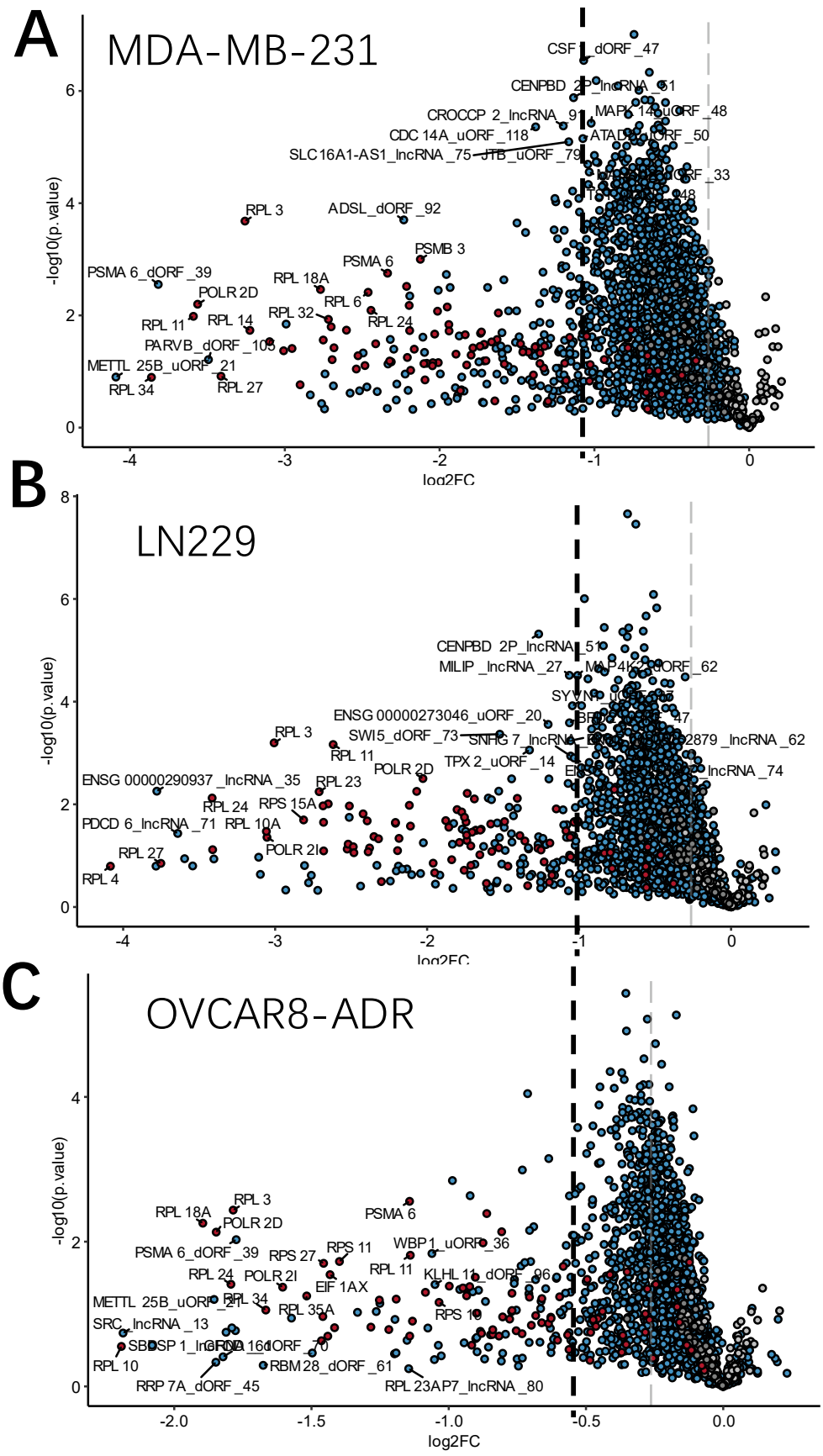


Figure 4-9 NGS results of CRISPR data from three cell lines, MDA-MB-231 (A),

LN229 (B), and OVCAR8-ADR (C). Red dots represent positive controls (essential genes), blue dots represent microproteins, and gray dots represent negative controls.

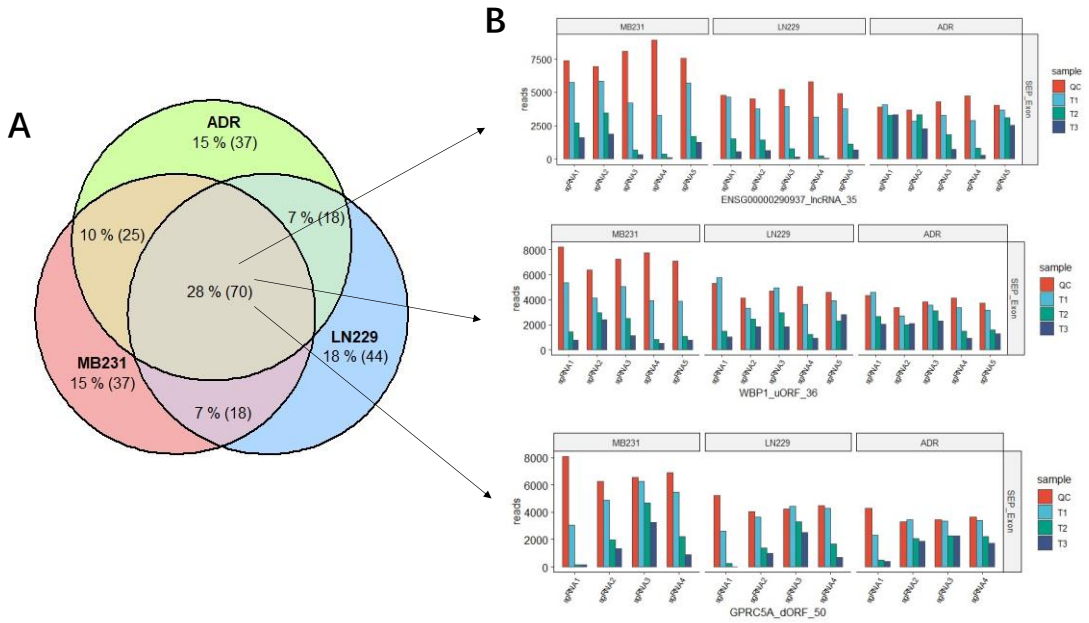


Figure 4-10 Comparison of CRISPR data across three cell lines. (A) Overlap of the top 150 significantly scored microproteins in each cell line. (B) Examples of CRISPR results for three microproteins.

4.3.7 Microproteins May Function Independently of Their Associated Canonical Proteins

To explore whether microproteins function independently or in conjunction with their associated canonical proteins, we compared our CRISPR screening results with data from the DepMap project,[185] where CRISPR screens target the entire canonical ORF. In contrast, our screening specifically targeted microproteins. As shown in **Figure 4-11B**, we plotted the depletion scores for microprotein-targeting sgRNAs (x-axis) against those for their corresponding canonical ORFs (y-axis). Interestingly, we observed two distinct patterns: for some genes, the knockout of their canonical ORFs in DepMap did not significantly impact cell proliferation, but the knockout of their associated microproteins showed strong depletion effects (Zone 1). This suggests that these microproteins may have functional roles independent of their canonical proteins. Conversely, we also observed cases where the knockout of canonical ORFs had significant effects on cell proliferation, whereas the knockout of their associated microproteins did not show similar effects (Zone 2). These findings suggest that the functions of certain microproteins are independent and not exerted by regulating the activity of nearby canonical proteins.

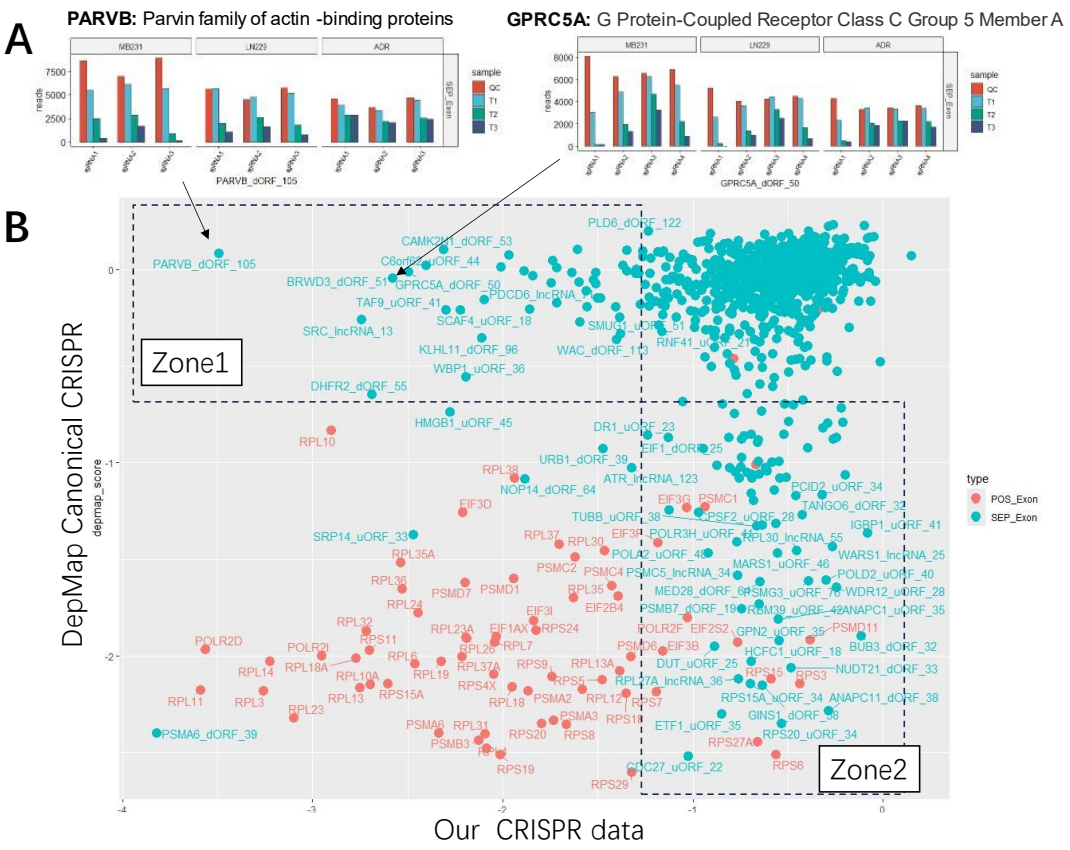


Figure 4-11 Comparison of our CRISPR results targeting microprotein KOs with public data from DepMap. (A) CRISPR results for two microproteins as examples. (B) Scatter plot comparing fold change between DepMap data and our CRISPR data. The y-axis represents the scores of canonical proteins in DepMap, while the x-axis represents the scores of the corresponding microproteins (e.g., uORF/dORF) in our data.

