UNVEILING ROLES OF sORF-ENCODED MICROPROTEINS IN LIVER DEVELOPMENT AND CANCER RESISTANCE BY PROTEOGENOMIC APPROACHES

YANG YING

PhD

THE HONG KONG POLYTECHNIC UNIVERSITY

2025

The Hong Kong Polytechnic University

Department of Applied Biology and Chemical Technology

Unveiling Roles of sORF-encoded Microproteins in Liver Development and Cancer Resistance by Proteogenomic Approaches

YANG YING

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

OCT 2024

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.

YANG Ying

Abstract

Recent advancements in computational, genomic, and proteomic techniques have revealed the potential of unannotated small open reading frames (sORFs) that are capable of encoding peptides. These small open reading frame-encoded microproteins (SEPs), also known as alternative proteins (AltProts) or microproteins, are directly translated from sORFs and exhibit no similarity to the canonical reference proteins (RefProts) of the same gene. Microproteins have been demonstrated to contribute to the progression of various diseases by affecting cellular signaling and disease progression. Despite the growing recognition of their biological significance, microproteins have historically been overlooked because of their short length and relatively low abundance, which complicates their identification through mass spectrometry (MS).

The liver, a vital organ during embryonic development, plays a crucial role in hepatic organogenesis and hematopoiesis. It is essential for cell proliferation, immune function, and the synthesis and transport of proteins and nucleic acids. Despite its significance, many biological processes involved in liver development remain poorly understood. Understanding the molecular mechanisms that govern liver development can facilitate the development of regenerative medical strategies to treat liver injuries and diseases.

Hepatocellular carcinoma (HCC), a major focus in hepatology, ranks among the most prevalent cancers worldwide, facing challenges due to limited treatment options and a high rate of drug resistance. Therefore, it is essential to develop therapeutic strategies that can effectively address drug resistance to improve patient outcomes and increase survival rates in patients with advanced HCC. Microproteins play vital roles as biological regulators and are involved in numerous biological processes. However, their roles in liver development and drug resistance have not been fully explored.

In this study, we focused on the role of microproteins in the field of hepatology, aiming to provide a robust foundation for the development of novel therapeutic strategies and diagnostic tools. We accomplished two key objectives: (i) to develop and refine a systematic methodological approach for the proteogenomic discovery of microproteins in liver tissue across various developmental stages, resulting in a comprehensive proteomic dataset for future functional studies, and (ii) to optimize both DIA and DDA MS-based proteomics methodologies to identify novel microproteins in lenvatinib-resistant HCC cells. This study allowed us to explore the microproteins associated with drug resistance in liver cancer, thereby enhancing our understanding of their mechanisms of action, particularly in relation to drug resistance. This has the potential to facilitate the development of novel strategies for future clinical interventions.

First, we reported an approach utilizing size-exclusion chromatography (SEC) for the simultaneous enrichment and fractionation of microproteins from complex proteomes. This method greatly simplified the variance of microprotein discovery by enriching proteins smaller than 40 kDa. In a systematic comparison between the ten methods, our approach facilitated the discovery of more microproteins with overall higher intensities, while requiring less time and effort compared to other workflows. By applying this approach, we successfully identified 89 novel microproteins in the mouse liver, with 39 showing differential expression between the embryonic and adult stages. During embryonic development, upregulated microproteins were mainly involved in biological pathways related to RNA splicing and processing, whereas microproteins involved in metabolism were more active in adult livers. Our study not only presents an effective approach for identifying microproteins but also highlights novel microproteins that are potentially important in developmental biology.

We also presented a novel approach combining Ribo-seq and multiple MS methods, identifying and quantifying 815 microproteins from human HCC cells. Notably, we found one microprotein PPGlue was downregulated in resistant cells. Functionally, PPGlue sensitized HCC to lenvatinib treatment both *in vitro* and *in vivo*, with enhanced apoptosis, suppressed proliferation and less cancer stemness. Mechanistic studies showed that PPGlue acted as a molecular glue to facilitate the assembly of the protein phosphatase complex PPP2R3C/PP5, reducing a drug exporter P-glycoprotein and subsequently increasing intracellular drug accumulation. Synthetic PPGlue also displayed a synergistic effect with P-glycoprotein's substrate such as lenvatinib, pazopanib and doxorubicin, highlighting its potential therapeutic value. Our study not only provides a practical proteogenomic methodology to identify microproteins in large scale, but also underscores the potential of microproteins as promising modalities in cancer treatment with PPGlue as a representative.

In summary, we presented a comprehensive methodological optimization of the microprotein discovery workflow, which includes both sample preparation and MS-based analytical identification. As a result, we identified new molecular participants that are crucial for liver development and cancer biology. Our work offers a comprehensive framework for exploring microproteins, highlighting their identification methodologies and functional characteristics within the context of hepatopathy, thereby paving the way for future research and therapeutic development.

List of publications

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

1. **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.

2. Chen, L. [#], **Yang, Y.[#],** Leung, C. O. N., Li, K., Zhang, Y., Yang. Y., Wang, H., Lee, T. K. W., Zhao, Q., Proteogenomic Discovery of PPGlue: A microprotein in restoring drug sensitivity via PPP2R3C/P-gp interactions. Nature cancer 2024. (In submission). (#co-first authors)

3. 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. Analytical Chemistry, 97 (3), 1517-1527.

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. Guo, H.[#]; Yang, Y.[#]; Zhang, Q. [#]; Deng, J. R.; **Yang, Y.**; Li, S.; So, P. K.; Lam, T.

C.; Wong, M. K.; Zhao, Q.*, Integrated mass spectrometry reveals celastrol as a novel catechol-Omethyltransferase inhibitor. ACS Chem Biol 2022, 17 (8), 2003-2009.

7. Zhang, Q., Yang, Y., **Yang, Y.,** Shang, J., Su, S., Gao, P., Li, X., Kao, R. Y., Ko, B. C., Thompson, B., Zhao, Q., A strategy to sensitize broad-spectrum gram-positive bacteria to oxazolidinone antibiotics by lysozyme. Science Advances 2024. (In submission).

8. Yang, Y., Tse, Y. C., Zhang, Q., Wong, K., Yang, C., **Yang, Y.,** Li, S., Lau, K., Charles, T. C., Lam, T. C., Zhao, Q., Multiplexed target profiling with integrated chemical genomics and chemical proteomics. Journal of Medicinal Chemistry 2024, 67 (19), 17542–17550.

Conference papers

1. **Yang, Y.#,** Chen, L. #, Leung, C. O. N., Li, K., Zhang, Y., Yang. Y., Wang, H., Lee, T. K. W., Zhao, Q., Proteogenomic Discovery of PPGlue: A microprotein in restoring drug sensitivity via PPP2R3C/P-gp interactions. The 18th Chinese International Peptide Symposium, 2024. (Poster presentation)

2. **Yang, Y.#,** Chen, L. #, Leung, C. O. N., Li, K., Zhang, Y., Yang. Y., Wang, H., Lee, T. K. W., Zhao, Q., Proteogenomic Discovery of PPGlue: A microprotein in restoring drug sensitivity via PPP2R3C/P-gp interactions. The 5th ABCT Research Postgraduate Symposium, 2024. (Oral Presentation)

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. The 2nd Chinese American Society for Mass Spectrometry (CASMS), 2022.

4. **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. The 3rd ABCT Research Postgraduate Symposium, 2022. (Oral Presentation & Poster)
5. **Yang, Y.,** Zhao, Q.* An integrated approach for discovering non-canonical MHC-I peptides encoded by small open reading frames. The 2nd ABCT Research Postgraduate Symposium, 2021. (Oral Presentation)

Acknowledgements

First, I would like to give my deepest thanks to my supervisor, Dr. Zhao Qian, for her professional guidance, great support, and encouraging words throughout my PhD journey. Her rich knowledge, innovative scientific research thinking, and commitment to rigorous standards greatly benefited me. In addition, I also extend my sincere appreciation to all the members of Dr. Zhao's lab: Dr. Chen Lei Alyssa, Dr. Faleti Oluwasijibomi Damola, Dr. Yang Yang, Dr. Zhang Qi, Mr. Zhang Yuanliang, Miss Tse Yin-suen Chloe, Miss Li Shuqi, Miss Shang Jin, Mr. Li Kecheng, Miss Yang Chenxi, Mr. Zhu Tianyang, for their academic support, collaboration and friendship. The harmonious and friendly atmosphere within Dr. Zhao's group has made my time truly enjoyable, and I am grateful for the opportunity to work alongside such talented individuals, learn from them, and share wonderful experiences.

I would like to especially thank Prof. Lee Kin-wah Terrence and Dr. Leung Oi-ning Carmen for providing substantial assistance for their significant contributions to my project on HCC cancer resistance. I also deeply appreciate Dr. Chen Lei for her invaluable assistance and insightful suggestions. Furthermore, I extend my sincere thanks to Dr. Yang Yang and Mr. Zhang Yuanliang from our group for their exceptional patience and carefulness in the maintenance of mass spectrometry.

I would like to express my sincere gratitude to the lab technicians, Dr. Wong Lai-king Iris and Dr. So Pui-kin, for their unwavering dedication and patience in ensuring the smooth operation of the tissue culture lab and the mass spectrometry laboratory. I also extend my appreciation to the administrative staff, including Ms. Kwok Fung-yee

Peggy, Miss Yuen Chui-ying Ailsa, and Ms. So Hiu Tung Esther, for their diligent efforts. Furthermore, I would like to acknowledge my ABCT colleagues and friends for their generous support and heartfelt friendship. We look forward to a bright future.

I would like to appreciate the Grant Support from the Research Grants Council-GRF (15305821, 15304819), CRF Equipment (C5033-19E), RGC-RIF (R5050-18), Laboratory for Synthetic Chemistry and Chemical Biology Limited (LSCCB), and Center for Eye and Vision Research (CEVR) under the Health@InnoHK Programme launched by ITC, HKSAR. We thank the PolyU Research Facilities ULS and UCEA, and research institute/center RiFood and RCMI for technical support.

In addition, I acknowledge the use of artificial intelligence tools for language refinement, grammar checking, and sentence modification, which assisted in improving the clarity and presentation of this work.

Finally, I would like to sincerely thank my parents and my boyfriend Dr. Tu Yanbei for their encouragement, support, and love.

Grant support

This study was generously financed by,

1. RGC-GRF (15304819, 15305821),

2. RGC Collaborative Research Fund Equipment (C5033-19E),

3. RGC-RIF R5050-18.

I sincerely thank the abovementioned grants for their generous support.

List of Abbreviations

Affinity purification coupled with mass spectrometry	AP-MS
Alternative proteins	AltProts
Bovine serum albumin	BSA
Colorectal cancer	CRC
Column volume	CV
Data-dependent acquisition	DDA
Data-independent acquisition	DIA
Dithiothreitol	DTT
Downstream ORFs	dORFs
Downstream overlapping ORFs	doORFs
Doxorubicin	DOX
Dulbecco's modified Eagle's medium	DMEM
Electrostatic repulsion-hydrophilic interaction chromatography	ERLIC
Empty vector control	EV
Fetal bovine serum	FBS
Hepatocellular carcinoma	HCC
High pH reverse phase	HpRP
Immunohistochemical	IHC
Internal ORFs	intORFs

Iodoacetamide	IAM
Isoelectric point	pI
Knock down	KD
Lenvatinib	Len
Lenvatinib-resistant	LR
Long non-coding RNAs	lncRNAs
Lysyl endopeptidase	Lys-C
Mass spectrometry	MS
Mass-to-charge	m/z
Molecular weight cut-off	MWCO
Non-coding RNA	ncRNA
Non-target control	NC
Overexpression	OE
Parallel reaction monitoring	PRM
Pazopanib	PAZ
Peptide-drug conjugates	PDCs
Reference proteins	RefProts
Renal cell carcinoma	RCC
Retention time	Rt
Ribosome profiling	Ribo-seq
Ribosome-protected fragments	RPFs
RNA sequencing	RNA-seq

PPGlue overexpression	PPGlue
PPGlue overexpression with a mutant construct	PPGlue^{mut}
Scramble siRNA as the non-targeting control	siNC
siRNAs targeting PPGlue	siPPGlue
Stable isotope labelling by amino acids in cell culture	SILAC
Size exclusion chromatography	SEC
Small open reading frame-encoded microproteins	SEPs
Small open reading frames	sORFs
Solid phase extraction	SPE
Strong cation exchange	SCX
Tandem Mass Tags	TMT
Tyrosine kinase inhibitors	TKIs
Untranslated regions	UTRs
Upstream ORFs	uORFs
Upstream overlapping ORFs	uoORFs

Table of Contents

List of Tables and Figures

Chapter 1. Overview

1.1 Overview of small open reading frame-encoded microproteins (SEPs)

According to the central dogma of molecular biology, DNA is transcribed into RNA and translated into proteins. However, less than 2% of the human genome is known to encode proteins, and a significant portion of detectable transcripts remains unannotated. These transcripts were previously considered non-functional or "junk"[1]. It was not until this decade that advances in computational, genomic, and proteomic techniques revealed the potential of non-annotated small open reading frames (sORFs) capable of encoding proteins[2]. Although sORFs have been detected in various species, they have historically been overlooked for several reasons. First, sORFs in non-coding RNA (ncRNAs) often lack traditional start codons and are not limited to AUG, which initially obscures their coding potential[3]. Second, the length of sORFs typically does not exceed 300 nucleotides, an arbitrary threshold defined as the minimum for open reading frames (ORFs)[3, 4]. Moreover, the translational role of sORFs was initially overlooked because of (1) their ability to be translated into non-canonical modes, (2) their instability and rapid degradation, and (3) their highly specific expression patterns that exhibit significant temporal and spatial variation[5].

Small open reading frame-encoded microproteins (SEPs), also known as AltProts or microproteins, are directly translated from sORFs. In contrast to peptide hormones and neuropeptides, which are derived from large precursor proteins through proteolysis[6], SEPs exhibit no similarity to canonical reference proteins (RefProts) of the same gene. Recent technological advancements, including ribosome profiling (Ribo-seq) and mass spectrometry (MS), have substantiated the existence of sORFs and SEPs[7, 8]. These discoveries have facilitated the integration of numerous characterized sORFs into databases such as SmProt and OpenProt[9, 10]. An analysis of these repositories reveals that SEPs encoded by long non-coding RNAs (lncRNAs) typically consist of approximately 54 amino acids, while those located within untranslated regions (UTRs) average about 39 amino acids[11]. SEPs are systematically classified into several categories based on the characteristics of their encoding transcripts: (i) upstream ORFs (uORFs) in the 5′-UTRs, (ii) upstream overlapping ORFs (uoORFs), (iii) internal ORFs (intORFs) that overlap canonical mRNAs but are translated into an alternative reading frame, (iv) downstream overlapping ORFs (doORFs), (v) downstream ORFs (dORFs) in the 3′-UTRs, and (vi) lncRNA-ORFs[7]. This classification underscores the diverse genomic origins of SEPs, as illustrated in **Figure 1-1**. Collectively, the accumulation of ribosome profiling studies and databases has provided compelling evidence for the prevalence of SEPs. The primary challenge is the functional characterization of these candidates.

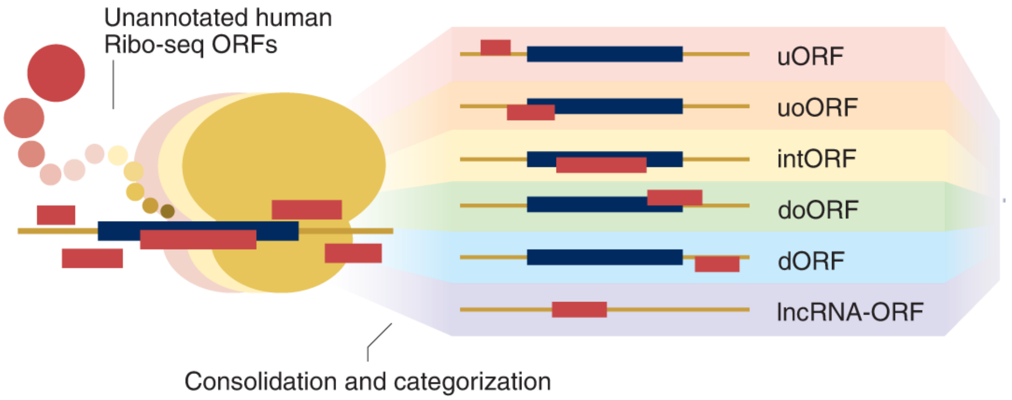

Unannotated human Ribo-seq ORFs
uORF
uoORF
intORF
doORF
dORF
lncRNA-ORF
Consolidation and categorization

Figure 1-1 Summary of various types of non-canonical sORFs[7].

The biological functions of various mammalian SEPs have been explored in humans and other vertebrates, revealing their involvement in DNA repair, mitochondrial activity, stress responses, and muscle development. Furthermore, SEPs have also been discovered in other organisms, such as bacteria, where they play critical biological roles. Given that sORFs make up at least 5-10% of genomes, a considerable number of functional SEPs remain to be identified and studied. The discovery of novel SEPs will deepen our understanding of the essential components within both the genome and proteome. Moreover, functional characterization of these SEPs is expected to provide valuable insights into fundamental biological processes, ultimately facilitating translational applications.

1.2 Methodologies for the prediction of SEPs

The small size, low abundance, and high environmental dependency of SEPs present significant challenges for their detection and functional analysis. Recently, researchers have successfully isolated various SEPs from a range of biological sources using diverse methodologies[12-14]. Currently, the reported methods for discovering new SEPs can be classified into three main approaches: ribosome profiling, computational approaches, and mass spectrometry (MS) (**Figure 1-2)**. Notably, the integration of sequencing results and computational predictions has significantly enhanced the efficacy of MS-based techniques for identifying SEPs within a comprehensive workflow. The following section summarizes and discusses the latest advancements in SEP discovery methodologies and provides insights into the future directions of this field.

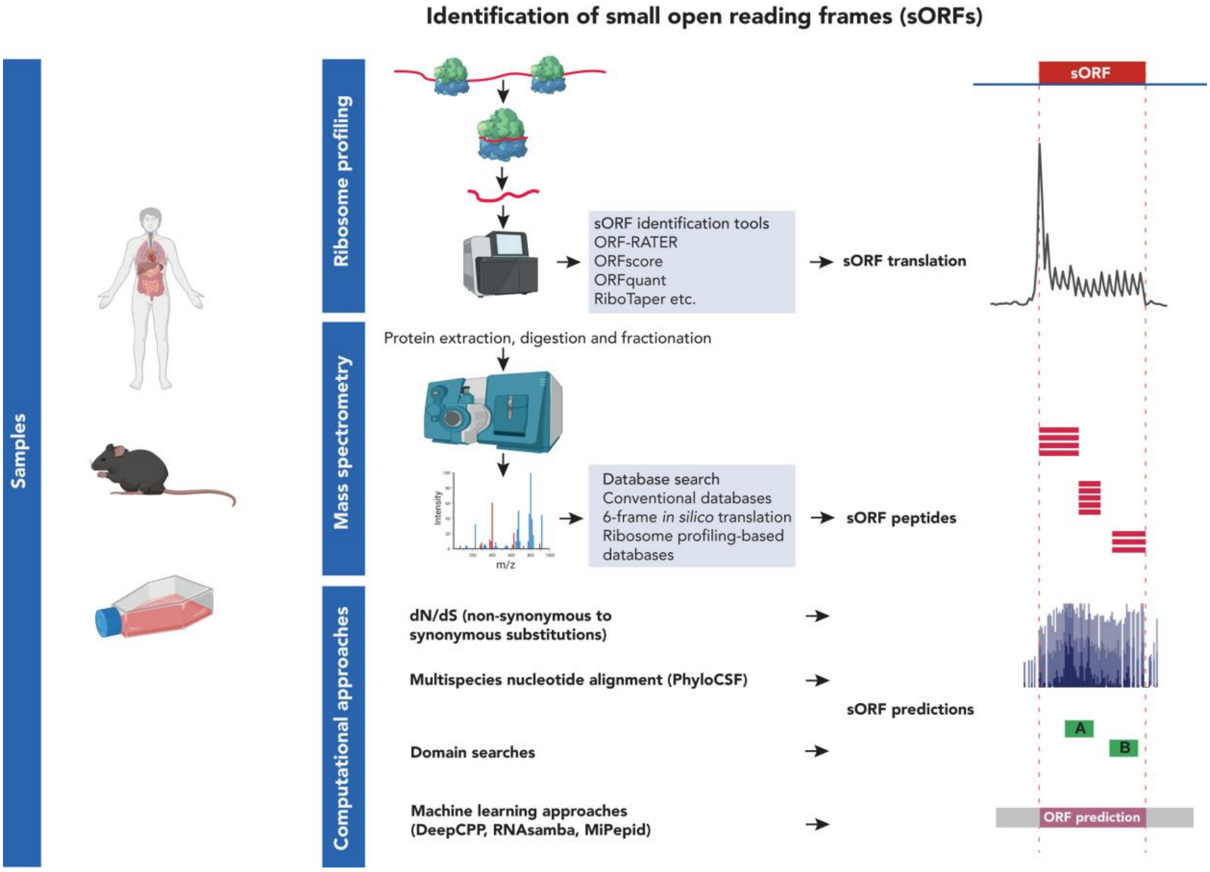

Identification of small open reading frames (sORFs)
Samples
Ribosome profiling
sORF identification tools
ORF-RATER
ORFscore
ORFquant
RiboTaper etc.
sORF translation
sORF
Mass spectrometry
Protein extraction, digestion and fractionation
Intensity
m/z
Database search
Conventional databases
6-frame *in silico* translation
Ribosome profiling-based databases
sORF peptides
Computational approaches
dN/dS (non-synonymous to synonymous substitutions)
Multispecies nucleotide alignment (PhyloCSF)
Domain searches
Machine learning approaches (DeepCPP, RNAsamba, MiPepid)
sORF predictions
A
B
ORF prediction

Figure 1-2 Overview of the workflow for the discovery of SEPs[15].

1.2.1 Ribosome profiling

Ribosomes serve as molecular machines responsible for protein translation within cells. They bind to processed mRNA transcripts and connect amino acids in a specific order according to the genetic code of the transcripts, resulting in protein products. Translation starts when the 40S ribosomal subunit binds to the 5'-m7GpppG cap of the mRNA and scans the 5' UTR from the 5' end to the 3' end until it finds the start codon AUG. At this point, the 60S ribosomal subunit joins to form an 80S ribosomal elongation complex, thereby initiating translation elongation. The translation process concludes when the ribosomal complex encounters a stop codon, leading to the termination of translation and the release of the ribosomal complex from the 3' UTR (**Figure 1-3**).

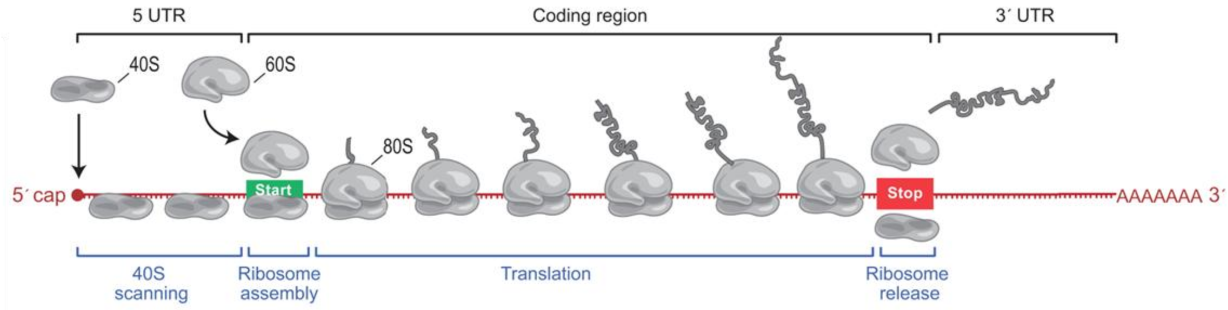


Figure 1-3 Overview of mRNA translation[16].

Ribosome profiling (Ribo-Seq) is an advanced technique developed to investigate ribosomal translation dynamics. This method involves the use of translation inhibitors to pause the translation process in cells, followed by the treatment of samples with nucleases. The robust structure of ribosomes allows them to remain attached to mRNA even after treatment, effectively protecting approximately 20 to 30 nucleotides from nuclease degradation.

This protection enables researchers to capture a snapshot of active translation, thereby providing valuable insights into gene expression and novel peptide identification[17]. Ribo-Seq results indicate that while most ribosomal footprints are located within known coding sequences, a significant number are associated with non-coding transcripts, including 3' UTRs, 5' UTRs, pseudogenes, and ncRNAs.

An analysis of 194,407 non-canonical ORF fragments from the OpenProt and SmProt databases revealed that 19,909 (10.2%) were derived from ncRNA-encoded sORFs, with an average length of approximately 54 amino acids. Additionally, 28,067 (14.4%) sORFs were found in 5' UTRs, with an average length of approximately 39 amino acids, whereas 5,509 (2.8%) sORFs were located in 3' UTRs[11]. The increasing development and application of analysis tools based on Ribo-Seq data, such as ORF-RATER[18] and RiboTaper[19], are continuously enhancing the data analysis ability and uncovering hidden ORFs within the genome. These advancements have significantly improved our ability to identify novel protein-coding regions and deepen our understanding of proteomes.

However, ribosomal occupancy does not necessarily correlate with protein translation. Many transcript translation processes may not result in stable functional peptides. Instead, they may influence the translation of downstream ORFs or simply be considered "transcriptional noise"[20].

1.2.2 Computational approaches

Bioinformatic analysis of genome sequences is crucial for predicting sORFs. Various methodologies have been employed for these predictions, including purifying selection identification, similarity comparisons with known protein sequences, and machine learning algorithms (**Table 1-1**). PhyloCSF is a widely used tool that identifies evolutionarily conserved coding ORFs by aligning transcripts from multiple species using phylogenetic codon models[21]. Other tools with similar functions include RNAcode[22], uPEPeroni[23], and micPDP[24]. Additionally, tools such as CRITICA[25], PhastCons[26], and sORF Finder[27] can predict coding ORFs by evaluating the nucleotide composition and considering sequence conservation. Tools such as BLAST[28], HMMER[29], and PFAM[30] assess the similarities between sORFs and known sequences. Emerging machine learning tools such as DeepCPP[31] and miPepid[32] are becoming prominent. Notably, miPepid achieves a remarkable 96% accuracy rate in predicting the coding potential of sORFs without requiring alignment[32].

Although bioinformatic tools are valuable, they have certain limitations. For instance, tools that focus on exon features may overlook SEPs with non-AUG initiation codons[33], whereas those dependent on phylogenetic conservation may suffer from poor-quality alignments. Moreover, tools that focus on known functional polypeptides may not identify newly emerging or tissue-specific SEPs[34]. Therefore, the use of a combination of tools based on different principles is crucial for the accurate and comprehensive identification of SEPs.

Table 1-1 Representative bioinformatic tools for SEP prediction.

Tools	Year	Functions
RNAcode	2011	Codon substitution
uPEPeroni	2014	Codon substitution
micPDP	2014	Codon substitution
CRITICA	1999	Nucleotide composition
PhastCons	2005	Nucleotide composition
sORF Finder	2009	Nucleotide composition
BLAST	1990	sequence similarity to known proteins
HMMER	1995	sequence similarity to known proteins
PFAM	2019	similarity to linear proteins
DeepCPP	2020	machine Learning algorithms
miPepid	2019	machine Learning algorithms
PhyloCSF	2011	Codon substitution, evolutionary conservation, multispecies transcript alignment

1.2.3 MS-based proteogenomic methods for identification of SEPs

Although ribosome profiling and computational prediction can reveal the prevalence of sORFs, they cannot directly detect SEPs. Currently, MS is the only method that can directly detect SEPs[35]. This method is effective for detecting low-abundance peptides and post-translational modifications, which significantly improves the identification of SEPs within complex biological samples.

The identification of SEPs follows a standard workflow of traditional bottom-up proteomics studies (**Figure 1-4**). Current efforts to optimize SEP identification focus primarily on key aspects, including sample extraction and enrichment, digestion and fractionation, mass spectrometry (MS) analysis, and data analysis.

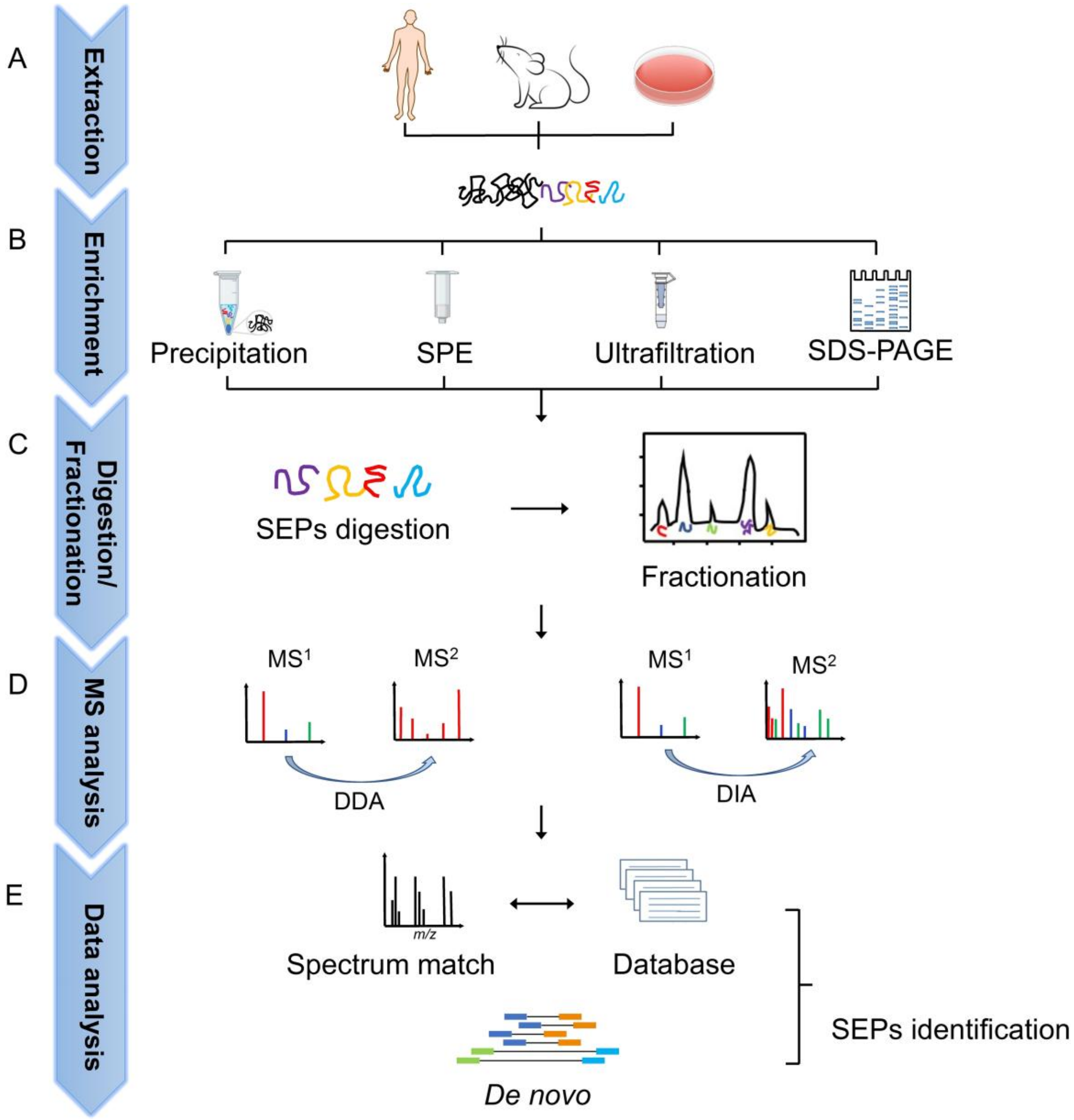


Figure 1-4 The MS-based workflow for the identification of SEPs contains several key steps. Initially, SEPs were extracted from complex biological samples and enriched in the total proteome. These peptides were digested with trypsin (or multiple enzymes), followed by fractionation. Subsequently, the tryptic peptides were subjected to MS data acquisition and analysis, enabling the accurate identification of SEPs.

1.2.3.1 Sample extraction and enrichment

The first crucial step in SEP identification using proteomic techniques is extraction from complex biological samples (**Figure 1-4A**). The extraction of SEPs is more challenging than that of canonical proteins because SEPs are particularly susceptible to hydrolysis by peptidases and may be masked by the degradation products of unwanted proteins. To maintain the integrity of SEPs, various strategies have been developed, such as heating in aqueous solutions or lysis buffers or using protease inhibitors to diminish protease activity[13, 36]. Nevertheless, such inhibition is not entirely effective. Another approach to prevent SEP degradation is the use of hydrochloric acid or acetic acid to induce protein precipitation, which inactivates both peptidases and proteases. The combination of these techniques has been widely employed for SEP extraction. A recent study by Cardon et al.[12] demonstrated that novel SEPs were successfully extracted using boiling water and RIPA lysis buffer and then enriched by acetate precipitation. This finding indicates that the blood marker AltEDARADD is associated with the diagnosis and prognosis of ovarian cancer. In summary, the selection of an appropriate extraction method prior to the enrichment of SEPs is critical and should be guided by specific research objectives and the stability of biological samples of interest.

Following the extraction of SEPs, it is crucial to enrich them with other proteins in the same sample (**Figure 1-4B**). This enrichment is typically achieved by employing a range of methodologies that utilize the diverse physical characteristics of SEPs, including size, hydrophobicity, and charge (**Figure 1-5**).

(1) Selective precipitation

It has been demonstrated that organic solvents such as methanol[12], acetonitrile[12, 37, 38], trichloroacetic acid[39], acetic acid[13], and chloroform[37] can remove larger proteins, while effectively retaining low-molecular-weight proteins, including SEPs, in the supernatant[13]. Cassidy et al.[38] employed acetonitrile for protein precipitation, which resulted in a reduction in sample complexity and the enrichment of small molecular weight proteins. This approach enabled the detection of 11 SEPs of *Methanosarcina mazei* with molecular weights lower than 15 kDa.

(2) Size selection

Ultrafiltration methods using membranes with a molecular weight cutoff (MWCO) of 10 or 30 kDa are commonly used for size selection. In this process, low-molecular-weight SEPs pass through the membrane, while higher-molecular-weight proteins are retained. Nevertheless, ultrafiltration has some limitations. First, concentrated macromolecules can clog the membrane pores, which diminishes the filtration efficiency. Second, nonspecific adsorption of proteins onto membranes is often unavoidable. Third, processing large sample volumes can be time consuming.

The other method for molecular weight-based separation is SDS-PAGE, which enables the excision of gel bands corresponding to the desired molecular weights for subsequent MS analysis. Ma et al. identified 90 and 94 SEPs using 30 kDa MWCO

Amicon filters and tricine gels, respectively[40]. He et al.[41] integrated four specific enrichment strategies to enhance the sequence coverage and identification rates of SEPs: HCl-Tricine, Urea-Tricine, HCl-MWCO, and Urea-MWCO. Among these, the urea-tricine method yielded the highest number of identified SEPs, and all four strategies demonstrated complementary benefits. Zhang et al. employed two complementary enrichment methods, including a 30 kDa MWCO filter and C8 solid-phase extraction columns[36], which successfully identified 762 novel SEPs across 19 different biological samples.

Size exclusion chromatography (SEC) is an important method for separating proteins according to their size, and has been successfully utilized in peptidomics[39, 42]. Harney et al. demonstrated that SEC can effectively remove large molecular interferences from plasma samples, thereby markedly enhancing the detection sensitivity of low-abundance proteins with molecular weights below 10 kDa[43]. This approach has resulted in the identification of novel biomarkers, including C5ORF46. In a lung squamous cell carcinoma study, the combination of SEC with acetonitrile precipitation streamlined the sample processing workflow, reduced the risk of protein degradation, and ensured the reliability of subsequent analytical methods[44]. Consequently, this combined approach successfully facilitated the identification of potential tumor-associated peptides.

(3) Solid phase extraction (SPE)

SEPs can be separated based on their hydrophilic and hydrophobic properties.

Although the C8 SPE method may result in the loss of hydrophilic proteins[36], the number of SEPs identified by combination with acetate precipitation exceeded the number identified by 30 kDa MWCO[13]. Zhang et al. employed C8 SPE and 30 kDa Amicon filters, either individually or in combination, and found that these methods can enhance the overall identification of SEPs[36]. In conclusion, each extraction and enrichment method have its own unique benefits and drawbacks, and no single approach has been proven to consistently outperform the others. Therefore, it is crucial to develop a systematic sample preparation strategy to effectively identify SEPs.

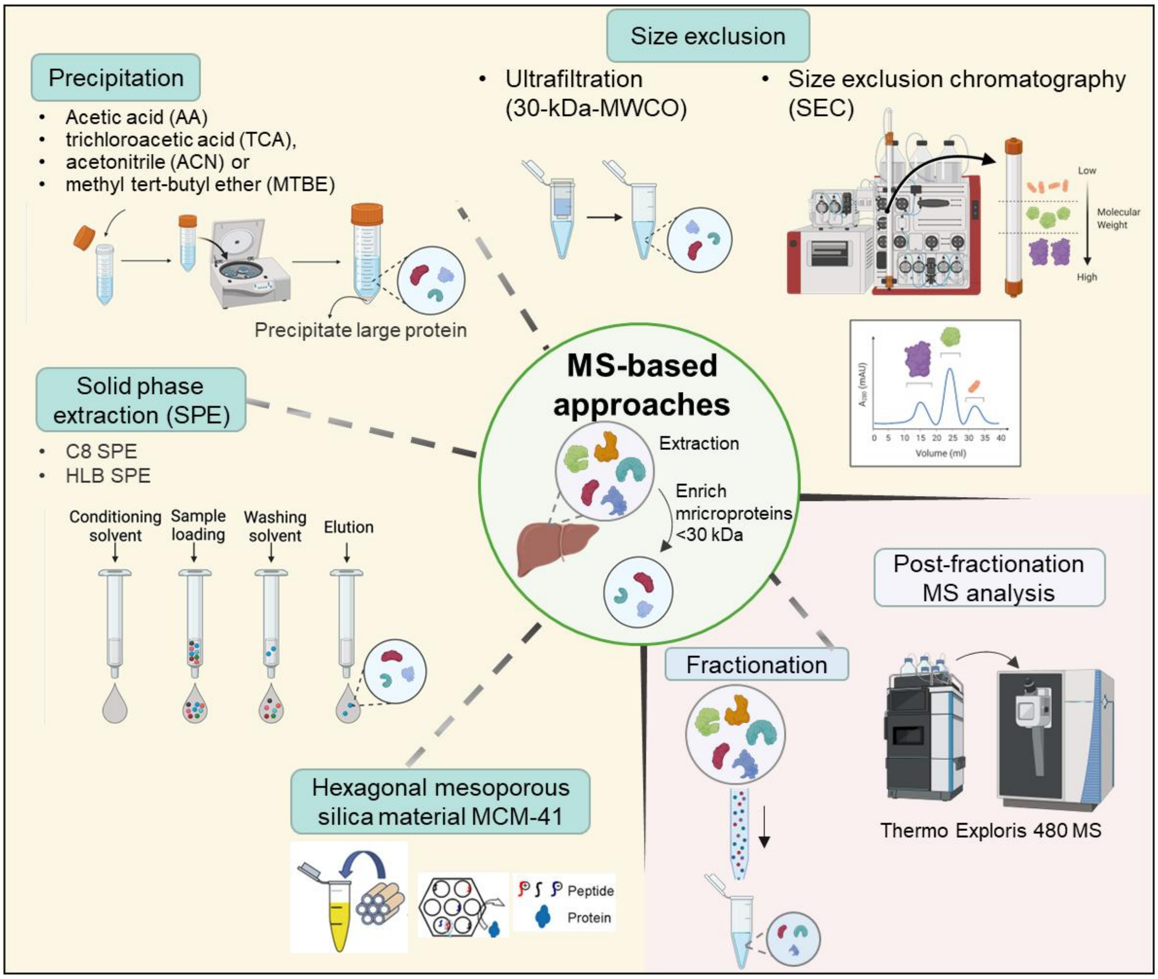


Figure 1-5 General enrichment methods for the identification of SEPs using a MS-based workflow. To isolate microproteins with a molecular weight of less than 30 kDa prior to MS analysis, several sample preparation methods were employed, including four distinct enrichment techniques: acid precipitation, solid-phase extraction, size exclusion, and hexagonal mesoporous silica material. Following the enrichment process, peptide fractionation was performed to reduce sample complexity, thereby enhancing sequence coverage and reducing background noise in mass spectrometry.

1.2.3.2 Enzyme digestion and fractionation

The choice of enzyme, whether using a single enzyme or a combination, plays a crucial role in the identification of SEPs (**Figure 1-4C**). Most studies used trypsin alone or in combination with Lys-C. However, if there are large or no lysine/arginine residues in the protein, trypsin may not provide a complete picture. Additionally, the intricate structure of SEPs may hinder trypsin accessibility, resulting in incomplete cleavage. Trypsin specifically hydrolyzes peptide bonds at the C-terminus of lysine and arginine, producing N-terminal tryptic peptides that are mainly double-charged, whereas C-terminal peptides are generally single-charged. This difference in charge makes the detection of C-terminal peptides challenging using MS[45]. Therefore, relying solely on trypsin may not provide a comprehensive identification of SEPs. A multi-protease digestion approach that incorporates trypsin, Lys-C, chymotrypsin, and Glu-C has been demonstrated to improve SEP identification, particularly in terms of the number of peptides identified, spectrum counts, and overall sequence coverage[46]. Given the complexity of the resultant peptide mixture after digestion, a variety of offline fractionation techniques have been employed to improve the sequencing depth before MS analysis. Techniques such as electrostatic repulsion-hydrophilic interaction chromatography (ERLIC)[40, 47], high-pH reverse-phase fractionation[37, 48-50], strong cation exchange (SCX)[51], and OFFGEL fractionation[52] have been shown to significantly enhance the identification of SEPs.

