UNIVERSITY OF CALIFORNIA SAN DIEGO

Application of proteogenomic techniques to the discovery and characterization of peptides and small proteins

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biology

by

Raymond Heng-Fai Mak

Committee in charge:

Professor Alan Saghatelian, Chair
Professor Steven P. Briggs
Professor Pieter C. Dorrestein
Professor Tony Hunter
Professor James T. Kadonaga
Professor Andres E. Leschziner

2020

The dissertation of Raymond Heng-Fai Mak is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California San Diego

2020

DEDICATION

I would like to dedicate this dissertation to the memory of Igor Shevelev. Igor was my first teacher at the bench and instilled me a great sense of passion, scientific rigor, discipline, technical proficiency, and thick skin (especially while spending countless hours teaching me how to purify proteins in the cold room wearing only shorts and a t-shirt). A few years ago, Igor lost his life to lung cancer.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A	adenine (deoxyribonucleic acid base)
A or Ala	alanine (amino acid)
AGC	automatic gain control
BCA	bicinchoninic acid
BSA	bovine serum albumin
C	cytosine (deoxyribonucleic acid base)
C or Cys	cysteine (amino acid)
C8	octylsilane
C18	octadecylsilane
cDNA	complementary deoxyribonucleic acid
CDS	coding sequence
CHO-S	Chinese hamster ovary cell line with suspension phenotype
CID	collision-induced dissociation
CM	conditioned media
CSF	cerebrospinal fluid
D or Asp	aspartate (amino acid)
DDA	data-dependent acquisition
DMEM	Dulbecco's modified Eagle medium
DNA	deoxyribonucleic acid
E or Glu	glutamate (amino acid)
EDTA	ethylenediaminetetraacetic acid
ERLIC	electrostatic repulsion hydrophilic interaction chromatography
ESI	electrospray ionization
F or Phe	phenylalanine (amino acid)
FBS	fetal bovine serum
FDR	false discovery rate
FLAG	FLAG-tag (DYKDDDDK) epitope
G	guanine (deoxyribonucleic acid base)
G or Gly	glycine (amino acid)
GELFrEE	gel-eluted liquid fraction entrapment electrophoresis
H or His	histidine (amino acid)
HCD	high-energy collisional dissociation
HEK293T	human embryonic kidney 293 cell line that expresses mutant SV40 large T antigen
I or Ile	isoleucine (amino acid)
IP	immunoprecipitation
K or Lys	lysine (amino acid)
kDa	kilodalton
L or Leu	leucine (amino acid)

LC-MS/MS	liquid chromatography coupled to electrospray ionization and tandem mass spectrometry
M or Met	methionine (amino acid)
MOPS	3-(N-morpholino)propanesulfonic acid
mRNA	messenger ribonucleic acid
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MWCO	molecular weight cutoff
N or Asn	asparagine (amino acid)
NA	numerical aperture
NEM	N-ethylmaleimide
ORF	open reading frame
P or Pro	proline (amino acid)
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline
PES	polyethersulfone
ppm	parts per million
PTM	post-translational modification
PVDF	polyvinylidene difluoride
R or Arg	arginine (amino acid)
Ribo-Seq	ribosome profiling and RNA sequencing
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RP	reverse phase
S or Ser	serine (amino acid)
SDS	sodium dodecyl sulfate
SEC	size-exclusion chromatography
SEP	small open reading frame-encoded polypeptide
smORF	small open reading frame
SPE	solid-phase extraction
T	thymine (deoxyribonucleic acid base)
T or Thr	threonine (amino acid)
TBST	tris-buffered saline with Tween-20
TCEP	tris(2-carboxyethyl)phosphine hydrochloride
TEAB	tetraethylammonium tetrahydroborate
TEAF	triethylammonium formate
U	uracil (ribonucleic acid base)
V or Val	valine (amino acid)
Var	variant
W or Trp	tryptophan (amino acid)
Y or Tyr	tyrosine (amino acid)

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL FILES

Supplemental Table 1: UniProt proteins detected in mouse brain

Supplemental Table 2: Neuropeptides detected in mouse brain

Supplemental Table 3: Non-UniProt ORFs detected in mouse brain

Supplemental Table 4: UniProt proteins detected in human cerebrospinal fluid

Supplemental Table 5: Neuropeptides detected in human cerebrospinal fluid

Supplemental Table 6: Non-UniProt ORFs detected in human cerebrospinal fluid

ACKNOWLEDGMENTS

Firstly, I would like to thank Alan Saghatelian for serving as my doctoral advisor. Alan's generous support and invaluable guidance were instrumental in helping me to complete my doctoral studies. I admire Alan's scientific acumen and all-round niceness, a quality for which he is truly unrivaled. Alan believed in me and saw potential in me when very few people did, including myself. His dedication to training the next generation of scientists and scholars is unparalleled. I am fortunate to be counted among them now. Thank you for everything, Alan.

I would also like to thank Steve Briggs, Pieter Dorrestein, Jim Kadonaga, Tony Hunter, and Andres Leschziner for serving on my doctoral committee. Their guidance has been vital in helping me to complete my degree. I would also like to thank Sue Ackerman and Matt Daugherty for their service as members of my initial doctoral committee. Aaron Coleman has been an outstanding teaching mentor to me. I could not have asked for anyone better in advising me as an instructional assistant and guiding me as a first-time instructor at the college level.

I am fortunate to have been surrounded by past and present members of PBL-A at the Salk Institute. I would like to thank Qian Chu, Cindy Donaldson, Jiao Ma, and Thomas Martinez for training me when I first joined the lab and Joan Vaughan for her invaluable technical assistance. Thanks to Matt Kolar and Annie Rathore for sharing their experiences through graduate school with me and keeping me company in the lab on weekends. I would also like to thank Jolene Diedrich and Jim Moresco for their insightful discussion, knowledgeable expertise, support, and advice on my research projects.

As the age-old African proverb goes, "it takes a village to raise a child." I would like to thank all my former academic mentors for their guidance and support, especially Ruedi

Aebersold, Eric Bennett, Jan-Michael Peters, and Igor Stagljar. I would also like to thank my former scientific colleagues and collaborators that I have had over the years: Tim Clausen, Tobias Dietschy, Daniel Gerlich, Franz Herzog, Alex Leitner, Marilyn Leonard, Nambi Sundaramoorthy, and Malene Urbanus. Jens Lykke-Andersen, Leonie Ringrose, Patrick Savaiano, Lillian Salcedo, and Debbie Yelon have counseled me through difficult situations. I would also like to thank my former high school teachers for providing me with a solid formative education without which I could not have succeeded academically: Les Damude, David Danter, Gerald Girouard, Sigrid Hynscht, Agnes Kalapun (née Meszaros), Mike Milhausen, Reed Needles, and Riko Oka. I would like to thank the rest of "the village," and I apologize profusely if your name has been inadvertently omitted.

I would like to acknowledge the Biological Sciences Graduate Program at UC San Diego for supporting me both academically and financially throughout my graduate studies. I would also like to acknowledge to the Mary K. Chapman Foundation for supporting me with a generous graduate award. Lastly, and most importantly, thank you to my family and friends for their endless encouragement and unwavering support. I could not have done any of this without you.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Mak, Raymond; Vaughan, Joan; Shokhirev, Max; Diedrich, Jolene; Saghatelian, Alan. "Proteogenomic discovery of open reading frames encoding peptides and small proteins in mouse brain". The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Mak, Raymond; Vaughan, Joan; Diedrich, Jolene; Saghatelian, Alan. "Biochemical characterization of C4ORF48 and GM1673 neuropeptides". The dissertation author was the primary investigator and author of this material.