4.4 Conclusion

In this study, we conducted a comprehensive analysis of microproteins, identifying a total of 36,494 microproteins using immunopeptidomics across 86 datasets and over 5,000 raw data files. By leveraging advanced proteomics tools and integrating classification methods, we demonstrated that microproteins originate from diverse sources, the majority of which are derived from lncRNAs. Our findings also reveal that many microproteins are located near oncogenes, suggesting potential roles in cancer-related pathways. This highlights the importance of further investigating microproteins as potential regulators in tumorigenesis and cancer progression.

We further validated the reliability of microprotein identification through detailed spectral quality assessments. With over 8,000 microproteins supported by at least two unique immunopeptides and over 18,000 microproteins associated with multiple PSMs, we demonstrated that the spectral quality of microprotein identification is comparable to that of canonical proteins. The case of the PTEN uORF microprotein,[63] supported by nine unique immunopeptides, provides compelling evidence of the active translation and functional existence of microproteins.

Our results highlight the advantages of immunopeptidomics over traditional tryptic digestion methods, particularly for identifying low-abundance and diverse peptides. The method's independence from trypsin cleavage patterns, combined with HLA diversity across samples, greatly enhances the coverage and discovery of novel

microproteins. Importantly, many previously reported microproteins with known biological functions were also identified in our dataset, serving as positive controls and further validating our approach.

We also characterized microproteins in detail, finding that they are significantly shorter than canonical proteins (average length of 66 amino acids) and often possess hydrophobic tails that may make them more susceptible to degradation via the BAG6 system. Interestingly, our analysis revealed a distinct length threshold of 29 amino acids, with microproteins preferentially exceeding this length, although the biological significance of this observation remains unclear. Conservation analysis further revealed that microproteins are more prevalent in mammals, with higher conservation in primates, suggesting that microproteins may have evolved alongside advanced biological functions in primates.

To investigate the functional roles of microproteins, we performed CRISPR screening experiments in three cancer cell lines (MDA-MB-231, LN229, and OVCAR8-ADR) using a library of over 18,000 sgRNAs. Our results demonstrated that many microproteins play critical roles in cell proliferation, with a subset exhibiting effects comparable to canonical essential genes. Notably, approximately 50% of the most significant microproteins were shared across all three cell lines, indicating a conserved set of microproteins essential for cancer cell survival. Examples such as ENSG00000290937_lncRNA_35, WBP1_uORF_36, and GPRC5A_dORF_50 were

supported by both CRISPR data and immunopeptide evidence, further emphasizing their critical roles.

Finally, our comparison of CRISPR results with DepMap KO data revealed that many microproteins exhibit independent functional roles that are not mediated by their nearby canonical proteins. This suggests that microproteins are not merely byproducts of canonical transcription and translation but rather represent a distinct layer of functional regulation within the proteome.

In summary, we provide extensive evidence that microproteins are not only actively translated but also play critical and independent roles in various biological processes, including cancer progression. Our study highlights the importance of microproteins as a previously underexplored class of biomolecules with significant functional and therapeutic potential. Future studies should further investigate the mechanisms by which microproteins exert their functions, particularly in cancer and other diseases, to unlock their full biological and clinical implications.

Chapter 5. Overall conclusions and future perspectives

My research project has systematically explored the hidden proteome, focusing on non-canonical proteins such as small open reading frame-encoded peptides (SEPs), alternative proteins (AltProts), and cryptic immunopeptides. By integrating advanced mass spectrometry techniques, including data-independent acquisition (DIA) proteomics and immunopeptidomics, with bioinformatics and functional screening tools like CRISPR, we have addressed key challenges in identifying and characterizing these understudied biomolecules. The work presented here not only expands the boundaries of proteomics but also provides novel insights into their biological roles and therapeutic potential, particularly in developmental biology, immunology, and cancer research.

In Chapter 2, we developed and optimized a DIA-based workflow for the enhanced discovery of AltProts in mouse cardiac development. Traditional data-dependent acquisition (DDA) methods have been instrumental in mapping the human proteome but often fall short in detecting low-abundance, non-canonical proteins like AltProts due to their stochastic precursor selection and higher rates of missing values. To overcome these limitations, we systematically compared four spectral library construction strategies—DDA-fractionated libraries, gas-phase fractionation (GPF)-based libraries using DIA-Umpire and DIA-NN, and fully predicted libraries—and evaluated their performance using three DIA search engines. My results demonstrated that DIA outperforms DDA by achieving up to a twofold increase in AltProt

identification while reducing missing values by 50%. The use of predicted spectral libraries, generated via machine learning tools like Prosit and DIA-NN, proved particularly effective, offering high accuracy in retention time and fragmentation spectra without the need for extensive experimental fractionation, which is resource-intensive.

Applying this optimized DIA workflow to mouse heart samples from embryonic (E15.5) and adult (P42) stages, we identified over 50 differentially expressed AltProts, many of which were enriched in pathways related to cardiac development, such as protein synthesis, mitochondrial function, and cellular homeostasis. A notable example is ASDURF, an upstream open reading frame (uORF)-encoded AltProt, which we validated through parallel reaction monitoring (PRM), Western blotting, and overexpression studies. ASDURF's differential expression between embryonic and adult hearts suggests its regulatory role in cardiac maturation, potentially through modulating translation efficiency or interacting with canonical proteins. This chapter establishes a robust, reproducible framework for AltProt analysis, highlighting DIA's superiority in quantifying dynamic proteomic changes during development and providing a reference for future studies in organogenesis and disease models.

Building on this foundation, Chapter 3 introduced a novel Pseudo-DIA Library Search Strategy to improve immunopeptidomics and neoantigen discovery. Immunopeptides, presented on major histocompatibility complex (MHC) molecules, are critical for

immune surveillance and represent promising targets for cancer immunotherapy. However, conventional DDA-based approaches suffer from limited sensitivity and depth, particularly for cryptic immunopeptides derived from non-canonical sources like sORFs. My Pseudo-DIA strategy integrates unrestricted DIA searches with predicted spectral libraries, enabling the identification of up to 3.8 times more immunopeptides than traditional methods across cell lines such as JY, 0D5P, and RA957.

Mechanistic evaluations confirmed the strategy's reliability, with q-value distributions, spectral correlations, and motif analyses aligning closely with established benchmarks. By applying this method to public datasets, we identified 1,223 cryptic immunopeptides from 1,083 OpenProt IDs, with one-third originating from long non-coding RNA (lncRNA) regions. Many of these were linked to tumor-associated genes, such as the uORF of MORF4L2 and the intORF of ZNF146, showing differential expression in The Cancer Genome Atlas (TCGA) data and potential as neoantigens. Validation using independent datasets from cell lines like HCT116 and H358 further demonstrated the strategy's ability to detect mutation-derived neoantigens, such as those from CHMP7 (p.A324T) and NAPA (p.A181V), with high spectral quality.

To facilitate broader adoption, we developed a user-friendly executable tool for spectral library generation from MSFragger results, streamlining the workflow for immunopeptidomics research. This chapter underscores the power of Pseudo-DIA in expanding the immunopeptide landscape, offering new opportunities for personalized

immunotherapy by uncovering cryptic neoantigens that traditional methods overlook.

In Chapter 4, we extended my analysis to large-scale identification and functional validation of microproteins using immunopeptidomics and CRISPR screening. By reanalyzing 86 datasets comprising over 5,000 raw mass spectrometry files, we identified 36,494 microproteins—one of the largest datasets to date—categorized by genomic origins (e.g., lncRNAs, uORFs, intORFs) and start codon usage (predominantly ATG, followed by CTG and GTG). Notably, around 3,000 microproteins were located near oncogenes, suggesting their involvement in cancer biology. Spectral quality assessments, including q-value distributions and peptide-spectrum matches (PSMs), confirmed the reliability of these identifications, with examples like the PTEN uORF supported by nine unique immunopeptides.

Characterization revealed that microproteins are significantly shorter (average 66 amino acids) than canonical proteins and often feature hydrophobic C-terminal tails, potentially facilitating degradation via the BAG6 system. A distinct length threshold at 29 amino acids was observed, warranting further investigation. Conservation analysis across 100 species showed higher prevalence in mammals, particularly primates, indicating evolutionary specialization.

To assess functionality, we constructed a CRISPR knockout (KO) library targeting over 2,000 microproteins, selected based on TCGA expression and survival data. Screening in three cancer cell lines (MDA-MB-231, LN229, OVCAR8-ADR) identified many

microproteins critical for proliferation, with ~50% overlap across lines, including ENSG00000290937_lncRNA_35 and GPRC5A_dORF_50. Comparison with DepMap data revealed that some microproteins function independently of associated canonical proteins, as their KO effects did not correlate with canonical ORF disruptions.