1.2.3.3 Data acquisition with mass spectrometry

Currently, MS is the exclusive technique for the direct detection and quantification of SEPs (**Figure 1-4D**). Among the MS methods employed, data-dependent acquisition (DDA), data-independent acquisition (DIA), and parallel reaction monitoring (PRM) are the most prominent. DDA operates on a shotgun approach, selectively fragmenting the top N most abundant precursor ions for analysis, making it suitable for both labeled and label-free quantification[53]. DIA, also known as SWATH[54], was developed by Aebersold's team and has become a key tool in proteomics. This method allows for the simultaneous fragmentation of all precursor ions by sequential isolation and fragmentation of specific mass-to-charge (m/z) ranges, which improves the identification rates and facilitates comprehensive data reanalysis using various spectral libraries. PRM focuses on the targeted fragmentation of precursor ions, producing unique spectra that are particularly beneficial for analyzing analytes, regardless of their concentration in the samples[55].

DDA is the most widely used label-free MS method and plays a crucial role in the identification of SEPs across various species, including humans[36, 56], *Escherichia coli*[57], and plants[58]. Although DDA is a relatively straightforward method with an efficient workflow, its stochastic nature means that it can only fragment ion peaks with the highest intensity, potentially leading to the omission of low-intensity ions. In contrast, DIA covers a wider dynamic range and exhibits enhanced sensitivity and reproducibility, making it highly effective for SEP identification and quantification. For instance, Pak et al.[59] reported a threefold increase in the number of immunopeptides identified using the DIA workflow.

Labeled quantitation is also crucial for the detection of SEPs. Zhu et al.[60] and Zhang et al.[48] have employed Tandem Mass Tags (TMT) to identify hundreds of SEPs, thereby highlighting the potential biological functions of specific peptides, including TATDN2P1 and BRAWNIN[48, 60].

PRM, known for its improved sensitivity and accuracy, is particularly suited for validating the presence of SEPs identified in preliminary studies[8, 36, 37]. It enables precise quantification and comparison of SEPs with synthetic peptides, utilizing prior data, such as m/z values and retention times. For example, Zhu et al.[60] validated 110 out of 117 SEPs using PRM MS. Additionally, Delcourt et al.[61] employed PRM and isotope labeling to quantify two translation products of the MIEFI gene, revealing AltMiD51 as a significant regulatory element. This demonstrates the capability of PRM in both the validation and functional analysis of SEPs.

1.2.3.4 Database search strategy and database construction

Effective database search strategies are crucial for the identification of proteins, including SEPs[62]. The successful identification of SEPs depends heavily on the quality of the reference databases. Ideally, these databases should encompass all relevant sample-specific SEPs, while minimizing the presence of irrelevant sequences to reduce false positives and enhance search efficiency[47, 63]. As most SEPs are not included in commonly used databases, such as RefSeq and UniProtKB, it is crucial to develop tailored reference databases to discover new SEPs[47, 63].

(1) Database construction from genomic and transcriptomic data

An efficient method for constructing reference databases is *the in silico* six-frame translation of whole-genome sequences. This method has been used to improve protein identification in organisms such as *Escherichia coli*[64], and has been applied to various species, including *Saccharomyces cerevisiae*[41] and *Arabidopsis thaliana*[65]. However, six-frame translation databases are typically limited to organisms with small intronless genomes, because the inclusion of non-canonical sequences can significantly increase database size and search times[66, 67]. To refine the database, integration of genomic annotations and transcriptomic data through three- or six-frame translations facilitates SEP discovery. Tools such as PeptideClassifier[68] and iPtgxDB[69] have successfully identified novel SEPs by combining six-frame genome translations with annotations. Further refinements have been achieved by incorporating evidence from bioinformatic predictions and Ribo-seq data, which improve the reliability of custom SEP databases and facilitate the discovery of novel SEPs across various species[58, 70, 71].

(2) Database construction from Ribo-seq data

Since its introduction in 2009, Ribo-seq has become a powerful tool for identifying actively translated regions in mRNAs, providing insights into the peptide-coding potential of sORFs at a single-codon resolution[72, 73]. Unlike traditional *in silico* methods, Ribo-seq does not rely exclusively on canonical start codons or transcript annotations, thus facilitating the inclusion of AUG- and non-AUG-initiated peptides in reference databases[74, 75].

Tools such as RiboTaper have utilized the 3-nucleotide periodicity of ribosome-protected fragments (RPFs) to identify actively translated regions, which have been incorporated into proteogenomic pipelines, leading to the discovery of hundreds of novel proteins[19, 76]. Subsequent tools, including PRICE, RibORF, and RiboCode, have further enhanced the prediction and validation of sORFs from Ribo-seq data, thereby facilitating the identification of numerous novel SEPs[77, 78].

Despite these advancements, challenges remain in the elimination of translation-irrelevant ribosome binding. To address this, a comprehensive approach that combines bioinformatics analysis, RNA sequencing, and genome-scale CRISPR screening is recommended[79]. Additionally, publicly available databases of sORFs predicted from Ribo-seq data, including OpenProt[9] and sORFs.org[80], serve as important resources for SEP identification, facilitating the construction of reference databases for SEP discovery in various samples.

1.3 Reported biological functions of SEPs

Recent technological advancements have enabled the discovery of an increasing number of SEPs, which display a wide range of biological functions. These functions include embryonic development[81-87], tumorigenesis[88-94], regulation of calcium homeostasis[95-97], mitochondrial metabolism[98-101], and immune modulation[102-107]. There is growing evidence that SEPs have broad potential applications in several areas, including disease management, diagnostics, and drug development.

1.3.1 Embryonic development

Toddler, a microprotein consisting of 58 amino acids encoded by lncRNA *LOC100506013*, has been discovered in zebrafish (*Danio rerio*). It acts as an activator of APJ/Apelin receptor signaling, which is crucial for cardiovascular development and various physiological processes, thereby facilitating the gastrulation process[81]. Pauli et al. demonstrated that zebrafish deficient in Toddler exhibit impaired heart and circulatory system development, highlighting its indispensable role in early embryonic development. Furthermore, microproteins associated with embryonic development have been identified in *Drosophila melanogaster* (fruit flies)[82, 83]. For instance, Kondo et al.[82] discovered that lncRNA *polished rice (pri)* in the epithelial tissue of fruit flies is transcribed into multi-exonic mRNA, which encodes a microprotein *Pri* of either 11 or 32 amino acids. This microprotein plays a crucial role in the regulation of F-actin and the development of epithelial morphology, conversely, the absence of *Pri* function leads to the deterioration of epidermal structure. Galindo et al.[83] identified the gene *tarsal-less (tal)* as essential for embryonic development and morphogenesis in fruit flies, which generates a microprotein as short as 11 amino acids that modulate gene expression and tissue folding. In other species, including vertebrates and Drosophila, a microprotein called Sarcolamba (consisting of 11 amino acids) influences the activity of the calcium transporter protein SERCA, which in turn affects muscle contraction and cardiac development[84]. Chng et al. identified an sORF, *ELABELA*, in human embryonic stem cells, which was present in the outer layer of embryonic cells in zebrafish zygotes. The 32-amino acid peptide hormone encoded by *ELABELA* interacts with the receptor APLNR, thereby establishing a critical signaling axis for early cardiovascular development[85]. Numerous review articles further emphasize that these newly identified SEPs play significant roles in plant growth, development, stress responses, and signal transduction[86, 87].

1.3.2 Cancer progression

The in-depth study of sORFs has led to the discovery of numerous SEPs that play a crucial role in cancer biology, as outlined in **Table 1-2**. These small proteins not only play significant roles in the metabolism and proliferation of cancer cells but also hold promise as novel biomarkers and therapeutic targets. For instance, in colorectal cancer (CRC) cells, lncRNA *HOXB-AS3* has been demonstrated to encode a conserved 53-amino acid microprotein, HOXB-AS3[88]. This microprotein, but not its lncRNA precursor, inhibits the splicing of pyruvate kinase M (PKM) by regulating hnRNPA1, which in turn suppresses aerobic glycolysis. Immunohistochemical (IHC) analysis revealed that CRC patients with lower HOXB-AS3 expression levels tended to have a poorer prognosis. Additionally, *LINC00675*, an lncRNA known to inhibit gastric cancer proliferation, encodes a 79-amino acid microprotein named FORCP (FOXA1 regulated conserved small protein)[89]. FORCP is highly expressed in normal human colon and stomach tissues and is found at elevated levels in more differentiated CRC cell lines. FORCP inhibited CRC cell growth by promoting apoptosis under endoplasmic reticulum stress. Conversely, FORCP depletion promotes cancer cell growth and tumorigenicity. Additionally, lncRNA *LOC90024* encodes a 130-amino acid microprotein that interacts with the splicing regulator SRSF3, thereby modulating mRNA splicing[90]. This interaction correlates positively with malignant phenotypes and poor outcomes in CRC patients, indicating its potential as a prognostic biomarker and therapeutic target.

Pang et al.[91] conducted RNA immunoprecipitation and sequencing analysis on RNAs directly bound to ribosomal protein RPS6 across four cancer cell types. This study led to the identification of *LINC00998*, a lncRNA that is highly expressed in liver cancer tissues and binds directly to ribosomal proteins. *LINC00998* encodes a peptide known as SMIM30 (Small integral membrane protein 30), which is localized to the cell membrane. The peptide SMIM30 promotes HCC cell proliferation, migration, and invasion by binding to the SRC/YES1 tyrosine kinase and activating the MAPK signaling pathway[91]. In contrast, LINC00278 is downregulated in male esophageal squamous cell carcinoma (ESCC). *LINC00278* encodes a microprotein, YY1BM, which interacts with Yin Yang 1 (YY1) and modulates the ESCC progression by inhibiting the interaction between YY1 and the androgen receptor (AR), leading to a reduction in the expression of eEF2K. The decreased expression of YY1BM is associated with smoking, thereby providing a novel mechanistic understanding of the interplay between smoking and the AR signaling pathway[92].

The lncRNA *TINCR* encodes a conserved microprotein known as pTINCR, which is widely expressed in epithelial tissues[93]. This microprotein enhances the SUMOylation modification of CDC42, thereby regulating epithelial cell differentiation and inhibiting the proliferation of epithelial tumors. As an anti-cancer agent in epithelial cells, low pTINCR expression correlates with poor prognosis in patients. Conversely, overexpression of pTINCR promotes cytoskeletal remodeling, strengthens intercellular connections, and upregulates epithelial differentiation markers. Conversely, knockout of pTINCR impedes epithelial cell differentiation[93]. The microprotein MIAC, encoded by lncRNA *AC025154.2*, is significantly downregulated in renal cell carcinoma (RCC). Its expression levels are notably correlated with tumor malignancy and patient prognosis[94]. Further research has revealed that MIAC plays a role in inhibiting the proliferation and metastasis of renal cancer cells by directly interacting with the AQP2 protein, thereby inhibiting the activation of the EREG/EGFR signaling pathway. Synthetic MIAC peptides have displayed considerable anti-tumor efficacy both *in vitro* and *in vivo*, outperforming the clinical agents sunitinib and axitinib. These results highlight the potential of MIAC as a novel approach for the diagnosis and treatment of RCC[94].

FBXW7-185aa plays a pivotal role in the pathophysiology of glioblastoma by inducing cell cycle arrest at the G0/G1 phase and inhibiting cellular proliferation[108]. This protein functions by competitive inhibition of the deubiquitinase USP28, promoting the degradation of c-Myc, which further suppresses tumor growth. Clinical evidence indicates that decreased levels of circ-FBXW7 and FBXW7-185aa in glioblastoma are positively correlated with patient survival, indicating their potential as prognostic biomarkers. Additionally, PINT-87aa, encoded by circPINTexon2, induces cell cycle arrest at the G1 phase and inhibits cell proliferation[109]. This protein functions by interacting with the polymerase-associated factor complex (PAF1c), which inhibits the transcriptional elongation of oncogenes, thereby decelerating glioblastoma progression. The low expression of PINT-87aa in brain tumor tissues further supports its potential as a prognostic indicator for glioblastoma.

In summary, SEPs such as HOXB-AS3, FORCP, FBXW7-185aa, and PINT-87aa play significant roles in cancer progression by regulating critical processes, such as cellular metabolism, proliferation, and apoptosis. These findings not only provide new insights into the molecular mechanisms of cancer but also lay the groundwork for the development of novel diagnostic and therapeutic strategies.

Table 1-2 Summary of microproteins with regulatory roles in human cancers.

Microproteins	**Transcript**	**Length (AA)**	**Subcellular location**	**Cancer**	**Peptide function**	**Ref**
APPLE	*ASH1L-AS1*	90	Endoplasmic reticulum	Acute myeloid leukemia	Regulates translocation initiation and promotes hematopoietic malignancy	[110]
ASAP	*LINC00467*	94	Mitochondria	Colorectal cancer	Modulates ATP synthesis activity, thereby promoting colorectal cancer proliferation	[111]
ASRPS	*LINC00908*	60	Cytoplasm	Breast cancer	Inhibits angiogenesis in triple-negative breast cancer	[112]
DIDO1	*CircDIDO1*	529	Nucleus	Gastric caner	Regulates PARP1 activity	[113]
FBXW7	*circ-FBXW7*	185	Cytoplasm	Glioma	Induces G0/G1 cell cycle arrest, thereby inhibiting proliferation in glioma	[108]
FORCP	*LINC00675*	79	Endoplasmic reticulum	Colorectal cancer	Inhibits proliferation and tumorigenicity in	[89]

					colorectal cancer cells	
HOXB-AS3	*HOXB-AS3*	53	Cytoplasm	Colorectal cancer	Regulates metabolic reprogramming, which in turn inhibits tumorigenesis	[88]
KRASIM	*NCBP2-AS2*	99	Cytoplasm	Liver cancer	Regulates the ERK signaling pathway, resulting in the suppression of hepatocarcinoma cell proliferation	[114]
MIAC	*AC025154.2*	51	Cytoplasm	Head and neck squamous cell carcinoma	Suppresses tumor growth and metastasis	[94]
NoBody	*LINC01420*	68	Cytoplasm	Lung cancer, breast cancer	Promotes tumor invasion by modulating the mRNA decapping process	[56]

PACMP	*Lnc15.2*	44	Nucleus	Breast cancer	Activates PARP1-dependent poly(ADP-ribosyl)ation, which promotes breast cancer growth and PARPi resistance	[115]
PINT87aa	*LINC-PINT* *(circPINT-exon2)*	87	Nucleus	Liver cancer	Inhibits transcriptional elongation in multiple oncogenes	[116]
pTINCR	*TINCR*	87	Nucleus	Epithelial tumors	Induces epithelial differentiation	[93]
RBRP	*LINC00266-1*	71	Nucleus	Colorectal cancer	Binds to m6A reader IGF2BP1 to increase c-Myc stability, thereby promoting colorectal cancer metastasis and tumorigenesis	[117]
SHPRH	*Circ- SHPRH*	144	Cytoplasm	Glioma	Maintains stable full-length SHPRH	[118]
SMIM22	*CASIMO1*	83	Endosome	Breast cancer	Induces SQLE protein accumulation and	[119]

					promotes cancer proliferation	
SMIM30	*LINC00998*	59	Cell membrane	Liver cancer	Activates the MAPK pathway and promotes hepatocellular carcinoma tumorigenesis	[91]
SRSP	*LOC90024*	130	Nucleus	Colorectal cancer	Interacts with the splicing regulator SRSF3, thereby modulating mRNA splicing	[90]
YY1BM	*LINC00278*	21	Cytoplasm	Esophageal squamous cell carcinoma	Under nutrient deficiency, promotes cancer cell apoptosis	[92]

1.3.3 Calcium homeostasis regulation

Calcium ions (Ca^{2+}) are essential regulators of muscle contraction and influence muscle growth, metabolism, and pathological remodeling[120]. Myoregulin (MLN), a microprotein consisting of 46 amino acids and encoded by a skeletal muscle-specific lncRNA, directly interacts with the sarcoplasmic reticulum Ca^{2+}-ATPase (SERCA).

This interaction reduces SERCA's affinity for Ca^{2+}, resulting in a reduced uptake of Ca^{2+} into the sarcoplasmic reticulum and a subsequent decrease in muscle contractility[95]. MLN has been identified as a SERCA-inhibitory microprotein within skeletal muscle. Conversely, a 34-amino acid microprotein encoded by the cardiac-specific lncRNA *dwarf open reading frame (DWORF)* enhances the sarcoplasmic reticulum's capacity for Ca^{2+} uptake by alleviating SERCA inhibition[96]. Furthermore, Anderson et al.[97] identified two additional SERCA-inhibitory microproteins in non-muscle cells: endoregulin (ELN) and another-regulin (ALN). These microproteins share structural and functional similarities with MLN, suggesting that Ca^{2+}-related microproteins are conserved across various cell types and play critical roles in the regulation of Ca^{2+}.

1.3.4 Mitochondrial metabolism

Mitochondria are essential organelles for metabolism and energy supply and contain sORFs within their DNA[98]. Makarewich et al.[99] identified a microprotein named MOXI (micropeptide regulator of β-oxidation) in the mitochondrial inner membrane. MOXI has been shown to interact with mitochondrial trifunctional protein (MTP), thereby enhancing the β-oxidation of long-chain fatty acids. Furthermore, Stein et al. discovered a mitochondrial transmembrane protein named Mitoregulin (Mtln), also encoded by the lncRNA *LINC00116*, in skeletal muscle and cardiac tissues[100]. Mtln serves as an adhesive molecule that improves mitochondrial respiratory efficiency by promoting the assembly and stability of mitochondrial protein complexes.

Additionally, a 16-amino acid microprotein, MOTS-c (mitochondrial open reading frame of the 12S rRNA-c), encoded by 12S rRNA, has been shown to inhibit the folate cycle and the *de novo* synthesis of purine nucleotides. This inhibition activates AMP-activated protein kinase (AMPK) and influences insulin sensitivity[101]. Collectively, these findings highlight the significant role of mitochondria in the regulation of metabolic homeostasis through microproteins at both cellular and organismal levels.

1.3.5 Immunomodulation and other functions

In recent years, SEPs encoded by sORFs have been increasingly recognized for their critical roles in immune regulation. A notable example is miPEP155[102], a microprotein consisting of 17 amino acids encoded by lncRNA *MIR155HG*. This microprotein is known to interact with the heat shock protein HSC70, which is crucial for the modulation of major histocompatibility complex class II (MHC II), thereby significantly contributing to antigen presentation and attenuation of autoimmune inflammation. Another microprotein, C15ORF48, consists of 83 amino acids and is encoded by *NMES1*, which is predominantly expressed in monocyte-derived macrophages in a mouse model[103]. This microprotein modulates mitochondrial function by competitively binding to the mitochondrial cytochrome c oxidase NDUFA4, facilitating NDUFA4 degradation under LPS stimulation, which subsequently affects mitochondrial energy metabolism and immune response. Additionally, Mm47, a 47-amino acid peptide encoded by 1810058I24Rik, is involved in the activation of the Nlrp3 inflammasome[104]. Its increased expression following LPS stimulation and its regulatory effect on interleukin-1 beta (IL-1β) release in mouse models highlight its importance in immune response. Aw112010 interacts with KDM5A to regulate the demethylation of IL-10, thereby influencing IL-6 expression and demonstrating its immunoregulatory function in bone marrow-derived macrophages[105]. Collectively, these findings highlight the significant biological importance of microproteins in immune regulation, laying the groundwork for the development of novel immunotherapeutic strategies.

Moreover, microproteins have been identified as key factors in the degradation of substances, particularly wastes and toxins. The small regulatory polypeptide for amino acid response (SPAR) is a conserved microprotein comprising 90 amino acids encoded by lncRNA *LINC00961*[106]. It is localized within late endosomes and lysosomes, where it interacts with the four subunits of the vacuolar ATPase (v-ATPase) complex on the lysosomal membrane. This interaction negatively regulates activation mTORC1, consequently inhibiting muscle regeneration. These findings suggest that microproteins exhibit a range of biological functions, thereby providing a foundation for the development of innovative therapeutic approaches.

1.4 Exploration on the role of SEPs in the development of novel cancer therapeutics

The pressing requirement in the field of clinical oncology is to discover more effective therapeutic agents for precision and personalized treatment approaches.

Peptides, which function as crucial biomolecular mediators, have several advantages over conventional chemotherapeutic agents, including improved efficacy, selective targeting, and reduced toxicity. SEPs may serve as novel repositories for screening anti-cancer peptides or protein-based therapeutics. Currently, there are approximately 80 peptide drugs on the global market[121], with an additional 150 undergoing clinical trials and 400 to 600 in preclinical development stages[122-124].

1.4.1 SEPs as diagnostic indicators of early disease and effective therapeutic targets for cancer

The objective of cancer research is to identify specific molecular targets for tumor treatment and to develop vaccines that can prevent the initiation and progression of cancer. Additionally, the field is seeking to develop innovative diagnostic tools and biomarkers for early detection of cancer. Recently, there has been growing recognition of the important role of SEPs in cancer progression and development, which has paved the way for the development of anti-cancer therapeutics[88-90]. These SEPs play a crucial role in regulating tumor energy metabolism, metastasis, and growth, thereby presenting themselves as promising candidates for use as biomarkers in cancer diagnosis and therapeutic targets. For instance, such as No body[125], RBRP[117], SMIM22[119], SMIM30[91], FBXW7-185aa[108], PINT-87aa[109] and SRSP[90] are aberrantly expressed in tumor tissues and are closely associated with prognosis, suggesting their potential as biomarkers and targets for inhibiting tumor growth. Specifically, FBXW7-185aa induces G0/G1 cell cycle arrest and inhibits cell proliferation by promoting c-Myc degradation. Clinical evidence suggests that lower levels of circ-FBXW7 and FBXW7-185aa in glioblastoma are associated with improved patient survival, highlighting their prognostic significance[108]. Similarly, PINT-87aa, which is encoded by circPINTexon2, arrests the cell cycle at the G1 phase and suppresses cell proliferation through its interaction with the polymerase-associated factor complex (PAF1c), thereby inhibiting oncogene transcription[109]. Its diminished expression in brain tumor tissues further supports its potential as a prognostic indicator for glioblastoma. Therefore, SEPs may serve as novel diagnostic markers for tumors, potentially facilitating early diagnosis, enhancing the efficiency of tumor detection, improving disease surveillance, and increasing the accuracy of cancer prognostic predictions.

1.4.2 The potential of SEPs as anti-cancer drug candidates

In addition to their noteworthy potential in tumor diagnosis and prognosis, SEPs may also function as promising cancer therapeutic agents owing to their small size, high specificity, and low cytotoxicity. These SEPs demonstrate notable anti-tumor activity and possess considerable potential as anti-cancer drugs by inhibiting tumor metabolic reprogramming, regulating oncogenic protein stability, and disrupting the epithelial-mesenchymal transition (EMT) process. For instance, microproteins such as HOXB-AS3[88], KRASIM[114], YY1BM[92], and ASRPS[112], which exhibit tumor-suppressive effects, can be developed as potential therapeutic agents. Furthermore, several SEPs, including Aw112010[126], miPEP155[127], and

C15ORF48[103], have shown significant efficacy in enhancing cancer immunogenicity, thereby enabling T lymphocytes to accurately identify tumor cells. In summary, SEPs represent a valuable source of targetable tumor-specific antigens, offering considerable promise for the development of tumor vaccines, and demonstrating significant potential as therapeutics to improve cancer immunogenicity.

However, the development of these drugs has several challenges. First, over 90% of peptide drugs are unsuitable for oral administration and must be injected, which is a limitation that impacts the development and application of specific peptide drugs (SEPs). Second, the low expression levels of SEPs complicate their isolation and purification, resulting in high production costs that hinder their clinical application. Moreover, the effective and safe delivery of peptides to targeted organs and cells presents a considerable challenge. To address these issues, researchers have proposed various strategies, including chemical modifications (cyclization, glycosylation, esterification, etc.), to improve the pharmacokinetic properties of peptides and enhance membrane penetration ability[128]. Second, the assembly of nanocarriers and recombinant adenoviral vectors, which are then injected into the patient and subsequently mixed with exosomes for delivery, can achieve a sustained release of peptides, thus enhancing the specificity and efficacy of the drug for cancer treatment. Such approaches can also be employed to improve the therapeutic efficacy, targeting capability, and bioavailability of SEPs. Furthermore, SEPs can be utilized in conjunction with traditional anti-cancer therapies, including radiotherapy and chemotherapy drugs, to increase treatment efficacy. Peptide-drug conjugates (PDCs) have emerged as novel targeted cancer therapies and attracted significant attention[129]. SEPs represent promising reservoirs for the discovery of novel peptides capable of forming PDCs with advantageous pharmacological properties.

In conclusion, SEPs have emerged as important players in the field of cancer research, serving as potential biomarkers, therapeutic targets, or therapeutic agents. These microproteins are characterized by their unique biological activities, high specificity, efficiency, and minimal adverse effects, thereby representing a valuable resource for the discovery of novel anti-cancer drugs. Undoubtedly, the exploration of novel and biologically active SEPs has paved the way for cancer diagnosis and treatment, thereby addressing the urgent need for more effective and precise therapeutic approaches in clinical practice. Further research and technological advances are expected to improve the role of these peptides in future cancer treatments, thereby expanding their applicability and effectiveness.

1.5 Research goals and objectives

In this study, we focused on the role of SEPs within the field of hepatology, aiming to provide a solid foundation for the development of new therapeutic strategies and diagnostic tools. The liver, an important organ during embryonic development, plays a critical role in hepatic organogenesis and hematopoiesis, and is essential for cell proliferation, immune function, and the synthesis and transport of proteins and nucleic acids[130, 131]. Despite its importance, many biological processes involved in liver development remain poorly understood. A comprehensive understanding of the molecular mechanisms underlying liver development may facilitate the development of novel regenerative medical strategies. SEPs are important regulators involved in various biological processes, but their role in liver development has not been fully explored. By developing an optimized proteomic method to identify novel SEPs during mouse liver development, we aimed to identify new molecular participants that are crucial for liver development and function.

Hepatocellular carcinoma (HCC) is a leading cause of cancer-related mortality and morbidity worldwide, with a particularly high incidence and mortality rate among men. It is the second leading cause of cancer-related death among men and the sixth leading cause among women[132]. Most new cases are diagnosed in developing countries, with over 400,000 new cases diagnosed annually in China alone. However, the incidence is also rapidly increasing in developed nations owing to multiple risk factors, such as cirrhosis, hepatitis C virus infections, and rising obesity rates[132, 133]. Owing to the absence of specific clinical symptoms in the early stages, over 70% of patients are diagnosed at an advanced stage[133]. HCC represents a significant public health challenge, particularly given the limited treatment options available for advanced stages, as existing therapies frequently result in drug resistance, which in turn leads to poor survival outcomes[134-136]. Consequently, it is essential to understand the underlying mechanisms of drug resistance and develop novel therapeutic strategies to enhance patient outcomes, mitigate drug resistance, and improve survival rates in individuals with advanced HCC.

SEPs have been shown to play a role in liver pathology by affecting cellular signaling and disease progression[137]. Despite the growing recognition of their biological importance, systematic studies of SEPs within liver tissue face difficulties owing to their low prevalence and the technical constraints associated with their detection. Consequently, the aims of this study are outlined as follows:

(1) Systematic development and optimization of proteomic methodologies for the discovery of SEPs in liver tissues at different developmental stages. By employing this methodology, we will discover novel SEPs that play key roles in liver development and provide a comprehensive proteomic dataset that will serve as a valuable resource for further functional studies on liver development.

(2) To optimize DIA and DDA MS-based proteomics methodological workflow while simultaneously integrating Ribo-seq technology. This integration aims to optimize the identification and quantification of novel SEPs in lenvatinib-resistant HCC cells. By employing these newly developed workflows, we aim to identify these SEPs and clarify their roles associated with drug resistance in liver cancer.

(3) To further clarify the mechanistic functions of these SEPs in liver cancer and their contribution to drug resistance, providing potential new strategies for future clinical interventions.

1.6 Overview of projects

This thesis is organized into four sections, as outlined below:

Chapter 1 introduces the essential concepts of small open reading frame-encoded microproteins (SEPs) and provides an overview of the current methods employed for their identification, including three key technologies. We provide a comprehensive overview of mass spectrometry (MS) detection methods and workflows, focusing on four main points. We summarize the functional SEPs that have been investigated thus far, emphasizing their associations with embryonic development, tumorigenicity, calcium homeostasis, and mitochondrial metabolism. Additionally, we delve into the role of SEPs in the development of anti-cancer drugs, underscoring their considerable potential in disease treatment and drug discovery, given their broad application prospects.

Chapter 2 describes the optimization of a MS-based workflow for the effective extraction and enrichment of SEPs from biological matrices. Given the typically low abundance of SEPs, it is necessary to use substantial quantities of homogenous protein samples, a condition that can be satisfied by employing liver tissues. This study endeavors to establish an optimized proteomic methodology to identify novel SEPs across various developmental stages of the mouse liver, with the objective of revealing new molecular participants that play a vital role in liver development. The details of this chapter are as follows.

Optimization of Protein Extraction: Comparison of various protein extraction methods to enhance overall efficiency.

Development of Protein Enrichment Methods: A comparison of ten methods from four categories was conducted to find the most efficient method for enriching microproteins from total proteins. Finally, we employ a novel size exclusion chromatography (SEC) method for SEPs enrichment, which separates proteins by size using gel filtration.

Optimization of Fractionation Processes: To improve protein fractionation techniques to effectively separate SEPs from other proteins and provide purified samples for MS analysis.

Construction of Databases: Use ribosomal sequencing data from the same tissue samples to customize a high-specificity database and enhance SEP discovery.

Comparison and Evaluation of Methods: To evaluate the efficacy of different protein enrichment and extraction methods for specific types of SEPs, which will inform and direct future research efforts.

In summary, we systematically optimized the entire workflow, which includes protein extraction, enrichment, fractionation, and database construction, leading to a notable increase in the identification efficiency of SEPs. We also identified novel molecular participants that are crucial for liver development and function.

Chapter 3 delves into the urgent need for more effective treatments, as our understanding of tumorigenicity mechanisms deepens. In recent years, the role of SEPs in tumor development and progression has become a focus of attention, offering new avenues for the development of novel anti-cancer therapeutic strategies. Therefore, in addition to the optimization of sample processing methods discussed in Chapter 2, we have refined the MS and Ribo-seq-based workflow for the SEP discovery, specifically screening for functional SEPs in lenvatinib-resistant liver cancer cells. The details of this chapter are outlined below:

Application of an Optimized Proteogenomic Workflow: Following the optimization of sample processing methods, as detailed in Chapter 2, we proceeded to further enhance the MS-based proteogenomic workflow. We employed a combination of two MS methodologies: DIA for comprehensive identification, and DDA for reliable identification and quantification. Furthermore, we integrated an experimentally generated spectral library with an *in silico* predicted spectral library to maximize SEP identification. Through these improvements, we were able to achieve comprehensive characterization and precise quantification of SEPs within liver cells, enabling in-depth identification and precise quantification and facilitating the discovery of SEPs that influence sensitivity to drugs.

Validation of the Expression of SEPs: Detection of the presence and variation levels of SEPs in drug-resistant liver cancer cells using PRM MS method, and validation of their translational potential through plasmid transfection.

Discovery of the Biological Functions of SEPs: Both *in vivo* and *in vitro* experiments to validate the functions of the identified SEPs, thereby confirming their involvement in drug resistance of liver cancer. This research aims to establish a scientific foundation for the identification of future therapeutic targets or agents.

Chapter 4 explores the molecular mechanisms by which SEPs contribute to drug resistance in HCC. Additionally, we explore the potential roles of SEPs across various cancer types and suggest possible applications for these peptides. The details of the chapter are outlined as follows:

Biological Significance of SEPs: Experimental validation of the functions of the identified SEPs, confirming their involvement in drug resistance in liver cancer.

Exploration of the Molecular Mechanisms of SEPs: A co-immunoprecipitation-based affinity purification coupled with mass spectrometry (AP-MS) workflow was employed to investigate the interactions and molecular mechanisms underlying the actions of SEPs.

Investigation of the Broad Applications: Application of the discovered SEPs to other disease models or drugs to evaluate their generalizability and effectiveness across diverse biological contexts.

In summary, we have presented a comprehensive methodological optimization of the SEP discovery workflow, which includes both sample preparation and MS-based analytical identification. Based on this, we have identified new molecular participants that are crucial for liver development and cancer biology. This thesis provides a comprehensive framework for the investigation of SEPs, detailing their identification methodologies and functional characteristics within the context of hepatopathy, thereby paving the way for future research and therapeutic development.

Chapter 2. The optimization of proteomic approaches reveals novel microproteins in mouse liver development

2.1 Introduction

Small open reading frames (sORFs) are translated sequences that have traditionally been excluded from genome annotation and are also known as AltORFs or smORFs, which contain any unannotated coding sequence of any reading frame of mRNA or alleged ncRNA[7, 138]. The translation products of sORFs are termed small open reading frame-encoded microproteins (SEPs), also known as AltProts or microproteins, which differ from canonical reference proteins (RefProts) of the same gene. Recently, microproteins have been shown to play essential roles in a variety of physiological processes and diseases[139-142], such as metabolism[143], transcriptional[144]/translational regulation[145], ion signaling[96], and development[81, 145].

However, the discovery of functional microproteins is mostly serendipitous. To date, we still lack a systematic approach to directly identify microproteins from biological specimens on a large scale[146]. The Ribosome profiling (Ribo-seq) technique sequences ribosome-protected RNA fragments and thus enables the prediction of thousands of sORFs using bioinformatics pipelines[147]. Currently, mass spectrometry (MS) is the only method that allows direct identification of microproteins[148]. However, only tens to hundreds of microproteins per sample can be identified using MS[138, 149-151]. The large difference in identification numbers between the two methods calls for urgent improvements in MS-based methodologies to detect microproteins. The discovery of microproteins by MS is challenging, partly because of their short length and interference from large canonical proteins[152]. Another major obstacle is the lack of well-established microprotein databases. The efficiency of microprotein discovery is far inferior to that of RefProts, which uses public databases that combine all translational products from various samples. Considering the high temporal/spatial specificity of microprotein translation, it is important to use a customized database of the same specific samples for mining novel microproteins. Although several prior works have improved the microprotein sample preparation procedures or database construction, there is still a vast room for improvement[138, 151, 153].

Protein translation plays a crucial role in embryonic development and is regulated precisely[145, 154]. Although many canonical proteins and their mechanisms in developmental biology have been thoroughly investigated, only a few microproteins have been studied[81, 145]. Considering that microproteins could also play pivotal roles in development, either independently or through the regulation of canonical proteins, large-scale and accurate identification of microproteins is crucial for understanding the mechanisms of embryonic development.

Herein, we report an optimized approach that integrates MS and Ribo-seq techniques to identify microproteins with improved depth and efficiency. Using this optimized approach, we discovered and quantified stage-dependent microproteins in embryonic and adult livers that were enriched in specific biological pathways. Our study not only provides a proteomics approach, but also novel microproteins as new players in liver development.

2.2 Materials and methods

2.2.1 Chemicals and reagents

Acetonitrile, methanol, formic acid, trichloroacetic acid, water (HPLC grade), 16% Tricine gel, and Tricine SDS running buffer were purchased from Thermo Fisher Scientific (Massachusetts, USA). Acetic acid, ethanol, and chloroform were from DUKSAN (Gyeonggi-do, Korea). Lysyl endopeptidase (Lys-C, mass spectrometry grade) and trypsin (sequencing grade) were purchased from Promega (Madison, USA). Ammonium formate, ammonium bicarbonate, DL-dithiothreitol (DTT), and iodoacetamide (IAM) were from Sigma Aldrich (Missouri, USA) and all other reagents were from Sigma Aldrich.

2.2.2 Animals and tissue collection

To compare microprotein enrichment and fractionation methods from liver total lysates, C57BL/6 mice weighing between 18 and 22 g were purchased from Centralized Animal Facilities, The Hong Kong Polytechnic University, Hong Kong. Adult mice were anesthetized and perfused with isotonic saline containing protease inhibitors (0.120 mM EDTA, 0.2 mM PMSF, and Roche Complete Protease Inhibitor tablets, pH 7.4) before decapitation. Livers were quickly dissected and immediately snap-frozen in liquid nitrogen. All animal experiments were approved by the Hong Kong Polytechnic University Animal Subjects Ethics Subcommittee (Approval No:

20-21/275-ABCT-R-STUDENT) and performed in accordance with the Institutional Guidelines and Animal Ordinance of the Department of Health.

For discovery of microproteins in embryonic liver development, livers were harvested separately from embryonic (E15.5) and adult (P42) C57BL/6 mice and immediately snap-frozen in liquid nitrogen. The mice were purchased from the Guangdong Medical Experimental Animal Center (Guangdong, China; License No: SCXK (YUE) 2018 0002). All experimental procedures were approved by the Animal Ethics Committee of the Zhongshan Ophthalmic Center, Sun Yat-sen University (Guangzhou, China; License No: SYXK (YUE) 2018 0189) and in accordance with the institutional animal welfare guidelines and Animal Protection Law of China.

2.2.3 Protein extraction and microprotein enrichment

Mouse liver tissues were obtained from The Hong Kong Polytechnic University25%ree different microprotein extraction methods were compared: (1) RIPA lysis buffer (50 mM Tris-HCl, 150mM Sodium chloride (NaCl), 2mM EDTA, 1% NP40, 1% Sodium Deoxycholate), (2) acid lysis buffer (50 mM hydrochloric acid (HCl), 0.1% β-mercaptoethanol; 0.05% Triton X-100)[151], and (3) boiling water[151]. The extracts were then centrifuged at 16,000 × g for 20 min at 4 °C to remove residual debris.

We tested 10 enrichment methods in triplicates from four categories, (1) precipitation, (2) size selection, (3) solid phase extraction (SPE) enrichment method, (4) hexagonal mesoporous silica materials, using equal amounts of lysates.

2.2.3.1 *Precipitation category*

Precipitation methods precipitate larger proteins to decrease supernatant complexity and enrich small proteins, according to previous protocols.

(1) Acetic acid (AA) at 0.25% or AA at 25% precipitation

AA (0.25%, v/v) [151] or (25%, v/v) in water[138] was added to the tissue homogenate supernatant followed by centrifugation at 16,000g for 20 min at 4 °C. The supernatant was then collected.

(2) Trichloroacetic acid (TCA) precipitation

TCA 20% was added to the samples as 1:1 (v/v), followed by chloroform ($CHCl_3$) 1:1 (v/v). The sample was centrifuged at 1,500 g for 10 min at 4 °C and the supernatant was transferred to a new tube. The lower sample was then washed with 100 μL of Milli-Q water and 100 μL of methanol, followed by vortexing and centrifugation at 1,500 g for 10 min at 4 °C. Subsequently, both supernatants were combined[42].

(3) Acetonitrile (ACN) precipitation

A 3.2-fold volume of ACN spiked with 0.1% trifluoroacetic acid was added to the sample and then vortexed for 30 s and incubated for 1 h. Then the sample was centrifuged at 16,000g for 20 min and the supernatant was transferred to a new tube[38].

(4) Methyl tert-butyl ether (MTBE)-based sequential precipitation

We used a single-phase buffer MTBE/methanol/water (5:3:1, v/v) and two-phase buffer MTBE/methanol/water (5:1:1, v/v) for precipitation and delipidation as described previously[155]. MTBE/methanol/water (5:3:1, v/v) was first added and incubated for 30 min at 4 °C for precipitation, and the supernatant was transferred into a new tube after centrifugation at 21,000 g for 20 min at 4 ℃. Then MTBE/methanol/water (5:1:1, v/v) was added. After vortex, the small proteins in the lower phase were collected by centrifugation at 1,000 g for 10 min at 4 °C. Subsequently, both were combined.

2.2.3.2 Size selection category

(1) 30-kDa-molecular weight cut-off ultrafiltration (30-kDa-MWCO)

The liver homogenate supernatant was loaded into a 30-kDa-MWCO (Millipore) for centrifugation at 12,000 g for 20 min, and the flow through was collected[151].

(2) Size-exclusion chromatography (SEC) enrichment

To isolate proteins <30 kDa from liver lysates, a GE AKTA Explorer FPLC System (GE Healthcare) was combined with a Sephadex 75 Increase 5/150 GL column (GE Healthcare) for both enrichment and fractionation. The column was equilibrated with 3 column volumes of SEC running buffer (ammonium bicarbonate) prior to sample analysis. Low molecular weight standards (GE Healthcare) were used for mass calibration. Each SEC separation run was performed at a flow rate of 0.2 mL/min at a wavelength of 254 nm for 15 min. Only fractions between 8 min and 15 min of retention time were collected into a low protein binding tube (Eppendorf). These fractions corresponded to proteins of molecular weight <30 kDa in a total volume of 1.6 mL. For SEC enrichment purpose, these fractions were combined into one tube and lyophilized before use.

2.2.3.3 *Solid phase extraction (SPE) category*

(1) C8 SPE-based enrichment

C8 SPE cartridges (Agilent Technologies) were activated with one column volume of methanol and then equilibrated with two-column volumes of triethylammonium formate (TEAF) buffer (pH 3.0) before the lysate was applied. The cartridges were then washed with two column volumes of TEAF buffer (pH 3.0) and the enriched proteins were eluted with ACN: TEAF buffer (3:1, pH 3.0)[151].

(2) Hydrophilic-lipophilic-balanced SPE (HLB SPE, Waters)-based enrichment

HLB SPE cartridges (Waters) were activated with methanol and then equilibrated with water before the lysate was applied. The cartridges were then washed with water and eluted with 60% ACN.

2.2.3.4 *Hexagonal mesoporous silica materials MCM-41 (MCM-41)*

MCM-41 beads were mixed with liver lysates and small proteins were extracted as described by Du et al[156]. After incubation and shaken for 30 min, the protein adsorbed was eluted with ACN/0.5 M HCl (1:1, v/v) by centrifugation at 12,000 g for 10 min.