VITA

2008	Honours Bachelor of Science, University of Toronto
2011	Master of Science, Eidgenössische Technische Hochschule Zürich
2011-2012	Research Assistant, Eidgenössische Technische Hochschule Zürich
2012-2014	Graduate Researcher, Institute of Molecular Pathology Vienna
2014-2015	Staff Research Associate, University of California San Diego
2016-2019	Graduate Instructional Assistant, University of California San Diego
2020	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Dietschy, T., Shevelev, I., Peña-Diaz, J., Hühn, D., Kuenzle, S., **Mak, R.**, Miah, M.F., Hess, D., Fey, M., Hottiger, M.O., Janscak, P., and Stagljar, I. p300-mediated acetylation of the Rothmund-Thomson-syndrome gene product RECQL4 regulates its subcellular localization. *Journal of Cell Science* **122**, 1258-1267 (2009).

Herzog, F., Kahraman, A., Böhringer, D., **Mak, R.**, Bracher, A., Walzthöni, T., Leitner, A., Beck, M., Hartl, F.-U., Ban, N., Mamström, L., and Aebersold, R. Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* **337**, 1348-1352 (2012).

Higgins, R., Gendron, J.M., Rising, L., **Mak, R.**, Webb, K., Kaiser, S.E., Zuzow, N., Riviere, P., Yang, B., Fenech, E., Tang, X., Lindsay, S.A., Christianson, J.C., Hampton, R.Y., Wasserman, S.A., and Bennett, E.J. The unfolded protein response triggers site-specific regulatory ubiquitylation of 40S ribosomal proteins. *Molecular Cell* **59**: 35-49 (2015).

Neal, S., **Mak, R.**, Bennett, E.J., and Hampton, R.Y. A Cdc48 "retrochaperone" function is required for the solubility of retrotranslocated, integral membrane endoplasmic reticulum-associated degradation (ERAD-M) substrates. *Journal of Biological Chemistry* **292**: 3112-3128 (2017).

Sundaramoorthy, E., Leonard, M., **Mak, R.**, Liao, J., Fulzele, A., and Bennett, E.J. ZNF598 and RACK1 regulate mammalian ribosome-associated quality control function by mediating regulatory 40S ribosomal ubiquitylation. *Molecular Cell* **65**: 751-760 (2017).

Reinke, A.W., **Mak, R.**, Troemel, E.R., and Bennett, E.J. In vivo mapping of tissue- and subcellular-specific proteomes in *Caenorhabditis elegans*. *Science Advances* **3**: e1602426 (2017).

Markmiller S., Soltanieh S., Server K.L., **Mak R.**, Jin W., Fang M.Y., Luo E.C., Krach F., Yang D., Sen A., Fulzele A., Wozniak J.M., Gonzalez D.J., Kankel M.W., Gao F.B., Bennett E.J., Lécuyer E., and Yeo G.W. Context-dependent and disease-specific diversity in protein interactions within stress granules. *Cell* **172**: 590-604 (2018).

AWARDS

2018 Excellence in Teaching Award, Division of Biological Sciences, University of California San Diego

2018 Mary K. Chapman Foundation Award, Salk Institute for Biological Studies

2019 Barbara J. and Paul D. Saltman Excellent Teaching Award, Graduate Student, University of California San Diego

FIELDS OF STUDY

Major Field: Biology

Studies in Biochemistry
Professors Ruedi Aebersold, Eric Bennett, Jan-Michael Peters, Alan Saghatelian, and Igor Stagljar

ABSTRACT OF THE DISSERTATION

Application of proteogenomic techniques to the discovery and characterization of peptides and small proteins

by

Raymond Heng-Fai Mak

Doctor of Philosophy in Biology

University of California San Diego, 2020

Professor Alan Saghatelian, Chair

Peptides are polymers of amino acids that constitute one of the major classes of molecules in biological organisms. Peptides and small proteins function in diverse biological processes. They frequently act to convey molecular signals by binding to cell-surface receptors and regulating intracellular signaling pathways. Peptide hormones, neurotransmitters, and neuropeptides are examples of these molecules that serve important signaling functions in organisms with central nervous systems. In the field of peptide biology, several outstanding questions remain. These questions include: How many biologically active peptides exist? How

are these biologically active peptides produced? What biological functions do these peptides serve? How are the functions of these biologically active peptides regulated? Which biological pathways do these peptides regulate? The answers to these questions will not only reveal the diversity of peptide and small protein effectors in biological systems, but also a deeper understanding of how these biological processes are regulated.

Recent proteogenomic studies that combine next-generation sequencing and proteomics have revealed the existence of hundreds of peptides and small proteins, also called microproteins or small open reading frame-encoded polypeptides, in *Escherichia coli*, *Saccharomyces cerevisiae*, *Mus musculus*, and *Homo sapiens*. As proteogenomic techniques have been successful at identifying and detecting small open reading frames (smORFs) and microproteins, we wondered if these techniques could also be applied to other classes of peptides and small proteins that have been historically challenging to detect: peptide hormones, neurotransmitters, and neuropeptides.

Firstly, we developed an integrated proteogenomics strategy that was optimized to detect peptides and proteins. We applied this strategy to mouse brain tissue and were able to identify known peptide hormones, neurotransmitters, and neuropeptides. We also identified microproteins from unannotated smORFs that might be candidates with similar biological function. We then applied this proteogenomic strategy to extracellular fluids and detected secreted microproteins that are encoded by unannotated smORFs. Finally, we characterized the biochemical structure of the human C4ORF48 neuropeptide and its mouse ortholog Gm1673. Taken together, our findings increased the diversity of the genomes and proteomes of both human and mouse. Our approach reported here can be used more generally to discover and characterize microproteins in other organisms.

Chapter 1

Strategies to detect peptides and small proteins in biological samples

1.1 Abstract

Peptides are short polymers of amino acids that constitute one of the major classes of molecules in biological organisms. Peptides participate in the regulation of nearly all biological processes, including pathogen defense, immune response, growth and development, metabolism, homeostasis, and behavior. The recent development of proteogenomic techniques has revealed the existence of small open reading frames that are translated into microproteins, which have confirmed that the peptidome and proteome are more diverse than previously thought. In this chapter, I will first give an overview of peptide biology and list the outstanding questions in the field. I will then describe some challenges in detecting peptides and small proteins and propose strategies to overcome these challenges.

1.2 Overview of peptide biology

As one of the four major categories of molecules in biological systems, peptides and proteins play essential roles in the function and regulation of biological cells. Peptides and proteins are polymers of amino acids (polypeptide), joined by a covalent bond between the carbonyl carbon of one amino acid and the nitrogen atom on another amino acid[1]. A water molecule is lost in the process to generate an amide or peptide bond linkage. While pools of free amino acids do exist inside cells[2], amino acids are also found in polymeric forms that range in several orders of magnitude in length: dipeptides with two amino acids to Titin, the longest

known protein with 38,138 amino acids in humans[3,4]. The distinction between peptide and protein is related to the number of repeating units of the polymer. Amino acid repeating units are commonly referred to as residues in a polypeptide. According to the definition in the "Gold Book" from the International Union of Pure and Applied Chemists, an oligopeptide is a polymer of between 3 and 10 amino acids, a peptide is a polymer of greater than 10 amino acids, and proteins are polymers of amino acids that have a molecular weight of greater than 10,000[1]. This limit would correspond to approximately 90 amino acids, although this limit is not precise.

The biosynthesis of peptides and proteins is an enzymatically catalyzed process. In eukaryotic, prokaryotic, and archaeal cells, the ribosome catalyzes the formation of polypeptide chains from a messenger ribonucleic acid (mRNA) template[5–8]. In some bacteria and fungi, modular enzyme complexes also produce nonribosomal peptides and polyketides in an mRNA-independent manner[9,10]. Functional peptides can also be formed from the enzyme-catalyzed hydrolysis of longer polypeptide chains by proteases[11,12]. Peptides can be chemically synthesized in the laboratory on a solid support, a technique first pioneered by Bruce Merrifield[13]. Synthetic peptides can then be joined together to generate small proteins using a process called native chemical ligation[14].

Peptides and small proteins function in diverse biological processes[15]. They frequently act to convey molecular signals by binding to cell-surface receptors and regulating intracellular signaling pathways. Peptide hormones, neurotransmitters, and neuropeptides are examples of these molecules that serve essential signaling functions in organisms with central nervous systems. Other pathways that are regulated by peptides include cellular proliferation[16], cellular identity maintenance in plants[17], exoskeleton shedding in insects[18], the adaptive immune response in vertebrates[19], bactericidal defenses[20], and other modes of pathogen defense[21].