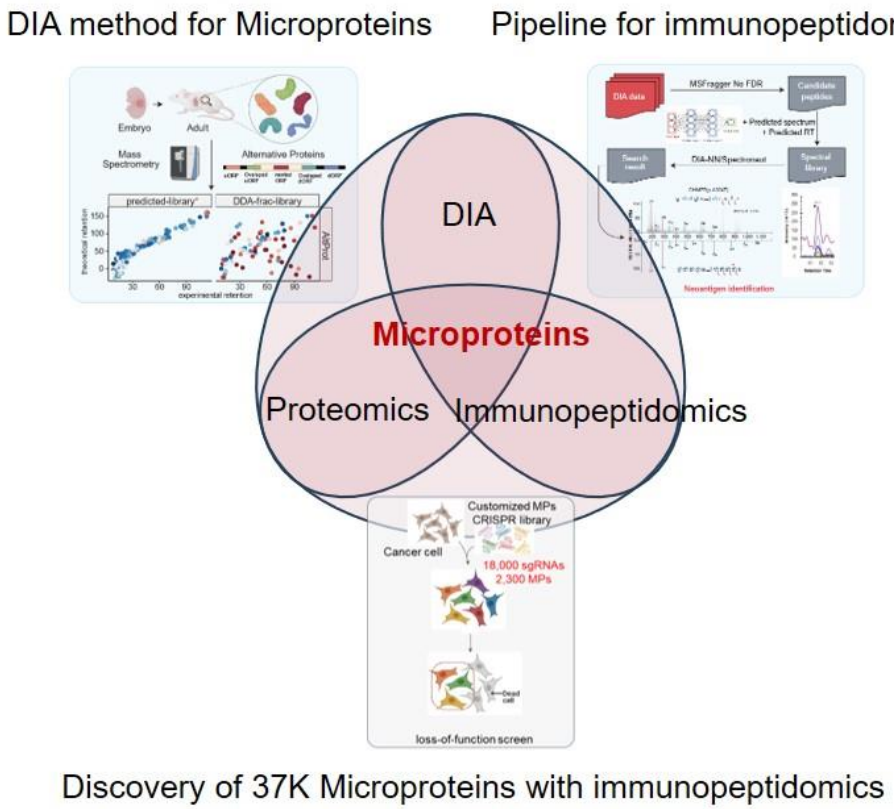


Figure 5-1 Diagrammatic Summary and Interconnections of the Three Research Projects.

Overall, my research project advances our understanding of the hidden proteome by developing innovative methodologies that enhance the detection and characterization of SEPs, AltProts, and immunopeptides. The optimized DIA workflows in Chapters 2 and 3 provide scalable tools for proteomics and immunopeptidomics, while Chapter 4's large-scale functional screening demonstrates the biological independence and relevance of microproteins. These findings challenge traditional proteome annotations

and highlight non-canonical proteins as key regulators in development, immunity, and disease.

Looking forward, the field of microprotein research holds immense promise but requires continued exploration to fully unravel their complexities. First, mining new microproteins remains a priority. While my study identified over 36,000 microproteins, this likely represents only a fraction of the hidden proteome. Future efforts should explore additional potential microproteins by integrating multi-omics data, such as ribosome profiling (Ribo-seq) with advanced proteomics, to predict and validate sORFs in unannotated genomic regions. Seeking evidence from diverse perspectives is essential; for instance, Western blotting (WB) with custom antibodies could confirm protein expression in specific tissues or cell types, complementing mass spectrometry's high-throughput nature. Investigating subcellular localization using confocal microscopy and fluorescent tagging (e.g., GFP-fusions) would reveal where microproteins exert their functions—whether in mitochondria, lysosomes, or the nucleus—providing clues to their mechanisms. Furthermore, characterizing molecular structures through techniques like X-ray crystallography or cryo-electron microscopy (cryo-EM) could elucidate 3D conformations and interaction interfaces, facilitating drug design targeting microproteins.

Second, delving deeper into microprotein functions is crucial. My CRISPR screens identified functional microproteins, but expanding these to more cell lines, tissues, and

disease models (e.g., patient-derived organoids) could uncover context-specific roles. High-throughput CRISPR activation/inhibition libraries might reveal gain-of-function phenotypes, complementing KO approaches. Studying clinical relevance, particularly with oncogenes, is imperative; for example, correlating microprotein expression with TCGA survival data or tumor progression markers could identify biomarkers for prognosis or therapy response. Given their proximity to oncogenes, microproteins may regulate pathways like RAS signaling or DNA repair, offering novel therapeutic targets. Finally, discovering neoantigens derived from microproteins represents a transformative opportunity in immunotherapy. Building on my Pseudo-DIA strategy, future work could integrate HLA typing with microprotein-derived peptide predictions to design personalized vaccines. Clinical trials testing microprotein neoantigens in cancer patients could validate their immunogenicity and efficacy, potentially revolutionizing treatments for immunotherapy-resistant tumors.

In conclusion, my research project lays a strong foundation for non-canonical proteome research, but the journey is far from complete. By pursuing these perspectives, we can unlock the full potential of microproteins, bridging fundamental biology with translational applications in health and disease.

References:

1. Mudge, J. M.; Ruiz-Orera, J.; Prensner, J. R.; Brunet, M. A.; Calvet, F.; Jungreis, I.; Gonzalez, J. M.; Magrane, M.; Martinez, T. F.; Schulz, J. F.; Yang, Y. T.; Albà, M. M.; Aspden, J. L.; Baranov, P. V.; Bazzini, A. A.; Bruford, E.; Martin, M. J.; Calviello, L.; Carvunis, A. R.; Chen, J.; Couso, J. P.; Deutsch, E. W.; Flicek, P.; Frankish, A.; Gerstein, M.; Hubner, N.; Ingolia, N. T.; Kellis, M.; Menschaert, G.; Moritz, R. L.; Ohler, U.; Roucou, X.; Saghatelian, A.; Weissman, J. S.; van Heesch, S., Standardized annotation of translated open reading frames. *Nat Biotechnol* **2022,** *40* (7), 994-999.
2. D'Lima, N. G.; Ma, J.; Winkler, L.; Chu, Q.; Loh, K. H.; Corpuz, E. O.; Budnik, B. A.; Lykke-Andersen, J.; Saghatelian, A.; Slavoff, S. A., A human microprotein that interacts with the mRNA decapping complex. *Nat Chem Biol* **2017,** *13* (2), 174-180.
3. Ray, S.; Rosenberg, M. I.; Chanut-Delalande, H.; Decaras, A.; Schwertner, B.; Toubiana, W.; Auman, T.; Schnellhammer, I.; Teuscher, M.; Valenti, P.; Khila, A.; Klingler, M.; Payre, F., The mlpt/Ubr3/Svb module comprises an ancient developmental switch for embryonic patterning. *Elife* **2019,** *8*.
4. Niu, L.; Lou, F.; Sun, Y.; Sun, L.; Cai, X.; Liu, Z.; Zhou, H.; Wang, H.; Wang, Z.; Bai, J.; Yin, Q.; Zhang, J.; Chen, L.; Peng, D.; Xu, Z.; Gao, Y.; Tang, S.; Fan, L.; Wang, H., A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Sci Adv* **2020,** *6* (21), eaaz2059.
5. Matsumoto, A.; Pasut, A.; Matsumoto, M.; Yamashita, R.; Fung, J.; Monteleone, E.; Saghatelian, A.; Nakayama, K. I.; Clohessy, J. G.; Pandolfi, P. P., mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **2017,** *541* (7636), 228-232.
6. Stein, C. S.; Jadiya, P.; Zhang, X.; McLendon, J. M.; Abouassaly, G. M.; Witmer, N. H.; Anderson, E. J.; Elrod, J. W.; Boudreau, R. L., Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep* **2018,** *23* (13), 3710-3720.e8.
7. Zhang, S.; Reljić, B.; Liang, C.; Kerouanton, B.; Francisco, J. C.; Peh, J. H.; Mary, C.; Jagannathan, N. S.; Olexiouk, V.; Tang, C.; Fidelito, G.; Nama, S.; Cheng, R. K.; Wee, C. L.; Wang, L. C.; Duek Roggli, P.; Sampath, P.; Lane, L.; Petretto, E.; Sobota, R. M.; Jesuthasan, S.; Tucker-Kellogg, L.; Reversade, B.; Menschaert, G.; Sun, L.; Stroud, D. A.; Ho, L., Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat Commun* **2020,** *11* (1), 1312.
8. Pauli, A.; Norris, M. L.; Valen, E.; Chew, G. L.; Gagnon, J. A.; Zimmerman, S.; Mitchell, A.; Ma, J.; Dubrulle, J.; Reyon, D.; Tsai, S. Q.; Joung, J. K.; Saghatelian, A.; Schier, A. F., Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **2014,** *343* (6172), 1248636.
9. Guo, B.; Zhai, D.; Cabezas, E.; Welsh, K.; Nouraini, S.; Satterthwait, A. C.; Reed, J. C., Humanin peptide suppresses apoptosis by interfering with Bax activation.