After enrichment, to compare the enrichment effect of different methods for small proteins, all samples need to be concentrated by speed-vac. Braford protein assay was used to measure protein concentration, and an equal amount of total protein was analyzed by Tricine 4-12% and BisTris SDS-PAGE gel.

2.2.4 Protein sample cleanup with the SP3 method

For each 20 μg sample, Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophobic and Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophilic (GE Healthcare) were gently combined in a ratio of 1:1 (v/v) and used as described by Hughes et al.[157]. The samples were then reduced and alkylated using DTT and IAM. The bead slurries were then transferred to samples. Subsequently, absolute ethanol was added at a final concentration of 50% (v/v). Beads were resuspended in 50 mM ammonium bicarbonate supplemented with Lys-C enzymes at an enzyme-to-protein ratio of 1:100 (w/w). After 4 h of incubation, trypsin was added at an enzyme-to-protein ratio of 1:20 (w/w), as 1:25 was recommended by Hughes et al., for complete digestion, and the sample was incubated at 37 °C for 12 h. The peptide concentration was determined using a Pierce Quantitative Fluorometric Peptide assay (Thermo Fisher Scientific). For each sample, the peptides were labeled with TMT6plex (including channels 126, 127N, 127C, 128N, 128C, and 129N; Thermo Fisher Scientific) according to the manufacturer's instructions.

2.2.5 SDS-PAGE tricine gel analysis of enriched microproteins samples

After enrichment, the protein content was quantified by the Bradford assay, and the same amount (25 μg) of protein was loaded onto each lane of the gel. The samples were analyzed using 16% tricine-SDS-PAGE and separated at a constant voltage of 60 V until they completely entered the separating gel from the stacking gel. Then, a constant voltage of 110 V was maintained until the tracking dye reached the gel bottom. Finally, the gel was stained with Coomassie Brilliant Blue R-250 (Bio-Rad).

2.2.6 Comparison of fractionation methods after SEC enrichment

2.2.6.1 SEC enrichment into 4 fractions (SEC-fraction)

Mouse liver samples were loaded onto the SEC column, and the final four fractions of the low-molecular-weight range were collected and injected separately into the MS for detection.

2.2.6.2 High-pH reversed-phase fractionation

After SEC enrichment, the obtained proteins were digested and the peptides were fractionated using a Waters Acquity UPLC Peptide BEH C18 column (2.1 × 100 mm, 1.7 μM, Waters) on an Agilent 1290 Infinity LC system (Agilent Technologies) operating at 50 μL/min. Buffer A consisted of 10 mM ammonium formate and buffer B consisted of 10 mM ammonium formate and 90% ACN, both of which were adjusted to pH 9 using ammonium hydroxide, as described previously[158]. Fractions were collected every 1 min from 6 min to 100 min retention time (96 fractions, finally concatenated into 8 fractions). Peptides were separated using a linear gradient as follows: 0-10 min, 1% B; 10-38 min, 1-8% B; 38-75 min, 8-62% B; 75-85 min, 62-95% B; 85-100 min, 95% B. The final eight fractions were concentrated and analyzed by LC−MS/MS.

2.2.6.3 ERLIC fractionation

After SEC enrichment, the obtained protein was digested, and the peptides were fractionated using an Agilent 1290 Infinity LC system equipped with a PolyWAX

ERLIC column (200 × 2.1 mm, 5 μM, 300 Å, PolyLC), as described previously[159]. Buffer A consisted of 90% acetonitrile and 0.1% acetic acid, and buffer B consisted of 30% acetonitrile and 0.1% formic acid. From 6 min to 100 min retention time, fractions were collected every 1 min (96 fractions, finally concatenated into 8 fractions). Peptides were separated by a stepwise gradient as follows: 0-10 min, 0% B; 10-22 min, 0-8% B; 22-38 min, 8-45% B; 38-50 min, 45-80% B; 50-68 min, 80-98% B; 68-100 min, 98% B. The final eight fractions were concentrated and analyzed by LC−MS/MS.

2.2.7 LC-MS/MS analysis

For data-dependent acquisition, all mass spectrometry data were collected on an Orbitrap Exploris 480 mass spectrometry equipped with the FAIMS interface and coupled with an Ultimate 3000 RSLC nano system (Thermo Fisher Scientific). The digested samples were re-dissolved in 0.1% FA and separated on a self-packed capillary column packed with Reprosil-Pur C18 1.9 μM particles (Dr. Maisch GmbH). Mobile phase A (0.1% formic acid) and mobile phase B (80% ACN and 0.1% formic acid) were used to separate the peptides with the following gradients: 2 min, 8–10% B; 2-120 min, 10−35% B, 120-140 min, 35-90% B; 140-150 min, 90%B in bottom-up proteomics, at a constant flow rate of 300 nL/min. Full-scan spectra were measured with a resolution of 120,000 within a maximum injection time of 50 ms, followed by MS2 scans with a resolution of 30,000 within a maximum injection time of 55 ms. The isolation window of the MS2 scan was set to 1.6 *m/z*, and only ions with 2-6 charges were triggered during the MS2 event. The normalized collision energy (NCE) was set to 32. The dynamic exclusion time was set to 45 s. The compensation voltages were set at -45 V and -65 V to remove singly charged ions.

2.2.8 Construction of a putative microproteins database

This study used the previously reported Ribo-seq dataset[145]. Briefly, preprocessing of Ribo-seq raw data included adaptor removal using Cutadapt[160] (v 2.4, with parameters "--minimum-length 6 --discard-untrimmed --match-read-wildcards --max-n=0.5"), low-quality trimming using Sickle[161] (v 1.33, with parameters "se -x -t sanger"). rRNA and tRNA contaminants were removed by aligning the trimmed reads to mouse tRNA and rRNA sequences (5S, 5.8S, 18S, and 28S) using Bowtie 2[162] (v1.0.1, with command "-q -L 20 --phred33 --end-to-end"). All remaining reads were mapped to the mouse reference genome GRCm 38 with a GTF annotation file (GENCODE vM25) using STAR (v 2.7.2 a)[163], and further unique mapped reads were extracted. Ten pipelines, RiboTISH (v 0.2.1)[164], ORFquant (v 0.99.0)[165], ORFRATER[18], RiboCode (v 1.2.11)[166], riboHMM[167], Ribotricer (v 1.3.1)[168], RiboWave (v 1.0)[169], RP-BP (v 2.0.0)[170], RibORF (v 1.0)[171] and PRICE (v 1.0.3b)[75], were used to perform ORF and sORF detection with the longest strategy under the default threshold setting (**Table 2-1**). The final set of actively translated ORFs with all near-cognate start codons (AUG, TUG, CUG, and GUG) followed by an in-frame stop codon in annotated transcripts was stringently filtered based on the requirement of a minimum length of 18 nucleotides and the expression of the ORF-containing gene at an above-background level, as described in a previous report[145]. ORFs that passed the above filtering criteria were classified into several categories based on their relative location with the nearest annotated CDS, as described previously[7]. In the classification results, ORFs were defined as annotated proteins. Upstream ORFs (uORFs) and downstream ORFs (dORFs) were defined as sORFs originating from the 5’UTRs and 3’UTRs of annotated protein-coding genes, respectively. Long non-coding RNA ORFs (lncRNA-ORFs) were defined as sORFs originating from transcripts currently annotated as long non-coding RNAs (lncRNAs). Upstream overlapping ORFs (uoORFs), downstream overlapping ORFs (doORFs), and internal out-of-frame ORFs (intORFs) were defined as sORFs located upstream, downstream, and intermediate of the CDS and out-frame overlapping with annotated CDSs, respectively. Finally, nucleic acid sequences of all actively translated sORFs were converted into amino acid sequences in FASTA format for the construction of protein databases.

Table 2-1 Summary of algorithms used for ORF prediction tools in Ribo-seq.

Methods	Prediction strategy	Start codon	Cutoff	PMID	Version	Installation: Language	Year	Source
Ribotricer v1.3.1	Low-dimensional projected vector	NTG	phase score > 0.4285	31750902	1.3.1	Conda: Python3	2020	https://github.com/smithlabcode/ribotricer
RiboCode v1.2.11	Wilcox single rank test	NTG	combined p-value < 0.05	29538776	1.2.11	Conda: Python2	2018	https://github.com/xryanglab/RiboCode
RiboWave v1.0	Wavelet transform	NTG	p-value < 0.05	29945224	1.0	Git-Hub: Shell and R	2018	https://github.com/lulab/Ribowave
Ribo-TISH v0.2.1	Wilcoxon rank-sum test	NTG	RiboQvalue (FrameQvalue) < 0.05	29170441	0.2.1	Git-Hub/Pip: Python2/3	2017	https://github.com/zhpn1024/ribotish
RP-BP v2.0.0	Unsupervised Bayesian	NTG	#bayes_factor_mean > 5	28126919	2.0.0	Git-Hub: Python3	2017	https://github.com/dieterich-lab/RP-BP

ORF-RATER	Linear regression and a random forest classifier	NTG	orfrating > 0.8	26638175	Git-Hub	Git-Hub: Python2	2015	https://github.com/alexfields/ORF-RATER
RibORF v1.0	Percentage maximum entropy	NTG	pred. pvalue > 0.7	26687005	1.0	Git-Hub: perl	2015	https://github.com/zhejilab/RibORF
riboHMM	Hidden Markov model (HMM)	NTG	posterior > 8000	27232982	Git-Hub	Git-Hub: python	2016	https://github.com/rajanil/riboHMM
ORFquant (SaTAnn) v0.99.0	Splicing reads and multitaper method	ATG only	pval_uniq < 0.05	32601440	0.99.0	Git-Hub: R	2020	https://github.com/lcalviell/ORFquant (https://github.com/lcalviell/SaTAnn)

PRICE v1.0.3b	EM algorithm and generalized binomial test	NTG, ANG and ATN only	FDR < 0.1	29529017	1.0.3b	Git-hub: java	2018	https://github.com/erhard-lab/gedi/wiki/Price

2.2.9 Identification of canonical proteins and microproteins

The LC-MS/MS raw data were analyzed using MSFragger (version 3.3). The common parameters were set as follows: precursor mass tolerance: 10 ppm, fragment mass tolerance: 0.02 Da; trypsin as enzyme; 2 missed cleavages; oxidation (methionine), acetyl (protein N-term), and TMT-6plex (N terminus) as variable modifications; carbamidomethylation (cysteine) and TMT-6plex (lysine) as fixed modification; the validation was performed using PeptideProphet; the FDR was set as 1%. Two different protein databases were used in this study: (1) Mouse OpenProt and sORF databases were used for comparison of enrichment methods. Mouse OpenProt protein database was derived from OpenProt (https://openprot.org, version number 1.6, 01 Sep 2020)[172] and contains 563,275 entries consisting of RefProts, novel isoforms and microproteins predicted from both Ensembl and RefSeq. There were 503,679 entries in the Mus musculus microprotein protein database from sORF.org (http://www.sorfs.org, downloaded on 01 Jun 2021)[173]; (2) In-house mouse microprotein database had 146,461 entries, which were used for microprotein discovery in TMT-labeled embryonic and adult livers. Identification of microproteins was always based on a peptide specific to the microprotein sequence and not common with the RefProts. The results from the custom database search were further filtered against the reference mouse proteins database (RefProt, containing Ensembl, NCBI RefSeq, and UniProtKB) using a stringent string-searching-based mapping algorithm to ensure that we did not report any known protein degradation, mutants, or isoforms.

We performed Gene Ontology (GO) analysis mainly based on annotated sORFs, which are in the same genes that encode the related upstream ORFs (uORFs), downstream ORFs (dORFs) and upstream overlapping ORFs (uoORFs), as well as lncRNA-ORFs that were encoded by the retained introns of protein coding genes with known functions. GO analysis was performed with R package clusterProfiler (v4.0.5).

2.2.10 Validation of novel microproteins with parallel reaction monitoring

For parallel reaction monitoring (PRM), the samples were separated on the same LC-MS system using a 150 min gradient. Full scan spectra were measured with a resolution of 120,000 within a 50 ms maximum injection time, followed by targeted peptide MS2 scans with a resolution of 30,000 within a 60 ms maximum injection time under the 1.2 *m/z* isolation window. The normalized collision energy was set to 30. PRM data (tier 3 level) were processed using Skyline (version 21.1) software, as described previously[174]. The predicted Rt and MS/MS spectra were calculated using two deep learning tools, DeepRT[175] and pDeep2[176], respectively.

2.2.11 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.

2.2.12 Experimental design and statistical rationale

To test the performance of different microprotein enrichment methods, we performed triplicate for each enrichment method using adult C57BL/6 mice liver samples. To investigate microprotein expression during liver development, the livers of embryonic (E15.5) and adult (P42) C57BL/6 mice were used in triplicate. Data were analyzed by a two-tailed unpaired Student's t-test (unless otherwise indicated), and $p < 0.05$ was selected as the statistical limit of significance. We selected * and ** for $p < 0.05$, and $p < 0.01$, respectively. Unless otherwise stated, all data in the graphs are expressed as arithmetic mean ± standard deviation (SD) from at least three repeated experiments.

2.3 Results

2.3.1 Optimization of microprotein extraction methods

Considering the distinct lengths and properties of canonical reference proteins (RefProts) and alternative proteins (microproteins)[177], the identification of microproteins using classical proteomic methods is analytically challenging. Therefore, we sought to improve the proteomics workflow at multiple steps, including protein extraction, microprotein enrichment, and peptide fractionation, by comparing various conditions.

First, three widely employed protein extraction methods, RIPA lysis buffer, acidic lysis buffer, and boiling water, were tested for extracting microproteins from mouse liver homogenates. Significant protein loss was observed with acid lysis buffer and boiling water, although they have been reported for extraction of small proteins by preferentially causing aggregation of high-molecular-weight proteins[151, 152]. In contrast, RIPA lysis buffer offered a much higher efficiency for total protein extraction and was therefore adopted in all subsequent experiments (**Figure 2-1**).

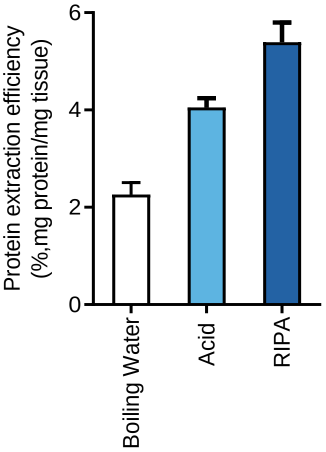


Figure 2-1 Comparison of protein extraction efficiency (%, mg protein/mg tissue) in the

three lysis buffers using mouse liver tissues.

2.3.2 Size-exclusion chromatograph is the most efficient method for microprotein enrichment

Next, we tested ten methods from four categories to find the most efficient method for enriching microproteins from total proteins. In the first category “precipitation”, organic solvents or acids precipitated high-molecular-weight proteins and subsequently enriched microproteins. In the second category “size selection”, ultrafiltration tubes and size-exclusion chromatograph (SEC) enabled separation of proteins by size. In the third category “solid phase separation”, the non-polar reversed-phase sorbent trapped large hydrophobic proteins, while small and polar proteins were eluted and enriched. The fourth category is hexagonal mesoporous silica materials MCM-41, which enabled selectively enrich peptides and small protein through size selectivity and adsorptive mechanism. The efficiency of the methods was compared side-by-side based on gel images and/or MS analysis of the enriched proteins. Based on tricine gel and glycine gel images, most methods were able to remove proteins larger than 40 kDa efficiently. However, the enriched proteins displayed vastly different profiles (**Figures 2-2**).

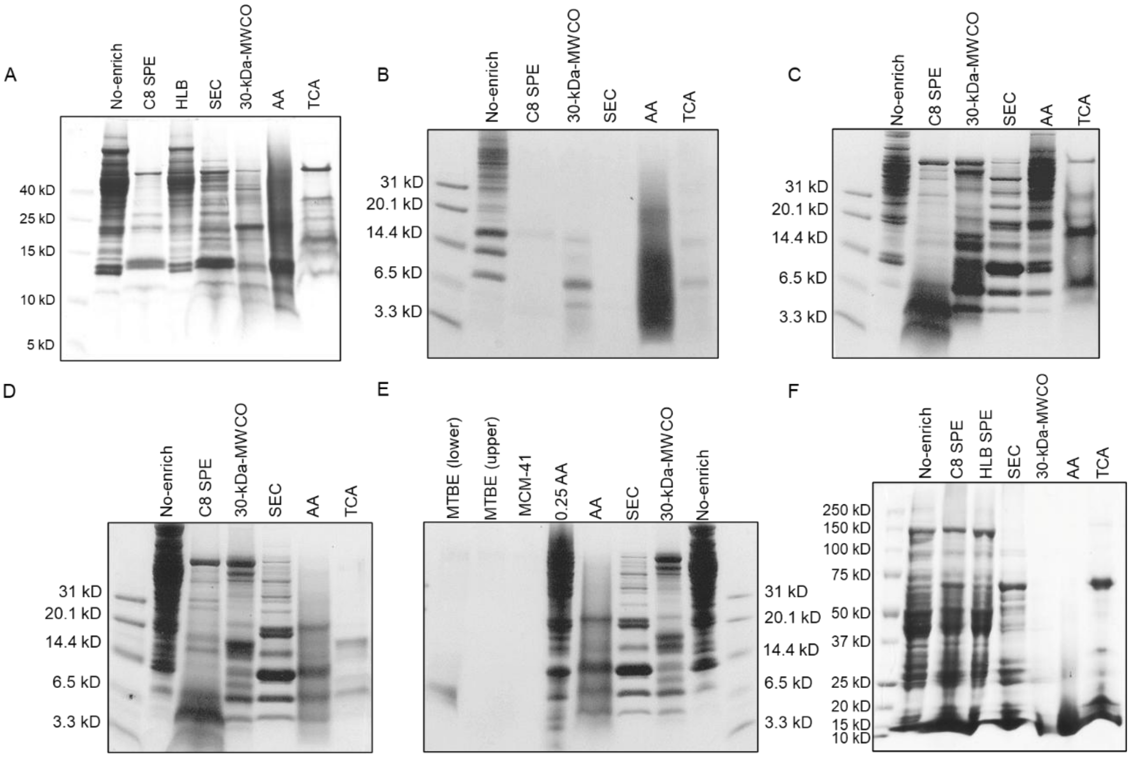


Figure 2-2 Comparison of different extraction and enrichment methods using mouse liver tissues. (A) Liver lysates were lysed in RIPA lysis buffer followed by enrichment, including C8 SPE, HLB SPE, SEC, 30-kDa MWCO, AA precipitation, or TCA precipitation. The results from these enrichments were analyzed by SDS-PAGE (Coomassie staining). (B) Tricine SDS-PAGE results of boiling water extraction combined with different enrichment methods. (C) Tricine SDS-PAGE results of acid lysis buffer extraction combined with different enrichment methods. (D) (E) Tricine SDS-PAGE results of RIPA lysis buffer extraction combined with different enrichment methods. (F) Glycine SDS-PAGE results of RIPA lysis buffer extraction combined with different enrichment methods.

Trichloroacetic acid (TCA) precipitation, acetic acid (AA) precipitation, C8 Solid-phase extraction (C8 SPE), hydrophilic-lipophilic-balanced solid-phase extraction (HLB SPE), 30-kDa-molecular weight cut-off ultrafiltration (30-kDa-MWCO), and SEC resulted in strong protein bands and therefore were chosen for the following comparison with MS. With equal protein amounts, the highest identification number was achieved using SEC enrichment, with an average of 51 microproteins identified, which was more than twice that of the other methods (**Figure 2-3A**). Meanwhile, although the intensity of RefProts was similar across all tested methods, the intensity of microproteins after SEC enrichment was five folds higher than that of other methods. SEC greatly reduced the difference between RefProts and microproteins in terms of MS intensity, which demonstrated its effectiveness in concentrating microproteins out of the total lysates (**Figure 2-3B**).

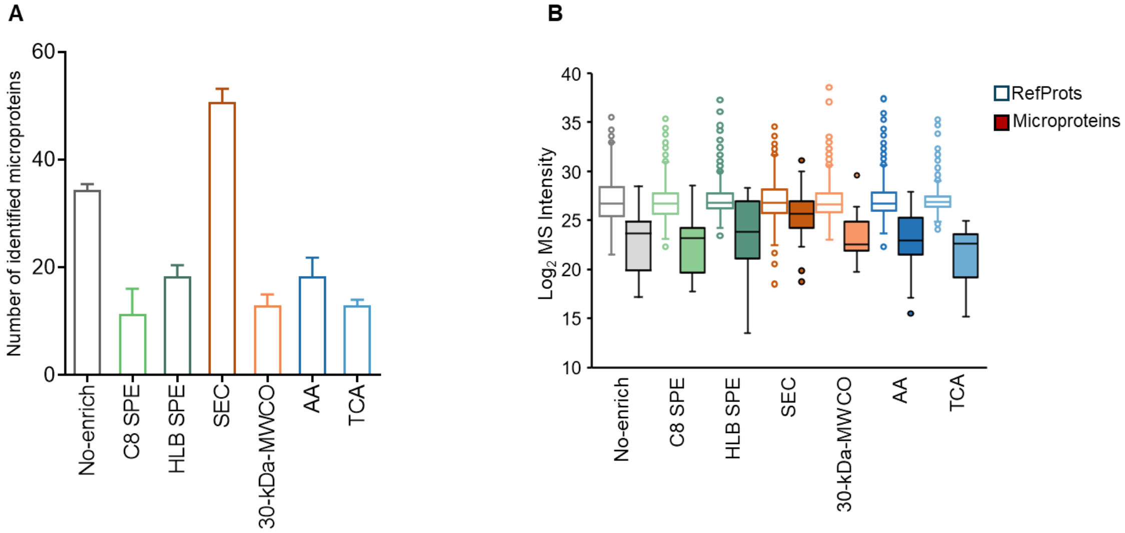


Figure 2-3 Comparison of different enrichment methods for microproteins in mouse livers. (A) Average number of microproteins detected using different enrichment methods. (B) MS intensity of the identified microproteins and RefProts in each enrichment method.

2.3.3 Characteristics of microproteins enriched by various methods

Given the complementary nature of these enrichment methods, there were only a few microproteins commonly identified by using different categories of methods (**Figure 2-4**). Although the individual methods did not yield a high number of microproteins, the methods collectively contributed a greater variety of microproteins. In our study, we found that No-enrich and SEC method were complementary in identifying different categories of microproteins. The reproducibility was higher within the same category than between categories. For example, over 60% of microproteins identified with TCA precipitation were reproducibly identified with acetic acid precipitation. Among all the methods, SEC was found to be the most comprehensive. For microproteins that were identified by multiple enrichment methods, SEC resulted in the highest intensities (highlighted in red in **Figure 2-4A**). Next, we analyzed the hydrophobicity and isoelectric point (pI) to investigate whether microprotein identification was associated with their biophysical properties (**Figures 2-4**). As expected, the acid precipitation methods enriched more hydrophilic microproteins with lower GRAVY scores (**Figures 2-4C**). TCA precipitation and acetic acid precipitation preferentially enriched more microproteins with a high pI compared to the other methods (**Figure 2-4D**). Such differential biophysical properties partially explained the observation that a complementary pool of microproteins was enriched with different methods. SEC-based method enriched microproteins with evenly distributed hydrophobicity and pI and therefore was the most efficient method.

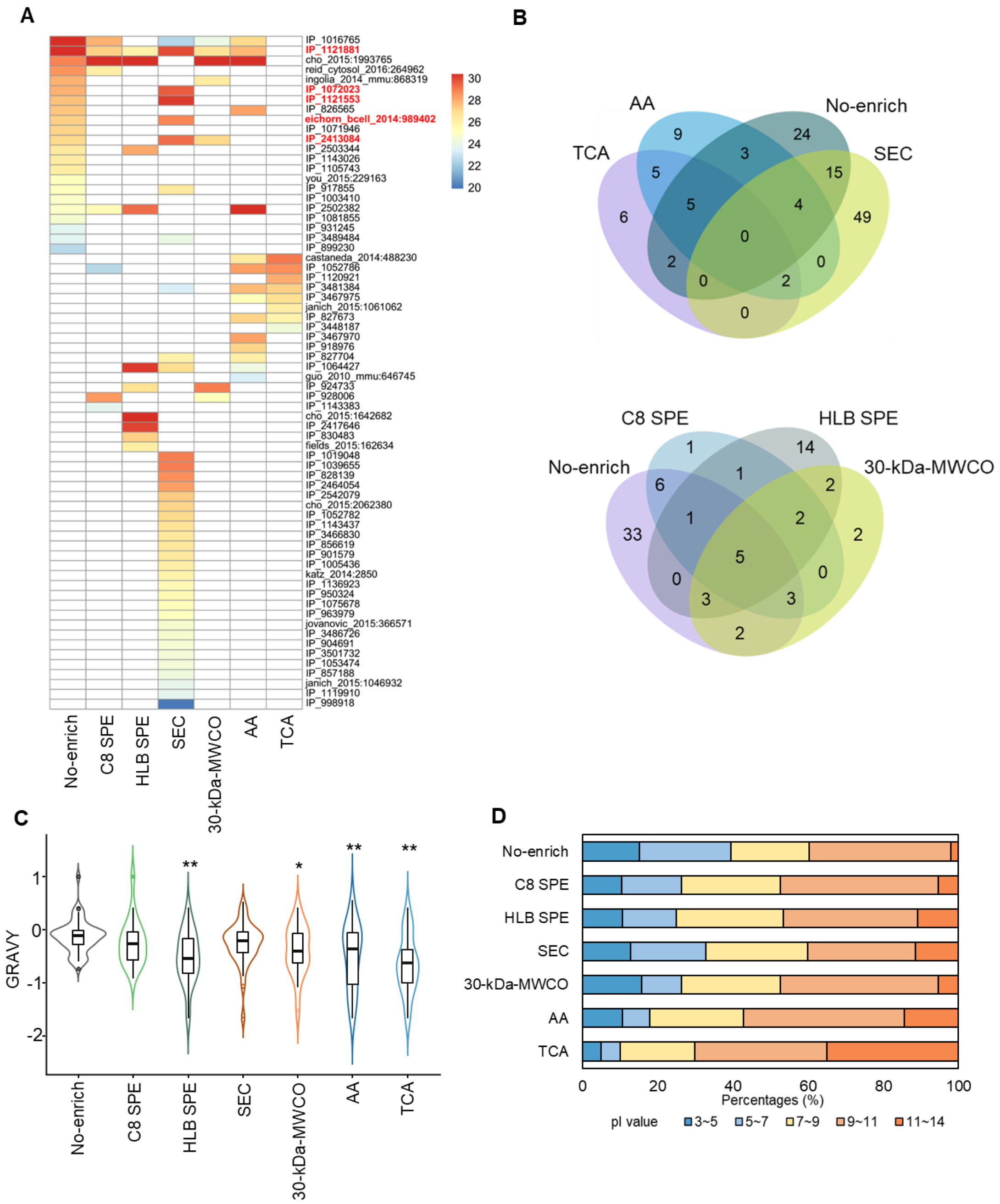


Figure 2-4 Performance of different approaches for enriching microproteins. (A) Distribution of MS intensities of the identified microproteins from different enrichment methods. (B) Comparison of the identified microproteins between different categories of methods. (C, D) Distribution of hydrophobicity (C), and isoelectric point (D) of identified microproteins from different enrichment methods. **p* < 0.05, ***p* < 0.01 compared to No-enrich by Student's t test.

2.3.4 Fractionation improves microprotein discovery

Peptide fractionation using electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) and high-pH reverse phase (HpRP) has been reported to improve the discovery of microproteins in prior studies[159, 178]. SEC, which was found to be the most efficient and unbiased method for enrichment of microproteins in our study, could also serve for protein fractionation to obtain four fractions of different molecular weight ranges. Therefore, we evaluated four fractionation methods for improving the depth of microprotein discovery, including SEC enrichment without fractionation (SEC), SEC enrichment into 4 fractions (SEC-fraction), SEC enrichment followed by ERLIC fractionation (SEC-ERLIC), and SEC enrichment followed by HpRP fractionation (SEC-HpRP) (**Figure 2-5**). SEC-fraction and ERLIC fractionation increased the number of microproteins by 1.4 to 1.6 folds. SEC-ERLIC led to the highest number of microproteins while SEC-fraction was the most time- and effort-effective, as it could enrich and fractionate microproteins simultaneously within 15 min (**Figure 2-5A**). The intensities of microproteins even showed a slight increase after SEC-fraction and SEC-ERLIC (**Figure 2-5B**).

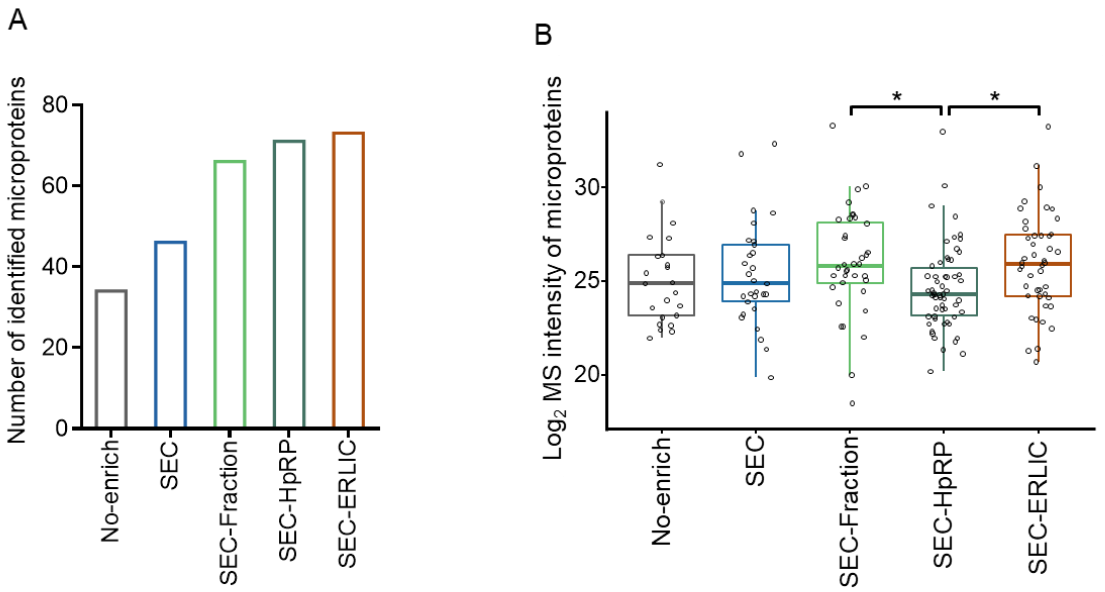


Figure 2-5 The effect of fractionation methods on microprotein discovery. The number of identified microproteins (*A*) and MS intensity of microproteins (*B*) before and after fractionation. * indicates $p < 0.05$ for comparison by Student's t test.

2.3.5 An optimized workflow enables discovery of microproteins in embryonic liver development

Next, we applied the optimized workflow in combination with TMT-based quantification to investigate microprotein expression during liver development (**Figure 2-6A**). Total protein lysates were extracted from the livers of embryonic (E15.5) or adult (P42) C57BL/6 mice in triplicate followed by SEC-ERLIC and TMT-based quantification. As we previously studied the protein translation landscape of mouse livers during development by using Ribo-seq, we were able to construct a liver-specific protein database based on Ribo-seq results and search the MS data against it. Although our customized database was much smaller than public databases[172, 173], we were able to reproducibly detect 5146 RefProts and 89 microproteins from embryonic and adult mouse livers. Even though MS and Ribo-seq were two completely different techniques, the measured fold change between embryonic and adult livers showed a positive correlation with R equals to 0.71 (**Figure 2-6B**), indicating that both techniques can precisely capture the overall changes in the proteome during development. A large majority of the microproteins identified were encoded by long non-coding RNA ORFs (lncRNA-ORFs,74%) and upstream overlapping ORFs (uoORFs, 22%), and some were encoded by upstream ORF (uORF), downstream ORF (dORF), and internal out-of-frame ORFs (intORFs) (**Figure 2-6C**). The identified microproteins showed similar hydrophobicity with RefProts (**Figure 2-6D**). Furthermore, 39 microproteins were differentially expressed (**Figure 2-6E**). Gene Ontology (GO) analysis of sORFs showed that microproteins upregulated in embryonic livers were involved in RNA splicing and processing, whereas microproteins upregulated in adult livers were enriched in metabolic pathways (**Figure 2-7A**). The biological pathways were consistent with those of RefProts, suggesting the functional importance of microproteins in liver development. We employed an alternative MS strategy, parallel reaction monitoring (PRM), to validate the identification and quantification of the novel microproteins. For example, the MS2 spectrum of the non-canonical peptide QLLLAGLQNAGR strongly agreed with the predicted spectrum (**Figure 2-7B**). The amount of this peptide was significantly downregulated in three embryonic livers compared to adult livers (**Figure 2-7C**). Finally, we sought to understand the relationship between microproteins and RefProts. We specifically searched for actively translated sORFs within the 5’- and 3’-UTRs of canonical ORFs. With stringent criteria, 6 pairs of microproteins and their primary RefProts from the same gene were detected by MS in the same experiment (**Figure 2-7D**, **Table 2-2**). Among them, dihydrofolate reductase (DHFR), ceruloplasmin (Cp), and beta-globin (Hbb-bs) and their corresponding microproteins were significantly changed between embryonic and adult mice, indicating a potential *cis* gene regulatory effect between sORFs and the corresponding primary ORFs.

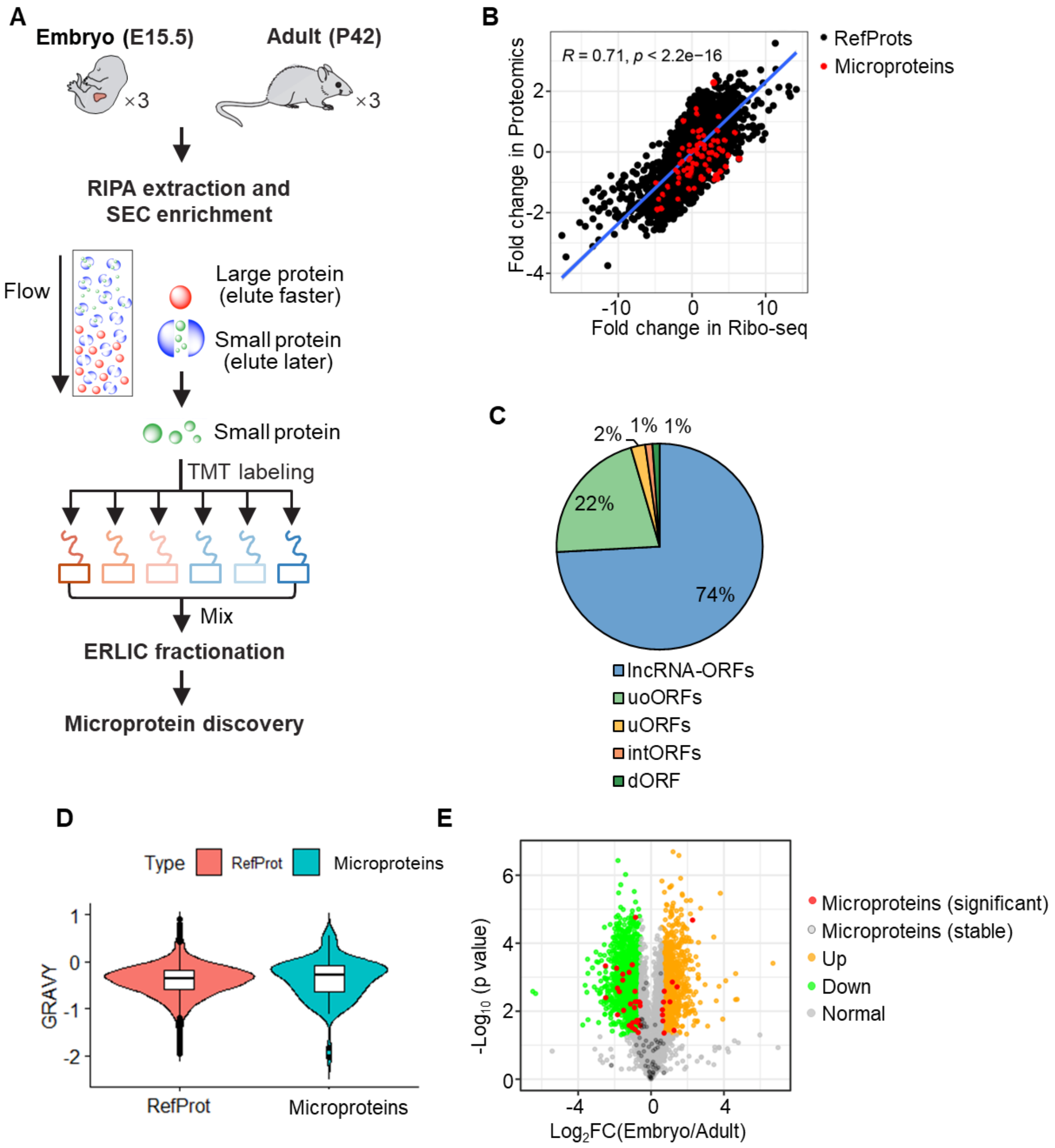


Figure 2-6 Microprotein profiling in adult and embryonic livers via SEC-ERLIC and TMT-MS workflows. (A) SEC-ERLIC workflow for the discovery of microproteins in adult and embryonic livers using a TMT-based MS approach. (B) Correlation of differences in protein expression between embryonic and adult mice detected using two techniques: Ribo-seq and MS-based proteomic approaches. (C) RNA type distribution of identified microproteins. (D) Hydrophobicity of the identified microproteins and RefProts in adult and embryonic livers. (E) Volcano plot of the identified RefProts and microproteins in adult and embryonic livers. The orange and green dots represent upregulated and downregulated

RefProts, respectively. Red and dark gray dots represent significantly changed microproteins and stable microproteins, respectively. (p value < 0.05; |fold change (FC)| > 1.5).

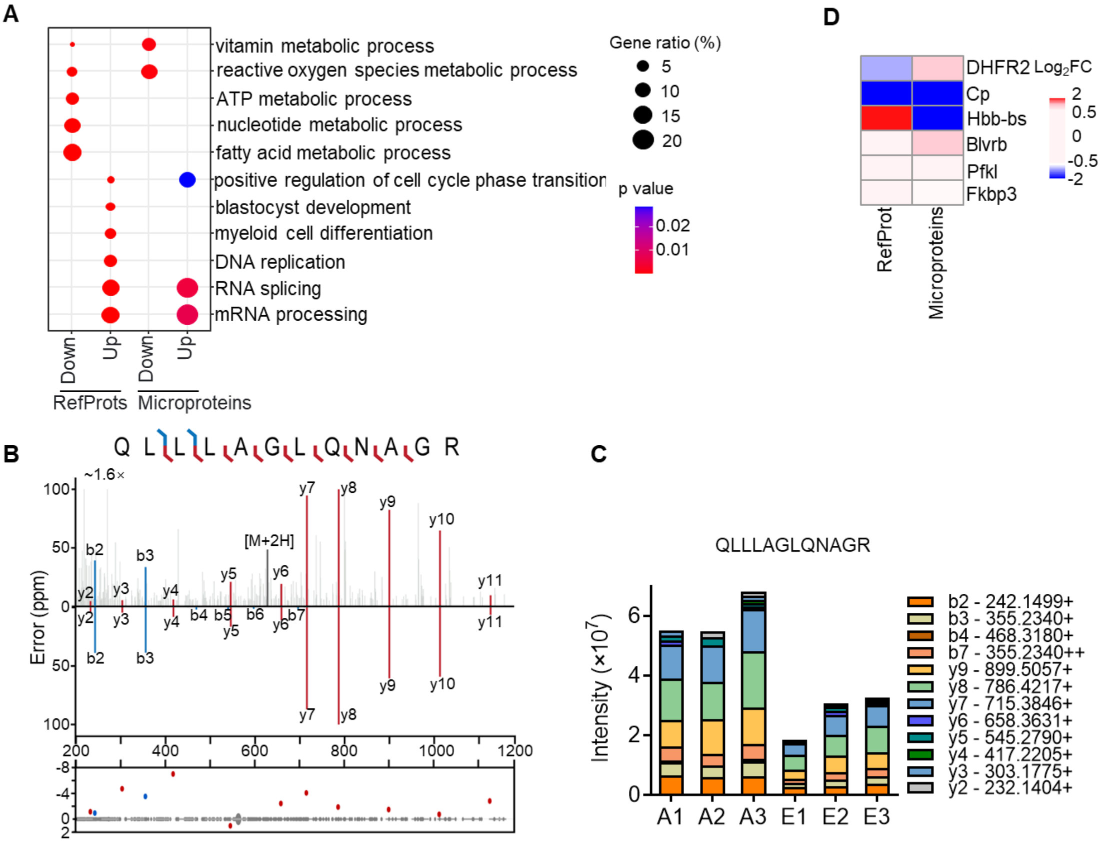


Figure 2-7 Discovery of microproteins in embryonic and adult livers. (A) GO analysis of significantly altered RefProts and microproteins. (B) Examples of the experimental and predicted spectra of the non-canonical peptide QLLLAGLQNAGR. (C) Corresponding peak areas of the representative non-canonical peptide QLLLAGLQNAGR in embryonic and adult livers using the PRM method. (D) Heatmap of MS intensity of pairs of microproteins and their primary RefProts from the same gene.

Table 2-2 List of 6 pairs of microproteins and their primary RefProts from the same gene that detected directly by TMT-based MS approach.