In the field of peptide biology, several outstanding questions remain (Figure 1). These questions include: How many biologically active peptides exist? How are these biologically active peptides produced? What biological functions do these peptides serve? How are the functions of these biologically active peptides regulated? Which biological pathways do these peptides regulate? The answers to these questions will not only reveal the diversity of peptide and small protein effectors in biological systems, but also a deeper understanding of how these biological processes are regulated.

For the remainder of this chapter, I will be focusing on peptides and small proteins (less than 150 amino acids in length) in vertebrate species. The functions of peptides and small proteins in microbes, insects, arthropods, and plants have been extensively reviewed elsewhere. The biological functions of larger proteins have been extensively described and reviewed in biochemistry textbooks[22–24].

Outstanding questions in the field of peptide biology.

- **How many biologically active peptides exist?**
- **How are these biologically active peptides produced?**
- **What biological functions do these peptides serve?**
- **How are the functions of these biologically active peptides regulated?**
- **Which biological pathways do these peptides regulate?**

Figure 1.1: Outstanding questions in the field of peptide biology.

1.3 Microproteins

The discovery and characterization of the Tarsal-less or polished rice (Tal/Pri) gene revealed an emerging class of protein-coding peptides and small proteins derived from small

open reading frames (smORFs)[25,26]. The actual number of genes that encode peptides and small proteins remains unexplored because many algorithms used to predict protein-coding genes used a lower length cutoff of 300-500 basepairs[27]. As a result, many polypeptides shorter than 100-150 amino acids are likely unannotated in the repertoire of protein-coding genes[28,29]. Indeed, a re-analysis of the mouse transcriptome revealed the existence of thousands of potential protein-coding sequences below this length cutoff[30]. Some smORFs act as upstream ORFs (uORFs) that both regulate the expression in cis and interact with a downstream ORF in trans encoded on a bicistronic mRNA[31]. This finding further supports the existence of smORFs that have functional and complex roles in the regulation of gene expression and proteome composition.

Recent proteogenomic studies that combine next-generation sequencing and proteomics have revealed the existence of hundreds of peptides and small proteins, also called microproteins or small open reading frame (smORF)-encoded polypeptides (SEPs), in human cell lines and tissues[32–34]. As more smORFs and microproteins have been found, the number with defined biological roles has grown as well. Several smORFs that encode peptides have been identified with roles in muscle biology, including the peptide minion, which is necessary for the proper fusion of muscle cells into multinucleated fibers[35]. CYREN is another newly characterized microprotein that regulates DNA repair pathway choice during the cell cycle by inhibiting non-homologous end-joining repair to favor the higher fidelity homology-directed repair[36]. Lastly, PIGBOS, an outer mitochondrial transmembrane microprotein, interacts with the endoplasmic reticulum to regulate the unfolded protein response[37].

1.4 Application of proteogenomics to accelerate the discovery of peptide hormones, neurotransmitters, and neuropeptides

As proteogenomic techniques have been successful at identifying and detecting smORFs and microproteins, we wondered if these techniques could also be applied to other classes of peptides and small proteins: peptide hormones, neurotransmitters, and neuropeptides. Historically, these peptides and small proteins were identified from large amounts of starting material and extensive fractionation of a protein extract. The goal was to identify a single factor from this extract that was responsible for a biochemical activity in an assay. For example, insulin was purified from kilograms of dog, ox, and calf pancreatic glands using several extraction and precipitation steps[38]. Thyrotropin-releasing hormone (TSH) was purified from 55 kg of sheep hypothalamus fragments with numerous sequential extraction and fractionation steps[39]. We reasoned that advances in genomics and proteomics technologies would allow for the identification of peptides and small proteins using much less input material. Combining both genomics and proteomics approaches would also allow for improved mapping of identified peptides and small proteins to the encoding genomics sequence. I will now outline the principles, challenges, and strategies used to identify peptides and small proteins.

1.5 Methods to detect peptides and small proteins

The sequence of polypeptides was historically determined by chemical derivatization and sequential release of individual amino acids. Two methods that were used successfully to sequence proteins were dinitrophenyl derivatization[40] and Edman degradation[41], although many other methods exist[42]. Once identified, the polypeptide sequence could then be matched to a complementary deoxyribonucleic acid (cDNA) clone to identify the encoding gene. Although Edman degradation can now be used in a high-throughput manner when combined with

fluorescent dyes and total internal reflection microscopy[43], mass spectrometers have now become commonplace in determining protein sequence.

High-resolution mass spectrometers are routinely used to identify thousands of proteins in biological samples. In a typical "discovery" proteomics experiment, where the precise identities of the constitutive peptides and proteins are unknown, the sample is first digested into smaller peptides using trypsin. These digested peptides are ionized, and two mass spectra of the unfragmented and fragmented peptide ions are acquired in tandem by a mass spectrometer. Trypsin is a serine hydrolase that preferentially cleaves amide linkages on the carboxy-terminal side of arginine or lysine residues in a polypeptide. Arginine and lysine residues occur at a frequency that generates peptides of optimal length (10-20 amino acids) for detection by mass spectrometers following trypsin digestion. Ideally, the polypeptide chains are fragmented between the amide linkages, creating a ladder series of sub-fragments that cover the length of the digested peptide. The identities of the proteins present are deduced from the ensemble of tryptic peptides that were identified.

Using specialized computer algorithms, the peptide sequence can be identified from the mass spectra. Two main approaches are currently used: (1) de novo peptide sequencing and (2) database searching. Both approaches determine the peptide sequence based on the fragmentation pattern of the peptide in the mass spectrometer. In de novo sequencing, the mass difference between each sub-fragment and the known mass of each amino acid residue are used to deduce the sequence of the peptide. De novo mass spectra identification can also identify PTMs and isoforms. De novo identification remains a specialized technique but is rapidly becoming more widespread in the field.

A database search strategy to identify proteins in a proteomics experiment is used routinely. In this technique, the acquired mass spectra are searched for matching spectra in a database compiled from the in silico digestion of all possible protein sequences that are likely to be present in the sample. Several computer algorithms have been developed for database searches of tandem mass spectrometry data[44–46]. UniProt KnowledgeBase[47] or International Protein Index[48] databases are commonly used. To reduce the likelihood that random matches occur between the acquired mass spectra to mass spectra in the search database, decoy sequences are often appended to the database. Reverse decoy sequences are typically used, although random decoy sequences can also be used. This database search strategy is referred to as "target-decoy." The spectrum matches are scored and compared with the scores to the decoy spectrum matches. A subset of all spectrum matches is typically reported, with a certain number of decoy matches, which defines the false discovery rate (FDR) of the search results. A commonly accepted FDR of 1-5% has been reported for most large-scale proteomics experiments. This workflow has two main shortcomings that will be addressed in the subsequent sections. Firstly, larger proteins are likely to generate more tryptic peptides that can lead to the undersampling of tryptic peptides generated from peptides and smaller proteins. Secondly, the database used to identify peptides must contain a mass spectrum against which the acquired mass spectrum can be matched.

1.6 Challenges for detection of peptides and small proteins

The proteome complexity of biological tissues can result in the undersampling of peptides and small proteins, thereby going undetected in a mass spectrometer[49]. Larger proteins contain, on average, more lysine and arginine residues than do smaller proteins and therefore

generate more tryptic peptides. The higher numbers of tryptic peptides introduces a detection bias towards more abundant species in the sample[50], which would mask the identification of smaller peptides and proteins. Peptides are smaller proteins that are also more likely to be found at lower abundance levels than larger proteins, resulting in further undersampling. Further adding complexity to the repertoire of peptides and small proteins in a biological sample are post-translational modifications (PTMs), which add mass to peptides and proteins. PTM modifications are often sub-stoichiometric, thereby increasing the number of peptide and protein isoforms in the sample. PTMs that add mass in a fixed manner (acetylation, phosphorylation) can be predicted and identified using search engines. PTMs that add mass in a variable manner (glycosylation, lipidation) require more specialized search engines and are not routinely performed. Removing the PTM prior to mass spectrometric analysis increases the likelihood of detection. For example, disulfides are removed using reducing agents and certain types of glycosylation can be removed enzymatically or chemically. The use of trypsin, while advantageous for creating peptide fragments that are amenable to detection in a mass spectrometer, precludes the identification of protein isoforms, regulatory protease sites, and binding interaction.