Nature **2003,** *423* (6938), 456-61.
10. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016,** *537* (7620), 347-55.
11. Dettmer, K.; Aronov, P. A.; Hammock, B. D., Mass spectrometry-based metabolomics. *Mass Spectrom Rev* **2007,** *26* (1), 51-78.
12. Cajka, T.; Fiehn, O., Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal Chem* **2016,** *88* (1), 524-45.
13. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** *422* (6928), 198-207.
14. McDonald, W. H.; Yates, J. R., 3rd, Shotgun proteomics and biomarker discovery. *Dis Markers* **2002,** *18* (2), 99-105.
15. Ludwig, C.; Gillet, L.; Rosenberger, G.; Amon, S.; Collins, B. C.; Aebersold, R., Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol* **2018,** *14* (8), e8126.
16. Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **2012,** *11* (6), O111.016717.
17. Poulos, R. C.; Hains, P. G.; Shah, R.; Lucas, N.; Xavier, D.; Manda, S. S.; Anees, A.; Koh, J. M. S.; Mahboob, S.; Wittman, M.; Williams, S. G.; Sykes, E. K.; Hecker, M.; Dausmann, M.; Wouters, M. A.; Ashman, K.; Yang, J.; Wild, P. J.; deFazio, A.; Balleine, R. L.; Tully, B.; Aebersold, R.; Speed, T. P.; Liu, Y.; Reddel, R. R.; Robinson, P. J.; Zhong, Q., Strategies to enable large-scale proteomics for reproducible research. *Nat Commun* **2020,** *11* (1), 3793.
18. Bekker-Jensen, D. B.; Martínez-Val, A.; Steigerwald, S.; Rüther, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V., A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol Cell Proteomics* **2020,** *19* (4), 716-729.
19. Bruderer, R.; Sondermann, J.; Tsou, C. C.; Barrantes-Freer, A.; Stadelmann, C.; Nesvizhskii, A. I.; Schmidt, M.; Reiter, L.; Gomez-Varela, D., New targeted approaches for the quantification of data-independent acquisition mass spectrometry. *Proteomics* **2017,** *17* (9).
20. Messner, C. B.; Demichev, V.; Bloomfield, N.; Yu, J. S. L.; White, M.; Kreidl, M.; Egger, A. S.; Freiwald, A.; Ivosev, G.; Wasim, F.; Zelezniak, A.; Jürgens, L.; Suttorp, N.; Sander, L. E.; Kurth, F.; Lilley, K. S.; Mülleder, M.; Tate, S.; Ralser, M., Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol* **2021,** *39* (7), 846-854.
21. Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Xuan, Y.; Sondermann, J.; Schmidt, M.; Gomez-Varela, D.; Reiter, L., Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol Cell Proteomics* **2017,** *16* (12), 2296-2309.
22. Decaestecker, T. N.; Vande Casteele, S. R.; Wallemacq, P. E.; Van Peteghem, C. H.; Defore, D. L.; Van Bocxlaer, J. F., Information-dependent acquisition-mediated LC-MS/MS screening procedure with semiquantitative potential. *Anal Chem* **2004,** *76*

(21), 6365-73.
23. Cai, X.; Ge, W.; Yi, X.; Sun, R.; Zhu, J.; Lu, C.; Sun, P.; Zhu, T.; Ruan, G.; Yuan, C.; Liang, S.; Lyu, M.; Huang, S.; Zhu, Y.; Guo, T., PulseDIA: Data-Independent Acquisition Mass Spectrometry Using Multi-Injection Pulsed Gas-Phase Fractionation. *J Proteome Res* **2021,** *20* (1), 279-288.
24. Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J., Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nat Protoc* **2015,** *10* (6), 887-903.
25. Meier, F.; Brunner, A. D.; Frank, M.; Ha, A.; Bludau, I.; Voytik, E.; Kaspar-Schoenefeld, S.; Lubeck, M.; Raether, O.; Bache, N.; Aebersold, R.; Collins, B. C.; Röst, H. L.; Mann, M., diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. *Nat Methods* **2020,** *17* (12), 1229-1236.
26. Derks, J.; Leduc, A.; Huffman, R. G.; Specht, H.; Ralser, M.; Demichev, V.; Slavov, N., Increasing the throughput of sensitive proteomics by plexDIA. *bioRxiv* **2021**, 2021.11.03.467007.
27. Guo, T.; Kouvonen, P.; Koh, C. C.; Gillet, L. C.; Wolski, W. E.; Röst, H. L.; Rosenberger, G.; Collins, B. C.; Blum, L. C.; Gillessen, S.; Joerger, M.; Jochum, W.; Aebersold, R., Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med* **2015,** *21* (4), 407-13.
28. Tabb, D. L., The SEQUEST family tree. *J Am Soc Mass Spectrom* **2015,** *26* (11), 1814-9.
29. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** *26* (12), 1367-72.
30. Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **2017,** *14* (5), 513-520.
31. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007,** *7* (5), 655-67.
32. Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L., Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin Appl* **2015,** *9* (7-8), 745-54.
33. Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B. C.; Payne, S. H.; Noble, W. S.; MacCoss, M. J., PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods* **2017,** *14* (9), 903-908.
34. Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villén, J.; MacCoss, M. J., Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat Commun* **2018,** *9* (1), 5128.
35. Lu, Y. Y.; Bilmes, J.; Rodriguez-Mias, R. A.; Villén, J.; Noble, W. S., DIAmeter: matching peptides to data-independent acquisition mass spectrometry data.

Bioinformatics **2021,** *37* (Suppl_1), i434-i442.
36. Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **2015,** *12* (3), 258-64, 7 p following 264.
37. Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H. C.; Aiche, S.; Kuster, B.; Wilhelm, M., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **2019,** *16* (6), 509-518.
38. Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **2020,** *17* (1), 41-44.
39. Yang, Y.; Liu, X.; Shen, C.; Lin, Y.; Yang, P.; Qiao, L., In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat Commun* **2020,** *11* (1), 146.
40. Zhou, X. X.; Zeng, W. F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S. M.; Zhang, Z., pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal Chem* **2017,** *89* (23), 12690-12697.
41. Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; Aebersold, R., OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **2014,** *32* (3), 219-23.
42. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-8.
43. Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M., Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* **2019,** *16* (1), 63-66.
44. Vitsios, D.; Dhindsa, R. S.; Middleton, L.; Gussow, A. B.; Petrovski, S., Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nature communications* **2021,** *12* (1), 1-14.
45. Statello, L.; Guo, C.-J.; Chen, L.-L.; Huarte, M., Gene regulation by long non-coding RNAs and its biological functions. *Nature reviews Molecular cell biology* **2021,** *22* (2), 96-118.
46. Lu, S.; Zhang, J.; Lian, X.; Sun, L.; Meng, K.; Chen, Y.; Sun, Z.; Yin, X.; Li, Y.; Zhao, J., A hidden human proteome encoded by 'non-coding'genes. *Nucleic acids research* **2019,** *47* (15), 8111-8125.
47. Kung, J. T.; Colognori, D.; Lee, J. T., Long noncoding RNAs: past, present, and future. *Genetics* **2013,** *193* (3), 651-69.
48. Sandmann, C. L.; Schulz, J. F.; Ruiz-Orera, J.; Kirchner, M.; Ziehm, M.; Adami, E.; Marczenke, M.; Christ, A.; Liebe, N.; Greiner, J.; Schoenenberger, A.; Muecke, M. B.; Liang, N.; Moritz, R. L.; Sun, Z.; Deutsch, E. W.;

Gotthardt, M.; Mudge, J. M.; Prensner, J. R.; Willnow, T. E.; Mertins, P.; van Heesch, S.; Hubner, N., Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol Cell* **2023,** *83* (6), 994-1011.e18.
49. Morgado-Palacin, L.; Brown, J. A.; Martinez, T. F.; Garcia-Pedrero, J. M.; Forouhar, F.; Quinn, S. A.; Reglero, C.; Vaughan, J.; Heydary, Y. H.; Donaldson, C.; Rodriguez-Perales, S.; Allonca, E.; Granda-Diaz, R.; Fernandez, A. F.; Fraga, M. F.; Kim, A. L.; Santos-Juanes, J.; Owens, D. M.; Rodrigo, J. P.; Saghatelian, A.; Ferrando, A. A., The TINCR ubiquitin-like microprotein is a tumor suppressor in squamous cell carcinoma. *Nat Commun* **2023,** *14* (1), 1328.
50. Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S., A draft map of the human proteome. *Nature* **2014,** *509* (7502), 575-581.
51. Thul, P. J.; Åkesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Björk, L.; Breckels, L. M., A subcellular map of the human proteome. *Science* **2017,** *356* (6340), eaal3321.
52. Adhikari, S.; Nice, E. C.; Deutsch, E. W.; Lane, L.; Omenn, G. S.; Pennington, S. R.; Paik, Y.-K.; Overall, C. M.; Corrales, F. J.; Cristea, I. M., A high-stringency blueprint of the human proteome. *Nature communications* **2020,** *11* (1), 1-16.
53. Zhang, Y.; Lin, Z.; Hao, P.; Hou, K.; Sui, Y.; Zhang, K.; He, Y.; Li, H.; Yang, H.; Liu, S.; Ren, Y., Improvement of Peptide Separation for Exploring the Missing Proteins Localized on Membranes. *J Proteome Res* **2018,** *17* (12), 4152-4159.
54. Zhang, Y.; Lin, Z.; Tan, Y.; Bu, F.; Hao, P.; Zhang, K.; Yang, H.; Liu, S.; Ren, Y., Exploration of Missing Proteins by a Combination Approach to Enrich the Low-Abundance Hydrophobic Proteins from Four Cancer Cell Lines. *J Proteome Res* **2020,** *19* (1), 401-408.
55. Zhang, Y.; Zhang, K.; Bu, F.; Hao, P.; Yang, H.; Liu, S.; Ren, Y., D283 Med, a Cell Line Derived from Peritoneal Metastatic Medulloblastoma: A Good Choice for Missing Protein Discovery. *J Proteome Res* **2020,** *19* (12), 4857-4866.
56. Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A., Tissue-based map of the human proteome. *Science* **2015,** *347* (6220), 1260419.
57. Wang, B.; Wang, Z.; Pan, N.; Huang, J.; Wan, C., Improved identification of small open reading frames encoded peptides by top-down proteomic approaches and de novo sequencing. *International journal of molecular sciences* **2021,** *22* (11), 5476.
58. Ma, J.; Ward, C. C.; Jungreis, I.; Slavoff, S. A.; Schwaid, A. G.; Neveu, J.; Budnik, B. A.; Kellis, M.; Saghatelian, A., Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *Journal of proteome research* **2014,** *13* (3), 1757-1765.
59. Zhu, Y.; Orre, L. M.; Johansson, H. J.; Huss, M.; Boekel, J.; Vesterlund, M.; Fernandez-Woodbridge, A.; Branca, R. M.; Lehtiö, J., Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nature communications* **2018,** *9* (1), 1-14.