Protein ID	Peptide	Gene ID	Transcript ID	Protein name	Gene name	Changes of RefProts	Changes of microproteins	RNA type
chr13,+,ENSMUSG00000021707.5_92354728_92355099_123,uORF,ENSMUST00000022218.5,5347,E15_liver_STAR_Rp-Bp	AGLLGAR	ENSMUSG00000021707	ENSMUST00000022218	Dihydrofolate reductase	Dhfr	Down	Up	uORF
chr3,+,ENSMUSG00000003617.16_19957284_19966254_53,novel,ENSMUST00000128615.7,1877,E15_liver_STAR_Ribo-TISH-longest:E15_liver_STAR_RiboWave	LISVDTSR	ENSMUSG00000003617	ENSMUST00000128615	Ceruloplasmin	Cp	Down	Down	lncRNA-ORFs
chr7,-,ENSMUSG00000052305.6_103826627_103826553_24,dORF,ENSMUST00000023934.7,628,P42_liver_STAR_RiboCode:P42_liver_ST	TFENLSSDK	ENSMUSG00000052305	ENSMUST00000023934	Beta-globin	Hbb-bs	Up	Down	dORF

AR_Rp-Bp								
chr10,-,ENSMUSG00000020277.10_78009683_78009573_36,novel,ENSMUST00000220304.1,2141,P42_liver_STAR_riboHMM:P42_liver_STAR_RiboWave	AIGVLTSGGDAQGDGTR	ENSMUSG00000020277	ENSMUST00000220304	ATP-dependent 6-phosphofructokinase, liver type	Pfkl	None	None	lncRNA-ORFs
chr12,-,ENSMUSG00000020949.9_65073889_65065772_69,nested_ORF,ENSMUST00000221913.1,600,E15_liver_STAR_Ribo-TISH-longest	DHLVNAYNHLFESKVQRR	ENSMUSG00000020949	ENSMUST00000221913	Peptidyl-prolyl cis-trans isomerase FKBP3	Fkbp3	None	None	intORFs
chr7,+,ENSMUSG00000040466.16_27448064_27465877_265,N_extension,ENSMUST00000108358.7,951,E15_liver_STAR_Ribo-TISH-bestframe:E15_liver_STAR_Ribo-TISH-longest:E15_liver_STAR_RiboWave	VLGVLSFSGPGPR	ENSMUSG00000040466	ENSMUST00000108358	Flavin reductase (NADPH)	Blvrb	None	Up	uoORFs

2.3.6 Integration of Ribo-seq and PRM for the discovery of additional microproteins

It is noteworthy that the number of sORFs predicted by Ribo-seq was dramatically higher than that detected by MS (**Figure 2-8A**). Therefore, we tested an alternative approach by integrating Ribo-seq with a targeted MS method to discover additional microproteins that were undetectable using conventional shotgun proteomics (**Figure 2-8B**). To provide a precise list of sORFs, we used ten different bioinformatics pipelines to predict possible translational sORFs and kept only those that were reproducibly reported with at least two pipelines. The full-length sequences of microproteins were subsequently generated by using 3-frame translation. Out of the 27 selected microproteins with unique peptides, 11 were detectable with PRM (**Table 2-3**). The retention time (Rt) showed a high correlation between theoretical and experimental values (R=0.84-0.88), indicating a high confidence of microprotein identification (**Figure 2-8C**). Even though the identification rate was 40% with this approach, it could serve as a supplement to traditional shotgun proteomics and possibly allow detection of microproteins with low abundance. For example, peptide LALGPAAR was from a novel microprotein with 50 amino acids. This microprotein was encoded by the 5'-UTR sequence of Hnrnpa0 gene encoding Heterogeneous nuclear ribonucleoprotein A0 (HnRNPA0) **(Figure 2-9A)**. Both this microprotein and RefProt HnRNPA0 were significantly upregulated in embryonic livers (**Figures 2-9 B-D**). Hnrnpa0 plays an important role in myeloid cell differentiation[179],as well as neurodevelopment[180]. The identification of its upstream sORF could lead to novel regulatory mechanisms of this important protein.

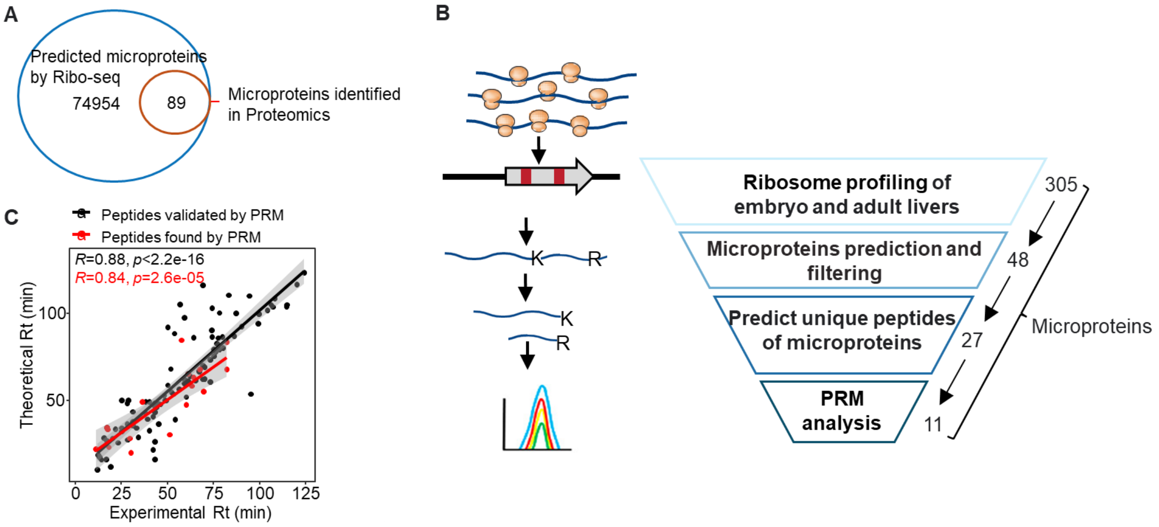


Figure 2-8 Discovery of additional microproteins using PRM method. (A) Venn diagram of non-canonical microproteins predicted by Ribo-seq and microproteins detected by targeted MS-based proteomics. (B) Flowchart of microprotein discovery using the PRM method. (C) Correlation between the experimental and theoretical retention times of the identified microproteins. Red dots represent the peptides found by targeted PRM, black dots represent the peptides found by SEC-ERLIC workflow and validated by PRM method.

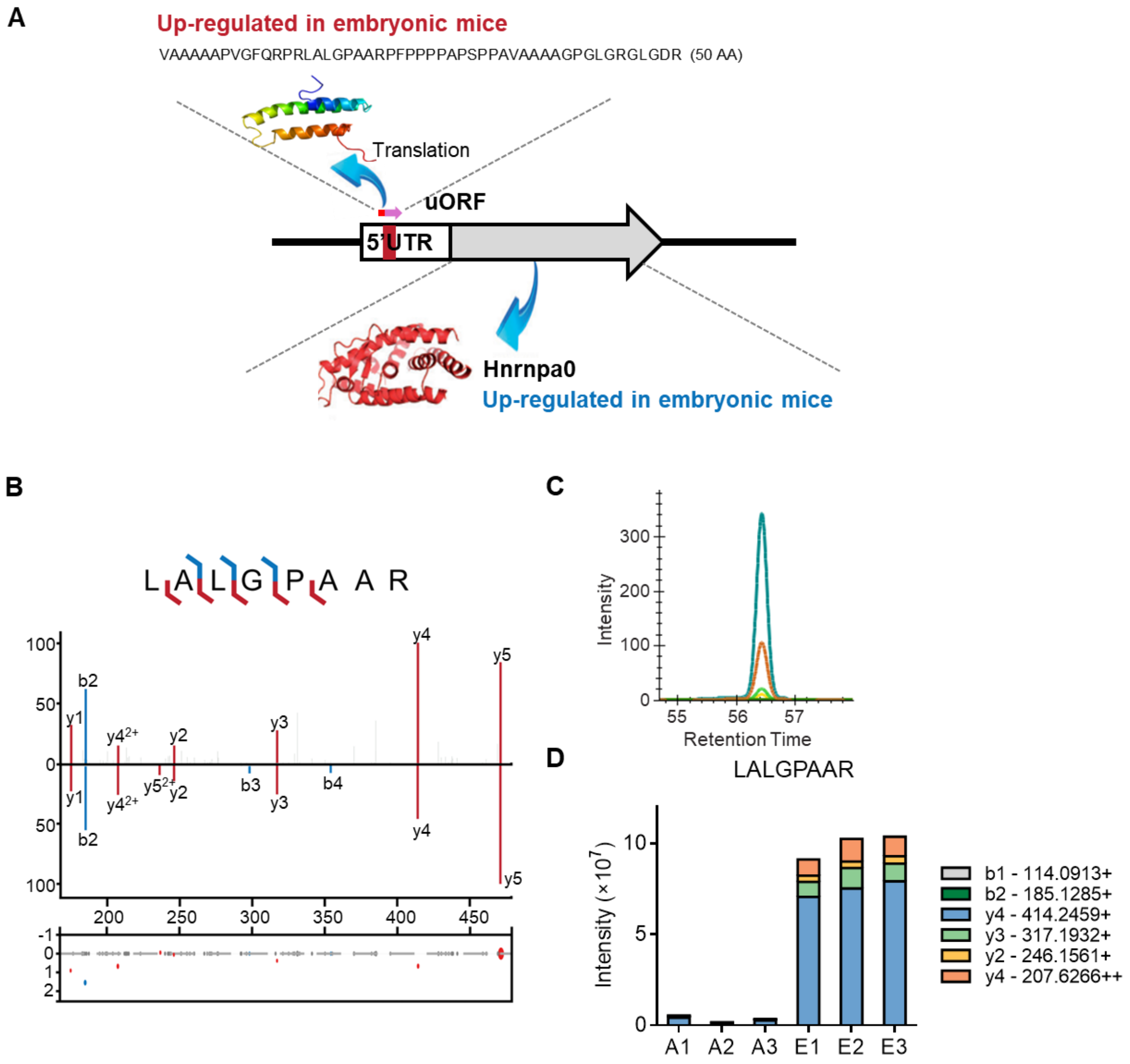


Figure 2-9 Characterization of uORF-encoded microproteins in Hnrnpa0 via PRM analysis. (A) A typical schematic diagram of a microprotein expressed in the uORF of the Hnrnpa0 gene encoding HnRNPA0. (B) Example of the experimental spectrum and predicted spectrum of the non-canonical peptide LALGPAAR. (C, D) The chromatograph (C) and peak areas (D) of the representative non-canonical peptide LALGPAAR using PRM method.

Table 2-3 List of microproteins that detected by integrating Ribo-seq with targeted MS method.

Protein ID	Peptide	Gene ID of RefProts	Transcript ID of RefProts	Protein name of RefProts	Gene name of RefProts	Changes of RefProts	Changes of microproteins	sORF type
chr13,-,ENSMUSG00000007836.6_58128523_58128371_50,uORF,ENSMUST00000007980.6,2678,E15_liver_STAR_RiboCode:E15_liver_STAR_Rp-Bp	LALGPAAR; VAAAAAPVGFQR	ENSMUSG00000007836	ENSMUST00000007980.6	Heterogeneous nuclear ribonucleoprotein A0	Hnrnpa0	up	up	uORF
chr2,-,ENSMUSG00000027185.15_103761261_103757791_71,uORF,ENSMUST00000028608.12,3879,P42_liver_STAR_ORFquant:P42_liver_STAR_RiboTaper:P42_liver_STAR_Ribo-TISH-longest	ASVVQLPGVGR	ENSMUSG00000027185	ENSMUST00000028608.12	RNA cytidine acetyltransferase	KRE33	Up	down	uORF

chr9,+,ENSMUSG00000010048.10_107587742_107587846_34,uORF,ENSMUST00000010192.10,2056,E15_liver_STAR_ORFquant:E15_liver_STAR_Ribo-TISH-longest:E15_liver_STAR_Rp-Bp	QSQLSTR	ENSMUSG00000010048	ENSMUST00000010192	Interferon-related developmental regulator 2	Ifrd2	Down	up	uORF
chr1,+,ENSMUSG00000037942.5_172699037_172699207_56,dORF,ENSMUST00000038495.4,1695,E15_liver_STAR_Ribo-TISH-bestframe:E15_liver_STAR_Ribo-TISH-longest	VSPEAPPGITFSPLSR	ENSMUSG00000037942	ENSMUST00000038495.4	C-reactive protein	CRP	Down	none	dORF
chr17,-,ENSMUSG00000040048.14_24724427_24724338_29,uORF,ENSMUST00000045602.8,732,E15_liver_STAR_Ribo-TISH-longest:E15_liver_STAR_Rp-Bp	GPASLAGPADPDVER	ENSMUSG00000040048	ENSMUST00000045602.8	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 10	Ndufb10	Down	down	uORF
chr7,+,ENSMUSG00000040824.3_19149734_19149871_45,uORF,ENSMUST00000049294.3,1420,E15_liver_STAR_Ribo	WYLVAQAPTEVSR	ENSMUSG00000040824	ENSMUST00000049294.3	Small nuclear ribonucleoprotein Sm	Snrpd2	Up	down	uORF

Code:E15_liver_STAR_Rp-Bp				D2				
chr8,+,ENSMUSG00000069922.12_105058116_105058307_63,dORF,ENSMUST00000093222.12,2046,P42_liver_STAR_Ribo-TISH-bestframe:P42_liver_STAR_Ribo-TISH-longest:P42_liver_STAR_Rp-Bp	LDLPDWANPR	ENSMUSG00000069922	ENSMUST00000093222.12	Carboxylesterase 3A	Ces3a	down	up	dORF
chr13,+,ENSMUSG00000021210.16_4457192_4457269_25,dORF,ENSMUST00000021630.14,1396,P42_liver_STAR_PRICE:P42_liver_STAR_RiboCode	LEVHFVPCAR	ENSMUSG00000021210	ENSMUST00000021630.14	Estradiol 17 beta-dehydrogenase 5	Akr1c6	down	none	dORF
chrX,-,ENSMUSG00000031264.13_134583124_134583023_33,uORF,ENSMUST00000033617.12,2540,E15_liver_STAR_Ribo-TISH-longest:E15_liver_STAR_Rp-Bp	SHLPSPGISR	ENSMUSG00000031264	ENSMUST00000033617.12	Tyrosine-protein kinase BTK	Btk	up	down	uORF
chr3,-,ENSMUSG00000027782.10_69074084_69073953_43,	VDPSSLAHGIC	ENSMUSG00000027	ENSMUST0000002935	Importin subunit	Kpna3	up	none	dORF

dORF,ENSMUST00000029353.8,8819,P42_liver_STAR_RiboTaper:P42_liver_STAR_Rp-Bp	CCIK	782	3.8	alpha-3				
chr9,+,ENSMUSG00000064225.6_95559739_95559942_67,uORF,ENSMUST00000079597.6,8367,E15_liver_STAR_RiboCode:E15_liver_STAR_Ribo-TISH-bestframe:E15_liver_STAR_Ribo-TISH-longest:E15_liver_STAR_Rp-Bp	SSSLPPGAPSPR; PAPAVASR	ENSMUSG00000064225	ENSMUSG00000064225	Membrane progesterone receptor epsilon	Paqr9	down	none	uORF

2.4 Discussion

In this study, we tested various methods and found “RIPA extraction/SEC enrichment/ERLIC fractionation” was the most efficient strategy for identifying microproteins with MS. Consequently, we established an optimal proteomics detection workflow for microproteins, as illustrated in **Figure 2-10**. With this strategy we investigated novel microproteins in embryonic and adult mouse livers.

Although a few elegant works using MS for microprotein detection have been reported in recent years, but the number of microprotein identified to our knowledge still varies widely, from tens[138, 149, 155] to hundreds[150, 151, 153]. This is probably explained by the different enrichment and analysis methods used, and the sample variation. In our study, 89 novel microproteins were identified and compared between embryonic and adult mice, although not the highest, it is based on only one sample type. Our study is so far the most comprehensive one to optimize multiple steps and various combinations for microprotein identification and we found that different workflows favor different types of microproteins. According to our results, the SEC-based enrichment outperformed other methods in terms of identification number, specificity, and reproducibility of low-abundant microproteins. In contrast, sample loss and batch-to-batch variability were observed in the 30-kDa-MWCO method, probably due to non-specific protein binding to the filter membrane[152]. SPE cartridges extracted a limited number of peptides[181] probably due to the undesired retention of relatively large and hydrophobic proteins by non-polar materials.

An additional advantage of the SEC approach is its capability to separate microproteins by size[158, 182], enabling analysis depth comparable to HpRP or ERLIC fractionation without extra cost of time and effort. Besides, SEC does not require specific buffer conditions and therefore is usually compatible with downstream experiments like top-down MS and functional characterization. SEC-based approach has great potential in future microproteins studies.

We also compared two types of SEC columns for microprotein enrichment, considering that flow rate, particle pore size, sample volume and column volume could all influence the separation efficiency. Conventional SEC requires relatively large amounts of proteins and more time due to the large column volume[158, 183]. Scaling-up the volumes would also dilute the proteins of interest, which impeded the detection sensitivity. We found that SEC column with smaller column volume (CV, 3 mL) outperformed the one with larger CV (24 mL). Smaller column is also more efficient to complete the enrichment and fractionation simultaneously within 15 min.

We acknowledged that the identification number and confidence of microproteins are highly dependent on the size and quality of database, and therefore decided to use only a non-inflated, customized database. In this study, 89 microproteins were identified from embryonic and adult mouse livers. Our results showed that many microproteins that were up regulated in embryonic livers were involved in RNA splicing, RNA processing and regulation of cell cycle transition (**Figure 2-7**). RNA splicing is a crucial to changes mature mRNA into functional protein, a process that is required during mammalian embryogenesis to generate a viable organism from a single cell[184]. RNA processing maintains protein synthesis during early developmental stages[185].

Cell cycle transition determines cell-fate transition and embryonic development[186]. All these biological pathways are important in embryonic development.

One of the important roles that sORFs play is to regulate the translation of downstream canonical ORFs[145, 187]. The translation of uORFs of GCN4 promoted the release of ribosomes from the same transcript, preventing ribosomes from reaching start codon and subsequent inhibiting translation of the GCN4 gene[188]. Some other uORFs positively regulated the translation of the downstream canonical ORFs[189]. In our study, two uORFs and corresponding canonical ORFs of hnRNPA0 and hnRNPA2/B1 showed significant activation in embryonic livers. The observation was highly consistent in both MS and Ribo-seq results. hnRNPA0 *was reported to affect myeloid cell differentiation and neurodevelopment*[179, 180]. hnRNPA2/B1 regulated mammalian embryonic development[190]. We speculate that microproteins encoded by uORF could promote the expression of downstream CDS, thereby regulating liver development. The detailed relationship in functions and mechanisms will be studied in due course.

Although we have discovered interesting microproteins involved in embryonic development with an optimized approach, the total identification number of microproteins was not comparable to that of RefProts. One possible reason is that we used a small, specific database and stringent cut-offs to filter the findings. However, the intrinsic short length and low abundance of microproteins are more important factors. Therefore, improvement in MS instrumentation with high sensitivity is needed in future studies of microproteins.

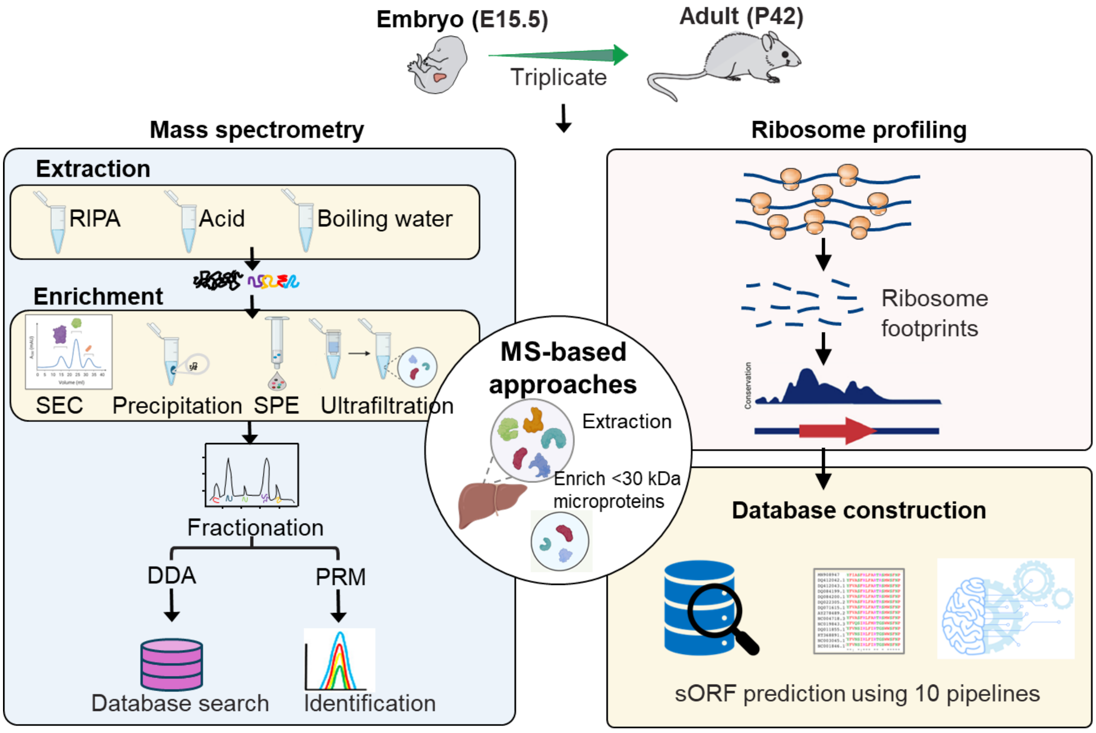


Figure 2-10 Schematic illustration of the workflow for MS-based discovery of sORF-encoded microproteins in mouse liver tissue.

Chapter 3. Discovery of a microprotein PPGlue to restore lenvatinib sensitivity in Hepatocellular Carcinoma by proteogenomics

3.1 Introduction

Hepatocellular carcinoma (HCC) is a leading cause of cancer-related deaths globally[191], particularly in developing countries such as China, which reports over 400,000 new cases annually. The rising incidence is also observed in developed nations, driven by risk factors such as cirrhosis, hepatitis C virus, and obesity[192-194]. Given the lack of specific clinical manifestations for early-stage HCC diagnosis, over 70% of patients are diagnosed at an advanced stage when symptoms become apparent[195].

Recent advancements in molecular targeted therapy and immunotherapy, including tyrosine kinase inhibitors (TKIs) such as sorafenib and lenvatinib, and the combination of atezolizumab and bevacizumab, offer new hope for patients with advanced HCC. Lenvatinib, the second authorized first-line medication for advanced HCC after sorafenib, is an orally administered multitarget tyrosine kinase receptor inhibitor. It inhibits VEGFR 1/2/3, FGFR 1/2/3/4, platelet-derived growth factor receptor α, and the proto-oncogenes KIT and RET. However, the relatively rapid emergence of resistance to lenvatinib treatment limits its overall therapeutic benefit[134-136], highlighting the urgent need to investigate the molecular mechanisms and identify new therapeutic strategies to overcome drug resistance. Owing to acquired drug resistance, a recent phase III clinical trial (NCT01761266) on lenvatinib in HCC patients showed that lenvatinib had a similar overall survival rate to sorafenib in untreated advanced HCC patients, but it did not meet the expected outcomes[196]. Therefore, identifying and developing therapeutic strategies or combination therapies that can overcome lenvatinib resistance are critical for improving patient outcomes.

Proteogenomics is an emerging field that integrates proteomics and genomics to provide a comprehensive understanding of the molecular mechanisms underlying various biological processes, including cancer. By combining mass spectrometry-based proteomics with genomic data, proteogenomics enables the identification and characterization of novel proteins, including those encoded by small open reading frames (sORFs), which are often overlooked in traditional genomic studies. The integration of proteomic and genomic data has led to the identification of novel therapeutic targets and candidates, including previously unannotated proteins and peptides encoded by sORFs. These targets can be exploited for the development of new cancer therapies.

In recent years, the pivotal role of peptides in the development of anti-tumor drugs has garnered increasing attention. Small open reading frames (sORFs) encode microproteins, also referred to as alternative proteins (AltProts) or microproteins. These microproteins represent a novel reservoir for discovering anti-tumor peptides and protein-based therapeutics. Although advancements in RNA sequencing (RNA-seq) and ribosome profiling (Ribo-seq) have facilitated the prediction of sORFs, the direct detection and functional characterization of these short, low-abundance microproteins remain challenging. Current large-scale microprotein identification efforts using mass spectrometry (MS) typically detect only 100-200 microproteins per study, highlighting significant limitations in this field.

Despite these challenges, recent studies have shown that microproteins play critical roles in various cellular processes, including proliferation, differentiation, metabolism, and immunity[56, 79, 106, 116, 126, 197-204]. For instance, NoBody is a microprotein that regulates mRNA decapping and degradation, implicating it in the regulation of gene expression in cancer cells, thereby influencing tumor growth and progression[56]. Similarly, HOXB-AS3, a peptide encoded by a long non-coding RNA (lncRNA), has been shown to inhibit colon cancer cell proliferation and metastasis by modulating glucose metabolism[204]. PINT87aa, another microprotein encoded by the lncRNA PINT, has been found to suppress tumor growth in glioblastoma by interacting with polycomb repressive complex 2 (PRC2)[116]. Additionally, SPAR is a microprotein that regulates autophagy and influences cancer cell survival and resistance to chemotherapy[106]. However, to date, only one microprotein, N1DARP, has been shown to inhibit chemoresistance in pancreatic cancer[205]. The vast number of microproteins associated with cancer drug resistance and their potential functions remain largely unexplored.

In this study, we established a robust proteogenomic approach for the large-scale identification and quantification of novel microproteins. By analyzing lenvatinib-sensitive and resistant cancer cells using this approach, we identified 815 novel microproteins encoded by various non-coding genomic sequences. After extensive validation and functional screening, we discovered a microprotein, PPGlue, that is downregulated in drug-resistant cells and patient tissues and can significantly sensitize cancer cells to lenvatinib treatment both *in vitro* and *in vivo*. Leveraging the function of PPGlue, we designed a strategy that effectively sensitized multiple cancer drugs in HCC, non-small cell lung cancer (NSCLC), and renal cell carcinoma (RCC) cells. Our study not only provided a practical approach for studying microproteins but also used PPGlue as a representative to illustrate the distinct landscape of microproteins in different cancer status, demonstrating the important functions of these new players in cell signaling.

3.2 Materials and methods

3.2.1 Study approval

The license to conduct experiments on animals was obtained from the Department of Health, Hong Kong SAR. Approval to conduct animal work at the Hong Kong Polytechnic University was obtained from the Animal Subjects Ethics Sub-Committee. The slides were from patient-derived tumor xenograft PY003 established previously [206]. The PY003 HCC tissues were obtained from patients undergoing hepatectomy at Pamela Youde Nethersole Eastern Hospital, Hong Kong. This process was conducted with approval from the Joint Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (UW 17-056), and with informed consent obtained from each patient.

3.2.2 Plasmid and reagents

Lenvatinib was purchased from Selleckchem (*in vitro* experiments) and LC Laboratories (*in vivo* experiments). Doxorubicin and verapamil were kind gifts from

Prof. Larry Chow (The Hong Kong Polytechnic University). Regorafenib, pazopanib, and MG132 were purchased from MedChemExpress.

3.2.3 Cell lines and cell culture

Huh7 cells were maintained in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS). Huh7-lenvatinib resistant (LR) and Huh7-sensitive cells were maintained in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) containing 10% FBS, 1% PS, and 10 μM lenvatinib (Huh7-LR), or an equivalent amount of DMSO (Huh7-sensitive). All cell lines were incubated at 37 °C in a 5% (v/v) CO_2 atmosphere.

3.2.4 Enrichment of microproteins using Single-pot solid-phase-enhanced sample preparation (SP3) method

The cell pellets were collected by centrifugation and lysed with RIPA lysis buffer (50 mM Tris-HCl, 150 mM sodium chloride, 2 mM EDTA, 1% NP-40, 1% sodium deoxycholate) supplemented with 1× Complete Protease Inhibitor (Roche) on ice for 15 min. The cell lysates were centrifuged, and proteins were collected. Protein concentration was determined by BCA assay, and 40 μg of protein was taken for subsequent reduction and alkylation using dithiothreitol and iodoacetamide, respectively. The protein sample was then subjected to SP3 beads clean-up to remove detergents as described previously[14]. In brief, Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophobic and Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophilic (GE Healthcare) were mixed in a ratio of 1:1 (v/v), and the mixed beads were added to the protein sample followed by the addition of absolute ethanol to induce protein binding to the beads. The beads were washed with 6.5 M urea in 67% ethanol and 80% ethanol three times. The beads were resuspended in 50 mM ammonium bicarbonate with Lys-C at an enzyme-to-protein ratio of 1:100 (w/w) and incubated at 37 °C for 4 h. Trypsin was added to the sample mixture at an enzyme-to-protein ratio of 1:20 (w/w) and the mixture was digested at 37 °C for 12 h. The samples were centrifuged, and the supernatant containing the digested peptides was collected and dried for further analysis.

3.2.5 High-pH reversed-phase fractionation

Following trypsin digestion, peptide fractionation was performed as previously described (34). Briefly, peptides were combined, concentrated, and separated using a Waters Acquity UPLC Peptide BEH C18 column (2.1×100 mm, 1.7 µm) on an Agilent 1290 Infinity LC system at 50 µL/min. The mobile phase consisted of buffers A (10 mM ammonium formate, pH 9) and B (10 mM ammonium formate in 90% acetonitrile, pH 9). Peptides were eluted using the following gradient: 0 - 10 min, 1% B; 10 - 38 min, 1 - 8% B; 38 - 75 min, 8 - 62% B; 75 - 85 min, 62 - 95% B; 85 - 100 min, 95% B. A total of 96 fractions were collected at each minute, starting from 6 min to 100 min. These small fractions were further combined into 8 fractions, concentrated, and analyzed by LC-MS/MS. For the sample-specific library, 10 µg tryptic peptides from each sample were combined and fractionated using the sample method to give the eight fractions for LC-MS/MS.

3.2.6 LC-MS/MS analysis (both DDA and DIA)

All mass spectrometry analysis were performed on an Orbitrap Exploris 480 mass spectrometry coupled with an Ultimate 3000 RSLC nano system (Thermo Fisher Scientific). The mobile phase consisted of buffer A (0.1% formic acid) and buffer B (0.1% formic acid in 80% ACN). The digested samples were resuspended in buffer A and separated on a self-packed C18 capillary column. Peptides were eluted at a flow rate of 300 nL/min using the following gradient: 0 min, 8% B; 2 min, 10% B, 120 min, 35% B, 140 min, 90% B; 150 min, 90% B; 151 min 8% B; 155 min, 8% B. For DDA, the full scan spectrum was measured with a resolution of 120,000 within 50 ms maximum injection time and the MS2 scans with a resolution of 30,000 within 120 ms maximum injection time. The isolation window of the MS2 scan was set to 1.6 m/z, and only ions with 2 to 6 charges were triggered for the MS2 event. The normalized collision energy was set to 30. For DIA, the same column and gradient were applied. The full scan spectra were measured in the range of 350–1200 m/z with a resolution of 60,000 within a maximum injection time of 50 ms, and the MS2 spectra were collected in the range of 150–2000 m/z with a resolution of 30,000. The isolation width was set to 10 m/z.

3.2.7 Identification of microproteins

The LC-MS/MS raw data were analyzed using DIA-NN. The parameters for searching were set as follows: a mass tolerance of 10 ppm for precursor ions; trypsin as enzyme with two missed cleavages allowed; oxidation (methionine), acetyl (protein N-term) as variable modifications; carbamidomethylation (cysteine) as fixed modification; the FDR was set as 1%. The raw data were searched against two databases: (1) a public database from OpenProt (v1.6) and sORFs.org (downloaded on June 01, 2021). The OpenProt protein database consisted of RefProts (Ref_), novel isoforms (II_), and microproteins (IP_) predicted from both Ensembl and NCBI RefSeq with a total of 698,963 entries. The sORFs.org consisted of 601,636 entries after removing duplicated peptide sequences; (2) in-house generated human sORF database predicted from Ribo-Seq which consisted of 409,009 entries. A two-step filtering was further applied to exclude (i) the peptides from the known coding sequence by mapping the identified peptide sequences against the human RefProts and (ii) the peptides with more than 80% sequence homology with annotated proteins.

3.2.8 Validation of identified microproteins with targeted MS

The same LC-MS system and elution gradient were applied for PRM validation. The full scan spectra were measured with a resolution of 120,000 within 50 ms maximum injection time and the targeted peptide MS2 scans with a resolution of 60,000 within 80 ms maximum injection time under the 1.2 m/z isolation window. The normalized collision energy was set to 30. After PRM data acquisition, the data were imported into a skyline for analysis as described before.

3.2.9 Establishment of stable cell lines expressing candidate microproteins

pLVX plasmids were co-transfected into HEK293T cells with the packaging plasmids psPAX2 and pMD2.G using polyethylenimine (Polysciences, 23966). The culture medium was collected after 48 h and centrifuged to remove any cell debris. The supernatant was further filtered through a 0.45 μM syringe filter to remove residual HEK293T cells. The filtered lentivirus-containing supernatant was supplemented with 8 μg/mL hexadimethrine bromide and added to the target cells. The cells were incubated for 24 h to allow viral transduction, and fresh medium with FBS was added to replace the old medium. Cells with overexpression were selected with puromycin for at least four passages to establish stable cell lines. A single clone of Huh7-LR-PPGlue cells was selected and maintained for biological assays.

3.2.10 Western blotting

Total cell proteins were extracted using RIPA buffer. Protein concentration was measured by a BCA protein assay kit (Thermo Fisher Scientific, USA). Equal amounts of protein (20 μg) were separated by SDS-polyacrylamide gel electrophoresis, followed by transfer to a PVDF membrane using a wet-transfer apparatus for 45 min at 70 V, followed by 90 min at 110 V. The membrane was blocked for 1 h in Tris-buffered saline/Tween 20 (TBST) containing 5% non-fat milk, and then were probed overnight at 4 °C with a specific primary antibody (1:1000 dilution). After that, the membrane was washed three times with TBST, followed by 1h incubation with horseradish peroxidase-conjugated secondary antibody (1:2000) at room temperature. Protein bands were detected using an enhanced chemiluminescence detection kit and visualized using Bio-Rad ChemiDoc^{TM} MP Imaging System (Hercules, CA, USA). The band intensity was quantified using

ImageJ software.

3.2.11 Immunofluorescence staining

Cells were seeded on a confocal dish, fixed in 4% paraformaldehyde (PFA) in PBS for 30 min, and permeabilized with 0.2% Triton-X100 for 20 min. Cells were blocked with 1% bovine serum albumin (BSA) in PBS containing 0.1% Tween-20 (PBST) for one hour. Anti-FLAG (D6W5B) (1:500; Cell Signaling Technology) was applied and stained at 4 °C overnight. Goat anti-rabbit secondary antibody conjugated with Alexa 568 (1:1000, Invitrogen) was applied for an hour at room temperature. Slides were counterstained with DAPI for nuclei, mounted, and subjected to Leica TCS SPE confocal microscope examination.

3.2.12 Cell proliferation assay

The various transfections of cells were seeded into 96-well plates at a density of 1,500 cells per well in a complete culture medium with six replicates. The cells were then incubated for 24, 48, 72, or 96 h. CCK-8 solution was added to each well according to the manufacturer’s instructions. The cells were incubated for 2 h at 37 °C in a 5% (v/v) CO_2 atmosphere before measurement. The absorbance at 450 nm was determined using a Varioskan LUX microplate reader (Thermo Fisher Scientific).

3.2.13 Cell viability assay

The cells were subjected to various transfections and treated with a series of concentrations of lenvatinib for 48 h at 37 °C, with three replicates. Subsequently, cell viability was assessed using a CCK-8 kit. Cell viability was quantified as a percentage of the absorbance measured in the control cells. IC50 values were defined by non-linear regression (curve fit) analysis using GraphPad Prism 9.0.0 (121) version.

3.2.14 Colony formation assay

Cells from different transfections were seeded into 6-well plates. After a 24-hour period for cell attachment, the cells were treated with either a low lenvatinib dose (3 μM) or low lenvatinib dose (6 μM), while control cells were exposed to 0.5% DMSO. Following a 2-week incubation, colony formation was quantified by staining with crystal violet, and the experiment was conducted in triplicate.

3.2.15 Cell apoptosis assay

The cells were seeded overnight in six-well plates and then treated with or without lenvatinib (20 or 40 μM) for another 48 h. Cells were then harvested, stained with an Annexin V-FITC/PI kit (Abcam, ab14085) according to the manufacturer's instructions, and analyzed using a BD Accuri C6 flow cytometer and FACSDiva software (BD Biosciences).

3.2.16 RNA extraction and quantitative PCR (qRT-PCR) analysis

Total RNA was extracted using TRI Reagent (T9424, Sigma-Aldrich) according to the manufacturer’s protocol. The extracted RNA was reverse-transcribed to cDNA using the PrimeScript^{TM} RT Reagent Kit (RR047A, Takara). The qRT-PCR amplifications of target genes and internal control β-actin (ACTB) were performed using SYBR Green PCR Master Mix (Applied Biosystems) with primers specific to the sequences of interest. Amplifications were performed with a QuantStudio^{TM} 5 Real-time PCR system (Thermo Fisher Scientific). Relative expression differences were calculated using the $2^{-\Delta\Delta CT}$ method with reference to ACTB.

PPGlue-forward, 5’-TTCTCGCTCAGCAGAGGTGG-3’;

PPGlue-reverse, 5’-GGCCCTGTGTAGGCACCTT-3’.

ABCB1-forward, 5’-GCTGTCAAGGAAGCCAATGCCT-3’;

ABCB1-reverse, 5’-TGCAATGGCGATCCTCTGCTTC-3’.

ACTB-forward, 5’-CACCATTGGCAATGAGCGGTTC-3’;

ACTB-reverse, 5’-AGGTCTTTGCGGATGTCCACGT-3’.

3.2.17 Transient transfection for PPGlue knockdown

Huh7 cells were transfected with 50 nM targeting siRNA or non-target control siRNA (NC) using Lipofectamine RNAiMAX transfection reagent (Thermo Fisher Scientific). After 10 h, the medium containing the transfection reagent was removed and replaced with fresh complete DMEM supplemented with 10% FBS and 1% PenStrep. Cells were subsequently cultured for lenvatinib-induced apoptosis assays and harvested for real-time qPCR.

siPPGlue-1, 5’-CTACAGAGATGGAATCTGA-3’.

siPPGlue-2, 5’-GTATCCTAGGTGCTTGGTA-3’.

3.2.18 Construct stable cells with PPGlue knockdown

To establish PPGlue-knockdown clones, lentiviral particles were generated by co-transfecting HEK293T cells with shRNA-1 or non-target control (NC) plasmids and a packaging plasmid mix. Viral supernatants were collected for infection of Huh7 cells.

Plasmids for silencing PPGlue were cloned using the pLKO.1-puro lentiviral vector. Stable knockdown cell lines were generated by puromycin selection, and cell viability and apoptosis assays were performed.

shPPGlue-1, 5’-CTACAGAGATGGAATCTGA-3’.

3.2.19 Animal experiments

All mice were housed in 12-hours light/dark cycles (8:00–20:00 light, 20:00–8:00 dark), with controlled room temperature (23 ± 2 °C) and humidity (30–70%), in groups according to stocking density as recommended. *In vivo* assays were performed in male BALB/c mice aged 4–6-week-old by induction of tumor xenografts. Cells were suspended in 1:1 ratio of serum-free DMEM and BD Matrigel Matrix (BD Biosciences) and subcutaneously injected into the nude mice, which were kept under observation. Briefly, each mouse received one injection of cells in the right flank, and cells from each experimental group (EV vs. PPGlue) were injected into different mice. Tumors were harvested at the end of the experiment for documentation. For the *in vivo* assay without lenvatinib treatment, either Huh7-LR-EV or -PPGlue (0.5×10^5 or 1×10^5) cells were injected, respectively. Tumor formation was confirmed when the tumor reached approximately 6 mm × 6 mm (length × width). Tumor volume and body weight were measured every three days. Tumor volume was calculated using the following formula: volume (mm^3) = length × width^2 × 0.5. The mice were monitored for 21 days before sacrifice, at which point the tumors were harvested for analysis. For the *in vivo* lenvatinib treatment assay, a total of either Huh7-LR-EV or -PPGlue (1×10^5) cells were injected. Once the tumors were established and reached approximately

6 mm × 6 mm (length × width), the mice were treated with lenvatinib (30 mg/kg) which was resuspended in 0.5% methylcellulose in saline. The mice were administered lenvatinib daily by oral gavage. Tumor volume and body weight were measured every three days. Tumor volume was calculated using the following formula: volume (cm^3) = length × $width^2$ × 0.5. The mice were treated for 30 days before sacrifice, at which point the tumors were harvested for analysis.

The study protocol was approved and performed in accordance with the guidelines for the Use of Live Animals in Teaching and Research at Hong Kong Polytechnic University. For the subcutaneous tumor model, the tumor volumes did not exceed 10% of normal body weight. Mice were sacrificed if the percentage of body weight loss was greater than 20%.

3.2.20 Immunohistochemistry in HCC specimens and resected mouse tumors

The sections were deparaffinized in xylene and rehydrated in graded alcohols and distilled water. Slides were processed for antigen retrieval using a standard microwave heating technique in Tris-EDTA buffer. Endogenous peroxidase activity was quenched using 3% hydrogen peroxide. The sections were immersed in serum-free-protein block solution (DAKO). Specimens were subsequently incubated with primary antibodies (PPGlue, 1:250, customized) overnight at 4 °C. The sections were then washed thoroughly and incubated with anti-rabbit Envision^{TM} HRP-conjugated secondary antibody (DAKO) for 1 h at room temperature. Positive signals were visualized using the Liquid DAB+ Substrate-Chromogen System (DAKO). Sections were counterstained with Mayer's hematoxylin followed by examination using a light microscope. The expression of PPGlue was quantified using ImageJ software.