1.7 Strategies to overcome challenges in detecting peptides and small proteins

I will now discuss three strategies to overcome challenges in detecting peptides and small proteins: (1) enrichment, (2) fractionation, (3) bioinformatics.

1.7.1 Enrichment strategies

Peptides and small proteins can be enriched by exploiting biochemical differences between these and larger proteins. One way to enrich for peptides is to exploit a difference in solubility using acetonitrile. Peptides, being smaller, are usually more soluble in an organic solvent like acetonitrile. Using a high concentration of acetonitrile leads to less soluble species (like proteins) precipitating and the more soluble species (like peptides) staying in solution. Another biochemical property that can be exploited is the molecular weight. Molecular weight cutoff filters (MWCO) made of cross-linked fibers of different pore sizes allow some analytes with a defined molecular weight to pass through, while others are retained. These filters can vary in terms of their effectiveness, and the pore size diameters are in a specific range; therefore, the cutoff limit is not absolute. Some of the peptides and proteins might adsorb to the inert filter material and thus depleting certain species regardless of the filter pore size. Hydrophobicity can also be used, by trapping less hydrophobic analytes like peptides on a solid support and then selectively eluting them in a solvent. This is the basis for solid-phase exchange (SPE), where the analytes are adsorbed and then eluted from a solid support. In a previous study, SPE outperformed MWCO filters to enrich and identify microproteins in tissues and cell lines[33]. Peptides can also be immunoprecipitated by using antibodies that are immobilized on a solid support.

Table 1.1 Biochemical methods used to enrich peptides and small proteins.

Biochemical Property	Biochemical Technique
Solubility	Precipitation using organic solvents or acid
Molecular weight	MWCO filters/dialysis
Hydrophobicity	Solid-phase exchange
Shape	Immunoprecipitation

1.7.2 Fractionation strategies

Another strategy to reduce sample complexity is fractionation. Prior studies that have focused on peptides and small proteins have used sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), isoelectric focusing[51], ion-exchange chromatography[52], electrostatic repulsion hydrophilic interaction chromatography (ERLIC), and reverse-phase chromatography to fractionate polypeptides before mass spectrometry analysis.

Complex mixtures of peptides, much like proteins, can be fractionated based on the following biochemical properties: molecular weight, conformation, density, hydrophobicity, net charge. The fractions containing peptides can be separated away from those containing proteins, thereby reducing the sample complexity and presence of interfering proteins. Prior studies that have focused on peptides and small proteins have used sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), isoelectric focusing[51], ion-exchange chromatography[52], electrostatic repulsion hydrophilic interaction chromatography (ERLIC)[33], and reverse-phase chromatography to fractionate polypeptides prior to mass spectrometry analysis.

Table 1.2 Biochemical methods used to fractionate peptides and small proteins

Biochemical Property	Biochemical Technique
Molecular Weight or Size	SDS-PAGE, GELFrEE
Conformation or Structure	PAGE, SEC
Density	Differential centrifugation
Hydrophobicity	ERLIC, Reverse-phase chromatography
Net Charge or Isoelectric Point	Ion exchange chromatography, Isoelectric focusing

1.7.3 Bioinformatic strategies

Using a target-decoy strategy to identify proteins implies that only the protein sequences contained within that database will be queried. For this reason, many microproteins have escaped

detection in typical proteomics experiments because the protein-coding sequences are unlikely to be annotated in the UniProtKB or IPI databases. Ribosome profiling has allowed for the annotation of protein-coding smORFs. Re-analysis of previously acquired mass spectrometry data with updated databases containing smORFs has revealed the existence of many of these microproteins. A proteogenomics approach can also be used where a protein database is generated from the in silico translation of a DNA or RNA dataset. The translation is done in all possible reading frames, thereby generating all theoretically possible protein sequences that can be produced. Tryptic peptides that are identified in the mass spectrometry data are then mapped back onto RNA transcripts, from which the genomic coordinates of the ORF can be annotated. As the databases generated in this approach are larger than the curated UniProtKB and IPI databases, this approach requires more computational resources, and it can also result in a higher probability of false peptide and protein identifications.

1.8 Overview and significance of this dissertation

In this dissertation, I describe the development of an integrated workflow for the identification of peptides and small proteins in tissues and extracellular fluids. This workflow combines several steps that are optimized for the detection of peptides and small proteins: extraction, fractionation, and proteogenomics. This workflow was applied to mouse brain tissue (Chapter 2) and the conditioned media of human cell lines and human cerebrospinal fluid (Chapter 3). Our results reveal the existence of microproteins and smORFs in mouse and human that have not yet been annotated. Some of these microproteins might act as peptide hormones, neurotransmitters, and neuropeptides. In Chapter 4, I then describe the characterization of the

biochemical structure of one such identified microprotein, human C4ORF48 and its mouse ortholog GM1673.

The results presented in this dissertation suggest that the genome and proteome of mouse and human are both more diverse than previously thought. The smORFs and microproteins here may represent peptide hormones, neurotransmitters, and neuropeptides that have not been previously detected. These peptides and small proteins may contribute essential roles to the function and regulation of vertebrate organisms and new avenues of research to pursue as biomarkers or therapeutic agents of disease. Further study at the molecular and physiological levels will be required to understand their functions.

1.9 References

1. Moss, G. P., Smith, P. A. S. & Tavernier, D. Glossary of class names of organic compounds and reactive intermediates based on structure (IUPAC recommendations 1995). *Pure Appl. Chem.* **67,** 1307–1375 (1995).

2. Monod, J., Pappenheimer, A. M. & Cohen-Bazire, G. La cinétique de la biosynthèse de la β-galactosidase chez E. coli considérée comme fonction de la croissance. *Biochim. Biophys. Acta* **9,** 648–660 (1952).

3. Labeit, S. & Kolmerer, B. Titins: Giant proteins in charge of muscle ultrastructure and elasticity. *Science* **270,** 293–296 (1995).

4. Bang, M. L., Centner, T., Fornoff, F., Geach, A. J., Gotthardt, M., McNabb, M., Witt, C. C., Labeit, D., Gregorio, C. C., Granzier, H. & Labeit, S. The complete gene sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ. Res.* **89,** 1065–1072 (2001).

5. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289,** 905–920 (2000).

6. Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* **407,** 340–348 (2000).

7. Wimberly, B. T., Brodersen, D. E., Jr, W. M. C., Morgan-warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. Structure of the 30S ribosomal subunit. 1–

13 (2008).

8. Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I. & Yonath, A. Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Angstroms Resolution. **102,** 615–623 (2000).

9. Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: Combinations, permutations, and mutations. *Science* **282,** 63–68 (1998).

10. Finking, R. & Marahiel, M. A. Biosynthesis of Nonribosomal Peptides. *Annu. Rev. Microbiol.* **58,** 453–488 (2004).

11. López-Otín, C. & Bond, J. S. Proteases: Multifunctional enzymes in life and disease. *J. Biol. Chem.* **283,** 30433–30437 (2008).

12. Klein, T., Eckhard, U., Dufour, A., Solis, N. & Overall, C. M. Proteolytic Cleavage - Mechanisms, Function, and 'omic' Approaches for a Near-Ubiquitous Posttranslational Modification. *Chem. Rev.* **118,** 1137–1168 (2018).

13. Merrifield, R. B. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *J. Am. Chem. Soc.* **85,** 2149–2154 (1963).

14. Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. H. Synthesis of proteins by native chemical ligation. *Science* **266,** 776–779 (1994).