60. Egertson, J. D.; Kuehn, A.; Merrihew, G. E.; Bateman, N. W.; MacLean, B. X.; Ting, Y. S.; Canterbury, J. D.; Marsh, D. M.; Kellmann, M.; Zabrouskov, V., Multiplexed MS/MS for improved data-independent acquisition. *Nature methods* **2013,** *10* (8), 744-746.
61. Barkovits, K.; Pacharra, S.; Pfeiffer, K.; Steinbach, S.; Eisenacher, M.; Marcus, K.; Uszkoreit, J., Reproducibility, specificity and accuracy of relative quantification using spectral library-based data-independent acquisition. *Molecular & Cellular Proteomics* **2020,** *19* (1), 181-197.
62. Müller, F.; Kolbowski, L.; Bernhardt, O. M.; Reiter, L.; Rappsilber, J., Data-independent Acquisition Improves Quantitative Cross-linking Mass Spectrometry*[S]. *Molecular & Cellular Proteomics* **2019,** *18* (4), 786-795.
63. Martinez, T. F.; Lyons-Abbott, S.; Bookout, A. L.; De Souza, E. V.; Donaldson, C.; Vaughan, J. M.; Lau, C.; Abramov, A.; Baquero, A. F.; Baquero, K.; Friedrich, D.; Huard, J.; Davis, R.; Kim, B.; Koch, T.; Mercer, A. J.; Misquith, A.; Murray, S. A.; Perry, S.; Pino, L. K.; Sanford, C.; Simon, A.; Zhang, Y.; Zipp, G.; Bizarro, C. V.; Shokhirev, M. N.; Whittle, A. J.; Searle, B. C.; MacCoss, M. J.; Saghatelian, A.; Barnes, C. A., Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. *Cell Metab* **2023,** *35* (1), 166-183.e11.
64. Wang, H.; Wang, Y.; Yang, J.; Zhao, Q.; Tang, N.; Chen, C.; Li, H.; Cheng, C.; Xie, M.; Yang, Y.; Xie, Z., Tissue- and stage-specific landscape of the mouse translatome. *Nucleic Acids Res* **2021,** *49* (11), 6165-6180.
65. Martinez, T. F.; Chu, Q.; Donaldson, C.; Tan, D.; Shokhirev, M. N.; Saghatelian, A., Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol* **2020,** *16* (4), 458-468.
66. Navarro, P.; Kuharev, J.; Gillet, L. C.; Bernhardt, O. M.; MacLean, B.; Röst, H. L.; Tate, S. A.; Tsou, C.-C.; Reiter, L.; Distler, U., A multicenter study benchmarks software tools for label-free proteome quantification. *Nature biotechnology* **2016,** *34* (11), 1130-1136.
67. Cloutier, P.; Poitras, C.; Faubert, D.; Bouchard, A.; Blanchette, M.; Gauthier, M. S.; Coulombe, B., Upstream ORF-Encoded ASDURF Is a Novel Prefoldin-like Subunit of the PAQosome. *J Proteome Res* **2020,** *19* (1), 18-27.
68. Pino, L. K.; Just, S. C.; MacCoss, M. J.; Searle, B. C., Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Molecular & Cellular Proteomics* **2020,** *19* (7), 1088-1103.
69. Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature methods* **2017,** *14* (5), 513-520.
70. Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* **2015,** *12* (3), 258-264.
71. Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature methods* **2020,** *17* (1), 41-44.

72. Ge, W.; Liang, X.; Zhang, F.; Hu, Y.; Xu, L.; Xiang, N.; Sun, R.; Liu, W.; Xue, Z.; Yi, X., Computational optimization of spectral library size improves DIA-MS proteome coverage and applications to 15 tumors. *Journal of Proteome Research* **2021,** *20* (12), 5392-5401.
73. Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* **2007,** *4* (11), 923-925.
74. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-968.
75. Boratyn, G. M.; Camacho, C.; Cooper, P. S.; Coulouris, G.; Fong, A.; Ma, N.; Madden, T. L.; Matten, W. T.; McGinnis, S. D.; Merezhuk, Y., BLAST: a more efficient report with usability improvements. *Nucleic acids research* **2013,** *41* (W1), W29-W33.
76. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L., clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2021,** *2* (3), 100141.
77. Toprak, U. H.; Gillet, L. C.; Maiolica, A.; Navarro, P.; Leitner, A.; Aebersold, R., Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Molecular & Cellular Proteomics* **2014,** *13* (8), 2056-2071.
78. Li, K.; Jain, A.; Malovannaya, A.; Wen, B.; Zhang, B., DeepRescore: leveraging deep learning to improve peptide identification in immunopeptidomics. *Proteomics* **2020,** *20* (21-22), 1900334.
79. Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods* **2019,** *16* (6), 509-518.
80. Van Puyvelde, B.; Daled, S.; Willems, S.; Gabriels, R.; Gonzalez de Peredo, A.; Chaoui, K.; Mouton-Barbosa, E.; Bouyssié, D.; Boonen, K.; Hughes, C. J., A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Scientific data* **2022,** *9* (1), 1-12.
81. Jiang, L.; Wang, M.; Lin, S.; Jian, R.; Li, X.; Chan, J.; Dong, G.; Fang, H.; Robinson, A. E.; Aguet, F., A quantitative proteome map of the human body. *Cell* **2020,** *183* (1), 269-283. e19.
82. Wang, H.; Wang, Y.; Yang, J.; Zhao, Q.; Tang, N.; Chen, C.; Li, H.; Cheng, C.; Xie, M.; Yang, Y., Tissue-and stage-specific landscape of the mouse translatome. *Nucleic acids research* **2021,** *49* (11), 6165-6180.
83. Brunet, M. A.; Lucier, J.-F.; Levesque, M.; Leblanc, S.; Jacques, J.-F.; Al-Saedi, H. R.; Guilloy, N.; Grenier, F.; Avino, M.; Fournier, I., OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Research* **2021,** *49* (D1), D380-D388.
84. Olexiouk, V.; Van Criekinge, W.; Menschaert, G., An update on sORFs. org: a

repository of small ORFs identified by ribosome profiling. *Nucleic acids research* **2018,** *46* (D1), D497-D502.
85. Hao, Y.; Zhang, L.; Niu, Y.; Cai, T.; Luo, J.; He, S.; Zhang, B.; Zhang, D.; Qin, Y.; Yang, F., SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Briefings in Bioinformatics* **2018,** *19* (4), 636-643.
86. Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villén, J.; MacCoss, M. J., Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature communications* **2018,** *9* (1), 1-12.
87. Consortium, U., UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **2019,** *47* (D1), D506-D515.
88. Yang, Y.; Liu, X.; Shen, C.; Lin, Y.; Yang, P.; Qiao, L., In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature communications* **2020,** *11* (1), 1-11.
89. Searle, B. C.; Swearingen, K. E.; Barnes, C. A.; Schmidt, T.; Gessulat, S.; Küster, B.; Wilhelm, M., Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature communications* **2020,** *11* (1), 1-10.
90. Lu, Y. Y.; Bilmes, J.; Rodriguez-Mias, R. A.; Villén, J.; Noble, W. S., DIAmeter: matching peptides to data-independent acquisition mass spectrometry data. *Bioinformatics* **2021,** *37* (Supplement_1), i434-i442.
91. Yang, K. L.; Yu, F.; Teo, G. C.; Li, K.; Demichev, V.; Ralser, M.; Nesvizhskii, A. I., MSBooster: improving peptide identification rates using deep learning-based features. *Nat Commun* **2023,** *14* (1), 4539.
92. Yang, Y.; Wang, H.; Zhang, Y.; Chen, L.; Chen, G.; Bao, Z.; Yang, Y.; Xie, Z.; Zhao, Q., An Optimized Proteomics Approach Reveals Novel Alternative Proteins in Mouse Liver Development. *Mol Cell Proteomics* **2023,** *22* (1), 100480.
93. Na, C. H.; Barbhuiya, M. A.; Kim, M.-S.; Verbruggen, S.; Eacker, S. M.; Pletnikova, O.; Troncoso, J. C.; Halushka, M. K.; Menschaert, G.; Overall, C. M., Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. *Genome research* **2018,** *28* (1), 25-36.
94. Sun, L.; Wang, W.; Han, C.; Huang, W.; Sun, Y.; Fang, K.; Zeng, Z.; Yang, Q.; Pan, Q.; Chen, T.; Luo, X.; Chen, Y., The oncomicropeptide APPLE promotes hematopoietic malignancy by enhancing translation initiation. *Mol Cell* **2021,** *81* (21), 4493-4508.e9.
95. Li, Y.; Zhang, J.; Sun, H.; Chen, Y.; Li, W.; Yu, X.; Zhao, X.; Zhang, L.; Yang, J.; Xin, W.; Jiang, Y.; Wang, G.; Shi, W.; Zhu, D., lnc-Rps4l-encoded peptide RPS4XL regulates RPS6 phosphorylation and inhibits the proliferation of PASMCs caused by hypoxia. *Mol Ther* **2021,** *29* (4), 1411-1424.
96. Cloutier, P.; Poitras, C.; Faubert, D.; Bouchard, A.; Blanchette, M.; Gauthier, M.-S.; Coulombe, B., Upstream ORF-encoded ASDURF is a novel prefoldin-like subunit of the PAQosome. *Journal of proteome research* **2019,** *19* (1), 18-27.
97. Coulombe, B.; Cloutier, P.; Pinard, M.; Forget, D.; Poitras, C.; Gauthier, M.-S., The PAQosome, a novel molecular chaperoning machine for assembly of human protein complexes and networks. *The FASEB Journal* **2020,** *34* (S1), 1-1.