3.3 Results

3.3.1 The approach to comprehensively profile microproteins in cancer cells

To achieve large-scale identification of novel microproteins from cancer cells, we established a proteogenomic approach combining multiple mass spectrometry (MS) methods and comprehensive sORF databases based on Ribo-seq datasets and public Ribo-seq databases (OpenProt and sORF.org) (**Figure 3-1A**). The MS workflow combined data independent acquisition (DIA) for in-depth microprotein identification and data dependent acquisition (DDA) for robust identification and quantitation. We employed two pairs of HCC cells, Huh7-sensitive and -LR cells, PLC/PRF/5-sensitive and -LR cells, for unbiased identifying microproteins associated with drug resistance. Microproteins were extracted and processed following previously reported protocol[14], and the resulting tryptic peptides were divided into two portions: one for off-line HPLC fractionation and data acquisition in DDA mode to build the experimental spectral library, the other one for MS data acquisition in DIA mode. To increase the depth of microprotein identification, we supplemented the DDA-based experimental spectral library with an *in-silico* predicted spectral library derived from the sORF databased derived from Ribo-seq datasets[207]. DIA data was searched against the two spectral libraries to identify and quantify microproteins that are differentially expressed between the paired cell lines. To distinguish the identified peptides from canonical proteins fragments, we first removed the sequences that align to known proteins. Secondly, we applied another stringent filtering criteria to the identified microproteins, requiring sequence similarity less than 80% compared to human canonical proteome. When we used canonical proteins as a benchmark to evaluate our workflow and data, the reproducibility between biological replicates was satisfactory, with Pearson correlation coefficients exceeded 0.94 (**Figure 3-2A & B)**. Using this comprehensive workflow, we identified 815 microproteins with reliable and reproducible quantification (**Figure 3-2C & D)**. In addition, we also used the retention time (RT) as another benchmark to assess the identification confidence of the workflow. The correlation between experimental and predicted RTs for the identified microproteins proteins was 0.95, which was as high as that for the identified canonical proteins (**Figure 3-1B & 3-2E**), indicating the identification of these microproteins were reliable and robust. Among the 815 identified microproteins, over 75% were commonly detected across the three replicates for each group and 273 were shared between the two cell types (**Figure 3-2C, D, & F**). The microproteins were encoded by a diversity of sORFs, with over 33% originating from long non-coding RNAs (**Figure 3-1C**). Most identified microproteins were typically initiated with the canonical ATG start codon (**Figure 3-1D**), and 97% of them were shorter than 300 amino acids (**Figure 3-1E**).

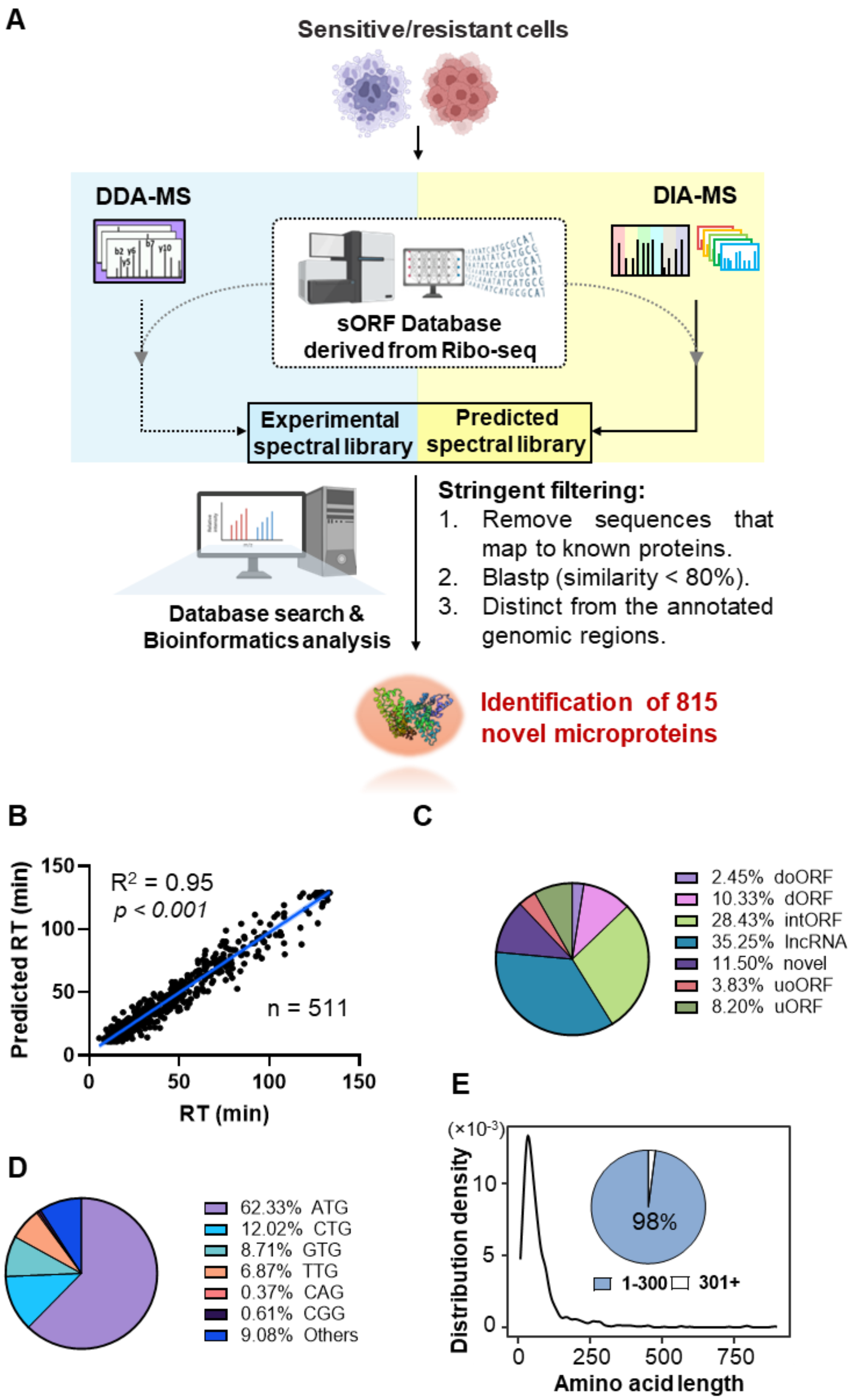


Figure 3-1 A proteogenomic approach customized for large-scale identification and quantification of microproteins. (A) Schematic diagram of an integrated workflow for

microprotein identification. (B) Correlation between the experimental and predicted retention times (RT) by DeepLC for the peptides that contributed to the identified microproteins. (C, D) Categorization of the sORF type (B) and start codon usage (C) for the identified microproteins. CDS, coding sequence; doORF, downstream overlapped ORF; dORF, downstream ORF; intORF, internal ORF; lncRNA, long non-coding RNA; uoORF, upstream overlapped ORF; uORF, upstream ORF. (E) Length distribution of the identified microproteins.

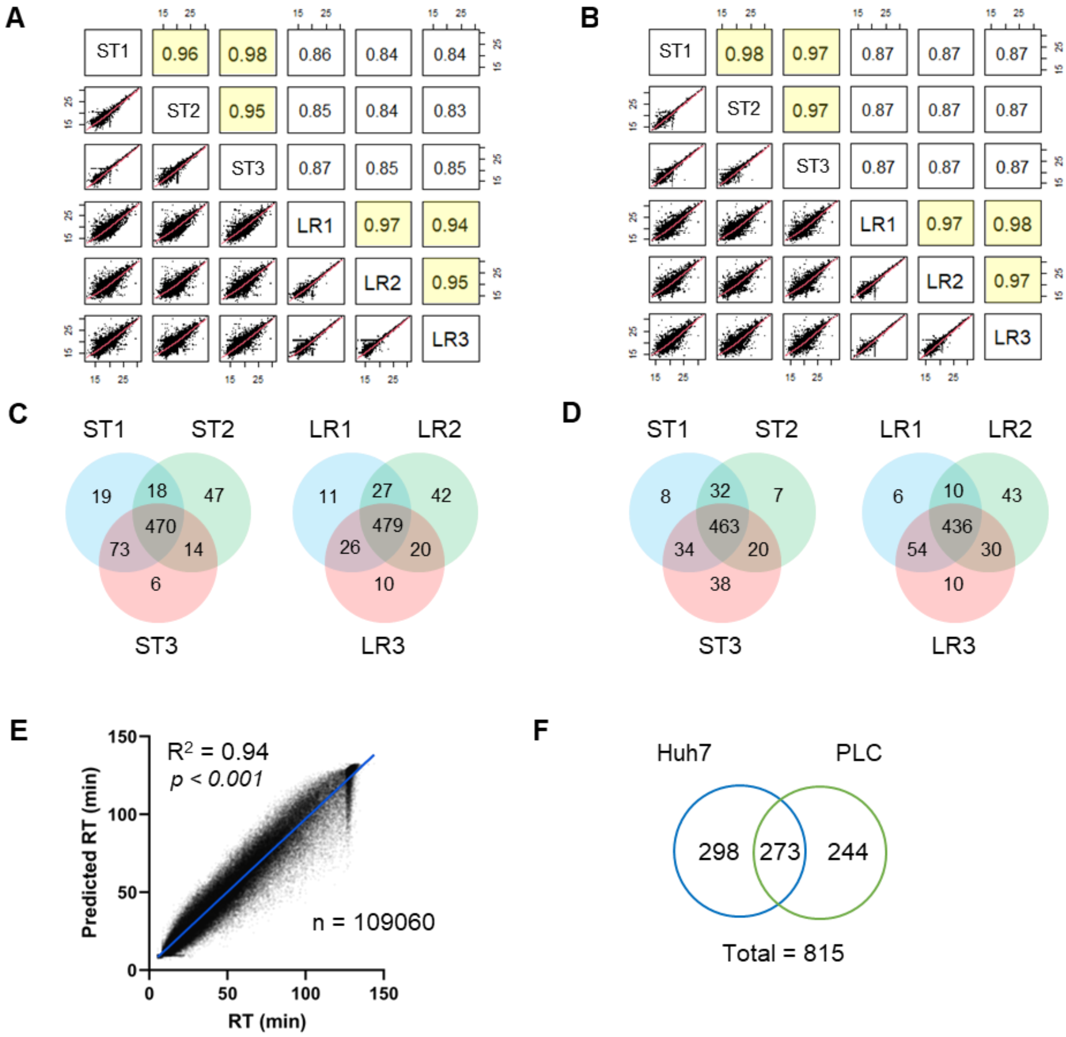


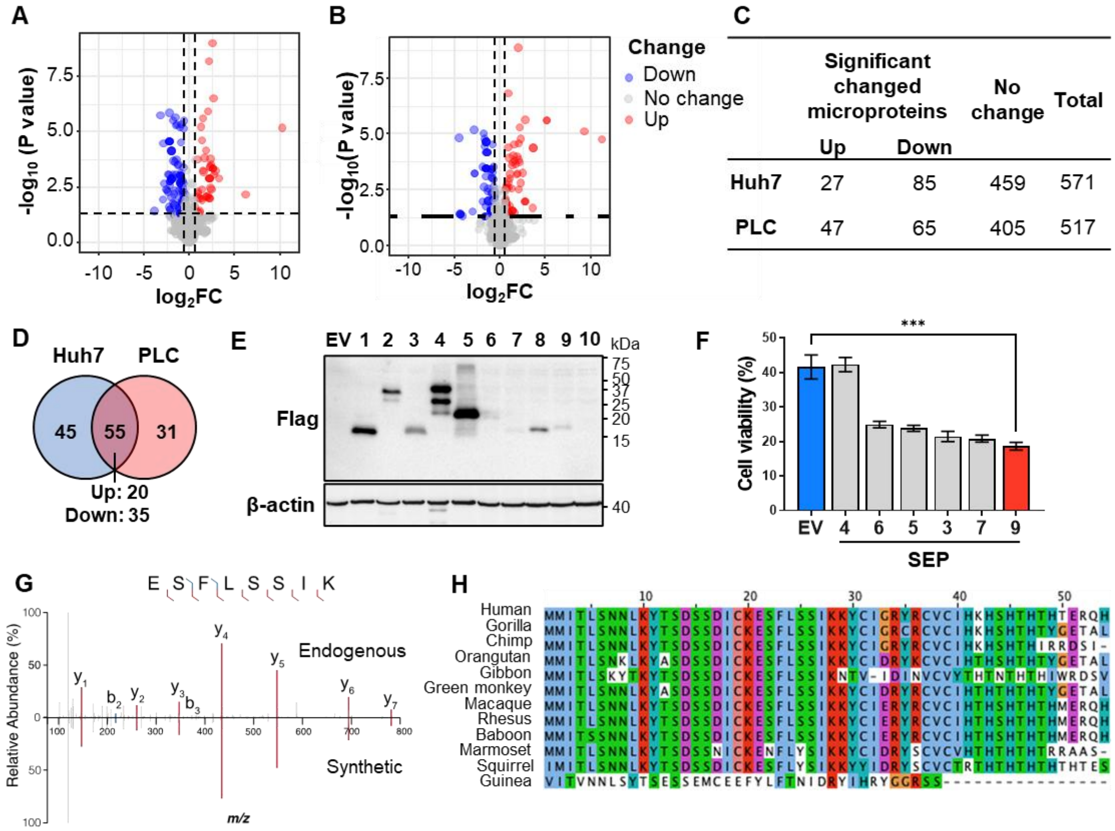
Figure 3-2 Evaluation of MS data reproducibility and confidence in microprotein

identification. (A, B) Pearson's correlation of the identified proteins in Huh7 (A) and PLC (B) cells as a benchmark. ST, sensitive cells; LR, lenvatinib-resistant cells. (C, D) Reproducibility of microproteins identified in triplicate from Huh7 (C) and PLC (D) cells. (E) Correlation between the experimental and predicted RT for the peptides which contributed to the identification of annotated proteins. (F) Venn diagram showing the microproteins that were identified in the two cell types.

3.3.2 Analysis of differentially expressed microproteins in lenvatinib-resistant HCC

There were 112 out of 571 microproteins in Huh7-LR cells, and 112 out of 517 microproteins in PLC/PRF/5-LR cells differentially expressed with fold change greater than 1.5 and p-value below 0.05, respectively (**Figure 3-3A-C**). To validate the quantification results based on shot-gun proteomics, we employed an alternative MS technique called parallel reaction monitoring (PRM) to specifically focus on microproteins of interest and calculate their relative abundance across samples in high accuracy. Over 76% of the differentially expressed microproteins were detected with PRM, including 26 upregulated and 35 downregulated ones detected in both cell types (**Figure 3-3D**). For instance, the novel peptide SIVIHTITK, which contributed to the identification of IP_581119, was downregulated in Huh7-LR cells. We used PRM to validate and re-quantify this peptide, not only supporting the existence of IP_581119 but also confirming its downregulation observed during initial profiling (**Figure 3-4**). Similarly, the peptide YSHESDWQWALR, which was validated using PRM, confirmed the existence and upregulation of gawron_2016:390621 (**Figure 3-4**). These findings demonstrate the widespread involvement of these microproteins in lenvatinib sensitivity.

To investigate the biological stability of these microproteins, we generated C-terminal Flag-tagged constructs for 10 microproteins (SEP1-10) with the most significant changes in abundance. Western blot analysis confirmed the expression of SEP1-9 as stable translation products in HEK293T (**Figure 3-3E)** and Huh7 cells **(Figure 3-5A**). The absence of SEP10 was likely due to protein degradation, making it difficult to detect; therefore, we were only able to detect it upon the addition of the proteasome inhibitor MG132 (**Figure 3-5B**). Next, we evaluated the potential impact of these microproteins on lenvatinib response in cells expressing the six microproteins (SEP3-7, 9) that were downregulated in both Huh7-LR and PLC/PRF/5-LR cells. Overexpression of SEP9 drastically reduced cell viability when combined with lenvatinib treatment (**Figure 3-3F**); therefore, we focused further investigations on this microprotein. SEP9, a 54 amino acid microprotein, can be detected in various cell types including Huh7, PLC/PRF/5-LR, MHCC97L, 786-O and A549 using PRM (**Figure 3-3G & 3-6A**). The endogenous presence of SEP9 can also be verified with MS data from normal thyroid tissue (**Figure 3-6B**), and it is highly conserved among primates (**Figure 3-3H**). Interestingly, SEP9 remained largely uncharacterized, as no functional domain or motif was detected within its sequence. Based on its association with drug sensitivity and mechanism of action, SEP9 was designated as PPGlue hereafter.



	Significant changed microproteins		No change	Total
	Up	Down		
Huh7	27	85	459	571
PLC	47	65	405	517

Figure 3-3 The microproteins differentially expressed in drug-resistant HCC cells. (A, B) Volcano plot of the identified microproteins in Huh7-LR (A) and PLC-LR cells (B). Fold change (FC) was calculated with the microprotein intensity in Huh7-sensitive cells over LR cells. Differentially expressed microproteins were determined with a p value < 0.05, FC > 1.5, or < -1.5). (C) Summary of the identified microproteins from Huh7 and PLC/PRF/5 cells. (D) Differentially expressed microproteins from the two HCC cell lines were thoroughly validated using a targeted MS method (PRM). (E) Expression of selected microproteins in HEK293T cells. EV, empty vector; 1-10, SEP1-10. (F) Cell viability assay of Huh7 cells expressing microproteins that were downregulated by MS. The cells were treated with 50 µM lenvatinib or DMSO as solvent controls. The experiments were conducted in triplicate. Data are represented as the mean ± s.d. ****p* < 0.001 by one-way ANOVA. Experiments were

conducted in triplicate. (G) Spectra of a synthetic peptide (ESFLSSIK) from SEP9 and its endogenous counterpart in Huh7 cells. (H) Conservation analysis by aligning SEP9 amino acid sequences across different primates.

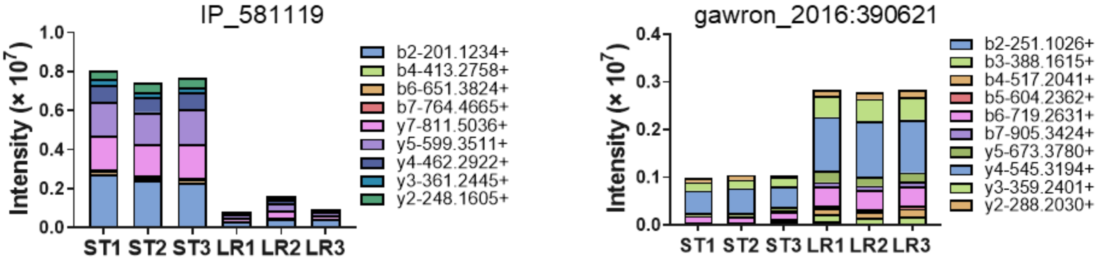


Figure 3-4 Representative PRM validation of the detected peptides from differentially expressed microproteins. SIVIHTITK and YSHESDWQWALR contributed to the identification of IP_581119 (SEP3) and gawron_2016:390621 (SEP8), respectively. ST, sensitive; LR, lenvatinib-resistant.

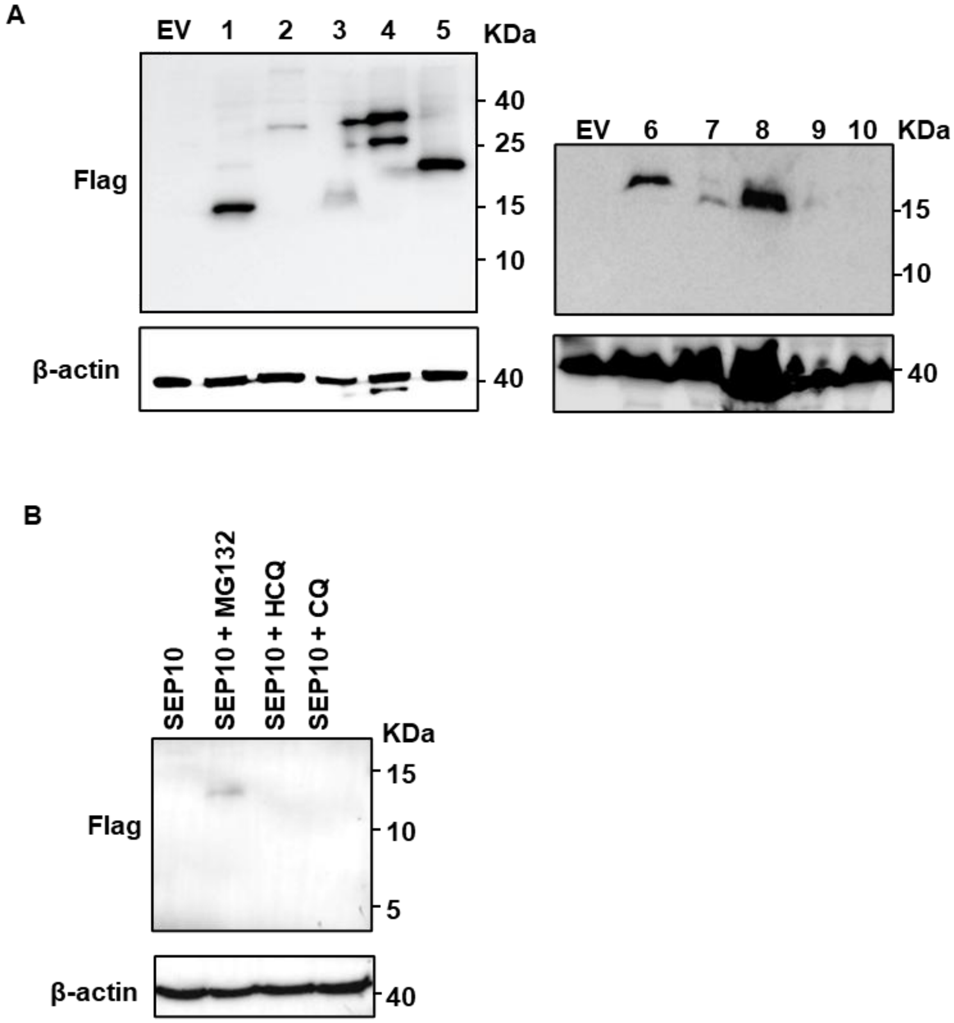


Figure 3-5 Expression validation of microproteins in Huh7 Cells. (A) Expression of the 10 microproteins in Huh7 cells. EV, empty vector; 1–10, SEP1-10. (B) Expression of SEP1-10 in Huh7 cells treated with protease inhibitors and lysosome protease inhibitors. HCQ, hydroxychloroquine; CQ, chloroquine.

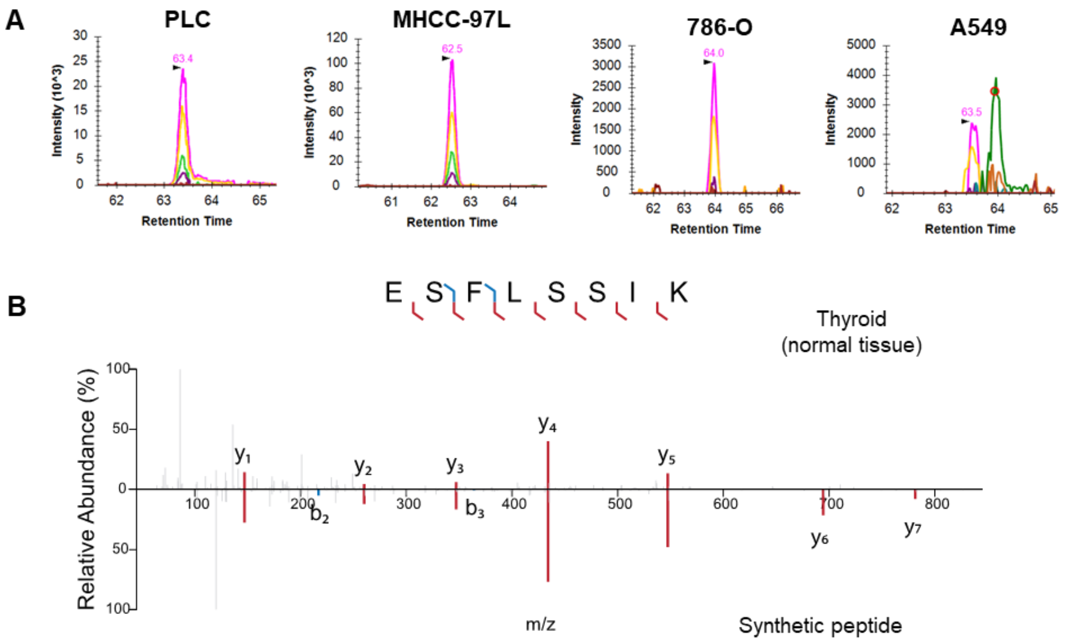


Figure 3-6 Validation and analysis of PPGlue across different cell types and conditions. (A) PRM analysis confirmed the presence of PPGlue in various cell types, including PLC/PRF/5, MHCC97L, 786-O, and A549 cells. (B) MS detection of endogenous PPGlue in normal thyroid tissue.

3.3.3 PPGlue overexpression restores HCC's sensitivity to lenvatinib

To explore the role of PPGlue in LR cells, we generated lentiviral vectors for PPGlue overexpression with a C-terminal Flag tag (PPGlue), as well as a mutant construct (PPGlue^{mut}) with a disrupted start codon (ATG to ATT) and an empty vector control (EV). These constructs were then transfected into Huh7-sensitive and -LR cells, followed by antibiotic selection. Western blot analysis and immunofluorescence staining confirmed the stable expression of the tagged microprotein (**Figure 3-7A-B, & 3-8A**), which was abolished by the start codon mutation, indicating the functional recognition of PPGlue's start codon. We generated a customized antibody for PPGlue, which was successfully validated in Huh7-LR cell lysates spiked with synthetic PPGlue peptides (**Figure 3-8B**). The overexpression of PPGlue was further validated using a customized PPGlue antibody, which was abolished by the start codon mutation (**Figure 3-8C**). Immunofluorescence staining further revealed the colocalization and distribution of PPGlue along the periphery of the cell membrane (**Figure 3-7B**).

Next, we conducted cell viability assays to investigate the potential impact of PPGlue on lenvatinib sensitivity and cell proliferation. The results demonstrated that PPGlue overexpression significantly enhanced the efficacy of lenvatinib in both Huh7-sensitive and -LR cells, leading to an 8-fold decrease in lenvatinib IC50 values (**Figure 3-7C & 3-8D**). Furthermore, PPGlue overexpression suppressed cell proliferation and reduced colony formation, and the impact was even exaggerated in the presence of lenvatinib **(Figure 3-7D–E)**. In Huh7-sensitive cells, PPGlue reduced colony formation with and without lenvatinib treatment (**Figure 3-8E**). Flow cytometry analysis revealed that PPGlue enhanced lenvatinib-induced apoptosis in different cell types in a dose-dependent manner, with 55% in Huh7-LR-PPGlue relative to 14% in Huh7-LR-EV cells and 91% Huh7-sensitive-PPGlue relative to 56% in Huh7-sensitive-EV cells, respectively (**Figure 3-7F & 3-8F**). All the sensitizing and suppressing effects of PPGlue required proper translation of this microprotein, which can be abolished by the start codon mutation in PPGlue^{mut}. The dependence on an effective start codon further indicates that the function of PPGlue should be derived from translation instead of transcription.

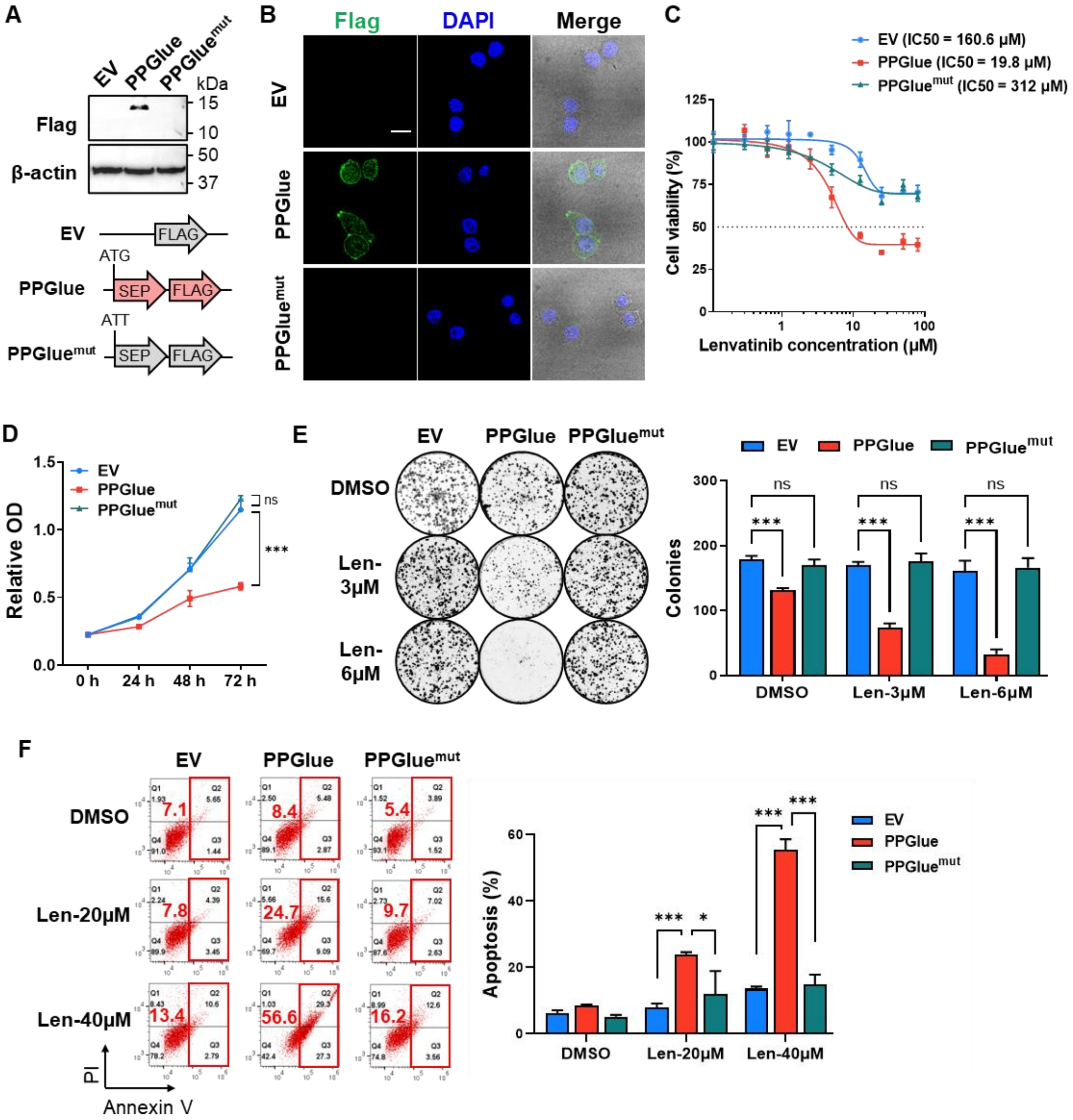


Figure 3- 7 PPGlue overexpression restored drug sensitivity. (A) Expression of PPGlue, and it could be abolished by mutating the start codon. Top, western blot of PPGlue overexpression; bottom, illustration of plasmids used to construct stable cells. (B) Membrane localization of PPGlue in Huh7-LR cells. Flag staining (green) and DAPI staining (blue). EV, empty vector. Scale bar = 25 μm. (C) Cell viability assay confirmed that overexpression of PPGlue sensitized the cells to lenvatinib treatment. (D, E) Overexpression of PPGlue suppressed cell proliferation (D) and enhanced the inhibitory effect of lenvatinib (Len) on colony formation (E) in Huh7-LR cells. The cells were treated with DMSO or lenvatinib (3 and 6 μM). ***$p < 0.001$ by two-way ANOVA. (F) Overexpression of PPGlue increased lenvatinib-induced apoptosis in Huh7-LR cells. Left, representative flow cytometry results; right, statistical analysis of lenvatinib-induced apoptosis rates in Huh7-LR cells. The experiments were conducted in triplicate. The experiments were conducted in triplicate. Data were presented as the mean ± s.d., *$p < 0.05$, ***$p < 0.001$ by two-way ANOVA.

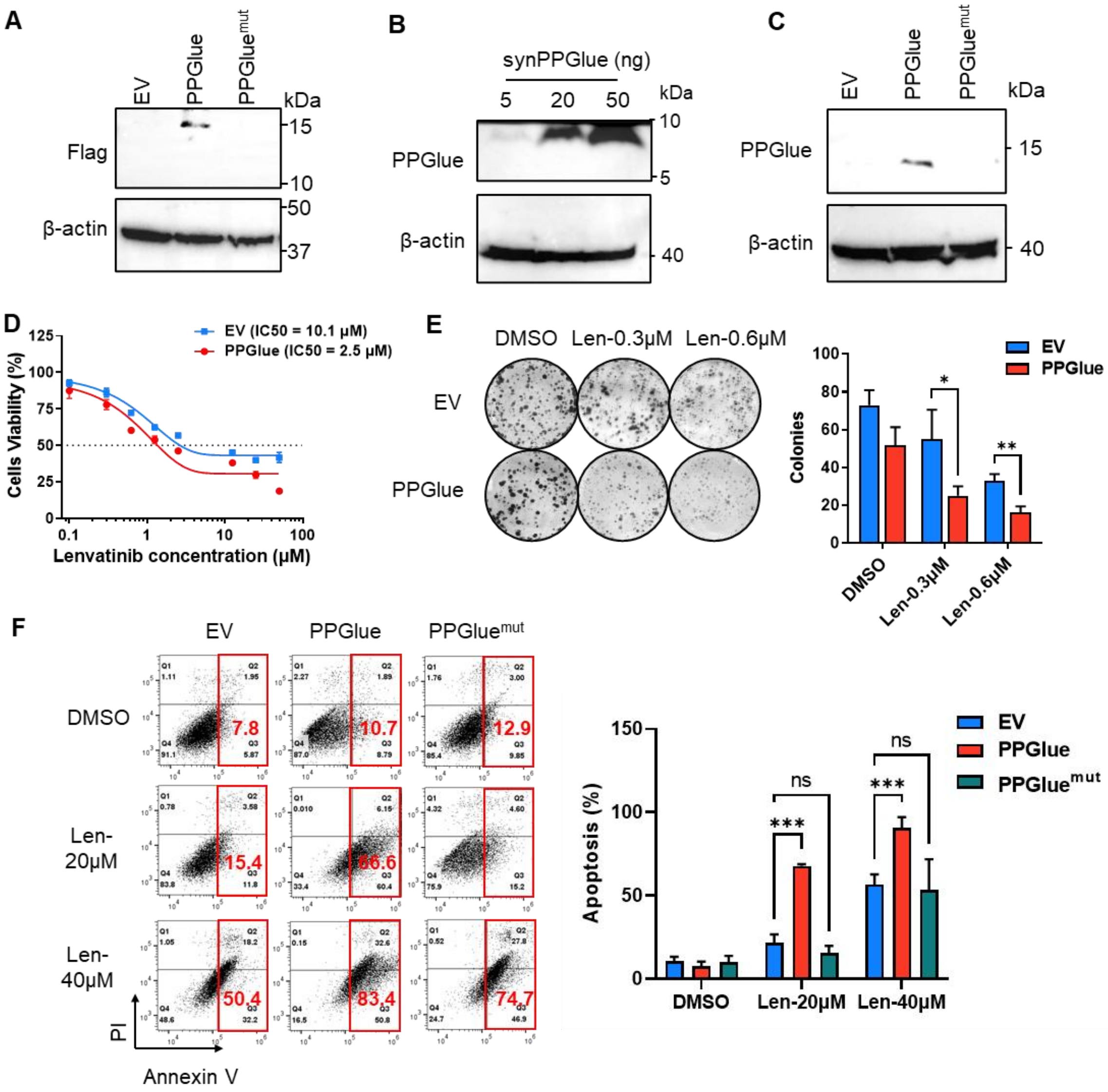


Figure 3-8 Overexpression of PPGlue restored drug sensitivity in Huh7 cells. (A)

Expression of PPGlue in Huh7 cells could be abolished by a start codon mutation, as verified by a Flag antibody. (B) Evaluation of the customized PPGlue antibody by spiking synthetic PPGlue peptides into the Huh7-LR cell lysate. synPPGlue, synthetic PPGlue. (C) Overexpression of PPGlue was validated using an anti-PPGlue antibody. (D) Cell viability assay confirmed that overexpression of PPGlue sensitized the cells to lenvatinib treatment. (E) Colony formation assay for Huh7-PPGlue cells. The experiments were conducted in triplicate, and data were presented as the mean ± s.d., $^{*}p < 0.05$, $^{**}p < 0.01$ by Student's t-test. (F) Lenvatinib-induced apoptosis assay in Huh7-PPGlue cells. Left, representative flow cytometry results; right, statistical analysis of the lenvatinib-induced apoptosis rate in Huh7-sensitive cells. The experiments were conducted in triplicate, and data were presented as the mean ± s.d., $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ by Student's t-test.

Lenvatinib plus pembrolizumab for the treatment of lung cancer is currently under phase 3 clinical trials[208], while it has been approved in combination with everolimus/pembrolizumab for the treatment of advanced renal cell carcinoma[209]. Therefore, we constructed cells stably expressing PPGlue in A549 (derived from lung adenocarcinoma) and 786-O (derived from renal carcinoma) cells to evaluate the potential drug sensitization effect of PPGlue in other cancer cells in addition to HCC (**Figure 3-9A & B**). Next, we conducted cell viability assays to investigate whether PPGlue also affects lenvatinib response and progression in cancers other than HCC. The results demonstrated that PPGlue overexpression significantly enhanced the efficacy of lenvatinib in both A549 (**Figure 3-9C**) and 786-O cells **(Figure 3-9D**), leading to a two-fold decrease in lenvatinib IC50 values. Furthermore, PPGlue overexpression suppressed cell proliferation in both A549 (**Figure 3-9E**) and 786-O cells (**Figure 3-9F**). Similarly, overexpression of PPGlue enhanced lenvatinib-induced apoptosis in 786-O cells (**Figure 3-9G**). These findings collectively validate that PPGlue exhibits tumor suppression functions by inhibiting cell growth, promoting apoptosis, and sensitizing cells to lenvatinib treatment in a broad spectrum of cell types.

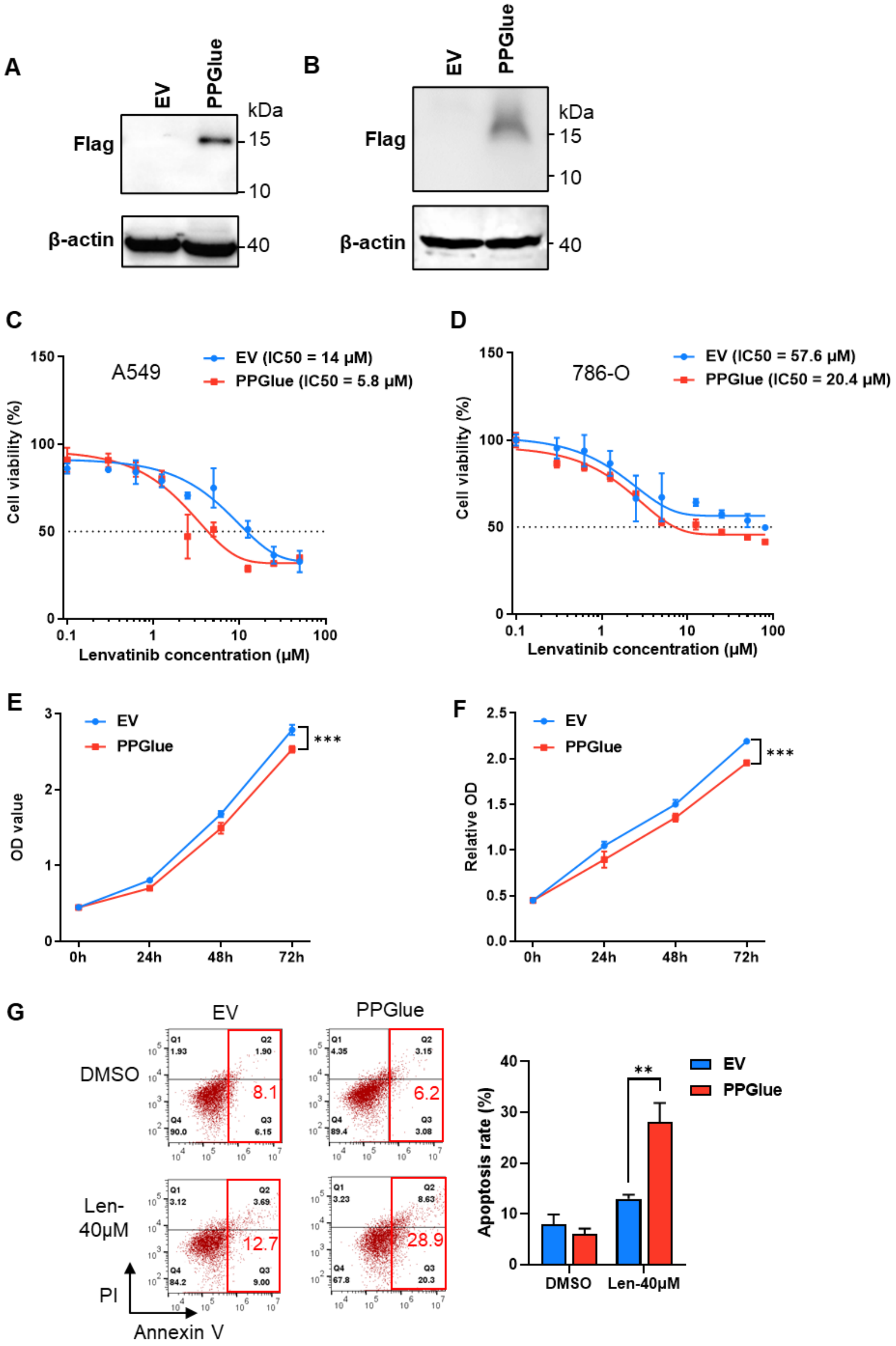


Figure 3-9 Overexpression of PPGlue restored drug sensitivity in other cancer cells. (A, B) Expression of PPGlue in A549 (A) and 786-O (B) cells. (C, D) Overexpression of PPGlue sensitized A549 (C) and 786-O (D) cells to lenvatinib. (E, F) Overexpression of PPGlue

inhibited cell proliferation in A549 (E) and 786-O (F) cells. ***$p < 0.001$ by Student's t-test. (G) Apoptosis assay of 786-O cells. Len, lenvatinib. The experiments were conducted in triplicate. Data were presented as the mean ± s.d., **$p < 0.01$, ***$p < 0.001$ by Student's t-test.