15. Kastin, A. *Handbook of Biologically Active Peptides*. (Academic Press, 2013).

16. Cohen, S. & Elliott, G. A. The stimulation of epidermal keratinization by a gland of the mouse. *J. Invest. Dermatol.* **40,** 1–5 (1963).

17. Opsahl-Ferstad, H. G., Deunff, E. Le, Dumas, C. & Rogowsky, P. M. ZmEsr, a novel endosperm-specific gene expressed in a restricted region around the maize embryo. *Plant J.* **12,** 235–246 (1997).

18. Zitnan, D., Kingan, T. G., Hermesman, J. L. & Adams, M. E. Identification of Ecdysis-Triggering Hormone from an Epitracheal Endocrine System. *Science* **271,** 88–91 (1996).

19. Honey, K. & Rudensky, A. Y. Lysosomal cysteine proteases regulate antigen presentation. *Nat. Rev. Immunol.* **3,** 472–482 (2003).

20. Cascales, E., Buchanan, S. K., Duche, D., Kleanthous, C., Lloubes, R., Postle, K., Riley, M., Slatin, S. & Cavard, D. Colicin Biology. *Microbiol. Mol. Biol. Rev.* **71,** 158–229 (2007).

21. Brogden, K. A. Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.* **3,** 238–250 (2005).

22. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. *Molecular Biology*

of the Cell, 4th Edition. (Garland Science, 2002).

23. Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. *Molecular Cell Biology, 4th Edition*. (W. H. Freeman, 2000).

24. Voet, D. & Voet, J. G. *Biochemistry, 4th Edition*. (Wiley, 2010).

25. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* **11,** 909–916 (2015).

26. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **18,** 575–589 (2017).

27. Burge, C. B. & Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8,** 346–354 (1998).

28. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: Beautiful needles in the haystack. *Genome Res.* **7,** 768–771 (1997).

29. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. Life with 6000 Genes. *Science* **274,** 546–567 (1996).

30. Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L. & Grimmond, S. M. The abundance of short proteins in the mammalian proteome. *PLoS Genet.* **2,** 515–528 (2006).

31. Chen, J., Brunner, A. D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D. & Weissman, J. S. Pervasive functional translation of noncanonical human open reading frames. *Science* **367,** 140–146 (2020).

32. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. & Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

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

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

35. Zhang, Q., Vashisht, A. A., O'Rourke, J., Corbel, S. Y., Moran, R., Romero, A., Miraglia, L., Zhang, J., Durrant, E., Schmedt, C., Sampath, S. C. & Sampath, S. C. The microprotein Minion controls cell fusion and muscle formation. *Nat. Commun.* **8,** (2017).

36. Arnoult, N., Correia, A., Ma, J., Merlo, A., Garcia-Gomez, S., Maric, M., Tognetti, M., Benner, C. W., Boulton, S. J., Saghatelian, A. & Karlseder, J. Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. *Nature* **549,** 548–552 (2017).

37. Chu, Q., Martinez, T. F., Novak, S. W., Donaldson, C. J., Tan, D., Vaughan, J. M., Chang, T., Diedrich, J. K., Andrade, L., Kim, A., Zhang, T., Manor, U. & Saghatelian, A. Regulation of the ER stress response by a mitochondrial microprotein. *Nat. Commun.* **10,** 1–13 (2019).

38. Best, C. H. & Scott, D. A. Preparation of Insulin solution. *J. Biol. Chem.* **57,** 709–723 (1923).

39. Guillemin, R., Sakiz, E. & Ward, D. N. Further Purification of TSH-Releasing Factor (TRF) from Sheep Hypothalamic Tissues, with Observations on the Amino Acid Composition. *Exp. Biol. Med.* **118,** 1132–1137 (1965).

40. Sanger, F. The free amino groups of insulin. *Biochem. J.* **39,** 507–515 (1945).

41. Edman, P. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem. Scand.* **4,** 283–293 (1950).

42. Fox, S. W. *Terminal Amino Acids in Peptides and Proteins*. *Adv. Protein Chem.* (1945).

43. Swaminathan, J., Boulgakov, A. A., Hernandez, E. T., Bardo, A. M., Bachman, J. L., Marotta, J., Johnson, A. M., Anslyn, E. V. & Marcotte, E. M. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **36,** 1076–1091 (2018).

44. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).

45. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).

46. Craig, R. & Beavis, R. C. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467 (2004).

47. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47,** D506–D515 (2019).

48. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. & Apweiler, R. The international protein index: An integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988 (2004).

49. Wang, H., Chang-Wong, T., Tang, H. Y. & Speicher, D. W. Comparison of extensive

protein fractionation and repetitive LC-MS/MS analyses on depth of analysis for complex proteomes. *J. Proteome Res.* **9,** 1032–1040 (2010).

50. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201 (2004).

51. Michel, P. E., Reymond, F., Arnaud, I. L., Josserand, J., Girault, H. H. & Rossier, J. S. Protein fractionation in a multicompartment device using Off-Gel^{TM} isoelectric focusing. *Electrophoresis* **24,** 3–11 (2003).

52. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247 (2001).

Chapter 2

Integrated proteogenomics strategy for the discovery of peptides and small proteins in mouse brain tissue

2.1 Abstract

Complex mixtures of peptides and proteins are generated from biological sources. This sample complexity can impede the identification and quantification of peptides and small proteins. Reducing the sample complexity in a sample can overcome some of the challenges for detecting peptides and small proteins. In this chapter, I will describe the development of an integrated proteogenomics strategy that is optimized for the detection of peptides and small proteins. The strategy combines extraction, fractionation, and proteogenomic methods. We applied this strategy to mouse brain tissue and identified 8,464 proteins, which included a subset of known peptide hormones, neurotransmitters, and neuropeptides. We also found evidence for translation from an additional 222 unannotated small open reading frames. These findings show that the composition and regulation of the mouse brain peptidome and proteome are more diverse than previously thought.

2.2 Introduction

Peptides and small proteins play diverse and essential roles in biological systems. Peptide hormones, neurotransmitters, and neuropeptides are examples of these molecules that serve important signaling functions in organisms with central nervous systems. The actual number of genes that encode peptides and small proteins remains unexplored because many gene prediction algorithms use a lower length cutoff of 300-500 basepairs to predict protein-coding sequences[1].

As a result, many polypeptides shorter than 100-150 amino acids are likely unannotated in the repertoire of protein-coding genes[2,3]. Indeed, a re-analysis of the mouse transcriptome revealed the existence of thousands of potential protein-coding sequences below this length cutoff[4]. Recent proteogenomic studies that combine next-generation sequencing and proteomics have revealed the existence of hundreds of peptides and small proteins, also called microproteins or small open reading frame (smORF)-encoded polypeptides (SEPs), in human cell lines and tissues[5–7]. We reasoned that a similar proteogenomic approach could be applied to discover other classes of polypeptides that fall within this length range: peptide hormones, neurotransmitters, and neuropeptides.

In this chapter, I will describe the development and application of a methodology that allows for the comprehensive identification of peptides and small proteins in mouse brain tissue (Figure 2.1). The methodology presented in this chapter for the identification of these polypeptides combines several distinct steps: enrichment, fractionation, and proteogenomics. I will first present a discussion of each of these steps, including the rationale of each step, followed by describing the application of this methodology to identify peptides and small proteins in mouse brain tissue comprehensively.

The first consideration was to find a biological source from which peptides and small proteins that function as peptide hormones, neurotransmitters, and neuropeptides could be identified, which limited the search to tissues from the central nervous system. We chose mouse brain tissue as a source, to provide the broadest coverage of unannotated peptides and small proteins. Cerebrospinal fluid, in which neurohormones and neuropeptides circulate[8], is not readily obtained in sufficient quantities from mice[9]. The conditioned media from primary cells of the mouse brain was also considered as a source, but these are costly to culture in sufficient

quantities and require special culture conditions. In contrast, mouse brain tissue is both readily available for purchase and can be acquired in large quantities.