98. Gauthier, M. S.; Cloutier, P.; Coulombe, B., Role of the PAQosome in Regulating Arrangement of Protein Quaternary Structure in Health and Disease. *Adv Exp Med Biol* **2018,** *1106*, 25-36.
99. Hofman, D. A.; Ruiz-Orera, J.; Yannuzzi, I.; Murugesan, R.; Brown, A.; Clauser, K. R.; Condurat, A. L.; van Dinter, J. T.; Engels, S. A. G.; Goodale, A.; van der Lugt, J.; Abid, T.; Wang, L.; Zhou, K. N.; Vogelzang, J.; Ligon, K. L.; Phoenix, T. N.; Roth, J. A.; Root, D. E.; Hubner, N.; Golub, T. R.; Bandopadhayay, P.; van Heesch, S.; Prensner, J. R., Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. *Mol Cell* **2024,** *84* (2), 261-276.e18.
100. Pishesha, N.; Harmand, T. J.; Ploegh, H. L., A guide to antigen processing and presentation. *Nat Rev Immunol* **2022,** *22* (12), 751-764.
101. Bassani-Sternberg, M.; Coukos, G., Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr Opin Immunol* **2016,** *41*, 9-17.
102. Prensner, J. R.; Abelin, J. G.; Kok, L. W.; Clauser, K. R.; Mudge, J. M.; Ruiz-Orera, J.; Bassani-Sternberg, M.; Moritz, R. L.; Deutsch, E. W.; van Heesch, S., What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome? *Mol Cell Proteomics* **2023,** *22* (9), 100631.
103. Bianchi, V.; Harari, A.; Coukos, G., Neoantigen-Specific Adoptive Cell Therapies for Cancer: Making T-Cell Products More Personal. *Front Immunol* **2020,** *11*, 1215.
104. Hu, Z.; Ott, P. A.; Wu, C. J., Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat Rev Immunol* **2018,** *18* (3), 168-182.
105. Schumacher, T. N.; Scheper, W.; Kvistborg, P., Cancer Neoantigens. *Annu Rev Immunol* **2019,** *37*, 173-200.
106. Zeng, W. F.; Zhou, X. X.; Willems, S.; Ammar, C.; Wahle, M.; Bludau, I.; Voytik, E.; Strauss, M. T.; Mann, M., AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat Commun* **2022,** *13* (1), 7238.
107. Chen, L.; Zhang, Y.; Yang, Y.; Yang, Y.; Li, H.; Dong, X.; Wang, H.; Xie, Z.; Zhao, Q., An Integrated Approach for Discovering Noncanonical MHC-I Peptides Encoded by Small Open Reading Frames. *J Am Soc Mass Spectrom* **2021,** *32* (9), 2346-2357.
108. Eng, J. K.; Hoopmann, M. R.; Jahan, T. A.; Egertson, J. D.; Noble, W. S.; MacCoss, M. J., A deeper look into Comet--implementation and features. *J Am Soc Mass Spectrom* **2015,** *26* (11), 1865-74.
109. Caron, E.; Kowalewski, D. J.; Chiek Koh, C.; Sturm, T.; Schuster, H.; Aebersold, R., Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol Cell Proteomics* **2015,** *14* (12), 3105-17.
110. Zhang, Y.; Yang, Y.; Li, K.; Chen, L.; Yang, Y.; Yang, C.; Xie, Z.; Wang, H.; Zhao, Q., Enhanced Discovery of Alternative Proteins (AltProts) in Mouse Cardiac Development Using Data-Independent Acquisition (DIA) Proteomics. *Anal Chem* **2025,** *97* (3), 1517-1527.

111. Lou, R.; Tang, P.; Ding, K.; Li, S.; Tian, C.; Li, Y.; Zhao, S.; Zhang, Y.; Shui, W., Hybrid Spectral Library Combining DIA-MS Data and a Targeted Virtual Library Substantially Deepens the Proteome Coverage. *iScience* **2020,** *23* (3), 100903.
112. Bichmann, L.; Gupta, S.; Röst, H., Data-Independent Acquisition Peptidomics. *Methods Mol Biol* **2024,** *2758*, 77-88.
113. Pak, H.; Michaux, J.; Huber, F.; Chong, C.; Stevenson, B. J.; Müller, M.; Coukos, G.; Bassani-Sternberg, M., Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition, and MS/MS Prediction. *Mol Cell Proteomics* **2021,** *20*, 100080.
114. Cai, Y.; Li, D.; Lv, D.; Yu, J.; Ma, Y.; Jiang, T.; Ding, N.; Liu, Z.; Li, Y.; Xu, J., MHC-I-presented non-canonical antigens expand the cancer immunotherapy targets in acute myeloid leukemia. *Sci Data* **2024,** *11* (1), 831.
115. Lin, M. J.; Svensson-Arvelund, J.; Lubitz, G. S.; Marabelle, A.; Melero, I.; Brown, B. D.; Brody, J. D., Cancer vaccines: the next immunotherapy frontier. *Nat Cancer* **2022,** *3* (8), 911-926.
116. Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; Nielsen, M., NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **2020,** *48* (W1), W449-w454.
117. Bassani-Sternberg, M.; Gfeller, D., Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions. *J Immunol* **2016,** *197* (6), 2492-9.
118. Gfeller, D.; Schmidt, J.; Croce, G.; Guillaume, P.; Bobisse, S.; Genolet, R.; Queiroz, L.; Cesbron, J.; Racle, J.; Harari, A., Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8(+) T-cell epitopes. *Cell Syst* **2023,** *14* (1), 72-83.e5.
119. Albert, B. A.; Yang, Y.; Shao, X. M.; Singh, D.; Smit, K. N.; Anagnostou, V.; Karchin, R., Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nat Mach Intell* **2023,** *5* (8), 861-872.
120. Leblanc, S.; Yala, F.; Provencher, N.; Lucier, J. F.; Levesque, M.; Lapointe, X.; Jacques, J. F.; Fournier, I.; Salzet, M.; Ouangraoua, A.; Scott, M. S.; Boisvert, F. M.; Brunet, M. A.; Roucou, X., OpenProt 2.0 builds a path to the functional characterization of alternative proteins. *Nucleic Acids Res* **2024,** *52* (D1), D522-d528.
121. Goldman, M. J.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A. N.; Zhu, J.; Haussler, D., Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* **2020,** *38* (6), 675-678.
122. Tate, J. G.; Bamford, S.; Jubb, H. C.; Sondka, Z.; Beare, D. M.; Bindal, N.; Boutselakis, H.; Cole, C. G.; Creatore, C.; Dawson, E.; Fish, P.; Harsha, B.; Hathaway, C.; Jupe, S. C.; Kok, C. Y.; Noble, K.; Ponting, L.; Ramshaw, C. C.; Rye, C. E.; Speedy, H. E.; Stefancsik, R.; Thompson, S. L.; Wang, S.; Ward, S.; Campbell, P. J.; Forbes, S. A., COSMIC: the Catalogue Of Somatic Mutations In Cancer.

Nucleic Acids Res **2019,** *47* (D1), D941-d947.
123. Schmidt, T.; Samaras, P.; Dorfer, V.; Panse, C.; Kockmann, T.; Bichmann, L.; van Puyvelde, B.; Perez-Riverol, Y.; Deutsch, E. W.; Kuster, B.; Wilhelm, M., Universal Spectrum Explorer: A Standalone (Web-)Application for Cross-Resource Spectrum Comparison. *J Proteome Res* **2021,** *20* (6), 3388-3394.
124. Ting, Y. S.; Egertson, J. D.; Payne, S. H.; Kim, S.; MacLean, B.; Käll, L.; Aebersold, R.; Smith, R. D.; Noble, W. S.; MacCoss, M. J., Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics* **2015,** *14* (9), 2301-7.
125. Bedran, G.; Gasser, H. C.; Weke, K.; Wang, T.; Bedran, D.; Laird, A.; Battail, C.; Zanzotto, F. M.; Pesquita, C.; Axelson, H.; Rajan, A.; Harrison, D. J.; Palkowski, A.; Pawlik, M.; Parys, M.; O'Neill, J. R.; Brennan, P. M.; Symeonides, S. N.; Goodlett, D. R.; Litchfield, K.; Fahraeus, R.; Hupp, T. R.; Kote, S.; Alfaro, J. A., The Immunopeptidome from a Genomic Perspective: Establishing the Noncanonical Landscape of MHC Class I-Associated Peptides. *Cancer Immunol Res* **2023,** *11* (6), 747-762.
126. Chong, C.; Müller, M.; Pak, H.; Harnett, D.; Huber, F.; Grun, D.; Leleu, M.; Auger, A.; Arnaud, M.; Stevenson, B. J.; Michaux, J.; Bilic, I.; Hirsekorn, A.; Calviello, L.; Simó-Riudalbas, L.; Planet, E.; Lubiński, J.; Bryśkiewicz, M.; Wiznerowicz, M.; Xenarios, I.; Zhang, L.; Trono, D.; Harari, A.; Ohler, U.; Coukos, G.; Bassani-Sternberg, M., Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* **2020,** *11* (1), 1293.
127. Raja, R.; Mangalaparthi, K. K.; Madugundu, A. K.; Jessen, E.; Pathangey, L.; Magtibay, P.; Butler, K.; Christie, E.; Pandey, A.; Curtis, M., Immunogenic cryptic peptides dominate the antigenic landscape of ovarian cancer. *Sci Adv* **2025,** *11* (8), eads7405.
128. Ruiz Cuevas, M. V.; Hardy, M. P.; Hollý, J.; Bonneil, É.; Durette, C.; Courcelles, M.; Lanoix, J.; Côté, C.; Staudt, L. M.; Lemieux, S.; Thibault, P.; Perreault, C.; Yewdell, J. W., Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep* **2021,** *34* (10), 108815.
129. Lodha, M.; Erhard, F.; Dölken, L.; Prusty, B. K., The Hidden Enemy Within: Non-canonical Peptides in Virus-Induced Autoimmunity. *Front Microbiol* **2022,** *13*, 840911.
130. Barczak, W.; Carr, S. M.; Liu, G.; Munro, S.; Nicastri, A.; Lee, L. N.; Hutchings, C.; Ternette, N.; Klenerman, P.; Kanapin, A.; Samsonova, A.; La Thangue, N. B., Long non-coding RNA-derived peptides are immunogenic and drive a potent anti-tumour response. *Nat Commun* **2023,** *14* (1), 1078.
131. Nagel, R.; Pataskar, A.; Champagne, J.; Agami, R., Boosting Antitumor Immunity with an Expanded Neoepitope Landscape. *Cancer Res* **2022,** *82* (20), 3637-3649.
132. Apavaloaei, A.; Zhao, Q.; Hesnard, L.; Cahuzac, M.; Durette, C.; Larouche, J. D.; Hardy, M. P.; Vincent, K.; Brochu, S.; Laverdure, J. P.; Lanoix, J.; Courcelles, M.; Gendron, P.; Lajoie, M.; Ruiz Cuevas, M. V.; Kina, E.;