3.3.4 PPGlue knockdown eliminates the sensitizing effect in LR HCC cells

Based on our findings that higher expression of PPGlue corresponded to higher drug sensitivity, we anticipated that depletion of PPGlue would confer lenvatinib resistance to the cells. To verify this hypothesis, we employed siRNA and shRNA to reduce PPGlue levels in Huh7-sensitive cells (**Figure 3-10A & C**) and evaluated the apoptosis rate induced by lenvatinib in these cells. Our results demonstrated that lenvatinib (at both 20 and 40 μM) significantly induced apoptosis in control cells, while only a marginal change in the apoptosis rate was detected after both transient and stable PPGlue knockdown (**Figure 3-10B & D**). Subsequently, we investigated whether loss of PPGlue would impact cell proliferation, viability, and clonogenicity in the presence of lenvatinib. We detected a 2-fold increase in the IC50 value due to decreased drug sensitivity in cells with PPGlue knockdown (**Figure 3-10E**). Additionally, loss of PPGlue not only intrinsically promoted cell proliferation and clonogenicity but also enhanced cell survival during colony formation in the presence of lenvatinib (**Figure 3-10F-G**). Taken together, these findings demonstrate that a decreased level of PPGlue facilitates the formation of lenvatinib resistance, highlighting its crucial role in restoring drug sensitivity in cancer treatment. Further research into the molecular mechanisms underlying the role of PPGlue in cancer progression and drug resistance may provide valuable insights for the development of targeted therapies.

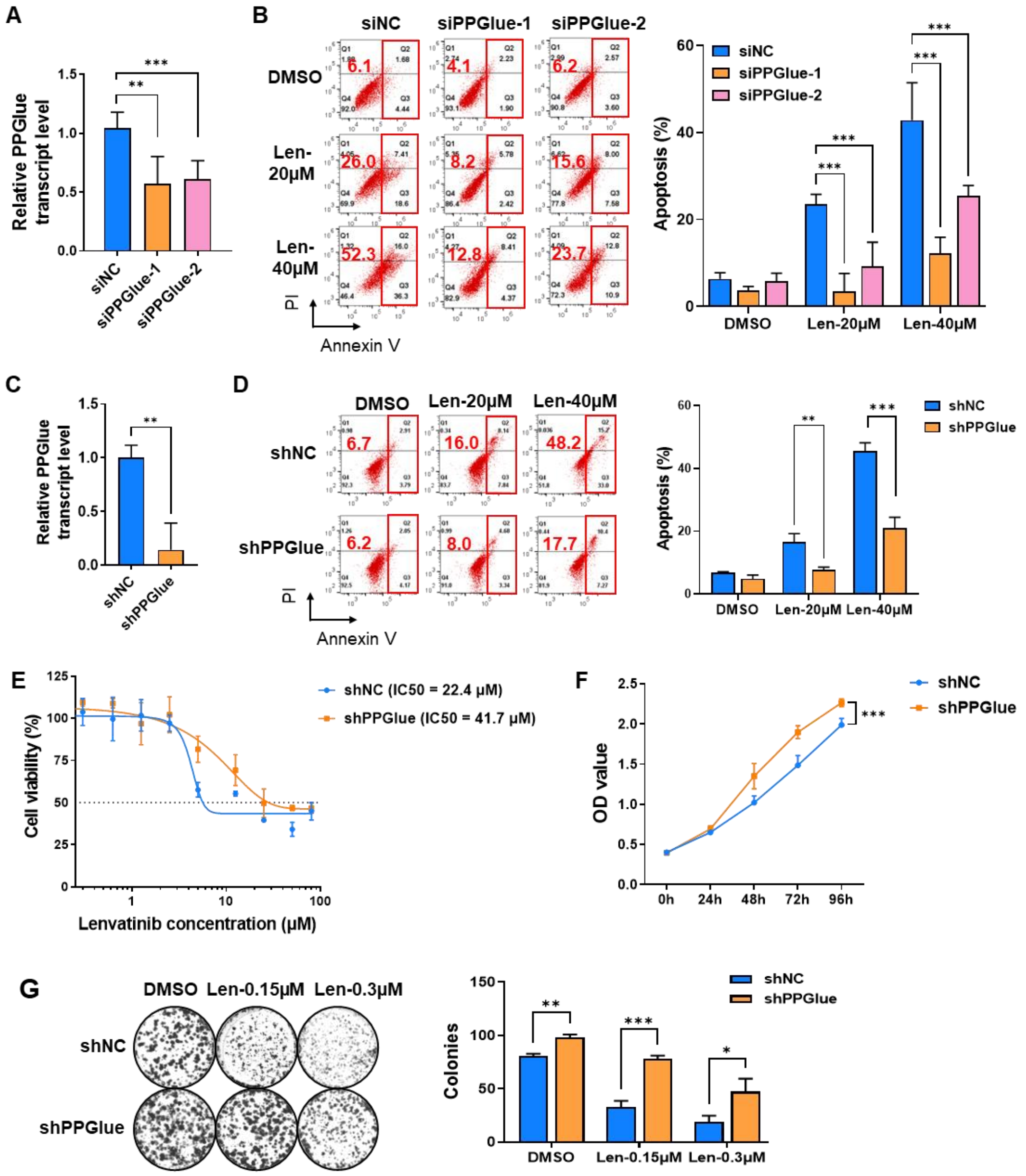


Figure 3-10 PPGlue knockdown conferred drug resistance in sensitive HCC cells. (A) Transient knockdown efficiency with siRNAs targeting PPGlue (siPPGlue-1 and -2) or scramble siRNA as the non-targeting control (siNC). $**p < 0.01$, $***p < 0.001$ by Student's

t-test. (B) Transient PPGlue knockdown abrogated lenvatinib-induced apoptosis in Huh7 cells. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ by two-way ANOVA. (C) Stable knockdown efficiency with shPPGlue targeting PPGlue in Huh7 parental cells or shNC as the non-targeting control. $**p < 0.01$ by Student's t-test. (D) Stable knockdown of PPGlue abrogated lenvatinib-induced apoptosis in Huh7 cells. $**p < 0.01$, $***p < 0.001$ by Student's t-test. (E) Cell viability assay confirmed that knockdown of PPGlue (shPPGlue) conferred lenvatinib resistance to the cells. Non-targeting shNC was used as the control. (F) Stable knockdown of PPGlue (shPPGlue) increased the proliferation of Huh7 cells. Non-targeting shNC was used as the control. $***p < 0.001$ by Student's t-test. (G) Stable knockdown of PPGlue (shPPGlue) alleviated the inhibitory effect of lenvatinib on colony formation in Huh7 parental cells. Non-targeting shNC was used as the control. The cells were treated with DMSO or lenvatinib (0.15 μM and 0.3 μM). $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ by Student's t-test.

3.3.5 PPGlue serves as a sensitizer of HCC to lenvatinib treatment *in vivo*

We examined the role of PPGlue in orchestrating tumor progression using a subcutaneous xenograft model derived from Huh7-LR cells in BALB/c nude mice (**Figure 3-11A**). We defined tumorgenicity as a size of 6 × 6 mm (length × width) in the corresponding xenografts. Tumor formation was confirmed on day 9 post inoculation with a mean tumor size of 158 mm^3 in the control group, but it was delayed by overexpression of PPGlue until day 15 with a mean tumor size of 122 mm^3 (**Figure 3-11B**). We also observed much slower tumor progression in the presence of PPGlue.

The tumor volumes and variations in mice body weight were continuously monitored for 24 days. The mice were euthanized, and the tumor tissues were weighed and photographed (**Figure 3-11C**). Immunohistochemical (IHC) analysis of mouse tumor tissues using a custom PPGlue antibody confirmed the presence of PPGlue (**Figure 3-11D**). PPGlue attenuated tumor growth and spheroid-forming ability by 2.4-fold, highlighting its tumor suppressive role both *in vitro* and *in vivo*.

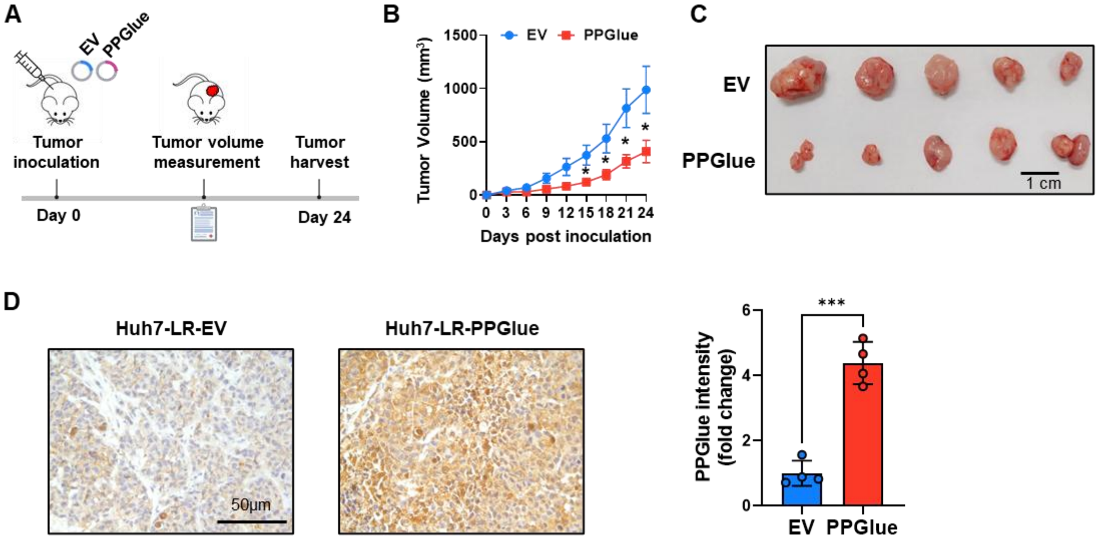


Figure 3-11 PPGlue overexpression slowed down tumor growth in lenvatinib resistant cell-derived xenografts in mice. (A) Schematic diagram of xenografted mice receiving subcutaneous injections of Huh7-LR cells overexpressing PPGlue. (B) Tumor progression curve of mice injected with 0.05 million of LR cells expressing PPGlue or empty vector (EV). Data were presented as the mean ± SEM (n = 5), *$p < 0.05$ by Student's t-test. (C) Representative tumor photographs. (D) IHC staining of mouse tumor tissues using a custom PPGlue antibody. ***$p < 0.001$ by Student's t-test.

To assess the effect of PPGlue on drug resistance *in vivo*, we examined tumor xenografts in response to lenvatinib administration (**Figure 3-12A)**. Mice were divided into two subgroups for treatment: (1) Huh7-LR-EV cells and lenvatinib (30 mg/kg, resuspended in 0.5% methylcellulose) and (2) Huh7-LR-PPGlue cells and lenvatinib. Treatment was initiated once the tumor reached approximately 6 × 6 mm (length × width). Tumor volumes were monitored and measured continuously for 30 days after drug treatment. We observed that lenvatinib treatment without PPGlue failed to suppress tumor progression due to the resistance of Huh7-LR cells (**Figure 3-12B & C**). However, overexpression of PPGlue sensitized the xenografts to lenvatinib treatment, resulting in a 5-fold and 4-fold reduction in tumor volume and tumor weight, respectively (**Figure 3-12B-D**). We observed that the tumor size increased drastically by more than 3-fold in each untreated mouse. This indicated that PPGlue overexpression alone could slow tumor progression, but it was unable to eliminate the tumor. Even in the mice treated with lenvatinib, the tumors were still growing 2.9-fold faster due to resistance to lenvatinib. However, mice bearing Huh7-LR-PPGlue xenograft tumors showed significant shrinkage in tumor size by 28% on average after receiving lenvatinib (**Figure 3-12E)**, suggesting that PPGlue exhibits a drug sensitizing effect greater than its effect on tumor progression. IHC staining of the tumor proliferation marker Ki-67 indicated that over 75% of the tumor cells were growing rapidly, while this level dropped to less than 10% in the presence of PPGlue, reflecting its potential in restoring drug sensitivity and decelerating tumor progression when combined with lenvatinib *in vivo* (**Figure 3-12F**). Meanwhile, terminal deoxynucleotidyl transferase dUTP nick-end labeling (TUNEL) showed that more DNA fragmentation was generated during lenvatinib-induced apoptosis in the presence of PPGlue, but not in the presence of the empty vector (**Figure 3-12G & H**). Collectively, these results demonstrate the tumor-suppressive properties of PPGlue and validate that PPGlue overexpression enhances the efficacy of lenvatinib against LR HCC xenografts, implying its potential therapeutic role in restoring drug sensitivity.

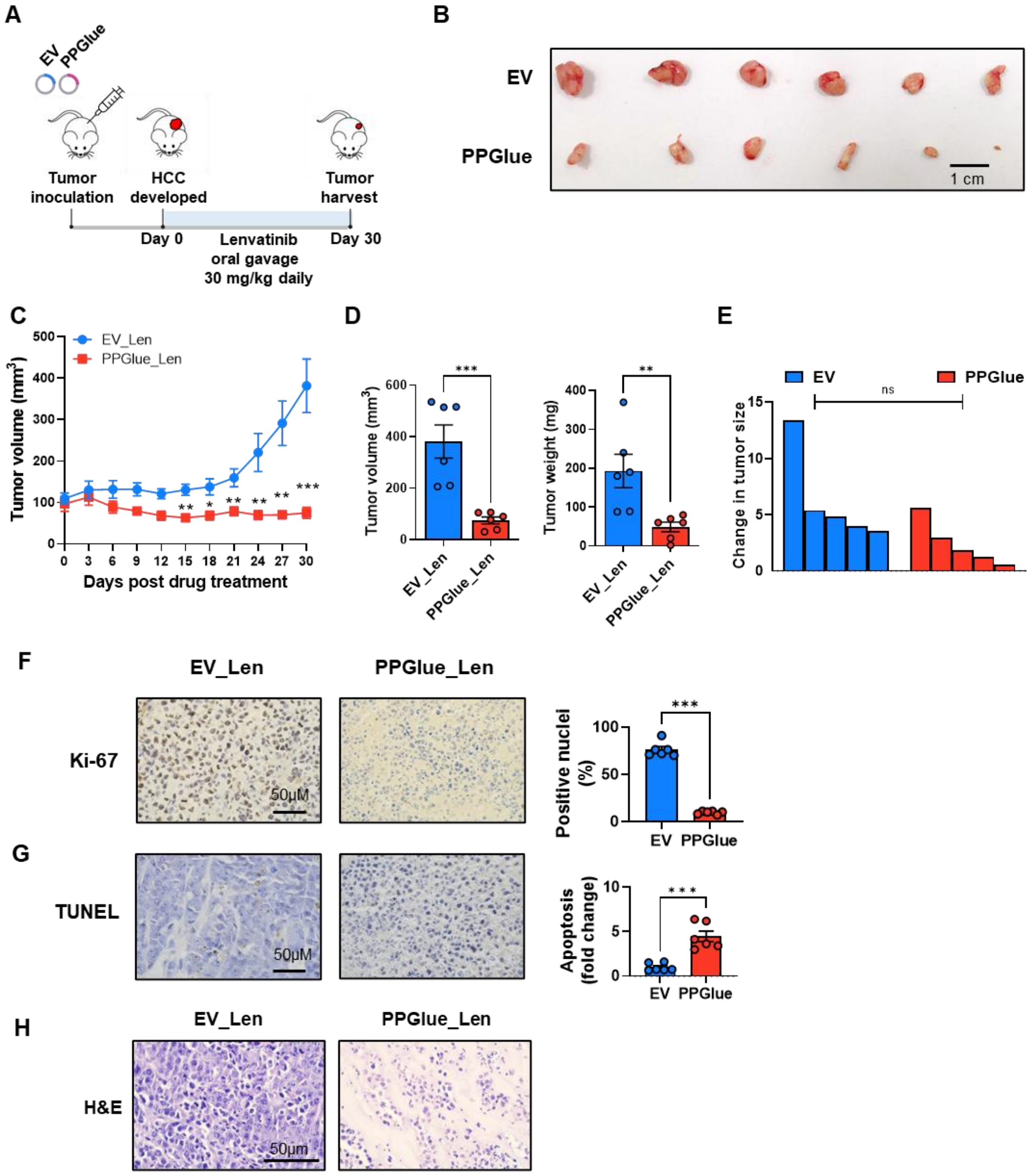


Figure 3-12 PPGlue overexpression enhanced the efficacy of lenvatinib against resistant cell-derived xenografts in mice. (A) Schematic diagram of xenografted mice receiving subcutaneous injections of Huh7-LR cells overexpressing PPGlue and subsequent lenvatinib treatment. (B) Representative tumor photographs on day 30 post lenvatinib treatment. (C, D) Tumor progression curve (C), tumor size (D, left), and weight (D, right) in xenografted mice (n=6). Data were presented as the mean ± SEM, **p* < 0.05, ***p* < 0.01, ****p* < 0.001 by Student's t-test. (E) Waterfall plot showing the response of each tumor to lenvatinib treatment

for 30 days. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ by Student's t-test. (F) TUNEL assay (left) and its quantification (right) indicated apoptosis induced by lenvatinib treatment. ***$p < 0.001$ by Student's t-test. (G) Proliferation in tumors with PPGlue overexpression after lenvatinib treatment. Left, representative IHC staining of Ki-67; right, quantification based on IHC. ***$p < 0.001$ by Student's t-test. (H) Haematoxylin and eosin stain of mouse tumor tissues after lenvatinib treatment at 30 mg/day for 30 days.

3.4 Discussion

HCC is recognized as one of the most aggressive and prevalent forms of cancer, resulting in approximately 780,000 deaths globally each year, according to the 2020 global cancer statistics[210]. The early diagnosis of HCC relies on the use of one or more costly and advanced imaging techniques, such as computed tomography (CT) scans and magnetic resonance imaging (MRI), which complicates the diagnosis of many cases at an early stage[211]. The prognosis for most HCC patients is poor, as approximately 80% are diagnosed at intermediate or advanced stages, thereby missing the optimal opportunity for curative surgical intervention[212]. Lenvatinib, a multi-receptor tyrosine kinase inhibitor that approved by FDA in 2018, serves as an alternative first-line treatment to sorafenib for unresectable HCC. Patients undergoing treatment with lenvatinib demonstrated improved tumor responses and survival outcomes compared to those treated with sorafenib. Nevertheless, patients receiving lenvatinib still have unfavorable prognoses, primarily owing to the emergence of drug resistance[213]. Recently, significant efforts have been made within the scientific community to understand the mechanisms underlying lenvatinib resistance and develop strategies to enhance its clinical effectiveness. Microproteins, which have been largely overlooked as potential therapeutic targets for lenvatinib resistance, may offer new avenues for treatment. Given that traditional protein-targeting approaches have shown limited success in overcoming lenvatinib resistance in HCC patients, there is an urgent need to investigate new targetable molecules within the largely unexplored realm of microproteins.

In Chapter 1, we explored the unique characteristics of microproteins compared to traditional proteins and the challenges associated with studying them. In Chapter 2, we describe a successful approach for identifying novel microproteins from mouse tissue. By optimizing a MS-based proteogenomic workflow, we tried to identify new targetable microproteins associated with lenvatinib resistance in HCC. The identification of microproteins associated with lenvatinib resistance was based on gathering evidence from various levels of microprotein expression. We used public databases and Ribo-seq data to generate a customized database, integrating translatomics and proteomics data for both DIA and DDA searches. Huh7 and PLC are two widely used liver cancer cell models for research on lenvatinib resistance. We identified 815 microproteins in two lenvatinib-resistant cell lines and their corresponding parental cell lines, supported by translatomics and proteomics. These newly identified microproteins exhibit distinct characteristics compared to canonical proteins, with many originating from lncRNA or being nested within main ORFs with different translation frames. This phenomenon has been widely reported in other research articles[214-216], indicating a huge unexplored group of proteins that could serve as therapeutic targets for lenvatinib resistance. Numerous studies have suggested that certain microproteins play significant roles in cancer development and progression[217]. Our results reveal widespread expression regulation of microproteins in the context of lenvatinib resistance in HCC, potentially filling a gap in understanding this resistance mechanism. Among the 815 microproteins identified, 61 candidates were further validated using the PRM method and were significantly altered in both Huh7 and PLC cell lines. To further validate the differentially expressed microproteins, we overexpressed the top 11 most significantly changed microproteins in HEK293T and Huh7 cell lines. Notably, Huh7 cells overexpressing PPGlue demonstrated the most substantial recovery in lenvatinib sensitivity compared to other candidates, whereas endogenous PPGlue expression was reduced in both Huh7-LR and PLC-LR cells. This suggests that PPGlue plays a crucial role in the development of lenvatinib resistance in HCC. PPGlue is a small protein composed of 54 amino acids, translated from lncRNA *XR_001750712*, which is derived from the uncharacterized gene *LOC105370440*, as annotated through automated computational analysis using the Gnomon gene prediction method[218]. Currently, there is a lack of literature regarding the function of XR_001750712 or its ability to encode PPGlue. Given the lack of prior research on PPGlue, we aimed to confirm its endogenous presentation. Our analysis revealed that the amino acid sequence of PPGlue is highly conserved among primates, suggesting its biological significance. Additional evidence supporting the existence of endogenous PPGlue was obtained from the A549 and

786-O cell lines, as well as thyroid tissue, through MS data. We believe that this evidence sufficiently substantiates the existence of endogenous PPGlue. Subsequently, we investigated the correlation between PPGlue expression and the lenvatinib-resistant phenotype in Huh7-sensitive and Huh7-LR cells. The results indicated that overexpression of PPGlue inhibited cell proliferation, increased lenvatinib-induced apoptosis, and sensitized Huh7 cells to lenvatinib treatment. Conversely, knockdown of PPGlue transcripts abrogated the sensitization effect of lenvatinib. The sensitizing effect of PPGlue overexpression was further validated in a patient-derived tumor xenograft model, in which we observed a significant tumor inhibitory effect of PPGlue in the lenvatinib treatment group. Collectively, these findings from both *in vivo* and *in vitro* studies suggest a strong correlation between PPGlue and lenvatinib resistance in HCC, positioning it as a promising therapeutic target for advanced HCC.

In summary, we utilized an optimized proteogenomic pipeline to identify novel targetable microproteins in advanced HCC patients with lenvatinib resistance. A total of 815 microproteins were identified in two HCC cell lines, Huh7 and PLC. Of these differentially changed microproteins, 11 high-confidence candidates were evaluated through overexpression and *in vitro* assays. Notably, PPGlue displayed the strongest correlation with lenvatinib resistance in HCC compared to the other ten candidates. Furthermore, we confirmed the endogenous expression of PPGlue across various cell lines and tissues and observed that overexpression of PPGlue could reverse lenvatinib resistance in HCC cell lines. These findings underscore the effectiveness of our proteogenomic pipeline for identifying novel microproteins as potential therapeutic targets, with PPGlue emerging as a significant candidate for restoring lenvatinib sensitivity in HCC. The mechanisms by which PPGlue mitigates lenvatinib resistance present an intriguing avenue for further research, which will be explored in a subsequent chapter. Additionally, other candidates identified within the cohort may also contribute to the development of lenvatinib resistance and could serve as promising targets for the HCC research community.

Chapter 4. Molecular mechanisms and therapeutic potential: PPGlue modulates P-gp through interaction with the PPP2R3C/PP5 complex

4.1 Introduction

Cancer is a leading cause of mortality globally, primarily owing to its considerable heterogeneity. Despite the continued use of chemotherapy and targeted therapies as primary treatments for cancer, the non-selective toxicity of mono-chemotherapy and drug resistance of targeted therapies have attracted significant criticism. Subsequently, several peptide-based therapeutics have been developed, which present notable advantages over conventional small-molecule targeted therapies. They exhibit higher specificity owing to their ability to precisely target tumor cells and receptors, thereby reducing off-target effects[121, 123]. Their small molecular size facilitates improved tissue penetration, particularly in solid tumors, and allows access to intracellular targets that are frequently unreachable by antibody therapies[219]. Moreover, peptide-drug conjugates facilitate enhanced targeting precision and therapeutic efficacy[129]. Furthermore, peptide synthesis is distinguished by enhanced flexibility and expediency, leading to shorter development timelines than the complex synthesis procedures necessitated by small molecules[220]. The high selectivity and reduced toxicity associated with peptide-based therapies frequently contribute to improved safety profiles and therapeutic outcomes, thereby rendering them a promising and increasingly utilized class of anti-cancer agents[122-124].

The microprotein family represents a substantial and largely untapped reservoir of potential peptide-based therapeutic agents. A complete understanding of the functions of microproteins is essential for the advancement of cancer research. To date, all the microproteins whose functions have been characterized are known to interact with other proteins. Affinity purification coupled with mass spectrometry (AP-MS) is the optimal approach for the unbiased identification of proteins that interact with a peptide or protein of interest[5]. The use of the AP-MS method allows us to effectively delineate the interaction networks of microproteins, thereby providing insight into their biological roles and mechanisms of action. A comprehensive understanding of these protein-protein interactions is vital for uncovering the therapeutic potential of microproteins.

P-glycoprotein (P-gp) is a member of the ATP-binding cassette (ABC) transporter family, encoded by the multidrug resistance protein 1 (MDR1) gene, and is expressed as a glycoprotein with a molecular weight of 170–180 kDa located in the plasma membrane. This transporter is capable of transporting an extensive range of substrates with molecular weights between 250 and 1,250 Da, including phospholipids, sterols, cholic acids, peptides, metabolites, and various drugs, against their concentration gradients through ATP hydrolysis. P-gp serves as a critical plasma membrane transporter and plays a significant role in the multidrug resistance observed in cancer[221]. Lenvatinib has been identified as a specific substrate for ABCB1 and is not excreted by other transporters including OAT1, OAT3, OATP1B1, OATP1B3, OCT1, OCT2, MATE1, MATE2-K, or bile salt export pumps. P-gp function is dependent on phosphorylation, and its activity is regulated by post-translational modifications, including glycosylation and phosphorylation[222-226]. A recent study identified PPP2R3C, a regulatory subunit of protein phosphatase 5 (PP5), as a binding partner for P-gp[227]. PPP2R3C is a regulatory subunit for protein phosphatase (PP) PP5[227], which is categorized as a serine/threonine-specific phosphoprotein phosphatase. PP5 plays a role in the regulation of several cellular processes including DNA replication, gene transcription, protein translation, metabolism, cell cycle progression, cell division, development, and apoptosis[227]. PPP2R3C is expressed in the fetal brain during various developmental stages[228] and is ubiquitously found across various tissues[229]. The structure of PP5 comprises three tetratricopeptide repeat (TPR) domains at the amino-terminus and a phosphatase domain at the carboxy-terminus. The low enzymatic activity of PP5 is attributed to the intramolecular interaction between the TPR domains and the carboxy terminus[230]. PPP2R3C enhances the activity of PP5 through its association with TPR domains, resulting in the formation of a PP5/PPP2R3C complex that exhibits phosphatase activity on substrates, such as casein, histone H1, and minichromosome maintenance complex component 3, similar to the PP2Ac/PR65/PPP2R3C complex[229]. It has been demonstrated that the PP5/PPP2R3C complex dephosphorylates P-gp, resulting in a reduction in P-gp functionality and, consequently, a decrease in drug efflux. Targeting this regulatory pathway represents a promising strategy for overcoming P-gp-mediated drug resistance in cancer therapy.

This study explored the molecular mechanisms by which PPGlue contributes to lenvatinib resistance in HCC. An AP-MS approach was employed to conduct quantitative analysis of the interactome of PPGlue in resistant cell lines. Our findings indicate that PPGlue functions as a molecular glue, stabilizing the protein complex composed of PPP2R3C, PP5, and P-glycoprotein (P-gp), which results in diminished P-gp activity and enhanced drug accumulation within cells. Consequently, PPGlue increases cellular sensitivity to pharmacological agents by facilitating intracellular drug retention. Furthermore, we synthesized a PPGlue peptide to mimic PPGlue overexpression in cellular models, thereby demonstrating its novel therapeutic potential in HCC. This may lead to improved efficacy of lenvatinib and other anti-cancer therapies.

4.2 Materials and methods

4.2.1 Study approval

License to conduct experiments on animals was obtained from the Department of Health, Hong Kong SAR. Approval to conduct animal work at the Hong Kong Polytechnic University was obtained from the Animal Subjects Ethics Sub-Committee.

4.2.2 Plasmids and reagents

Lenvatinib was purchased from Selleckchem (*in vitro* experiments) and LC Laboratories (*in vivo* experiments). Doxorubicin and verapamil were kind gifts from Prof. Larry Chow (The Hong Kong Polytechnic University). Regorafenib, pazopanib, and MG132 were purchased from MedChemExpress.

4.2.3 Cell lines and cell culture

Huh7 cells were maintained in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin (PS). Huh7-LR and Huh7-sensitive cells were maintained in

Dulbecco's modified Eagle's medium (DMEM, Invitrogen) containing 10% FBS, 1% PS, and 10 μM lenvatinib (Huh7-LR), or an equivalent amount of DMSO (Huh7-sensitive). All cell lines were incubated at 37 °C in a 5% (v/v) CO_2 atmosphere.

4.2.4 Stable isotope labeling by amino acid in cell culture (SILAC)

Huh7-LR cells were cultured in arginine- and lysine-deficient DMEM (Thermo Fisher Scientific), and the medium was supplemented with either natural L-arginine and L-lysine or their stable isotope-labeled counterparts (Arg10, [$^{13}C_6{}^{15}N_4$] arginine and Lys8, [$^{13}C_6{}^{15}N_2$] lysine; both from Thermo Fisher Scientific). After at least 5 passages for SILAC labeling, cells were harvested to determine labeling efficiency. The incorporation rate of individual peptides was calculated manually using the formula (heavy intensity/total intensity) × 100%. Only cells with incorporation percentages exceeding 98% were considered for the subsequent pull-down MS analysis.

4.2.5 Immunoprecipitation

To identify the proteins that interact with PPGlue, Huh7-LR cells overexpressing an empty vector (EV) and PPGlue with SILAC were used. In brief, the cells cultured with isotope-labeled (heavy or light) lysine and arginine were lysed and combined in a 1:1 ratio for reciprocal pulldown MS, including a forward sample (EV-light, PPGlue-heavy) and a reverse sample (EV-heavy, PPGlue-light). The protein lysate was incubated with anti-FLAG (D6W5B) antibody (Cell Signaling Technology) cross-linked to Protein A Mag Sepharose beads (Cytiva) at room temperature for 60 min with gentle agitation. The immobilized proteins were subjected to on-bead trypsin digestion at 37 °C overnight. The resulting tryptic peptides were collected, dried down and analyzed using MS.

To validate the proteins that interact with PPGlue, Huh7-LR cells overexpressing empty vector (EV) and PPGlue were used for anti-FLAG pulldown. The immobilized proteins were eluted by boiling in a 2× Tricine loading buffer for 10 min and then subjected to SDS-PAGE and immunoblotting analyses. To validate the effect of PPGlue in recruiting PP5 and PPP2R3C, 293T-PPP2R3C cells overexpressing an empty vector (EV) and PPGlue were used for anti-HA pulldown. Cell lysate was incubated with magnetic beads-conjugated anti-HA VHH Single Domain antibody (Abclonal, AE109) at room temperature for 60 min with gentle agitation. The immobilized proteins were eluted by boiling in a 2× Tricine loading buffer for 10 min and then subjected to SDS-PAGE and immunoblotting analyses.

4.2.6 Drug accumulation assay

Drug accumulation assay was performed as previously described with minor modifications [231]. Briefly, 1×10^6 cells in culture medium containing 20 μM of lenvatinib or pazopanib was added into an Eppendorf tube. DMSO was used as a solvent control. Subsequently, the cells were incubated at 37 °C for 2 h with gentle shaking at 250 rpm. Then, the cells were washed with cold PBS and lysed with lysis buffer (50 mM Tris-HCl, 150 mM sodium chloride, 2 mM EDTA, 1% NP-40, and 1% sodium deoxycholate). Intracellular drugs were extracted from cell lysates following the same protocol as previously described [232]. Briefly, each lysate (100 μL) was immediately mixed with 20 μL of methanol, 10 μL of trichloroacetic acid aqueous solution, and 70 μL of acetonitrile. The resulting mixture was vortexed for 1 min and then centrifuged at 15,000 g for 10 min. The supernatant was subsequently injected into the UPLC-MS/MS system (AB SCIEX QTRAP 6500) and analyzed using an HSS T3 C18 column (2.1 mm × 150 mm, 1.8 μm, Waters) maintained at 40 ℃. The mobile phase consisted of 0.1% v/v formic acid in water (A) and acetonitrile (B), with a flow rate of 0.32 mL/min. The elution was carried out with a gradient as follows: 0-1 min, 20% B; 1-4 min, 20-85% B; 4-4.1 min, 85-95% B; 4.1-7 min, 95% B. Lenvatinib and pazopanib were detected using tandem mass spectrometry in electrospray ion (ESI) positive mode via multiple reaction monitoring (MRM) mode. The MRM transitions for lenvatinib and pazopanib were 427.0/369.9 and 438.0/357.2, respectively.

Regarding the parameters of the mass spectrometer, the ion spray voltage was set at 5500 V, and the temperature was set at 550 °C. The curtain gas, nebulizer gas, and heater gas were all ultrahigh-purity (UHP) nitrogen, with pressures adjusted to 55, 55, and 30 psi, respectively. After the LC-MS/MS run, the acquired data were processed using Analyst® software version 1.4.2 (AB SCIEX).

4.2.7 Delivery of synthetic PPGlue into cells

The synthetic microprotein was diluted and mixed with PULSin$^{®}$ (Polyplus) according to the manufacturer's instructions. Huh7-LR cells were transfected with 25 μg of synthetic PPGlue for 6 h in serum-free medium. Following this incubation period, cells were prepared for lenvatinib-induced apoptosis and cell viability tests.

4.3 Results

4.3.1 PPGlue physically interacts with PPP2R3C/PP5 complex, leading to decreased P-gp levels and the increased intracellular drug concentration

Subsequently, we aimed to understand the mechanism by which PPGlue restores drug sensitivity. Microproteins exert biological functions through protein-protein interactions with annotated proteins. For example, EMBOW regulates the cell cycle by binding to WDR5 and alters the WDR5 interactome (9). Therefore, we employed stable isotope labeling by amino acids in cell culture (SILAC) and anti-Flag pulldown MS to identify the interactome of PPGlue in LR cells (**Figure 4-1A**). In brief, the cells were cultured in media containing isotope-labeled (heavy or light) lysine and arginine. Thereafter, the cells were lysed and combined in a 1:1 ratio based on the total protein amount to create replicates for pull-down MS, including a forward sample (EV-light and PPGlue-heavy) and a reverse sample (EV-heavy and PPGlue-light). Anti-Flag immunoprecipitation was used to pull down PPGlue and its binding partners. The enriched proteins were analyzed by MS to determine the isotopic ratios (heavy/light). The top 10 hits showing substantial abundance changes are presented in the upper-right quadrant of the corresponding scatter plot (**Figure 4-1B, Table 4-1**). STRING analysis revealed that three of the identified interacting proteins were predominantly clustered together based on their physical interactions (**Figure 4-1B**). Among the three, protein phosphatase 2A regulatory subunit B gamma (PPP2R3C) was the most specifically enriched as it has never been reported in the contaminant repository for affinity purification–mass spectrometry data (CRAPome)[233]. PPP2R3C forms a complex with protein phosphatase 5 (PP5), which negatively regulates P-gp abundance and activity, thereby regulating drug efflux[227].

To validate the protein interactions involving PPGlue, PPP2R3C, and PP5, we conducted anti-Flag co-immunoprecipitation, and the results demonstrated that PPP2R3C, PP5, and P-gp were accompanied by PPGlue-Flag (**Figure 4-1C**), confirming the physical association among these proteins. Co-immunoprecipitation against HA-tagged PPP2R3C in HEK293T cells also validated the association of P-gp with PPP2R3C/PP5 complex and PPGlue (**Figure 4-1D & 4-2**). In addition, PPGlue increased the amount of PP5 and P-gp that could be co-immunoprecipitated with HA-tagged PPP2R3C, although P-gp levels were reduced in cells overexpressing PPGlue (**Figure 4-1D**). Therefore, we anticipated that PPGlue may act as a "molecular glue" that facilitates the interaction between PP5 and PPP2R3C, potentially by stabilizing the protein complex (**Figure 4-1E**). Therefore, we named this microprotein PPGlue based on its molecular mechanism of action.

We hypothesized that PPGlue restores drug sensitivity by downregulating P-gp, since PPGlue overexpression correlated with a reduction in P-gp expression in both LR cells and xenografted tumors (**Figure 4-1F-G**). The observed change caused by PPGlue could be rescued by the proteasome inhibitor MG132 (**Figure 4-3A**), suggesting that PPGlue may play a role in the degradation of P-gp. Loss of PP5 has been reported to upregulate the expression and function of P-gp[227]. Since PPGlue is involved in the PPP2R3C/PP5 complex with P-gp, we investigated whether inhibiting

PP5 by okadaic acid (OA) would abolish the downregulation of P-gp by PPGlue. We detected a significant elevation in P-gp levels in cells overexpressing PPGlue after OA treatment (**Figure 4-3B**). Moreover, we evaluated both endogenous PPGlue and P-gp abundance in patient tumor-derived mouse xenografts. PPGlue was downregulated while P-gp levels were upregulated in LR tumors compared to lenvatinib-sensitive tumors (**Figure 4-1H & I**). Taken altogether, these results confirm that PPGlue mediated P-gp through the interaction with the PPP2R3C/PP5 complex.

Table 4-1 List of binding proteins identified in SILAC-labeled cells by anti-flag immunoprecipitation.

Protein ID	Gene	Protein Description	Log_2 (Forward, H/L)	$-Log_2$ (Reverse, H/L)	CRA Pome
Q5VT06	CEP350	Centrosome-associated protein 350	1.52	2.05	0.04
O95684	CEP43	Centrosomal protein 43	1.50	1.92	0.08
PPGlue-Flag	PPGlue	PPGlue	0.60	1.78	0.00
Q969Q6	PPP2R3C	Serine/threonine-protein phosphatase 2A regulatory subunit B" subunit gamma	1.75	1.06	0.00
Q03135	CAV1	Caveolin-1	2.20	0.86	0.04
Q9H223	EHD4	EH domain-containing protein 4	1.38	0.80	0.02
Q9H490	PIGU	Phosphatidylinositol glycan anchor biosynthesis class U protein	0.86	0.76	0.02
Q9BRU9	UTP23	rRNA-processing protein UTP23 homolog	0.64	0.74	0.01
Q9UNQ2	DIMT1	Probable dimethyladenosine transferase	0.63	0.65	0.07
Q6NZI2	CAVIN1	Caveolae-associated protein 1	0.78	0.64	0.04

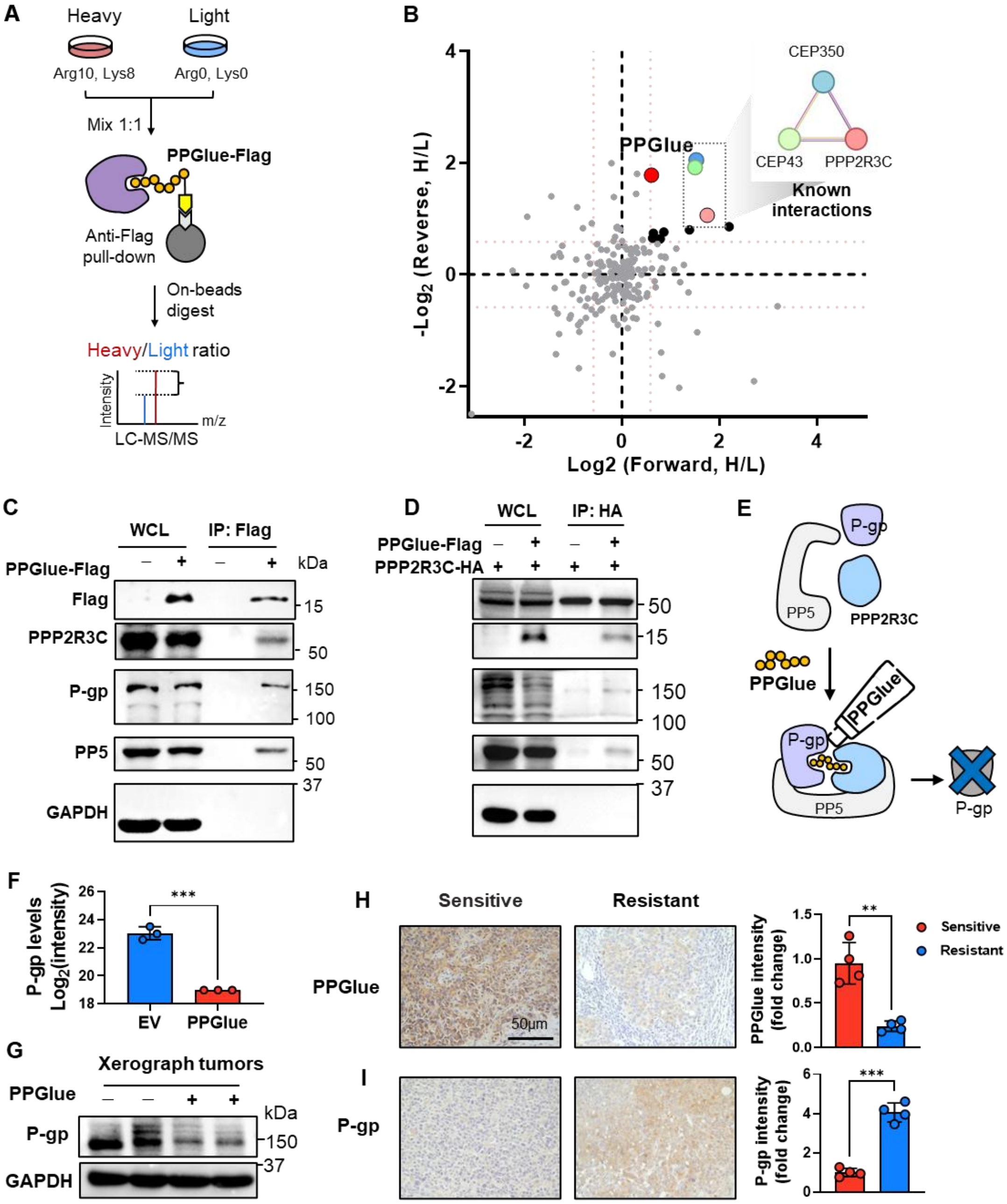


Figure 4-1 The molecular mechanism through which PPGlue enhanced the drug sensitivity of HCC cells. (A) Schematic workflow for identifying PPGlue-binding proteins. (B) The identified PPGlue interacting proteins using the SILAC pulldown method, and the STRING analysis of the identified proteins with CRAPome frequency less than 10%. Log$_2$(isotopic ratios) values of zero indicate no enrichment, while ratios exceeding log$_2$1.5 were considered significant. (C, D) The interacting proteins were validated by performing

forward pulldown assays against PPGlue-Flag (C) and reverse pulldown assays with PPP2R3C-HA (D). WCL, whole cell lysate. (E) Proposed mechanism for PPGlue to enhance lenvatinib sensitivity. (F) Protein levels of P-gp quantified by MS in Huh7-LR-EV and -PPGlue cells. ***$p < 0.001$ by Student's t-test. (G) Expression of P-gp in mouse tumors. EV, empty vector; PG, PPGlue. (H, I) Representative IHC staining of PPGlue (H, left) and P-gp (I, left) in patient tumor-derived xenografts (PY003). The levels of PPGlue (H, right) and P-gp (I, left) were quantified based on the IHC images. **$p < 0.01$, ***$p < 0.001$ by Student's t-test.