As biological tissues consist of other cellular proteins, a strategy was devised to remove larger proteins such that smaller peptides and proteins could be enriched and thereby more readily identified. Solid-phase extraction (SPE) can also be used to enrich peptide hormones[10,11] and microproteins[7] from tissues. Acid precipitation is compatible with the acidic conditions under which SPE is carried out. We, therefore, combined acid extraction with SPE, which has the advantage of removing impurities that might not have been removed by acid extraction alone. Several sorbent materials are commonly used for peptides, including C8 and C18 silica. Very polar and hydrophobic peptides will preferentially interact with alkyl chain sorbents such as C8 and C18[12]. Differences between sorbent interaction have been reported[13], prompting us to include both materials to capture as broad a repertoire of peptides and small proteins as possible.

Next, we considered and implemented several steps to optimize the C8- and C18-extracted mouse brain samples for analysis by mass spectrometry. Tissues are composed of complex proteomes with thousands of proteins. Post-translational modifications (PTMs) on proteins further increase proteome complexity. One common type of PTM found on peptide hormones and neuropeptides is glycosylation. Various glycan moieties can be added to asparagine (N-linked) and serine or threonine (O-linked) residues as secreted proteins, such as peptide hormones and neuropeptides, transit through the secretory pathway[14]. The diversity of glycans can add mass to peptides and proteins in an unpredictable manner, thereby precluding analysis by mass spectrometry[15]. To decrease the likelihood that a glycosylation event would prevent the identification of a polypeptide sequence, an enzymatic deglycosylation step was included. All N-linked glycans can be hydrolyzed and released from the anchoring asparagine

residue by the amidase PNGase F, converting the asparagine residue to aspartate. O-linked glycans are more structurally diverse and require specific enzymes to remove the glycan moiety. A mixture of N-linked and O-linked glycosylases is available for purchase commercially and has been used to in other glycoproteomic studies[16]. The mixture of O-linked glycosylases targets a broad range of glycoprotein structures, but the possibility remains that some O-linked glycans remain at least partially intact after this enzymatic deglycosylation treatment.

The proteome complexity of biological tissues can result in the undersampling of peptides in a mass spectrometer[17]. This introduces a detection bias towards more abundant species in the sample[18], which would mask the identification of biologically active polypeptides found at lower abundance. To reduce, but not wholly eliminate, the possibility of undersampling, we implemented a biochemical fractionation approach where the enzymatically deglycosylated C8- or C18-extracted sample was divided into a series of fractions. We chose two fractionation methods that predictably separate polypeptides based on molecular weight: gel-eluted liquid fraction entrapment electrophoresis (GELFrEE)[19] and size-exclusion chromatography using a Superdex 30 column. These two methods allows for the collection of fractions in the liquid phase and prevent the loss of material required in downstream extraction steps, such as in the case of cutting SDS-PAGE bands. In addition to reducing sample complexity, fractionation by molecular weight also allowed larger proteins to be separated from the rest of the sample. We only retained fractions with proteins less than 25 kDa, which includes peptide hormones, neurotransmitters, neuropeptides, and microproteins. As polypeptides show preferential interaction with sorbent materials used in solid-phase exchange, each fractionation technique will also be more effective at fractionating a particular sub-population of polypeptides in a sample. GELFrEE uses the same principle as SDS-PAGE to separate polypeptides. Polypeptides with lower molecular weights

will elute in earlier fractions, whereas polypeptides with larger molecular weights will elute in later fractions.

In contrast, size-exclusion chromatography uses a porous matrix in which polypeptides can selectively enter based on their size and shape. Larger polypeptides enter the pores less frequently and elute in earlier fractions. Smaller polypeptides enter the pores more frequently and are retarded in their mobility, thereby eluting in later fractions. Like the use of both C8 and C18 silica in the extraction steps, both GELFrEE and size-exclusion chromatography were included as complementary fractionation approaches.

Finally, we considered an optimal strategy to detect and identify small peptides and proteins that were extracted from the mouse brain tissue. High-resolution mass spectrometers are routinely used to identify thousands of proteins in biological samples. In a "discovery proteomics" experiment, the acquired mass spectra are searched against a database containing all possible protein sequences that might be in the sample. Because ORFs that encode peptides and small proteins are not likely to be annotated in genomes, these peptides and proteins will be absent from most proteome databases and not detected using this approach[20]. To overcome this limitation, we generated a custom sequence database from the in silico translation of mRNA transcripts that were detected in mouse brain tissue by RNA-Seq[21]. The mRNA transcripts were translated in all three reading frames to generate all theoretically possible protein-coding sequences, thereby ensuring that peptides and small proteins encoded by unannotated smORFs would be detected. The integration of genomics and transcriptomics data and in proteomics experiments is termed "proteogenomics"[20].

We applied our integrative proteogenomic strategy to identify peptides and small proteins in mouse brain tissue. We identified 8,464 proteins that were annotated in UniProtKB. We

further identified 242 ORFs that are not currently annotated as protein-coding in UniProtKB. Of these, 222 were less than 150 amino acids in length. Fractionation by GELFrEE and Superdex 30 improved the number of proteins identified nearly 2-fold and was essential to identify peptides and small proteins encoded by smORFs. Our results show that the translational landscape of the mouse brain is more diverse than previously thought and highlights new potential modes of regulation of the Pde1b gene.

2.3 Methods and Materials

Peptide extraction from mouse brain tissue

Peptide extraction was performed as previously described[10] with minor modifications[22]. Brains from 20 male mice of unspecified strains were heated in a mixture of 1.0 N acetic acid and 0.1 N hydrochloric acid heated at >90 °C for 5 minutes. After cooling to room temperature, the tissues were homogenized using a Polytron homogenizer, centrifuged at 30,000 × *g* for 30 minutes at 4 °C. Supernatants were filtered through Millex-SV 5 µm polyvinylidene fluoride syringe filters. Peptides were enriched from the flow-through by solid-phase extraction using Agilent Bond Elut 1 gram C8 or C18 silica cartridges on a vacuum manifold. Cartridges were pre-wet with one column volume of methanol and equilibrated with one column volume of triethylammonium formate (TEAF) buffer, pH 3.0. Samples were applied to the pre-equilibrated column, washed with one column volume of TEAF, and eluted with 2 ml of a mixture of 75% (v/v) acetonitrile and 25% (v/v) TEAF. Samples were evaporated to dryness in a vacuum centrifuge overnight and stored at -20 °C until further processing.

Enzymatic deglycosylation

C8- and C18-extracted samples were equilibrated to room temperature, resuspended in water, vortexed, and centrifuged at 17,000 × *g* for 10 minutes in a benchtop microcentrifuge. Protein concentration was measured using a Pierce BCA protein assay kit using bovine serum albumin standards. *N*- and *O*-linked glycans were removed enzymatically under denaturing reaction conditions using an NEB protein deglycosylation kit according to the manufacturer's procedures. Briefly, 800 µg of C8- or C18-extracted sample was dissolved in 225 µL water, to which 12.5 µL of deglycosylation buffer 2 was added, and the reaction was incubated at 75 °C for 10 minutes. After cooling to room temperature, 12.5 µL protein deglycosylation mix II was added, incubated at room temperature for 30 minutes, then at 37 °C for 1 hour.