Perrault, J.; Humeau, J.; Ehx, G.; Lemieux, S.; Watson, I. R.; Speiser, D. E.; Bassani-Sternberg, M.; Thibault, P.; Perreault, C., Tumor antigens preferentially derive from unmutated genomic sequences in melanoma and non-small cell lung cancer. *Nat Cancer* **2025**.
133. Kovalchik, K. A.; Hamelin, D. J.; Kubiniok, P.; Bourdin, B.; Mostefai, F.; Poujol, R.; Paré, B.; Simpson, S. M.; Sidney, J.; Bonneil, É.; Courcelles, M.; Saini, S. K.; Shahbazy, M.; Kapoor, S.; Rajesh, V.; Weitzen, M.; Grenier, J. C.; Gharsallaoui, B.; Maréchal, L.; Wu, Z.; Savoie, C.; Sette, A.; Thibault, P.; Sirois, I.; Smith, M. A.; Decaluwe, H.; Hussin, J. G.; Lavallée-Adam, M.; Caron, E., Machine learning-enhanced immunopeptidomics applied to T-cell epitope discovery for COVID-19 vaccines. *Nat Commun* **2024,** *15* (1), 10316.
134. Cao, Y.; Di, X.; Zhang, Q.; Li, R.; Wang, K., RBM10 Regulates Tumor Apoptosis, Proliferation, and Metastasis. *Front Oncol* **2021,** *11*, 603932.
135. Wortel, I. M. N.; van der Meer, L. T.; Kilberg, M. S.; van Leeuwen, F. N., Surviving Stress: Modulation of ATF4-Mediated Stress Responses in Normal and Malignant Cells. *Trends Endocrinol Metab* **2017,** *28* (11), 794-806.
136. Creamer, K. M.; Larsen, E. C.; Lawrence, J. B., ZNF146/OZF and ZNF507 target LINE-1 sequences. *G3 (Bethesda)* **2022,** *12* (3).
137. Liu, J.; Zhan, Y.; Wang, J.; Wang, J.; Guo, J.; Kong, D., Long noncoding RNA LINC01578 drives colon cancer metastasis through a positive feedback loop with the NF-κB/YY1 axis. *Mol Oncol* **2020,** *14* (12), 3211-3233.
138. Zhu, H.; Liu, Q.; Yang, X.; Ding, C.; Wang, Q.; Xiong, Y., LncRNA LINC00649 recruits TAF15 and enhances MAPK6 expression to promote the development of lung squamous cell carcinoma via activating MAPK signaling pathway. *Cancer Gene Ther* **2022,** *29* (8-9), 1285-1295.
139. Chen, X.; Chen, S., LINC00649 promotes bladder cancer malignant progression by regulating the miR-15a-5p/HMGA1 axis. *Oncol Rep* **2021,** *45* (4).
140. Zhang, J.; Du, C.; Zhang, L.; Wang, Y.; Zhang, Y.; Li, J., LncRNA LINC00649 promotes the growth and metastasis of triple-negative breast cancer by maintaining the stability of HIF-1α through the NF90/NF45 complex. *Cell Cycle* **2022,** *21* (10), 1034-1047.
141. Zeng, L.; Liao, Q.; Zeng, X.; Ye, J.; Yang, X.; Zhu, S.; Tang, H.; Liu, G.; Cui, W.; Ma, S.; Cui, S., Noncoding RNAs and hyperthermic intraperitoneal chemotherapy in advanced gastric cancer. *Bioengineered* **2022,** *13* (2), 2623-2638.
142. Li, K.; Jain, A.; Malovannaya, A.; Wen, B.; Zhang, B., DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* **2020,** *20* (21-22), e1900334.
143. Minegishi, Y.; Kiyotani, K.; Nemoto, K.; Inoue, Y.; Haga, Y.; Fujii, R.; Saichi, N.; Nagayama, S.; Ueda, K., Differential ion mobility mass spectrometry in immunopeptidomics identifies neoantigens carrying colorectal cancer driver mutations. *Commun Biol* **2022,** *5* (1), 831.
144. Aubel, M.; Buchel, F.; Heames, B.; Jones, A.; Honc, O.; Bornberg-Bauer, E.; Hlouchova, K., High-throughput Selection of Human de novo-emerged sORFs with High Folding Potential. *Genome Biol Evol* **2024,** *16* (4).

145. Fields, A. P.; Rodriguez, E. H.; Jovanovic, M.; Stern-Ginossar, N.; Haas, B. J.; Mertins, P.; Raychowdhury, R.; Hacohen, N.; Carr, S. A.; Ingolia, N. T.; Regev, A.; Weissman, J. S., A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* **2015,** *60* (5), 816-827.
146. Ma, J.; Ward, C. C.; Jungreis, I.; Slavoff, S. A.; Schwaid, A. G.; Neveu, J.; Budnik, B. A.; Kellis, M.; Saghatelian, A., Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **2014,** *13* (3), 1757-65.
147. Chen, L.; Yang, Y.; Zhang, Y.; Li, K.; Cai, H.; Wang, H.; Zhao, Q., The Small Open Reading Frame-Encoded Peptides: Advances in Methodologies and Functional Studies. *Chembiochem* **2022,** *23* (8), e202100534.
148. Chen, J.; Brunner, A. D.; Cogan, J. Z.; Nuñez, 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* (6482), 1140-1146.
149. Schlesinger, D.; Elsässer, S. J., Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *Febs j* **2022,** *289* (1), 53-74.
150. Nichols, C.; Do-Thi, V. A.; Peltier, D. C., Noncanonical microprotein regulation of immunity. *Mol Ther* **2024,** *32* (9), 2905-2929.
151. Yang, H.; Li, Q.; Stroup, E. K.; Wang, S.; Ji, Z., Widespread stable noncanonical peptides identified by integrated analyses of ribosome profiling and ORF features. *Nat Commun* **2024,** *15* (1), 1932.
152. Kadam, P. S.; Mueller, S. C.; Ji, H.; Liu, J.; Pai, A. V.; Ma, J.; Speth, R. C.; Sandberg, K., Modulation of the rat angiotensin type 1a receptor by an upstream short open reading frame. *Peptides* **2021,** *140*, 170529.
153. Li, M.; Shao, F.; Qian, Q.; Yu, W.; Zhang, Z.; Chen, B.; Su, D.; Guo, Y.; Phan, A. V.; Song, L. S.; Stephens, S. B.; Sebag, J.; Imai, Y.; Yang, L.; Cao, H., A putative long noncoding RNA-encoded micropeptide maintains cellular homeostasis in pancreatic β cells. *Mol Ther Nucleic Acids* **2021,** *26*, 307-320.
154. Cheng, R.; Li, F.; Zhang, M.; Xia, X.; Wu, J.; Gao, X.; Zhou, H.; Zhang, Z.; Huang, N.; Yang, X.; Zhang, Y.; Shen, S.; Kang, T.; Liu, Z.; Xiao, F.; Yao, H.; Xu, J.; Yan, C.; Zhang, N., A novel protein RASON encoded by a lncRNA controls oncogenic RAS signaling in KRAS mutant cancers. *Cell Res* **2023,** *33* (1), 30-45.
155. Wright, B. W.; Yi, Z.; Weissman, J. S.; Chen, J., The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol* **2022,** *32* (3), 243-258.
156. Azam, S.; Yang, F.; Wu, X., Finding functional microproteins. *Trends Genet* **2025,** *41* (2), 107-118.
157. Hofman, D. A.; Prensner, J. R.; van Heesch, S., Microproteins in cancer: identification, biological functions, and clinical implications. *Trends Genet* **2025,** *41* (2), 146-161.
158. Thibault, P.; Perreault, C., Immunopeptidomics: Reading the Immune Signal That Defines Self From Nonself. *Mol Cell Proteomics* **2022,** *21* (6), 100234.