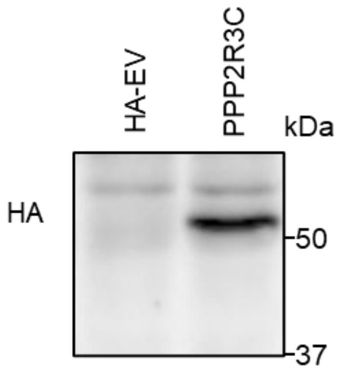


Figure 4-2 Overexpression of PPP2R3C in HEK293T cells.

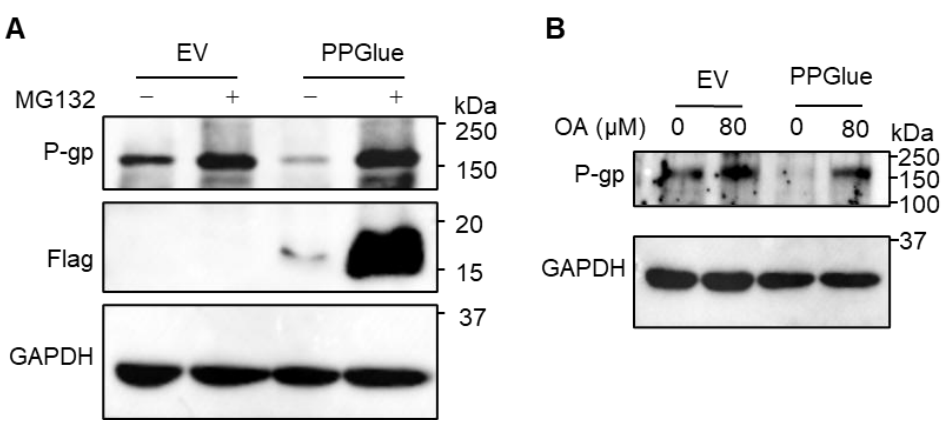


Figure 4-3 Effects of PPGlue on P-gp protein levels and degradation. (A) Protease

inhibitors (MG132) abrogated P-gp protein degradation in the cells with PPGlue overexpression. (B) Phosphatase inhibitors increased P-gp protein levels in cells overexpressing PPGlue. OA, okadaic acid.

4.3.2 The effects of PPGlue on the enhanced efficacy of multiple drug molecules for the inhibition of various cancers

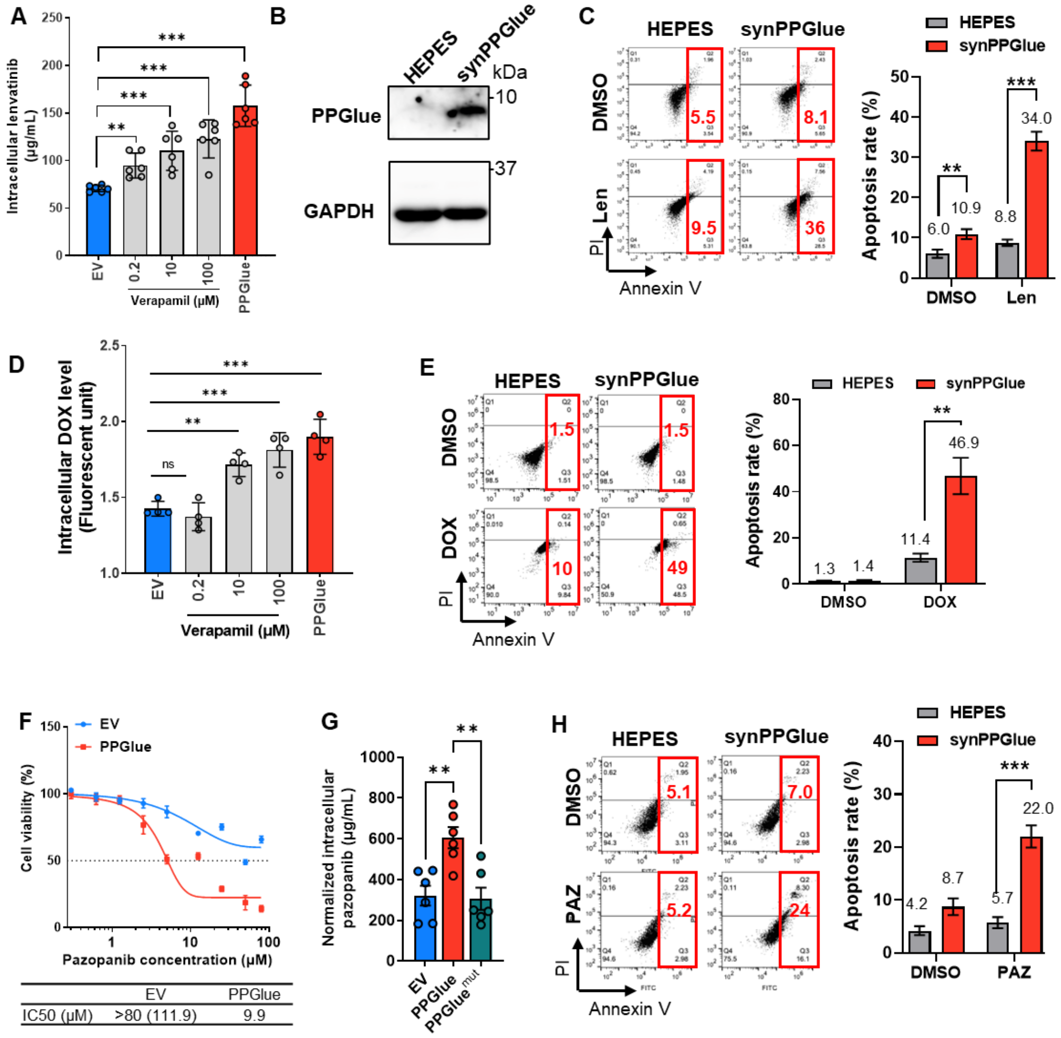
We investigated whether the enhanced assembly of the PPP2R3C/PP5 complex with P-gp by PPGlue would prohibit drug efflux and revive drug sensitivity. We conducted drug accumulation assays, which revealed a significant increase in intracellular lenvatinib concentration in cells overexpressing PPGlue (**Figure 4-4A**). Inhibition of P-gp by verapamil led to a dose-dependent elevation in lenvatinib concentration in cells without PPGlue. Nonetheless, PPGlue overexpression resulted in the highest drug accumulation (**Figure 4-4A**). This finding indicated that PPGlue can modulate drug efflux through P-gp, emphasizing its importance in maintaining drug sensitivity and influencing intracellular lenvatinib levels.

In addition to PPGlue overexpression, we also delivered exogenous synthetic PPGlue (synPPGlue) into LR parental cells and monitored its stability (**Figure 4-4B**). Our findings revealed that synPPGlue was gradually consumed after delivery, while it remained functional to downregulate P-gp levels and augmented lenvatinib-induced apoptosis in these cells (without exerting any influence on its transcription) (**Figure 4-4C & 4-5**). These observations corresponded to our earlier findings where stable cells were used, indicating that synPPGlue also has the potential to resume drug sensitivity in LR cells.

Since P-gp is known to transport a number of anti-cancer chemotherapeutic drugs, we investigated how broad its substrate spectrum can be tuned by PPGlue through its regulation of P-gp. Our results showed that PPGlue could sensitize LR cells to doxorubicin, a P-gp substrate, by increasing the drug concentration and decreasing the drug IC50 (**Figure 4-6A**). Inhibition of P-gp by verapamil led to intracellular accumulation of doxorubicin, whereas overexpression of PPGlue alone resulted in the greatest accumulation without verapamil (**Figure 4-4D**). Moreover, flow cytometry results demonstrated that synPPGlue synergized with doxorubicin to enhance apoptosis in LR cells (**Figure 4-4E**). These data highlight the important role of PPGlue in modulating drug accumulation beyond lenvatinib.

Drug resistance has been identified as a common challenge in targeted therapy – the emergence of innate or acquired resistance to TKIs, such as lenvatinib. Therefore, we assessed the impact of PPGlue on the efficacy of structurally similar drugs, pazopanib and regorafenib. According to the manufacturer, pazopanib is a substrate of P-gp, whereas regorafenib is not[205, 234, 235]. PPGlue overexpression induced a marginal change in the IC50 of regorafenib in the cell viability assays, while a striking 10-fold reduction in the IC50 of pazopanib was detected (**Figure4-4F**). In line with the difference in IC50 value, drug accumulation assays revealed minimal changes in intracellular regorafenib concentration, while a significant increase in pazopanib concentration was observed in the cells overexpressing PPGlue (**Figure 4-4G & 4-6B**). Pazopanib is transported by P-gp, resulting in a greater synergistic effect with PPGlue.

Consistently, flow cytometry analysis demonstrated that synPPGlue could lead to a higher apoptosis rate induced by pazopanib in LR cells (**Figure 4-4H**). Therefore, PPGlue is able to enhance the efficacy of drugs as long as they are P-gp substrates, and this offers an innovative strategy to combat drug resistance in cancer therapy.



	EV	PPGlue
IC50 (μM)	>80 (111.9)	9.9

Figure 4-4 Effects of PPGlue on multiple cancer types in response to various drugs. (A) Intracellular lenvatinib accumulation in Huh7-LR-EV, -PPGlue, and -PPGlue^{mut} cells. Huh7-LR-EV cells treated with the indicated concentration of verapamil were used as positive controls. ***p* < 0.01, ****p* < 0.001 by Student's t-test. (B) WB confirmed the delivery

of synthetic PPGlue (synPPGlue) into Huh7-LR cells. (C) Synthetic PPGlue enhanced lenvatinib-induced apoptosis (40 μM). $**p < 0.01$, $***p < 0.001$ by Student's t-test. (D) Inhibition of P-gp by verapamil increased intracellular doxorubicin accumulation. $**p < 0.01$, $***p < 0.001$ by Student's t-test. ns, no significance. (E) Synthetic PPGlue enhanced apoptosis induced by doxorubicin in Huh7-LR cells. $**p < 0.01$ by Student's t-test. (F) Cell viability of Huh7-LR-PPGlue cells with doxorubicin (0.35 μM) treatment. (G) Pazopanib concentration in Huh7-LR-EV, -PPGlue, and -PPGlue^{mut} cells. $**p < 0.01$ by Student's t-test. (H) Synthetic PPGlue enhanced pazopanib (40 μM) induced apoptosis in Huh7-LR cells. HEPES was used as a solvent control for peptide delivery. PAZ, pazopanib. $***p < 0.001$ by Student's t-test.

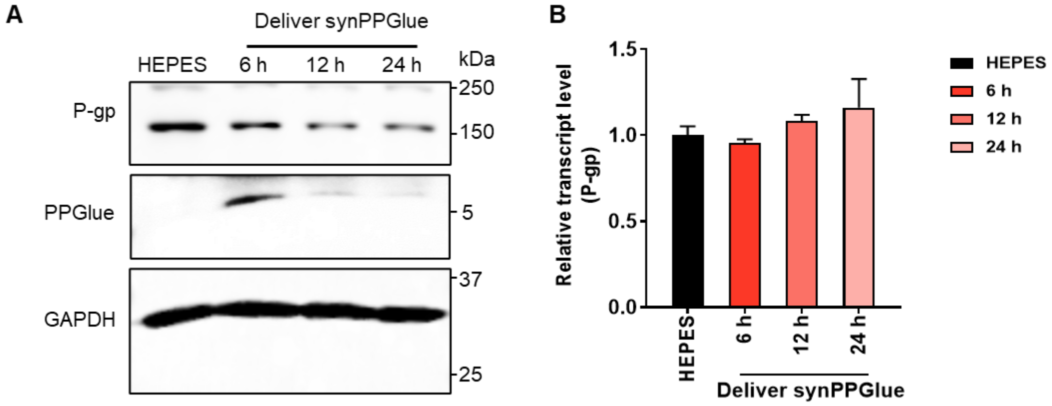


Figure 4-5 Effects of PPGlue on P-gp protein levels and degradation. The synthetic PPGlue peptide modulated P-gp at the protein level (A) but had no effect on P-gp transcript level (B). The experiments were conducted in triplicate, and data were presented as the mean ± s.d.

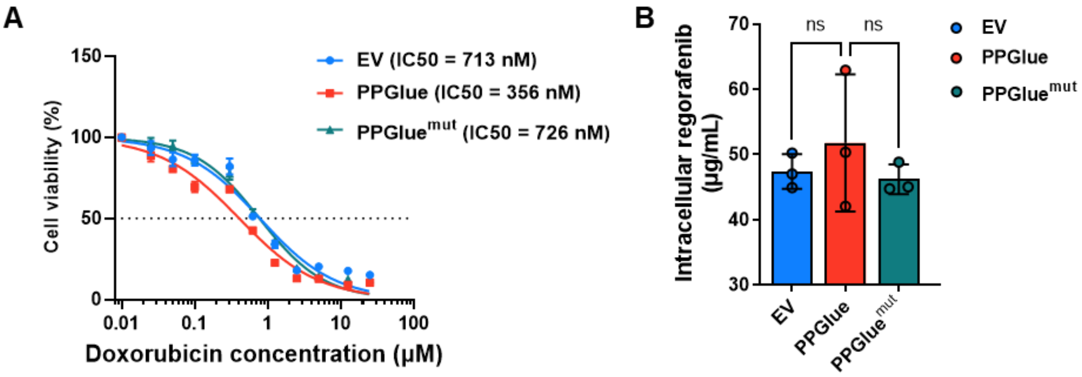


Figure 4-6 Impact of PPGlue overexpression on drug sensitivity and accumulation. (A) Cell viability of doxorubicin in Huh7-LR cells with PPGlue overexpression. (B) Regorafenib accumulation assay assessed using MRM. ns, no significance. The experiments were conducted in triplicate and data were presented as the mean ± s.d.

4.4 Discussion

Thus far, the molecular mechanisms underlying lenvatinib resistance in HCC remain elusive with multiple contributing factors, including altered signaling pathways, drug efflux, and tumor environment. While annotated proteins like EGFR[236], DUSP9[237], ITGB8[238], and CDK6[239] have been implicated, the role of microproteins in this context has not been explored. Our study represents an advancement in this field using an integrated proteogenomic approach to identify 815 microproteins in HCC cells. Among these, PPGlue, a novel microprotein, plays a crucial role in lenvatinib resistance. Overexpression of PPGlue restored lenvatinib sensitivity in LR cells, inhibited proliferation, and hampered colony formation. This discovery not only emphasizes the importance of comprehensive proteomic profiling in understanding cancer resistance mechanisms, but also offers a fresh perspective on restoring drug sensitivity in HCC. We have thoroughly validated the functionality of PPGlue, demonstrating its ability to sensitize both lenvatinib-sensitive and-resistant HCC cells to the drug. This finding suggests that the therapeutic potential of PPGlue extends beyond its impact on lenvatinib resistance. Furthermore, our observation of lower PPGlue levels in LR patient-derived xenografts underscores its clinical relevance.

Microproteins exert biological functions through protein-protein interactions with annotated proteins. For example, EMBOW regulates the cell cycle by binding to WDR5 and altering the WDR5 interactome[201]. Therefore, we anticipated that the mechanism of PPGlue would require a partner protein with relevant functions. Using SILAC pull-down, we identified PPGlue to be involved in the interaction between the PPP2R3C/PP5 complex and P-gp. A recent study has unveiled that PPP2R3C/PP5 dephosphorylates P-gp and knocking down the complex leads to increased levels of the transporter. P-gp is a drug efflux pump that confers resistance to anti-cancer drugs, and modulating P-gp has emerged as a strategy to enhance the effectiveness of chemotherapy drugs. However, whether P-gp renders lenvatinib resistance in HCC has not been reported. In the present study, we revealed elevated P-gp levels in LR cells, leading to insufficient intracellular accumulation of lenvatinib. Furthermore, the inhibitory effect of PPGlue on P-gp also increased the accumulation of other P-gp substrate drugs in cells, such as pazopanib and doxorubicin, highlighting its broad therapeutic potential.

Peptide therapy has emerged as a new and promising modality for the treatment of cancers with multidrug resistance. Shao et al. raised an effective strategy using a synergistic combination of membranolytic anti-tumor β-peptide polymer and doxorubicin to promote drug uptake by multidrug-resistant cancer cells[240]. Bioactive peptides such as PPGlue possess desirable properties for therapeutic development, including good biocompatibility and target binding affinity[241]. This study investigated the therapeutic efficacy of synthetic PPGlue combined with lenvatinib, pazopanib, and doxorubicin. Future studies should focus on determining the core PPGlue sequence that is essential for interacting with the PPP2R3C/PP5 complex and inhibiting drug efflux. Moreover, mRNA therapy offers another exciting avenue for combating diseases. Recent advances in biotechnology and molecular medicine have enabled the production of almost all functional proteins/peptides in the human body as therapeutic agents[242]. mRNA nanoparticles encoding tumor suppressors have shown promise in inducing rapid tumor-suppressive protein expression. In particular, delivery of synthetic PTEN mRNA particles can restore functional proteins in PTEN-null human prostate tumor-bearing mice[243]. TP53 mRNA nanoparticles can markedly improve the sensitivity of tumor cells to rapamycin (mTOR) inhibitors for potent combinatorial cancer treatment[244]. Given PPGlue's tumor suppressive activity in both lenvatinib-sensitive and resistant cells, we propose that PPGlue and its transcript could also be used for developing mRNA-based therapeutics to target P-gp-mediated multidrug resistance in HCC.

In conclusion, we identified PPGlue using an integrated proteogenomic workflow. We revealed a novel mechanism by which PPGlue acts as a molecular glue that enhances the efficacy of lenvatinib by potentially stabilizing the PPP2R3C/PP5/P-gp protein complex, leading to reduced P-gp function and increased drug accumulation. Consequently, PPGlue restored lenvatinib sensitivity in drug-resistant HCC cells, inhibited their proliferation, and hampered colony formation, as illustrated in **Figure 4-7**. Our findings suggest that targeting P-gp-mediated multidrug resistance with synthetic PPGlue could be an innovative therapeutic strategy for HCC, potentially improving the efficacy of lenvatinib and other anti-cancer drugs.

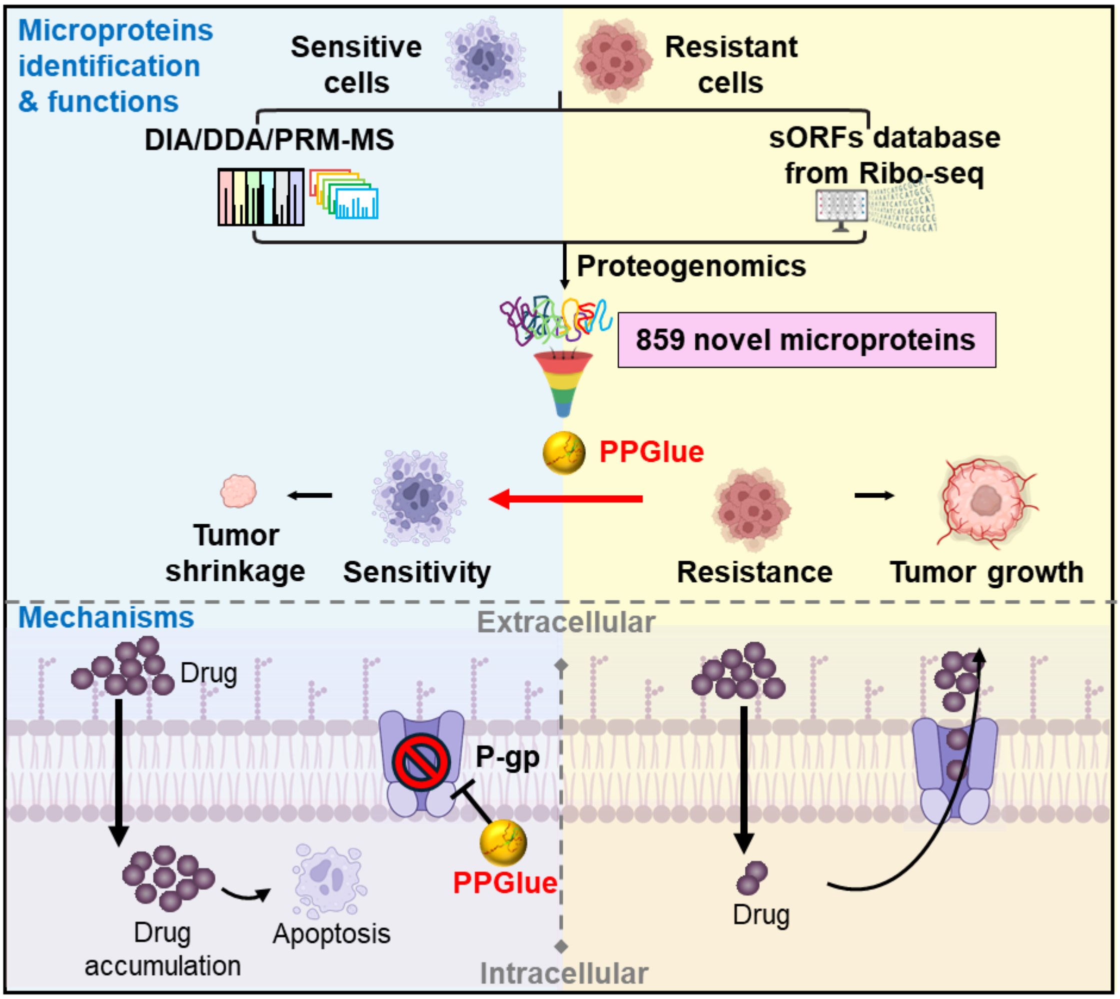


Figure 4-7 Schematic illustration of PPGlue discovery using a proteogenomic workflow. PPGlue acts as a molecular glue that enhances the efficacy of lenvatinib by potentially stabilizing the PPP2R3C/PP5/P-gp protein complex, leading to reduced P-gp function and

increased drug accumulation.

Chapter 5. Conclusion and future perspective

5.1 Conclusion

In this study, we have successfully developed and optimized proteogenomic methodologies to reveal the roles of sORF-encoded microproteins in liver development and HCC. By integrating Ribo-seq with advanced MS-based proteomics, we established a comprehensive framework for the systematic identification and functional characterization of microproteins. This study not only broadens the microprotein landscape but also underscores their potential as key regulators in both developmental biology and cancer therapy.

5.1.1 Discovery of microproteins involved in developmental regulation

We have refined a workflow that integrates SEC and MS to enhance the enrichment and identification of microproteins within complex proteomes. This workflow enabled the discovery of 89 novel microproteins in mouse liver, with 39 exhibiting differential expression between embryonic and adult stages. These findings revealed that microproteins upregulated during embryonic development are primarily involved in RNA splicing and processing, while those active in adult livers are enriched in metabolic pathways. These results underscore the functional importance of microproteins in liver development and provide a foundation for future studies into their regulatory roles.

5.1.2 PPGlue: a novel therapeutic agent in HCC

Utilizing an integrated proteogenomic approach that incorporates multiple mass spectrometry techniques, we successfully identified 815 microproteins in human HCC cells, including PPGlue, a previously uncharacterized microprotein. Functional studies demonstrated that PPGlue sensitizes lenvatinib-resistant HCC cells to treatment by enhancing apoptosis, suppressing proliferation, and increasing intracellular drug concentration. Mechanistically, PPGlue acts as a molecular glue, facilitating the assembly of the PPP2R3C/PP5 protein phosphatase complex, which downregulates P-gp, a key drug efflux transporter. This leads to increased intracellular drug accumulation and enhanced sensitivity to P-gp substrate drugs, such as lenvatinib, pazopanib, and doxorubicin. Additionally, synthetic PPGlue exhibited synergistic effects with these drugs, underscoring its potential as a therapeutic agent for overcoming multidrug resistance.

5.2 Future perspective

5.2.1 Exploring the potential of PPGlue as a therapeutic target in HCC

This study represents a significant step forward in the discovery and characterization of microproteins, highlighting their important roles in liver development and cancer resistance. By combining advanced proteogenomic approaches with functional studies, we have established a comprehensive framework for exploring the therapeutic and diagnostic potential of microproteins. PPGlue, as a representative microprotein, exemplifies the untapped potential of this class of molecules in addressing unmet medical requirements.

5.2.2 Exploring the potential of PPGlue-derived drugs and delivery systems as novel therapeutic strategies

PPGlue represents a promising therapeutic candidate for restoring drug resistance in HCC. Its ability to modulate P-gp and improve the effectiveness of multiple anti-tumor drugs positions it as a novel therapeutical strategy in cancer treatment. Nevertheless, several challenges remain before PPGlue can be translated into clinical applications. Future studies should focus on the following areas:

Structural characterization of PPGlue: As a noncanonical protein, PPGlue lacks structural information, which presents a significant challenge for understanding its action mechanism and optimizing its therapeutic potential. To address this, bioinformatics tools such as AlphaFold3 can be employed to predict its three-dimensional structure with high accuracy. These predictions can provide valuable insights into its functional domains, potential interaction sites, and overall stability. Nonetheless, it is imperative that computational predictions undergo experimental validation to confirm their accuracy. Techniques such as cryo-electron microscopy (cryo-EM) or nuclear magnetic resonance (NMR) spectroscopy can be used to determine the high-resolution structure. Cryo-EM is particularly suitable for studying protein complexes, which could reveal how PPGlue facilitates the assembly of the PPP2R3C/PP5 protein phosphatase complex. NMR, on the other hand, well-suited for smaller proteins, providing detailed insights into their dynamic characteristics. These structural studies will not only deepen our understanding of PPGlue’s mechanism of action but also inform the rational design of synthetic variants with enhanced therapeutic properties.

Structural optimization: Synthetic PPGlue requires further optimization to improve its potency, specificity, and stability. Employing rational design approaches, such as peptide stapling, cyclization, or chemical modifications (e.g., N-terminal acetylation), could enhance its resistance to proteolytic degradation and prolong its half-life in circulation. Furthermore, optimizing its structure characteristics could improve its pharmacokinetics properties. It is particularly essential to enhance PPGlue's capacity to penetrate cell membranes in order to maximize its intracellular activity. Utilizing strategies such as the conjugation of cell-penetrating peptides (CPPs) or the modification of hydrophobic residues could significantly augment its cellular uptake and therapeutic effectiveness.

Delivery systems: Developing efficient delivery systems is crucial for ensuring that PPGlue arrives at its target site in a stable and bioactive form. Utilizing nanoparticle-based delivery platforms, such as liposomes or polymeric nanoparticles, can safeguard PPGlue from enzymatic degradation and promote its accumulation in tumor tissues through the enhanced permeability and retention (EPR) effect. Alternatively, conjugating PPGlue with tumor-targeting ligands, such as antibodies or small molecules that recognize tumor-specific markers, could improve its selectivity and minimize off-target effects.

Peptide-drug conjugates (PDCs): PPGlue has the potential to be developed as part of a PDC in conjunction with existing anti-tumor agents such as doxorubicin or lenvatinib. By chemically linking PPGlue to these drugs, the resulting PDC could enhance tumor targeting and produce synergistic therapeutic effects. For example, the PDC could utilize PPGlue's ability to downregulate P-gp, increasing intracellular drug retention and overcoming multidrug resistance. This strategy represents a powerful avenue for combining the unique properties of PPGlue with the established efficacy of existing drugs.

5.2.3 Exploring the potential applications of PPGlue in disease diagnosis

In addition to its therapeutic applications, PPGlue may also function as a prognostic biomarker for drug resistance in HCC. Its expression levels could indicate the tumor's sensitivity to lenvatinib and other drugs that are substrates of P-glycoprotein. Future research should explore the relationship between PPGlue expression and clinical outcomes in HCC patients. This could pave the way for the development of diagnostic tools that guide personalized treatment strategies.

References:

1. Kung, J. T., Colognori, D., and Lee, J. T. (2013) Long noncoding RNAs: past, present, and future. *Genetics* 193, 651-669

2. Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J.,

Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E., Trout, D., Varley, K. E., Gasper, C., The, E. P. C., Overall, c., Data production, l., Lead, a., Writing, g., management, N. p., Principal, i., Boise State, U., University of North Carolina at Chapel Hill Proteomics, g., Broad Institute, G., Cold Spring Harbor, U. o. G. C. f. G. R. B. R. S. I. U. o. L. G. I. o. S. g., Data coordination center at, U. C. S. C., Duke University, E. B. I. U. o. T. A. U. o. N. C.-C. H. g., Genome Institute of Singapore, g., and HudsonAlpha Institute, C. U. C. I. S. g. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74

3. Yang, X., Tschaplinski, T. J., Hurst, G. B., Jawdy, S., Abraham, P. E., Lankford, P. K., Adams, R. M., Shah, M. B., Hettich, R. L., Lindquist, E., Kalluri, U. C., Gunter, L. E., Pennacchio, C., and Tuskan, G. A. (2011) Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res* 21, 634-641
4. Sieber, P., Platzer, M., and Schuster, S. (2018) The Definition of Open Reading Frame Revisited. *Trends Genet* 34, 167-170
5. Fabre, B., Combier, J.-P., and Plaza, S. (2021) Recent advances in mass spectrometry–based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. *Current Opinion in Chemical Biology* 60, 122-130
6. Hook, V., Funkelstein, L., Lu, D., Bark, S., Wegrzyn, J., and Hwang, S.-R. (2008) Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. *Annu Rev Pharmacol Toxicol* 48, 393-423
7. 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., and van Heesch, S. (2022) Standardized annotation of translated open reading frames. *Nat. Biotechnol.* 40, 994-999
8. 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., and Bassani-Sternberg, M. (2020) Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* 11, 1293
9. Brunet, M. A., Lucier, J. F., Levesque, M., Leblanc, S., Jacques, J. F., Al-Saedi, H. R. H., Guilloy, N., Grenier, F., Avino, M., Fournier, I., Salzet, M., Ouangraoua, A., Scott, M. S., Boisvert, F. M., and Roucou, X. (2021) OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res* 49, D380-d388

10. Li, Y., Zhou, H., Chen, X., Zheng, Y., Kang, Q., Hao, D., Zhang, L., Song, T., Luo, H., Hao, Y., Chen, R., Zhang, P., and He, S. (2021) SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genomics Proteomics Bioinformatics* 19, 602-610

11. Neville, M. D. C., Kohze, R., Erady, C., Meena, N., Hayden, M., Cooper, D. N., Mort, M., and Prabakaran, S. (2021) A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. *Genome Res* 31, 327-336

12. Cardon, T., Hervé, F., Delcourt, V., Roucou, X., Salzet, M., Franck, J., and Fournier, I. (2020) Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins. *Analytical Chemistry* 92, 1122-1129

13. Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates, J. R., and Saghatelian, A. (2016) Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Analytical Chemistry* 88, 3967-3975

14. Yang, Y., Wang, H., Zhang, Y., Chen, L., Chen, G., Bao, Z., Yang, Y., Xie, Z., and Zhao, Q. (2023) An Optimized Proteomics Approach Reveals Novel Alternative Proteins in Mouse Liver Development. *Molecular & Cellular Proteomics* 22, 100480

15. Valdivia-Francia, F., and Sendoel, A. (2024) No country for old methods: New tools for studying microproteins. *Iscience* 27

16. Guttman, M., Russell, P., Ingolia, Nicholas T., Weissman, Jonathan S., and Lander, Eric S. (2013) Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* 154, 240-251

17. Ingolia, N. T. (2016) Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* 165, 22-33

18. 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., and Weissman, J. S. (2015) A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* 60, 816-827

19. Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nature methods* 13, 165-170

20. Johnstone, T. G., Bazzini, A. A., and Giraldez, A. J. (2016) Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO Journal* 35, 706-723

21. Lin, M. F., Jungreis, I., and Kellis, M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27, i275-282

22. Washietl, S., Findeiss, S., Müller, S. A., Kalkhof, S., von Bergen, M., Hofacker, I. L., Stadler, P. F., and Goldman, N. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna* 17, 578-594

23. Skarshewski, A., Stanton-Cook, M., Huber, T., Al Mansoori, S., Smith, R., Beatson, S. A., and Rothnagel, J. A. (2014) uPEPperoni: An online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics* 15, 36

24. Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., and Giraldez, A.

J. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* 33, 981-993

25. Badger, J. H., and Olsen, G. J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 16, 512-524

26. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 1034-1050

27. Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., and Shiu, S.-H. (2009) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* 26, 399-400

28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* 215, 403-410

29. Eddy, S. R. (1995) Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol* 3, 114-120

30. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* 47, D427-d432

31. Zhang, Y., Jia, C., Fullwood, M. J., and Kwoh, C. K. (2020) DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. *Briefings in Bioinformatics* 22, 2073-2084

32. Zhu, M., and Gribskov, M. (2019) MiPepid: MicroPeptide identification tool using machine learning. *BMC Bioinformatics* 20, 559

33. Ivanov, I. P., Firth, A. E., Michel, A. M., Atkins, J. F., and Baranov, P. V. (2011) Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Research* 39, 4220-4234

34. Chugunova, A., Navalayeu, T., Dontsova, O., and Sergiev, P. (2018) Mining for Small Translated ORFs. *J Proteome Res* 17, 1-11

35. Peeters, M. K. R., and Menschaert, G. (2020) The hunt for sORFs: A multidisciplinary strategy. *Experimental Cell Research* 391, 111923

36. Zhang, Q., Wu, E., Tang, Y., Zhang, L., Wang, J., Hao, Y., Zhang, B., Zhou, Y., Guo, X., Luo, J., Cai, T., Chen, R., and Yang, F. (2021) Deeply Mining a Universe of Peptides Encoded by Long Noncoding RNAs. *Molecular & Cellular Proteomics*, 100109

37. Wang, B., Wang, Z., Pan, N., Huang, J., and Wan, C. (2021) Improved Identification of Small Open Reading Frames Encoded Peptides by Top-Down Proteomic Approaches and De Novo Sequencing. *Int J Mol Sci* 22

38. Cassidy, L., Kaulich, P. T., and Tholey, A. (2019) Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes. *J Proteome Res* 18, 1725-1734

39. Yoshikawa, H., Larance, M., Harney, D. J., Sundaramoorthy, R., Ly, T.,

Owen-Hughes, T., and Lamond, A. I. (2018) Efficient analysis of mammalian polysomes in cells and tissues using Ribo Mega-SEC. *Elife* 7

40. Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., Budnik, B. A., Kellis, M., and Saghatelian, A. (2014) Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of Proteome Research* 13, 1757-1765

41. He, C., Jia, C., Zhang, Y., and Xu, P. (2018) Enrichment-Based Proteogenomics Identifies Microproteins, Missing Proteins, and Novel smORFs in Saccharomyces cerevisiae. *J Proteome Res* 17, 2335-2344

42. Dingess, K. A., van den Toorn, H. W. P., Mank, M., Stahl, B., and Heck, A. J. R. (2019) Toward an efficient workflow for the analysis of the human milk peptidome. *Anal Bioanal Chem* 411, 1351-1363

43. Harney, D. J., Hutchison, A. T., Su, Z., Hatchwell, L., Heilbronn, L. K., Hocking, S., James, D. E., and Larance, M. (2019) Small-protein Enrichment Assay Enables the Rapid, Unbiased Analysis of Over 100 Low Abundance Factors from Human Plasma. *Mol Cell Proteomics* 18, 1899-1915

44. Liu, Y., Xun, X.-H., Yi, J.-M., Xiang, Y., and Hua, J. (2017) Discovery of lung squamous carcinoma biomarkers by profiling the plasma peptide with LC/MS/MS. *Chinese Chemical Letters* 28, 1093-1098

45. Huesgen, P. F., Lange, P. F., Rogers, L. D., Solis, N., Eckhard, U., Kleifeld, O., Goulas, T., Gomis-Rüth, F. X., and Overall, C. M. (2015) LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nature methods* 12, 55-58

46. Kaulich, P. T., Cassidy, L., Bartel, J., Schmitz, R. A., and Tholey, A. (2021) Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides. *J Proteome Res* 20, 2895-2903

47. 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., and Saghatelian, A. (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat Chem Biol* 9, 59-64

48. 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., and Ho, L. (2020) Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nature Communications* 11, 1312

49. Bartel, J., Varadarajan, A. R., Sura, T., Ahrens, C. H., Maaß, S., and Becher, D. (2020) Optimized Proteomics Workflow for the Detection of Small Proteins. *J Proteome Res* 19, 4004-4018

50. Yang, M., Shang, X., Zhou, Y., Wang, C., Wei, G., Tang, J., Zhang, M., Liu, Y., Cao, J., and Zhang, Q. (2021) Full-Length Transcriptome Analysis of Plasmodium falciparum by Single-Molecule Long-Read Sequencing. *Frontiers in Cellular and Infection Microbiology* 11

51. Cassidy, L., Helbig, A. O., Kaulich, P. T., Weidenbach, K., Schmitz, R. A., and

Tholey, A. (2021) Multidimensional separation schemes enhance the identification and molecular characterization of low molecular weight proteomes and short open reading frame-encoded peptides in top-down proteomics. *J Proteomics* 230, 103988

52. Prabakaran, S., Hemberg, M., Chauhan, R., Winter, D., Tweedie-Cullen, R. Y., Dittrich, C., Hong, E., Gunawardena, J., Steen, H., Kreiman, G., and Steen, J. A. (2014) Quantitative profiling of peptides from RNAs classified as noncoding. *Nat Commun* 5, 5429

53. McDonald, W. H., and Yates, J. R., 3rd (2003) Shotgun proteomics: integrating technologies to answer biological questions. *Curr Opin Mol Ther* 5, 302-309

54. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012) 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* 11, O111.016717

55. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., and Coon, J. J. (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* 11, 1475-1488

56. D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., Budnik, B. A., Lykke-Andersen, J., Saghatelian, A., and Slavoff, S. A. (2017) A human microprotein that interacts with the mRNA decapping complex. *Nature chemical biology* 13, 174-180

57. Hemm, M. R., Weaver, J., and Storz, G. (2020) Escherichia coli Small Proteome. *EcoSal Plus* 9

58. Fesenko, I., Kirov, I., Kniazev, A., Khazigaleeva, R., Lazarev, V., Kharlampieva, D., Grafskaia, E., Zgoda, V., Butenko, I., Arapidi, G., Mamaeva, A., Ivanov, V., and Govorun, V. (2019) Distinct types of short open reading frames are translated in plant cells. *Genome Res* 29, 1464-1477

59. Pak, H., Michaux, J., Huber, F., Chong, C., Stevenson, B. J., Müller, M., Coukos, G., and Bassani-Sternberg, M. (2021) Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition, and MS/MS Prediction. *Mol Cell Proteomics* 20, 100080

60. Zhu, Y., Orre, L. M., Johansson, H. J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R. M. M., and Lehtiö, J. (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat Commun* 9, 903

61. Delcourt, V., Brunelle, M., Roy, A. V., Jacques, J. F., Salzet, M., Fournier, I., and Roucou, X. (2018) The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1. *Mol Cell Proteomics* 17, 2402-2411

62. Chen, C., Hou, J., Tanner, J. J., and Cheng, J. (2020) Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *International journal of molecular sciences* 21, 2873

63. Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., Fenyö, D., Zhang, B., and Mani, D. R. (2017) Methods, Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteomics* 16, 959-981

64. Dainese, P., Staudenmann, W., Quadroni, M., Korostensky, C., Gonnet, G., Kertesz, M., and James, P. (1997) Probing protein function using a combination of gene knockout and proteome analysis by mass spectrometry. *Electrophoresis* 18, 432-442

65. Wang, S., Tian, L., Liu, H., Li, X., Zhang, J., Chen, X., Jia, X., Zheng, X., Wu, S., Chen, Y., Yan, J., and Wu, L. (2020) Large-Scale Discovery of Non-conventional Peptides in Maize and Arabidopsis through an Integrated Peptidogenomic Pipeline. *Molecular Plant* 13, 1078-1093

66. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., and Macek, B. (2013) Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular & cellular proteomics : MCP* 12, 3420-3430

67. Fabre, B., Combier, J. P., and Plaza, S. (2021) Recent advances in mass spectrometry-based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. *Curr Opin Chem Biol* 60, 122-130

68. Qeli, E., and Ahrens, C. H. (2010) PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat Biotechnol* 28, 647-650

69. Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., Nikolayeva, O., Québatte, M., Patrignani, A., Dehio, C., Frey, J. E., Robinson, M. D., Wollscheid, B., and Ahrens, C. H. (2017) An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Res* 27, 2083-2095

70. Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., and Obermayer, B. (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome biology* 16, 179-179

71. Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., and Menschaert, G. (2013) Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* 14, 648-648

72. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 324, 218-223

73. Zhang, P., He, D., Xu, Y., Hou, J., Pan, B.-F., Wang, Y., Liu, T., Davis, C. M., Ehli, E. A., Tan, L., Zhou, F., Hu, J., Yu, Y., Chen, X., Nguyen, T. M., Rosen, J. M., Hawke, D. H., Ji, Z., and Chen, Y. (2017) Genome-wide identification and differential analysis of translational initiation. *Nature Communications* 8, 1749

74. Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J., Jackson, S. E., Wills, M. R., and Weissman, J. S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* 8, 1365-1379

75. Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., Stevanovic, S., Zimmer, R., and Dölken, L. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nature methods* 15, 363-366

76. Li, S., Cha, S. W., Heffner, K., Hizal, D. B., Bowen, M. A., Chaerkady, R., Cole, R.

N., Tejwani, V., Kaushik, P., Henry, M., Meleady, P., Sharfstein, S. T., Betenbaugh, M. J., Bafna, V., and Lewis, N. E. (2019) Proteogenomic Annotation of Chinese Hamsters Reveals Extensive Novel Translation Events and Endogenous Retroviral Elements. *Journal of proteome research* 18, 2433-2445

77. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020) Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol* 16, 458-468

78. Liang, Y., Zhu, W., Chen, S., Qian, J., and Li, L. (2021) Genome-Wide Identification and Characterization of Small Peptides in Maize. *Front Plant Sci* 12, 695439-695439

79. 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., and Weissman, J. S. (2020) Pervasive functional translation of noncanonical human open reading frames. *Science* 367, 1140-1146

80. Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res* 46, D497-d502

81. 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., and Schier, A. F. (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* 343, 1248636

82. Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., and Kageyama, Y. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* 9, 660-665

83. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., and Couso, J. P. (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS biology* 5, e106

84. Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., and Couso, J. P. (2013) Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science* 341, 1116-1120

85. Chng, S. C., Ho, L., Tian, J., and Reversade, B. (2013) ELABELA: a hormone essential for heart development signals via the apelin receptor. *Developmental cell* 27, 672-680

86. Hsu, P. Y., and Benfey, P. N. (2018) Small but mighty: functional peptides encoded by small ORFs in plants. *Proteomics* 18, 1700038

87. Makarewich, C. A., and Olson, E. N. (2017) Mining for Micropeptides. *Trends in Cell Biology* 27, 685-696

88. Huang, J.-Z., Chen, M., Chen, D., Gao, X.-C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.-R. (2017) A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Molecular Cell* 68, 171-184.e176

89. Li, X. L., Pongor, L., Tang, W., Das, S., Muys, B. R., Jones, M. F., Lazar, S. B., Dangelmaier, E. A., Hartford, C. C. R., Grammatikakis, I., Hao, Q., Sun, Q., Schetter, A., Martindale, J. L., Tang, B., Jenkins, L. M., Robles, A. I., Walker, R. L., Ambs, S., Chari, R., Shabalina, S. A., Gorospe, M., Hussain, S. P., Harris, C. C., Meltzer, P. S.,

Prasanth, K. V., Aladjem, M. I., Andresson, T., and Lal, A. (2020) A small protein encoded by a putative lncRNA regulates apoptosis and tumorigenicity in human colorectal cancer cells. *eLife* 9, e53734

90. Meng, N., Chen, M., Chen, D., Chen, X. H., Wang, J. Z., Zhu, S., He, Y. T., Zhang, X. L., Lu, R. X., and Yan, G. R. (2020) Small Protein Hidden in lncRNA LOC90024 Promotes "Cancerous" RNA Splicing and Tumorigenesis. *Adv Sci (Weinh)* 7, 1903233

91. Pang, Y., Liu, Z., Han, H., Wang, B., Li, W., Mao, C., and Liu, S. (2020) Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. *Journal of Hepatology* 73, 1155-1169

92. Wu, S., Zhang, L., Deng, J., Guo, B., Li, F., Wang, Y., Wu, R., Zhang, S., Lu, J., and Zhou, Y. (2020) A novel micropeptide encoded by Y-linked LINC00278 links cigarette smoking and AR signaling in male esophageal squamous cell carcinoma. *Cancer research* 80, 2790-2803

93. Boix, O., Martinez, M., Vidal, S., Giménez-Alejandre, M., Palenzuela, L., Lorenzo-Sanz, L., Quevedo, L., Moscoso, O., Ruiz-Orera, J., and Ximénez-Embún, P. (2022) pTINCR microprotein promotes epithelial differentiation and suppresses tumor growth through CDC42 SUMOylation and activation. *Nature communications* 13, 6840

94. Li, M., Liu, G., Jin, X., Guo, H., Setrerrahmane, S., Xu, X., Li, T., Lin, Y., and Xu, H. (2022) Micropeptide MIAC inhibits the tumor progression by interacting with AQP2 and inhibiting EREG/EGFR signaling in renal cell carcinoma. *Molecular cancer* 21, 181

95. Anderson, D. M., Anderson, K. M., Chang, C. L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., Kasaragod, P., Shelton, J. M., Liou, J., Bassel-Duby, R., and Olson, E. N. (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160, 595-606

96. Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., Reese, A. L., McAnally, J. R., Chen, X., Kavalali, E. T., Cannon, S. C., Houser, S. R., Bassel-Duby, R., and Olson, E. N. (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* 351, 271-275

97. Anderson, D. M., Makarewich, C. A., Anderson, K. M., Shelton, J. M., Bezprozvannaya, S., Bassel-Duby, R., and Olson, E. N. (2016) Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Science Signaling* 9, ra119-ra119

98. Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J., Kim, S. J., Mehta, H., Hevener, A. L., de Cabo, R., and Cohen, P. (2015) The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab* 21, 443-454

99. Makarewich, C. A., Baskin, K. K., Munir, A. Z., Bezprozvannaya, S., Sharma, G., Khemtong, C., Shah, A. M., McAnally, J. R., Malloy, C. R., and Szweda, L. I. (2018) MOXI is a mitochondrial micropeptide that enhances fatty acid β-oxidation. *Cell reports* 23, 3701-3709

100. Stein, C. S., Jadiya, P., Zhang, X., McLendon, J. M., Abouassaly, G. M.,

Witmer, N. H., Anderson, E. J., Elrod, J. W., and Boudreau, R. L. (2018) Mitoregulin: a lncRNA-encoded microprotein that supports mitochondrial supercomplexes and respiratory efficiency. *Cell reports* 23, 3710-3720. e3718

101. Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J., Kim, S.-J., Mehta, H., Hevener, A. L., and de Cabo, R. (2015) The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell metabolism* 21, 443-454

102. Niu, L., Lou, F., Sun, Y., Sun, L., Cai, X., Liu, Z., Zhou, H., Wang, H., Wang, Z., and Bai, J. (2020) A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Science advances* 6, eaaz2059

103. Lee, C. Q., Kerouanton, B., Chothani, S., Zhang, S., Chen, Y., Mantri, C. K., Hock, D. H., Lim, R., Nadkarni, R., and Huynh, V. T. (2021) Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity. *Nature communications* 12, 2130

104. Bhatta, A., Atianand, M., Jiang, Z., Crabtree, J., Blin, J., and Fitzgerald, K. A. (2020) A mitochondrial micropeptide is required for activation of the Nlrp3 inflammasome. *The Journal of Immunology* 204, 428-437

105. Yang, X., Bam, M., Becker, W., Nagarkatti, P. S., and Nagarkatti, M. (2020) Long noncoding RNA AW112010 promotes the differentiation of inflammatory T cells by suppressing IL-10 expression through histone demethylation. *The Journal of Immunology* 205, 987-993

106. Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K. I., Clohessy, J. G., and Pandolfi, P. P. (2017) mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* 541, 228-232

107. Pueyo, J. I., Magny, E. G., Sampson, C. J., Amin, U., Evans, I. R., Bishop, S. A., and Couso, J. P. (2016) Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PLoS biology* 14, e1002395

108. Yang, Y., Gao, X., Zhang, M., Yan, S., Sun, C., Xiao, F., Huang, N., Yang, X., Zhao, K., Zhou, H., Huang, S., Xie, B., and Zhang, N. (2018) Novel Role of FBXW7 Circular RNA in Repressing Glioma Tumorigenesis. *J Natl Cancer Inst* 110, 304-315

109. Xiang, X., Fu, Y., Zhao, K., Miao, R., Zhang, X., Ma, X., Liu, C., Zhang, N., and Qu, K. (2021) Cellular senescence in hepatocellular carcinoma induced by a long non-coding RNA-encoded peptide PINT87aa by blocking FOXM1-mediated PHB2. *Theranostics* 11, 4929

110. Sun, L., Wang, W., Han, C., Huang, W., Sun, Y., Fang, K., Zeng, Z., Yang, Q., Pan, Q., and Chen, T. (2021) The oncomicropeptide APPLE promotes hematopoietic malignancy by enhancing translation initiation. *Molecular cell* 81, 4493-4508. e4499

111. Ge, Q., Jia, D., Cen, D., Qi, Y., Shi, C., Li, J., Sang, L., Yang, L.-j., He, J., and Lin, A. (2021) Micropeptide ASAP encoded by LINC00467 promotes colorectal cancer progression by directly modulating ATP synthase activity. *The Journal of Clinical Investigation* 131

112. Wang, Y., Wu, S., Zhu, X., Zhang, L., Deng, J., Li, F., Guo, B., Zhang, S., Wu, R.,

and Zhang, Z. (2019) LncRNA-encoded polypeptide ASRPS inhibits triple-negative breast cancer angiogenesis. *Journal of Experimental Medicine* 217, e20190950

113.Zhang, Y., Jiang, J., Zhang, J., Shen, H., Wang, M., Guo, Z., Zang, X., Shi, H., Gao, J., and Cai, H. (2021) CircDIDO1 inhibits gastric cancer progression by encoding a novel DIDO1-529aa protein and regulating PRDX2 protein stability. *Molecular cancer* 20, 101

114.Xu, W., Deng, B., Lin, P., Liu, C., Li, B., Huang, Q., Zhou, H., Yang, J., and Qu, L. (2020) Ribosome profiling analysis identified a KRAS-interacting microprotein that represses oncogenic signaling in hepatocellular carcinoma cells. *Science China Life Sciences* 63, 529-542

115.Zhang, C., Zhou, B., Gu, F., Liu, H., Wu, H., Yao, F., Zheng, H., Fu, H., Chong, W., and Cai, S. (2022) Micropeptide PACMP inhibition elicits synthetic lethal effects by decreasing CtIP and poly (ADP-ribosyl) ation. *Molecular Cell* 82, 1297-1312. e1298

116.Zhang, M., Zhao, K., Xu, X., Yang, Y., Yan, S., Wei, P., Liu, H., Xu, J., Xiao, F., and Zhou, H. (2018) A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nature communications* 9, 4475

117.Zhu, S., Wang, J.-Z., Chen, D., He, Y.-T., Meng, N., Chen, M., Lu, R.-X., Chen, X.-H., Zhang, X.-L., and Yan, G.-R. (2020) An oncopeptide regulates m6A recognition by the m6A reader IGF2BP1 and tumorigenesis. *Nature communications* 11, 1685

118.Begum, S., Yiu, A., Stebbing, J., and Castellano, L. (2018) Novel tumour suppressive protein encoded by circular RNA, circ-SHPRH, in glioblastomas. *Oncogene* 37, 4055-4057

119.Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S. E., Hildenbrand, C., Rom, J., Aulmann, S., Sinn, H.-P., and Vandesompele, J. (2018) The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* 37, 4750-4768

120. Dufresne, S. S., Dumont, N. A., Boulanger-Piette, A., Fajardo, V. A., Gamu, D., Kake-Guena, S. A., David, R. O., Bouchard, P., Lavergne, É., Penninger, J. M., Pape, P. C., Tupling, A. R., and Frenette, J. (2016) Muscle RANK is a key regulator of Ca2+ storage, SERCA activity, and function of fast-twitch skeletal muscles. *Am J Physiol Cell Physiol* 310, C663-672

121. Cabri, W., Cantelmi, P., Corbisiero, D., Fantoni, T., Ferrazzano, L., Martelli, G., Mattellone, A., and Tolomelli, A. (2021) Therapeutic peptides targeting PPI in clinical development: Overview, mechanism of action and perspectives. *Frontiers in Molecular Biosciences* 8, 697586

122. Muttenthaler, M., King, G. F., Adams, D. J., and Alewood, P. F. (2021) Trends in peptide drug discovery. *Nature reviews Drug discovery* 20, 309-325

123. Fosgerau, K., and Hoffmann, T. (2015) Peptide therapeutics: current status and future directions. *Drug discovery today* 20, 122-128

124. Lau, J. L., and Dunn, M. K. (2018) Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & medicinal chemistry* 26, 2700-2707

125. Wu, P., Mo, Y., Peng, M., Tang, T., Zhong, Y., Deng, X., Xiong, F., Guo, C., Wu, X., Li, Y., Li, X., Li, G., Zeng, Z., and Xiong, W. (2020) Emerging role of

tumor-related functional peptides encoded by lncRNA and circRNA. *Molecular Cancer* 19, 22

126. Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A. G., Khan, O. M., Brewer, J. R., Skadow, M. H., Duizer, C., Harman, C. C. D., Chang, L., Bielecki, P., Solis, A. G., Steach, H. R., Slavoff, S., and Flavell, R. A. (2018) The translation of non-canonical open reading frames controls mucosal immunity. *Nature* 564, 434-438

127. 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., and Wang, H. (2020) A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. *Science Advances* 6, eaaz2059

128. Setrerrahmane, S., Li, M., Zoghbi, A., Lv, X., Zhang, S., Zhao, W., Lu, J., Craik, D. J., and Xu, H. (2022) Cancer-related micropeptides encoded by ncRNAs: Promising drug targets and prognostic biomarkers. *Cancer letters* 547, 215723

129. Cooper, B. M., Iegre, J., O'Donovan, D. H., Halvarsson, M. Ö., and Spring, D. R. (2021) Peptides as a platform for targeted therapeutics for cancer: Peptide–drug conjugates (PDCs). *Chemical society reviews* 50, 1480-1494

130. Li, Y., Nie, Y., Yang, X., Liu, Y., Deng, X., Hayashi, Y., Plummer, R., Li, Q., Luo, N., and Kasai, T. (2024) Integration of Kupffer cells into human iPSC-derived liver organoids for modeling liver dysfunction in sepsis. *Cell Reports* 43

131. Gómez-Salinero, J. M., Izzo, F., Lin, Y., Houghton, S., Itkin, T., Geng, F., Bram, Y., Adelson, R. P., Lu, T. M., and Inghirami, G. (2022) Specification of fetal liver endothelial progenitors to functional zonated adult sinusoids requires c-Maf induction. *Cell Stem Cell* 29, 593-609. e597

132. Rumgay, H., Arnold, M., Ferlay, J., Lesi, O., Cabasag, C. J., Vignat, J., Laversanne, M., McGlynn, K. A., and Soerjomataram, I. (2022) Global burden of primary liver cancer in 2020 and predictions to 2040. *Journal of hepatology* 77, 1598-1606

133. Yang, J. D., and Roberts, L. R. (2010) Hepatocellular carcinoma: a global view. *Nature reviews Gastroenterology & hepatology* 7, 448-458

134. Huang, A., Yang, X.-R., Chung, W.-Y., Dennison, A. R., and Zhou, J. (2020) Targeted therapy for hepatocellular carcinoma. *Signal Transduction and Targeted Therapy* 5, 146-158

135. Forner, A., Reig, M., and Bruix, J. (2018) Hepatocellular carcinoma. *The Lancet* 391, 1301-1314

136. Tang, W., Chen, Z., Zhang, W., Cheng, Y., Zhang, B., Wu, F., Wang, Q., Wang, S., Rong, D., Reiter, F. P., De Toni, E. N., and Wang, X. (2020) The mechanisms of sorafenib resistance in hepatocellular carcinoma: theoretical basis and therapeutic aspects. *Signal Transduction and Targeted Therapy* 5, 87

137. Zhang, Q., and Liu, L. (2024) Novel insights into small open reading frame-encoded micropeptides in hepatocellular carcinoma: A potential breakthrough. *Cancer Letters*, 216691

138. Cardon, T., Hervé, F., Delcourt, V., Roucou, X., Salzet, M., Franck, J., and Fournier, I. (2020) Optimized Sample Preparation Workflow for Improved

Identification of Ghost Proteins. *Anal. Chem.* 92, 1122-1129

139. Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.-P., Gendron, P., Courcelles, M., Hardy, M.-P., Côté, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., Haddad, E., Lemieux, S., Thibault, P., and Perreault, C. (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Science Translational Medicine* 10, eaau5516

140. Laumont, C. M., Daouda, T., Laverdure, J. P., Bonneil, É., Caron-Lizotte, O., Hardy, M. P., Granados, D. P., Durette, C., Lemieux, S., Thibault, P., and Perreault, C. (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun* 7, 10238

141. 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., and Yewdell, J. W. (2021) Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep* 34, 108815

142. Yin, X., Hu, J., and Xu, H. (2018) Distribution of micropeptide-coding sORFs in transcripts. *Chin. Chem. Lett.* 29, 1029-1032

143. 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., and Ho, L. (2020) Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat. Commun.* 11, 1312

144. Koh, M., Ahmad, I., Ko, Y., Zhang, Y., Martinez, T. F., Diedrich, J. K., Chu, Q., Moresco, J. J., Erb, M. A., Saghatelian, A., Schultz, P. G., and Bollong, M. J. (2021) A short ORF-encoded transcriptional regulator. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2021943118

145. Wang, H., Wang, Y., Yang, J., Zhao, Q., Tang, N., Chen, C., Li, H., Cheng, C., Xie, M., Yang, Y., and Xie, Z. (2021) Tissue- and stage-specific landscape of the mouse translatome. *Nucleic Acids Res.* 49, 6165-6180

146. Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., Lilley, K. S., Lundberg, E., Marcotte, E. M., Ralser, M., and Rappsilber, J. (2022) Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods*, https://doi.org/10.1038/s41592-41022-01454-x

147. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020) Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* 16, 458-468

148. Hu, D. C., Liu, Q., Xu, H. B., Cui, H. R., Yu, S. W., Yang, X. D., and Wang, K. (2005) A novel protein found in selenium-rich silkworm pupas. *Chin. Chem. Lett.* 16, 1347-1350

149. 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., and Saghatelian, A. (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat. Chem. Biol.* 9, 59-64

150. Fabre, B., Choteau, S. A., Duboé, C., Pichereaux, C., Montigny, A., Korona,

D., Deery, M. J., Camus, M., Brun, C., Burlet-Schiltz, O., Russell, S., Combier, J.-P., Lilley, K. S., and Plaza, S. (2022) In Depth Exploration of the Alternative Proteome of Drosophila melanogaster. *Front. Cell Dev. Biol.* 10

151. Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates, J. R., 3rd, and Saghatelian, A. (2016) Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* 88, 3967-3975

152. Cassidy, L., Kaulich, P. T., Maaß, S., Bartel, J., Becher, D., and Tholey, A. (2021) Bottom-up and top-down proteomic approaches for the identification, characterization, and quantification of the low molecular weight proteome with focus on short open reading frame-encoded peptides. *Proteomics* 21, e2100008

153. Zhang, Q., Wu, E., Tang, Y., Cai, T., Zhang, L., Wang, J., Hao, Y., Zhang, B., Zhou, Y., and Guo, X. (2021) Deeply Mining a Universe of Peptides Encoded by Long Noncoding RNAs. *Mol. Cell. Proteomics* 20, 100109

154. Buszczak, M., Signer, Robert A. J., and Morrison, Sean J. (2014) Cellular Differences in Protein Synthesis Regulate Tissue Homeostasis. *Cell* 159, 242-251

155. Li, N., Zhou, Y., Wang, J., Niu, L., Zhang, Q., Sun, L., Ding, X., Guo, X., Xie, Z., Zhu, N., Zhang, M., Chen, X., Cai, T., and Yang, F. (2020) Sequential Precipitation and Delipidation Enables Efficient Enrichment of Low-Molecular Weight Proteins and Peptides from Human Plasma. *J Proteome Res* 19, 3340-3351

156. Du, Y., Wu, D., and Guan, Y. (2016) Further investigation of a peptide extraction method with mesoporous silica using high-performance liquid chromatography coupled with tandem mass spectrometry. *J. Sep. Sci.* 39, 2156-2163

157. Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., and Krijgsveld, J. (2019) Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* 14, 68-85

158. Harney, D. J., Hutchison, A. T., Su, Z., Hatchwell, L., Heilbronn, L. K., Hocking, S., James, D. E., and Larance, M. (2019) Small-protein Enrichment Assay Enables the Rapid, Unbiased Analysis of Over 100 Low Abundance Factors from Human Plasma. *Mol. Cell. Proteomics* 18, 1899-1915

159. Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., Budnik, B. A., Kellis, M., and Saghatelian, A. (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* 13, 1757-1765

160. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17, 10-12

161. Joshi, N., and Fass, J. (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle. Accessed 14 Feb 2020.

162. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25

163. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21

164. Zhang, P. A.-O., He, D., Xu, Y., Hou, J., Pan, B. F., Wang, Y. A.-O., Liu, T. A.-O., Davis, C. M., Ehli, E. A.-O., Tan, L., Zhou, F., Hu, J., Yu, Y., Chen, X., Nguyen, T. M., Rosen, J. M., Hawke, D. A.-O., Ji, Z., and Chen, Y. (2017) Genome-wide identification and differential analysis of translational initiation. 8, 1749

165. Calviello, L., Hirsekorn, A., and Ohler, U. (2020) Quantification of translation uncovers the functions of the alternative transcriptome. *Nat Struct Mol Biol* 27, 717-725

166. Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018) De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res* 46, e61

167. Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* 5, e13328

168. Choudhary, S., Li, W., and A, D. S. (2020) Accurate detection of short and long active ORFs using Ribo-seq data. *Bioinformatics* 36, 2053-2059

169. Xu, Z., Hu, L., Shi, B., Geng, S., Xu, L., Wang, D., and Lu, Z. J. (2018) Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. *Nucleic Acids Res* 46, e109

170. Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H., and Dieterich, C. (2017) Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res* 45, 2960-2972

171. Ji, Z., Song, R., Regev, A., and Struhl, K. (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 4, e08890

172. Brunet, M. A., Brunelle, M., Lucier, J. F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J. D., Dufour, P., Jacques, J. F., Fournier, I., Ouangraoua, A., Scott, M. S., Boisvert, F. M., and Roucou, X. (2019) OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* 47, D403-D410

173. Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* 46, D497-D502

174. Guo, H., Yang, Y., Zhang, Q., Deng, J.-R., Yang, Y., Li, S., So, P.-K., Lam, T. C., Wong, M.-k., and Zhao, Q. (2022) Integrated Mass Spectrometry Reveals Celastrol As a Novel Catechol-O-methyltransferase Inhibitor. *ACS Chem. Biol.*, https://doi.org/10.1021/acschembio.1022c00011

175. Ma, C., Ren, Y., Yang, J., Ren, Z., Yang, H., and Liu, S. (2018) Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* 90, 10881-10888

176. Zeng, W.-F., Zhou, X.-X., Zhou, W.-J., Chi, H., Zhan, J., and He, S.-M. (2019) MS/MS spectrum prediction for modified peptides using pDeep2 trained by transfer learning. *Anal. Chem.* 91, 9724-9731

177. Samandi, S., Roy, A. V., Delcourt, V., Lucier, J.-F., Gagnon, J., Beaudoin, M. C., Vanderperre, B., Breton, M.-A., Motard, J., Jacques, J.-F., Brunelle, M.,

Gagnon-Arsenault, I., Fournier, I., Ouangraoua, A., Hunting, D. J., Cohen, A. A., Landry, C. R., Scott, M. S., and Roucou, X. (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* 6, e27860

178. Cassidy, L., Prasse, D., Linke, D., Schmitz, R. A., and Tholey, A. (2016) Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon Methanosarcina mazei. *J. Proteome Res.* 15, 3773-3783

179. Young, D. J., Stoddart, A., Nakitandwe, J., Chen, S.-C., Qian, Z., Downing, J. R., and Le Beau, M. M. (2014) Knockdown of Hnrnpa0, a del (5q) gene, alters myeloid cell fate in murine cells through regulation of AU-rich transcripts. *Haematologica* 99, 1032-1040

180. Gillentine, M. A., Wang, T., Hoekzema, K., Rosenfeld, J., Liu, P., Guo, H., Kim, C. N., De Vries, B., Vissers, L. E., and Nordenskjold, M. (2021) Rare deleterious mutations of HNRNP genes result in shared neurodevelopmental disorders. *Genome Med.* 13, 63

181. Kononikhin, A. S., Starodubtseva, N. L., Bugrova, A. E., Shirokova, V. A., Chagovets, V. V., Indeykina, M. I., Popov, I. A., Kostyukevich, Y. I., Vavina, O. V., Muminova, K. T., Khodzhaeva, Z. S., Kan, N. E., Frankevich, V. E., Nikolaev, E. N., and Sukhikh, G. T. (2016) An untargeted approach for the analysis of the urine peptidome of women with preeclampsia. *J. Proteomics* 149, 38-43

182. Liu, Y., Xun, X.-H., Yi, J.-M., Xiang, Y., and Hua, J. (2017) Discovery of lung squamous carcinoma biomarkers by profiling the plasma peptide with LC/MS/MS. *Chin. Chem. Lett.* 28, 1093-1098

183. Müller, S. A., Findeiß, S., Pernitzsch, S. R., Wissenbach, D. K., Stadler, P. F., Hofacker, I. L., von Bergen, M., and Kalkhof, S. (2013) Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. *J. Proteomics* 86, 27-42

184. Revil, T., Gaffney, D., Dias, C., Majewski, J., and Jerome-Majewska, L. A. (2010) Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* 11, 399

185. Macaulay, A., Scantland, S., and Robert, C. (2011) RNA Processing during early embryogenesis: managing storage, utilisation and destruction. *RNA Processing*, pp. 307-557, IntechOpen

186. Zhao, J., Lu, P., Wan, C., Huang, Y., Cui, M., Yang, X., Hu, Y., Zheng, Y., Dong, J., Wang, M., Zhang, S., Liu, Z., Bian, S., Wang, X., Wang, R., Ren, S., Wang, D., Yao, Z., Chang, G., Tang, F., and Zhao, X.-Y. (2021) Cell-fate transition and determination analysis of mouse male germ cells throughout development. *Nat. Commun.* 12, 6839

187. Zhang, H., Wang, Y., and Lu, J. (2019) Function and Evolution of Upstream ORFs in Eukaryotes. *Trends Biochem. Sci.* 44, 782-794

188. Gunišová, S., Beznosková, P., Mohammad, M. P., Vlčková, V., and Valášek, L. S. (2016) In-depth analysis of cis-determinants that either promote or inhibit

reinitiation on GCN4 mRNA after translation of its four short uORFs. *RNA* 22, 542-558

189. Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N., and Walter, P. (2016) Translation from the 5' untranslated region shapes the integrated stress response. *Science* 351, aad3867

190. Kwon, J., Jo, Y.-J., Namgoong, S., and Kim, N.-H. (2019) Functional roles of hnRNPA2/B1 regulated by METTL3 in mammalian embryonic development. *Sci. Rep.* 9, 8640

191. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 68, 394-424

192. Kulik, L., and El-Serag, H. B. (2019) Epidemiology and Management of Hepatocellular Carcinoma. *Gastroenterology* 156, 477-491.e471

193. Ghouri, Y. A., Mian, I., and Rowe, J. H. (2017) Review of hepatocellular carcinoma: Epidemiology, etiology, and carcinogenesis. *Journal of carcinogenesis* 16

194. Balogh, J., Victor III, D., Asham, E. H., Burroughs, S. G., Boktour, M., Saharia, A., Li, X., Ghobrial, R. M., and Monsour Jr, H. P. (2016) Hepatocellular carcinoma: a review. *Journal of hepatocellular carcinoma*, 41-53

195. Ayuso, C., Rimola, J., Vilana, R., Burrel, M., Darnell, A., García-Criado, Á., Bianchi, L., Belmonte, E., Caparroz, C., Barrufet, M., Bruix, J., and Brú, C. (2018) Diagnosis and staging of hepatocellular carcinoma (HCC): current guidelines. *European Journal of Radiology* 101, 72-81

196. Kudo, M., Finn, R. S., Qin, S., Han, K.-H., Ikeda, K., Piscaglia, F., Baron, A., Park, J.-W., Han, G., and Jassem, J. (2018) Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. *The Lancet* 391, 1163-1173

197. Pang, Y., Liu, Z., Han, H., Wang, B., Li, W., Mao, C., and Liu, S. (2020) Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. *J Hepatol* 73, 1155-1169

198. Zhang, H., Liao, Z., Wang, W., Liu, Y., Zhu, H., Liang, H., Zhang, B., and Chen, X. (2023) A micropeptide JunBP regulated by TGF-beta promotes hepatocellular carcinoma metastasis. *Oncogene* 42, 113-123

199. Cao, X., Khitun, A., Harold, C. M., Bryant, C. J., Zheng, S. J., Baserga, S. J., and Slavoff, S. A. (2022) Nascent alt-protein chemoproteomics reveals a pre-60S assembly checkpoint inhibitor. *Nat Chem Biol* 18, 643-651

200. Boix, O., Martinez, M., Vidal, S., Gimenez-Alejandre, M., Palenzuela, L., Lorenzo-Sanz, L., Quevedo, L., Moscoso, O., Ruiz-Orera, J., Ximenez-Embun, P., Ciriaco, N., Nuciforo, P., Stephan-Otto Attolini, C., Alba, M. M., Munoz, J., Tian, T. V., Varela, I., Vivancos, A., Ramon, Y. C. S., Munoz, P., Rivas, C., and Abad, M. (2022) pTINCR microprotein promotes epithelial differentiation and suppresses tumor growth through CDC42 SUMOylation and activation. *Nat Commun* 13, 6840

201. Chen, Y., Su, H., Zhao, J., Na, Z., Jiang, K., Bacchiocchi, A., Loh, K. H., Halaban, R., Wang, Z., Cao, X., and Slavoff, S. A. (2023) Unannotated microprotein EMBOW regulates the interactome and chromatin and mitotic functions of WDR5.

Cell Reports 42, 113145

202. Nichols, C., Do-Thi, V. A., and Peltier, D. C. Noncanonical microprotein regulation of immunity. *Mol. Ther.*

203. 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., and Yewdell, J. W. (2021) Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Reports* 34, 108815

204. Huang, J.-Z., Chen, M., Chen, D., Gao, X.-C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.-R. (2017) A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Molecular cell* 68, 171-184. e176

205. Kort, A., Durmus, S., Sparidans, R. W., Wagenaar, E., Beijnen, J. H., and Schinkel, A. H. (2015) Brain and Testis Accumulation of Regorafenib is Restricted by Breast Cancer Resistance Protein (BCRP/ABCG2) and P-glycoprotein (P-GP/ABCB1). *Pharmaceutical Research* 32, 2205-2216

206. Mok, E. H. K., Leung, C. O. N., Zhou, L., Lei, M. M. L., Leung, H. W., Tong, M., Wong, T. L., Lau, E. Y. T., Ng, I. O. L., Ding, J., Yun, J. P., Yu, J., Zhu, H. L., Lin, C. H., Lindholm, D., Leung, K. S., Cybulski, J. D., Baker, D. M., Ma, S., and Lee, T. K. W. (2022) Caspase-3–Induced Activation of SREBP2 Drives Drug Resistance via Promotion of Cholesterol Biosynthesis in Hepatocellular Carcinoma. *Cancer research* 82, 3102-3115

207. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., and Ralser, M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 17, 41-44

208. Yang, J. C.-H., Han, B., De La Mora Jiménez, E., Lee, J.-S., Koralewski, P., Karadurmus, N., Sugawara, S., Livi, L., Basappa, N. S., Quantin, X., Dudnik, J., Ortiz, D. M., Mekhail, T., Okpara, C. E., Dutcus, C., Zimmer, Z., Samkari, A., Bhagwati, N., and Csőszi, T. (2024) Pembrolizumab With or Without Lenvatinib for First-Line Metastatic NSCLC With Programmed Cell Death-Ligand 1 Tumor Proportion Score of at least 1% (LEAP-007): A Randomized, Double-Blind, Phase 3 Trial. *Journal of Thoracic Oncology* 19, 941-953

209. Romero, D. (2023) First-line pembrolizumab plus lenvatinib is effective in non-clear-cell RCC. *Nature Reviews Clinical Oncology* 20, 661-661

210. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* 71, 209-249

211.Bialecki, E. S., and Di Bisceglie, A. M. (2005) Diagnosis of hepatocellular carcinoma. *HPB (Oxford)* 7, 26-34

212. Wang, Y., Jiang, M., Zhu, J., Qu, J., Qin, K., Zhao, D., Wang, L., Dong, L., and Zhang, X. (2020) The safety and efficacy of lenvatinib combined with immune checkpoint inhibitors therapy for advanced hepatocellular carcinoma. *Biomed Pharmacother* 132, 110797

213. Hiraoka, A., Kumada, T., Kariyama, K., Takaguchi, K., Itobayashi, E., Shimada, N., Tajiri, K., Tsuji, K., Ishikawa, T., Ochi, H., Hirooka, M., Tsutsui, A.,

Shibata, H., Tada, T., Toyoda, H., Nouso, K., Joko, K., Hiasa, Y., and Michitaka, K. (2019) Therapeutic potential of lenvatinib for unresectable hepatocellular carcinoma in clinical practice: Multicenter analysis. *Hepatol Res* 49, 111-117

214. Couso, J.-P., and Patraquim, P. (2017) Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology* 18, 575-589

215. Guerra-Almeida, D., and Nunes-da-Fonseca, R. (2020) Small Open Reading Frames: How Important Are They for Molecular Evolution? *Frontiers in Genetics* 11

216. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., and Saghatelian, A. (2020) Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology* 16, 458-468

217. Wu, P., Mo, Y., Peng, M., Tang, T., Zhong, Y., Deng, X., Xiong, F., Guo, C., Wu, X., and Li, Y. (2020) Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Molecular cancer* 19, 1-14

218. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44, D733-745

219. Sachdeva, S., Joo, H., Tsai, J., Jasti, B., and Li, X. (2019) A Rational Approach for Creating Peptides Mimicking Antibody Binding. *Scientific Reports* 9, 997

220. Wang, L., Wang, N., Zhang, W., Cheng, X., Yan, Z., Shao, G., Wang, X., Wang, R., and Fu, C. (2022) Therapeutic peptides: current applications and future directions. *Signal transduction and targeted therapy* 7, 48

221. Gottesman, M. M., Fojo, T., and Bates, S. E. (2002) Multidrug resistance in cancer: role of ATP–dependent transporters. *Nature Reviews Cancer* 2, 48-58

222. Chambers, T. C., Pohl, J., Glass, D. B., and Kuo, J. F. (1994) Phosphorylation by protein kinase C and cyclic AMP-dependent protein kinase of synthetic peptides derived from the linker region of human P-glycoprotein. *Biochemical Journal* 299, 309-315

223. Xie, Y., Burcu, M., Linn, D. E., Qiu, Y., and Baer, M. R. (2010) Pim-1 Kinase Protects P-Glycoprotein from Degradation and Enables Its Glycosylation and Cell Surface Expression. *Molecular Pharmacology* 78, 310-318

224. Wojtal, K. A., de Vries, E., Hoekstra, D., and van Ijzendoorn, S. C. (2006) Efficient trafficking of MDR1/P-glycoprotein to apical canalicular plasma membranes in HepG2 cells requires PKA-RIIalpha anchoring and glucosylceramide. *Mol Biol Cell* 17, 3638-3650

225. Mattaloni, S. M., Kolobova, E., Favre, C., Marinelli, R. A., Goldenring, J. R., and Larocca, M. C. (2012) AKAP350 Is involved in the development of apical

"Canalicular" structures in hepatic cells HepG2. *Journal of Cellular Physiology* 227, 160-171

226. Hardy, S. P., Goodfellow, H. R., Valverde, M. A., Gill, D. R., Sepúlveda, V., and Higgins, C. F. (1995) Protein kinase C-mediated phosphorylation of the human multidrug resistance P-glycoprotein regulates cell volume-activated chloride channels. *The EMBO Journal* 14, 68-75

227. Katayama, K., Yamaguchi, M., Noguchi, K., and Sugimoto, Y. (2014) Protein phosphatase complex PP5/PPP2R3C dephosphorylates P-glycoprotein/ABCB1 and down-regulates the expression and function. *Cancer Letters* 345, 124-131

228. Kamnasaran, D., Chen, C.-P., Devriendt, K., Mehta, L., and Cox, D. W. (2005) Defining a holoprosencephaly locus on human chromosome 14q13 and characterization of potential candidate genes. *Genomics* 85, 608-621

229. Kono, Y., Maeda, K., Kuwahara, K., Yamamoto, H., Miyamoto, E., Yonezawa, K., Takagi, K., and Sakaguchi, N. (2002) Mcm3-binding ganp DNA-primase is associated with a novel phosphatase component g5pr. *Genes to Cells* 7, 821-834

230. Hinds Jr, T. D., and Sánchez, E. R. (2008) Protein phosphatase 5. *The international journal of biochemistry & cell biology* 40, 2358-2362

231. Sun, W., Wong, I. L., Law, H. K.-W., Su, X., Chan, T. C., Sun, G., Yang, X., Wang, X., Chan, T. H., and Wan, S. (2023) In vivo reversal of P-glycoprotein-mediated drug resistance in a breast cancer xenograft and in leukemia models using a novel, potent, and nontoxic epicatechin EC31. *International Journal of Molecular Sciences* 24, 4377

232. Yang, Y., Hu, N., Gao, X.-J., Li, T., Yan, Z.-X., Wang, P.-P., Wei, B., Li, S., Zhang, Z.-J., and Li, S.-L. (2021) Dextran sulfate sodium-induced colitis and ginseng intervention altered oral pharmacokinetics of cyclosporine A in rats. *Journal of Ethnopharmacology* 265, 113251

233. Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J.-P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z.-Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J. R., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A.-C., and Nesvizhskii, A. I. (2013) The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nature Methods* 10, 730-736

234. Minocha, M., Khurana, V., Qin, B., Pal, D., and Mitra, A. K. (2012) Enhanced brain accumulation of pazopanib by modulating P-gp and Bcrp1 mediated efflux with canertinib or erlotinib. *Int J Pharm* 436, 127-134

235. (2014) Center for Drug Evaluation and Research of the US Department of Health and Human Services, Food and Drug Administration., Clinical pharmacology and biopharmaceutics review(s)

236. Jin, H., Shi, Y., Lv, Y., Yuan, S., Ramirez, C. F. A., Lieftink, C., Wang, L., Wang, S., Wang, C., Dias, M. H., Jochems, F., Yang, Y., Bosma, A., Hijmans, E. M., de Groot, M. H. P., Vegna, S., Cui, D., Zhou, Y., Ling, J., Wang, H., Guo, Y., Zheng, X.,

Isima, N., Wu, H., Sun, C., Beijersbergen, R. L., Akkari, L., Zhou, W., Zhai, B., Qin, W., and Bernards, R. (2021) EGFR activation limits the response of liver cancer to lenvatinib. *Nature* 595, 730-734

237. Lu, Y., Shen, H., Huang, W., He, S., Chen, J., Zhang, D., Shen, Y., and Sun, Y. (2021) Genome-scale CRISPR-Cas9 knockout screening in hepatocellular carcinoma with lenvatinib resistance. *Cell Death Discovery* 7, 359

238. Hou, W., Bridgeman, B., Malnassy, G., Ding, X., Cotler, Scott J., Dhanarajan, A., and Qiu, W. (2022) Integrin subunit beta 8 contributes to lenvatinib resistance in HCC. *Hepatology Communications* 6, 1786-1802

239. Leung, C. O. N., Yang, Y., Leung, R. W. H., So, K. K. H., Guo, H. J., Lei, M. M. L., Muliawan, G. K., Gao, Y., Yu, Q. Q., Yun, J. P., Ma, S., Zhao, Q., and Lee, T. K. W. (2023) Broad-spectrum kinome profiling identifies CDK6 upregulation as a driver of lenvatinib resistance in hepatocellular carcinoma. *Nature Communications* 14, 6699

240. Shao, N., Yuan, L., Liu, L., Cong, Z., Wang, J., Wu, Y., and Liu, R. (2024) Reversing Anticancer Drug Resistance by Synergistic Combination of Chemotherapeutics and Membranolytic Antitumor β-Peptide Polymer. *J. Am. Chem. Soc.* 146, 11254-11265

241. Qin, L., Cui, Z., Wu, Y., Wang, H., Zhang, X., Guan, J., and Mao, S. (2023) Challenges and Strategies to Enhance the Systemic Absorption of Inhaled Peptides and Proteins. *Pharmaceutical Research* 40, 1037-1055

242. Liu, C., Shi, Q., Huang, X., Koo, S., Kong, N., and Tao, W. (2023) mRNA-based cancer therapeutics. *Nature Reviews Cancer* 23, 526-543

243. Islam, M. A., Xu, Y., Tao, W., Ubellacker, J. M., Lim, M., Aum, D., Lee, G. Y., Zhou, K., Zope, H., Yu, M., Cao, W., Oswald, J. T., Dinarvand, M., Mahmoudi, M., Langer, R., Kantoff, P. W., Farokhzad, O. C., Zetter, B. R., and Shi, J. (2018) Restoration of tumour-growth suppression in vivo via systemic nanoparticle-mediated delivery of PTEN mRNA. *Nat Biomed Eng* 2, 850-864

244. Kong, N., Tao, W., Ling, X., Wang, J., Xiao, Y., Shi, S., Ji, X., Shajii, A., Gan, S. T., Kim, N. Y., Duda, D. G., Xie, T., Farokhzad, O. C., and Shi, J. (2019) Synthetic mRNA nanoparticle-mediated restoration of p53 tumor suppressor sensitizes p53-deficient cancers to mTOR inhibition. *Sci Transl Med* 11