Peptide fractionation

Deglycosylated samples were divided into two equal portions and fractionated using either an Expedeon GELFrEE 8100 fractionation station or by size-exclusion chromatography. For GELFrEE fractionation, samples were resuspended in 1× tris-acetate sample buffer, dithiothreitol was added to 53 mM and incubated at 50 °C for 10 minutes according to the manufacturer's instructions. Samples were cooled to room temperature and loaded on a 12% tris-acetate cartridge and electrophoresed for 20 minutes at 50 V until the sample entered the cartridge, after which the sample chamber was rinsed out with the supplied 1× HEPES running buffer, electrophoresed for an additional 50 minutes at 50 V, after which the first fraction was collected. The second fraction was collected after an additional 5 minutes at 50 V. Subsequently, the voltage was raised to 85 V and fractions were collected after the following times in minutes:

2.5, 2.8, 3, 3.5, 4, 6, 8, 10, 12, 15. A total of 12 fractions were collected and stored at -20 °C. For fractionation by size-exclusion chromatography, samples were diluted to 0.5 mL with a mixture of 30% acetonitrile and 0.1% trifluoroacetic acid in water. The sample was manually loaded onto a 0.5 mL sample loop attached to a GE AKTA protein purification system. The sample was automatically injected and fractionated on a 10 × 300 mm GE Superdex 30 Increase column pre-equilibrated with a mixture of 30% acetonitrile and 0.1% trifluoroacetic acid at a flow rate of 0.5 mL/minute. Fractions were automatically collected every minute after sample injection over 60 minutes. Protein elution from the column was monitored using an in-line ultraviolet (UV) detector at 280 nm. Fractions 6-57 were pooled in a pairwise manner (26 fraction pairs) to cover all peaks in the elution profile in the UV chromatogram. An aliquot of 50 µL was withdrawn and evaporated to dryness separately in a vacuum centrifuge overnight and stored at -20 °C until analysis by SDS-PAGE. The remainder of the sample was similarly dried and stored for LC-MS/MS analysis.

SDS-PAGE & Silver Staining

An amount corresponding to 5% of the fraction volume or input material was separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). GELFrEE fractions were combined with 4× loading buffer (250 mM Tris-HCl pH 6.8, 8% (w/v) SDS, 0.2% (w/v) bromophenol blue, 40% glycerol, 10% (v/v) 2-mercaptoethanol) and the dried Superdex 30 fractions were dissolved directly in 1× loading buffer. All samples were incubated at 95 °C for 3 minutes and cooled to room temperature. Samples were loaded on a pre-cast Novex 10-20% tris-glycine polyacrylamide gel in a running buffer at pH 8.3 (25 mM Tris, 192 mM glycine, 0.1% SDS) and electrophoresed at 200 V until the dye-front migrated to the bottom of the gel. Proteins

were visualized by silver staining using a Pierce silver stain kit according to the manufacturer's instructions and imaged on a Bio-Rad ChemiDoc XRS+ gel imaging system using a white light transilluminator.

LC-MS/MS analysis

Fractions for liquid chromatography coupled to electrospray ionization and tandem mass spectrometry (LC-MS/MS) analysis were selected based on the presence of protein bands under 25 kDa after silver staining. LC-MS/MS analysis was carried out as previously described[7]. Each fraction was analyzed once. Samples were precipitated by methanol/chloroform and re-dissolved in 8 M urea/100 mM TEAB, pH 8.5. Proteins were reduced with 5 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP, Sigma-Aldrich) and alkylated with 10 mM chloroacetamide (Sigma-Aldrich). Proteins were digested overnight at 37 °C in 2 M urea/100 mM TEAB, pH 8.5, with trypsin (Promega). Digestion was quenched with formic acid, 5 % final concentration. The digested samples were analyzed on a Fusion Orbitrap tribrid mass spectrometer (Thermo Fisher Scientific). The digest was injected directly onto a 30 cm, 75 μm inner diameter column packed with BEH 1.7 μm C18 resin (Waters). Samples were separated at a flow rate of 300 nL/min on an nLC 1000 chromatography system (Thermo Fisher Scientific). Buffer A and B were 0.1% formic acid in water and 0.1% formic acid in 90% acetonitrile, respectively. A gradient of 1-35% B over 110 min, an increase to 50% B over 10 min, an increase to 90% B over 10 min and held at 90% B for a final 10 min was used for a 140 min total run time. The column was re-equilibrated with 20 μL of buffer A prior to the injection of the sample. Peptides were eluted directly from the tip of the column and nanosprayed directly into the mass spectrometer by application of 2.5 kV voltage at the back of the column. The Orbitrap

Fusion was operated in a data-dependent mode. Full MS scans were collected in the Orbitrap at 120,000 resolution with a mass range of 400 to 1600 m/z and an automatic gain control (AGC) target of 5×10^5. The cycle time was set to 3 sec, and within this 3 sec the most abundant ions per scan were selected for CID MS/MS in the ion trap with an AGC target of 10^4 and minimum intensity of 5000. Maximum fill times were set to 50 ms, and 100 ms for MS and MS/MS scans, respectively. Quadrupole isolation at 1.6 m/z was used, monoisotopic precursor selection was enabled, and dynamic exclusion was used with a duration of 5 sec.

Data processing of UniProt-annotated proteins

Data analysis was performed as previously described[7] with minor modifications. Tandem mass spectra data were extracted and deconvolved from .raw files using RawConverter[23] 1.1.0.23 and searched using a target-decoy strategy with ProLuCID[24] and the Integrated Proteomics Pipeline (Integrated Proteomics Applications) data analysis platform. The UniProt mouse reference proteome (downloaded on February 15, 2019) and common contaminant proteins were compiled into a single protein sequence database. Reverse decoy sequences were generated and appended to the database. The following search parameters were used: CID/HCD fragmentation mode, monoisotopic mass, 50 parts per million (ppm) precursor ion mass tolerance, 600 ppm fragment ion mass tolerance, 600-6000 mass range, trypsin was set as the enzyme, requiring at least one tryptic end, up to two missed cleavages allowed, carbamidomethylation on cysteine as a static modification, and +0.984016 on asparagine as a differential modification. Mass spectra matches were filtered to 10 ppm precursor ion mass tolerance and evaluated with DTASelect 2.0[25] using XCorr and Zscore as the primary and secondary score types, respectively. Protein

identifications required at least one matched peptide per protein and were reported at a false discovery rate of 1%.

Data processing of unannotated proteins

Data analysis was performed as described previously[6] with minor modifications. The analysis procedure is the same as those used in the previous section, except that a custom protein database generated from the in silico translation in 3-frames of strand-specific RNA-Seq data from mouse brain tissue[21] was used. The in-silico-translated database was generated as previously described[6]. Reverse decoy sequences were generated and appended to the database. The same search parameters were used, except that no differential modifications were specified. Peptide sequences matching to reviewed and non-reviewed entries in the UniProt mouse reference proteome (downloaded on March 03, 2020), common laboratory contaminant proteins, deglycosylases, and all reverse decoy sequences were filtered out using custom string-searching scripts. For the remaining peptide sequences, an RNA transcript was identified in the NCBI Mouse Reference Sequence (RefSeq) Database[26] using tBlastn. Where possible, known RefSeq transcript sequences ("NM_" prefix) were used for annotation. If no reviewed transcript was found, then transcripts from model transcript sequences ("XM_" prefix) were used. Open reading frames were annotated using the nearest in-frame start (AUG) and stop codons (UAG, UGA, UAA). For RNA sequences lacking an upstream AUG start codon, the furthest upstream near-cognate codon (ACG, AAG, CUG, etc.) in a Kozak sequence in the RNA transcript was assigned as the start codon. If a near-cognate start in a Kozak sequence was not identified, then one of three codons was assigned as the translation initiation site in the transcript: (1) the furthest upstream near-cognate codon, (2) the codon after the nearest upstream stop codon, or (3) the

furthest upstream in-frame codon in the transcript. Open reading frames of less than 150 amino acids were designated as microproteins.

2.4 Results and Discussion

Integrated proteogenomic strategy for polypeptide discovery in mouse brain tissue

To comprehensively identify peptides and microproteins in the mouse brain, we devised an integrated proteogenomic strategy that combines peptide enrichment, peptide fractionation, and discovery proteomics (Figure 2.1). The goal was to increase the number of peptide and microprotein identifications, including those that might be biologically active in the mouse brain but are currently unannotated. In the remainder of the text, "peptides and proteins" will simply be referred to as "polypeptides." To extract polypeptides, tissue from 20 mouse brains was homogenized in strong acid at 90 °C. This step also served to inactivate proteases and avoid generating more complex peptide mixtures from non-specific proteolysis.