159. Chong, C.; Coukos, G.; Bassani-Sternberg, M., Identification of tumor antigens with immunopeptidomics. *Nat Biotechnol* **2022,** *40* (2), 175-188.
160. Ferreira, H. J.; Stevenson, B. J.; Pak, H.; Yu, F.; Almeida Oliveira, J.; Huber, F.; Taillandier-Coindard, M.; Michaux, J.; Ricart-Altimiras, E.; Kraemer, A. I.; Kandalaft, L. E.; Speiser, D. E.; Nesvizhskii, A. I.; Müller, M.; Bassani-Sternberg, M., Immunopeptidomics-based identification of naturally presented non-canonical circRNA-derived peptides. *Nat Commun* **2024,** *15* (1), 2357.
161. Deutsch, E. W.; Kok, L. W.; Mudge, J. M.; Ruiz-Orera, J.; Fierro-Monti, I.; Sun, Z.; Abelin, J. G.; Alba, M. M.; Aspden, J. L.; Bazzini, A. A.; Bruford, E. A.; Brunet, M. A.; Calviello, L.; Carr, S. A.; Carvunis, A. R.; Chothani, S.; Clauwaert, J.; Dean, K.; Faridi, P.; Frankish, A.; Hubner, N.; Ingolia, N. T.; Magrane, M.; Martin, M. J.; Martinez, T. F.; Menschaert, G.; Ohler, U.; Orchard, S.; Rackham, O.; Roucou, X.; Slavoff, S. A.; Valen, E.; Wacholder, A.; Weissman, J. S.; Wu, W.; Xie, Z.; Choudhary, J.; Bassani-Sternberg, M.; Vizcaíno, J. A.; Ternette, N.; Moritz, R. L.; Prensner, J. R.; van Heesch, S., High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. *bioRxiv* **2024**.
162. Meng, K.; Li, Y.; Yuan, X.; Shen, H. M.; Hu, L. L.; Liu, D.; Shi, F.; Zheng, D.; Shi, X.; Wen, N.; Cao, Y.; Pan, Y. L.; He, Q. Y.; Zhang, C. Z., The cryptic lncRNA-encoded microprotein TPM3P9 drives oncogenic RNA splicing and tumorigenesis. *Signal Transduct Target Ther* **2025,** *10* (1), 43.
163. Huang, N.; Li, F.; Zhang, M.; Zhou, H.; Chen, Z.; Ma, X.; Yang, L.; Wu, X.; Zhong, J.; Xiao, F.; Yang, X.; Zhao, K.; Li, X.; Xia, X.; Liu, Z.; Gao, S.; Zhang, N., An Upstream Open Reading Frame in Phosphatase and Tensin Homolog Encodes a Circuit Breaker of Lactate Metabolism. *Cell Metab* **2021,** *33* (1), 128-144.e9.
164. Procter, J. B.; Carstairs, G. M.; Soares, B.; Mourão, K.; Ofoegbu, T. C.; Barton, D.; Lui, L.; Menard, A.; Sherstnev, N.; Roldan-Martinez, D.; Duce, S.; Martin, D. M. A.; Barton, G. J., Alignment of Biological Sequences with Jalview. *Methods Mol Biol* **2021,** *2231*, 203-224.
165. Xu, H.; Xiao, T.; Chen, C. H.; Li, W.; Meyer, C. A.; Wu, Q.; Wu, D.; Cong, L.; Zhang, F.; Liu, J. S.; Brown, M.; Liu, X. S., Sequence determinants of improved CRISPR sgRNA design. *Genome Res* **2015,** *25* (8), 1147-57.
166. Nuñez Pedrozo, C. N.; Peralta, T. M.; Olea, F. D.; Locatelli, P.; Crottogini, A. J.; Belaich, M. N.; Cuniberti, L. A., In silico performance analysis of web tools for CRISPRa sgRNA design in human genes. *Comput Struct Biotechnol J* **2022,** *20*, 3779-3782.
167. Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nat Methods* **2012,** *9* (4), 357-9.
168. Zheng, C.; Wei, Y.; Zhang, P.; Lin, K.; He, D.; Teng, H.; Manyam, G.; Zhang, Z.; Liu, W.; Lee, H. R. L.; Tang, X.; He, W.; Islam, N.; Jain, A.; Chiu, Y.; Cao, S.; Diao, Y.; Meyer-Gauen, S.; Höök, M.; Malovannaya, A.; Li, W.; Hu, M.; Wang, W.; Xu, H.; Kopetz, S.; Chen, Y., CRISPR-Cas9-based functional interrogation of unconventional translatome reveals human cancer dependency on

cryptic non-canonical open reading frames. *Nat Struct Mol Biol* **2023,** *30* (12), 1878-1892.
169. Shalem, O.; Sanjana, N. E.; Hartenian, E.; Shi, X.; Scott, D. A.; Mikkelson, T.; Heckl, D.; Ebert, B. L.; Root, D. E.; Doench, J. G.; Zhang, F., Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **2014,** *343* (6166), 84-87.
170. Li, Y.; Zhou, H.; Chen, X.; Zheng, Y.; Kang, Q.; Hao, D.; Zhang, L.; Song, T.; Luo, H.; Hao, Y.; Chen, R.; Zhang, P.; He, S., SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genomics Proteomics Bioinformatics* **2021,** *19* (4), 602-610.
171. Olexiouk, V.; Van Criekinge, W.; Menschaert, G., An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* **2018,** *46* (D1), D497-d502.
172. Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B., How many human proteoforms are there? *Nat Chem Biol* **2018,** *14* (3), 206-214.
173. van Heesch, S.; Witte, F.; Schneider-Lunitz, V.; Schulz, J. F.; Adami, E.; Faber, A. B.; Kirchner, M.; Maatz, H.; Blachut, S.; Sandmann, C. L.; Kanda, M.; Worth, C. L.; Schafer, S.; Calviello, L.; Merriott, R.; Patone, G.; Hummel, O.; Wyler, E.; Obermayer, B.; Mücke, M. B.; Lindberg, E. L.; Trnka, F.; Memczak, S.; Schilling, M.; Felkin, L. E.; Barton, P. J. R.; Quaife, N. M.; Vanezis, K.; Diecke, S.; Mukai, M.; Mah, N.; Oh, S. J.; Kurtz, A.; Schramm, C.; Schwinge, D.; Sebode, M.; Harakalova, M.; Asselbergs, F. W.; Vink, A.; de Weger, R. A.; Viswanathan, S.; Widjaja, A. A.; Gärtner-Rommel, A.; Milting, H.; Dos Remedios, C.; Knosalla, C.; Mertins, P.; Landthaler, M.; Vingron, M.; Linke, W. A.; Seidman, J. G.; Seidman, C. E.; Rajewsky, N.; Ohler, U.; Cook, S. A.; Hubner, N., The Translational Landscape of the Human Heart. *Cell* **2019,** *178* (1), 242-260.e29.
174. Akimoto, C.; Sakashita, E.; Kasashima, K.; Kuroiwa, K.; Tominaga, K.; Hamamoto, T.; Endo, H., Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochim Biophys Acta* **2013,** *1830* (3), 2728-38.
175. Yu, R.; Hu, Y.; Zhang, S.; Li, X.; Tang, M.; Yang, M.; Wu, X.; Li, Z.; Liao, X.; Xu, Y.; Li, M.; Chen, S.; Qian, W.; Gong, L. Y.; Song, L.; Li, J., LncRNA CTBP1-DT-encoded microprotein DDUP sustains DNA damage response signalling to trigger dual DNA repair mechanisms. *Nucleic Acids Res* **2022,** *50* (14),

8060-8079.
176. Xu, W.; Liu, C.; Deng, B.; Lin, P.; Sun, Z.; Liu, A.; Xuan, J.; Li, Y.; Zhou, K.; Zhang, X.; Huang, Q.; Zhou, H.; He, Q.; Li, B.; Qu, L.; Yang, J., TP53-inducible putative long noncoding RNAs encode functional polypeptides that suppress cell proliferation. *Genome Res* **2022,** *32* (6), 1026-1041.
177. Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M. N.; Ma, J.; Levin, J. Z.; Karger, A. D.; Budnik, B. A.; Rinn, J. L.; Saghatelian, A., Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **2013,** *9* (1), 59-64.
178. Halasz, H.; Malekos, E.; Covarrubias, S.; Yitiz, S.; Montano, C.; Sudek, L.; Katzman, S.; Liu, S. J.; Horlbeck, M. A.; Namvar, L.; Weissman, J. S.; Carpenter, S., CRISPRi screens identify the lncRNA, LOUP, as a multifunctional locus regulating macrophage differentiation and inflammatory signaling. *Proc Natl Acad Sci U S A* **2024,** *121* (22), e2322524121.
179. Occhi, G.; Regazzo, D.; Trivellin, G.; Boaretto, F.; Ciato, D.; Bobisse, S.; Ferasin, S.; Cetani, F.; Pardi, E.; Korbonits, M.; Pellegata, N. S.; Sidarovich, V.; Quattrone, A.; Opocher, G.; Mantero, F.; Scaroni, C., A novel mutation in the upstream open reading frame of the CDKN1B gene causes a MEN4 phenotype. *PLoS Genet* **2013,** *9* (3), e1003350.
180. Guo, B.; Wu, S.; Zhu, X.; Zhang, L.; Deng, J.; Li, F.; Wang, Y.; Zhang, S.; Wu, R.; Lu, J.; Zhou, Y., Micropeptide CIP2A-BP encoded by LINC00665 inhibits triple-negative breast cancer progression. *Embo j* **2020,** *39* (1), e102190.
181. Shi, C.; Liu, F.; Su, X.; Yang, Z.; Wang, Y.; Xie, S.; Xie, S.; Sun, Q.; Chen, Y.; Sang, L.; Tan, M.; Zhu, L.; Lei, K.; Li, J.; Yang, J.; Gao, Z.; Yu, M.; Wang, X.; Wang, J.; Chen, J.; Zhuo, W.; Fang, Z.; Liu, J.; Yan, Q.; Neculai, D.; Sun, Q.; Shao, J.; Lin, W.; Liu, W.; Chen, J.; Wang, L.; Liu, Y.; Li, X.; Zhou, T.; Lin, A., Comprehensive discovery and functional characterization of the noncanonical proteome. *Cell Res* **2025,** *35* (3), 186-204.
182. Kesner, J. S.; Chen, Z.; Shi, P.; Aparicio, A. O.; Murphy, M. R.; Guo, Y.; Trehan, A.; Lipponen, J. E.; Recinos, Y.; Myeku, N.; Wu, X., Noncoding translation mitigation. *Nature* **2023,** *617* (7960), 395-402.
183. Blanchette, M.; Kent, W. J.; Riemer, C.; Elnitski, L.; Smit, A. F.; Roskin, K. M.; Baertsch, R.; Rosenbloom, K.; Clawson, H.; Green, E. D.; Haussler, D.; Miller, W., Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **2004,** *14* (4), 708-15.
184. Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M., The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **2013,** *45* (10), 1113-20.
185. Meyers, R. M.; Bryan, J. G.; McFarland, J. M.; Weir, B. A.; Sizemore, A. E.; Xu, H.; Dharia, N. V.; Montgomery, P. G.; Cowley, G. S.; Pantel, S.; Goodale, A.; Lee, Y.; Ali, L. D.; Jiang, G.; Lubonja, R.; Harrington, W. F.; Strickland, M.; Wu, T.; Hawes, D. C.; Zhivich, V. A.; Wyatt, M. R.; Kalani, Z.; Chang, J. J.; Okamoto, M.; Stegmaier, K.; Golub, T. R.; Boehm, J. S.; Vazquez,

F.; Root, D. E.; Hahn, W. C.; Tsherniak, A., Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* **2017,** *49* (12), 1779-1784.