We then used both C8 and C18 silica in pre-packed columns to extract and concentrate polypeptides from the mouse brain. An enzymatic deglycosylation step was added to remove N- and O-linked glycans, which can preclude identification by mass spectrometry by adding mass to polypeptides in a variable manner[15]. We used both GELFrEE and size-exclusion chromatography to fractionate polypeptides by molecular weight, choosing only to retain fractions that contained polypeptides of less than 25 kDa. Polypeptides from each of these fractions were digested with trypsin and analyzed by LC-MS/MS. Trypsin digestion reduces the length of polypeptides that fall outside the optimal range of 10-20 amino acids for detection by mass spectrometry[27–29].

Our proteogenomic strategy identified 8,464 UniProt-annotated proteins (Supplemental Table 1) in the mouse brain. We cross-referenced these proteins with entries in NeuroPep[30],

NeuroPedia[31], and UniProtKB[32] mouse databases. Of these, we found 86 known neuropeptides, neurotransmitters, and peptide hormones in our dataset (Supplemental Table 2), suggesting that our experimental strategy was able to identify biologically active polypeptides in the mouse brain with well-characterized signaling functions. To identify polypeptides that are not currently unannotated in UniProt, we searched our proteomics data against a protein sequence database generated from the in silico translation of strand-specific RNA-Seq data[21]. The translation was done in all three reading frames to capture all potentially translated protein sequences. We identified 6,890 RNA transcripts with matching tryptic peptides in our dataset. Of these, we were able to identify 242 open reading frames (ORFs) from tryptic peptides in our dataset that contain sequences that do not match any entry in the UniProt mouse reference proteome (Supplemental Table 3). Our integrated proteogenomic strategy shows that the mouse brain proteome is more diverse than previously thought.

Comparison of C8 and C18 solid-phase extraction to enrich peptides

We implemented a peptide enrichment step to enrich for smaller peptides and proteins of interest while simultaneously excluding larger and more abundant proteins from being included in our samples. We first wanted to compare the use of two extraction materials, C8 and C18 silica, in terms of enriching the small proteome fraction from a tissue source. Both extraction methods showed an enrichment for polypeptides with lengths of 100-250 amino acids (Figure 2.2a). Polypeptides that were less than 100 amino acids seemed to be less well represented relative to all proteins annotated in UniProt. Of the 8,464 UniProt-annotated proteins that were identified in the mouse brain, 5,001 proteins (59%) were found in both C8- and C18-extracted

samples (Figure 2.2b). Of the remaining proteins, 1,341 were in the C8-extracted sample only, and 2,122 were in the C18-extracted sample only.

In terms of the 242 non-UniProt ORFs that were detected, the proportion of overlapping identifications between C8 and C18 extractions was less than the UniProt pool (32 or 13%, Figure 2.2c). While most proteins in our dataset were extracted by both C8 and C18, there is a large proportion of the proteome that is extracted using one of the two materials.

We wanted to see if any biochemical characteristics could distinguish proteins extracted by the two materials. Both C8 and C18 preferentially interact with analytes through the hydrophobic effect in reverse-phase chromatography. We reasoned that the population of proteins captured on the shorter C8 material would be less hydrophobic than those captured by the longer C18 material.

To test this, we calculated a hydropathy index for each protein using Kyte & Doolittle amino acid scale values[33]. We then compared the distribution of hydropathy index values of the proteins that were extracted by C8 and C18. The proteins that were extracted by C8 showed a lower mean hydropathy index value (-0.44) than the proteins that were extracted only by C18 (-0.47), which indicates that the proteins extracted by C18 were slightly more hydrophobic. However, there was considerable overlap between the distribution of hydropathy indices between each of the C8 and C18 pools, indicating that the hydropathy index cannot be used as a predictor of which peptides would interact preferentially with either extraction material.

A similar analysis was done for protein length, showing that proteins extracted by C18 had a slightly longer mean length (475 residues) than the C8 pool (457 residues). The distribution of lengths was considerable between both pools. These findings indicate that while the proteins extracted by C18 might be more hydrophobic and longer, neither the hydropathy

index or length can be used to predict which proteins would be preferentially extracted by C8 or C18. These findings led us to conclude that both C8 and C18 extracts distinct pools of polypeptides. Still, other biochemical factors were at play to determine which extraction material would preferentially capture a specific polypeptide. We concluded that using both extraction materials would be complementary in enriching the microprotein fraction of the proteome.

Fractionation increased the number and depth of identifications

To reduce the likelihood that sample complexity would preclude the identification of peptides and small proteins, we used two biochemical fractionation techniques after C8 or C18 extraction that separate based on molecular weight: gel-eluted liquid fraction entrapment electrophoresis (GELFrEE)[19] and size-exclusion chromatography using a Superdex 30 column. Fractions from GELFrEE and Superdex 30 separations were analyzed by SDS-PAGE, and proteins were visualized by silver staining (Figures 2.5, 2.6, 2.7, 2.8). Fractions containing proteins less than 25 kDa were digested with trypsin and analyzed by liquid chromatography coupled to electrospray ionization and tandem mass spectrometry (LC-MS/MS). The fractions from Superdex 30 covered the majority of peaks in the elution chromatogram (Figures 2.9, 2.10). To evaluate whether fractionation improved the number of protein identifications, we compared the GELFrEE- and Superdex 30-separated samples to an unfractionated sample that was also digested by trypsin and analyzed by LC-MS/MS.

In the unfractionated C8-extracted sample, 1,728 proteins were identified (Figure 2.11a). With GELFrEE fractionation, the number of protein identifications increased to 4,016. With Superdex 30 fractionation, the number of protein identifications increased to 5,413. Some 3,141 proteins were detected in both GELFrEE and Superdex fractions, while 875 were detected with

GELFrEE fractionation only, and 2,326 were detected with Superdex 30 fractionation only. In the unfractionated C18-extracted sample, 1,813 proteins were identified (Figure 2.11b). With GELFrEE fractionation, the number of protein identifications increased to 4,299. With Superdex 30 fractionation, the number of protein identifications increased to 6,203. Some 3,379 proteins were detected in both GELFrEE and Superdex 30 fractions, while 920 were detected with GELFrEE fractionation only, and 2,824 were detected with Superdex 30 fractionation only. Overall, both fractionation methods did not affect the distribution of lengths of the proteins that were identified in the C8-extracted (Figure 2.12) or C18-extracted (Figure 2.13) samples. These findings indicate that our fractionation strategy was able to increase the number of protein identifications, especially for those with lengths of less than 150 amino acids.

Fractionation increased the identification of non-UniProt ORFs

We were interested in finding peptides and small proteins that might be biologically active in the mouse brain but were not yet annotated in the mouse proteome. To accomplish this goal, we took a proteogenomics approach by searching our proteomics dataset against a custom protein database generated from the in silico translation of mouse brain RNA-Seq data[21]. The translation was done in all three reading frames to capture all theoretically possible protein sequences that could be generated in the mouse brain. In our dataset, we mapped tryptic peptides to 6,890 RNA transcripts found in the mouse brain RNA-Seq dataset. We filtered out sequences that matched those found in the UniProt mouse reference proteome.

For the remaining sequences, open reading frames (ORFs) were annotated using in-frame start (AUG) and stop (UAG, UGA, UAA) codons. If an upstream AUG was not found, then the furthest upstream near-cognate start codon was used instead. If an upstream near-cognate codon

was not present, then the furthest upstream codon after a stop codon or the furthest upstream in-frame codon was designated as the translation start site. In all cases, we assigned the translation start site to the furthest upstream codon to avoid introducing biases towards shorter lengths. As a result, we identified 242 ORFs that were not annotated in UniProt. We identified 17 transcripts for which no downstream stop was present, and an ORF could not be annotated. These transcripts might have harbor downstream frameshift mutations that place the stop codon out-of-frame. Another 15 transcripts were excluded from further consideration as their peptides were in-frame with an existing ORF, as in the case of a 5' or 3' extension product from a splice isoform.

As with the number of UniProt-annotated protein identifications, the number of non-UniProt ORF identifications also increased with sample fractionation. In the C8-extracted sample, 136 non-UniProt ORFs were identified (Figure 2.14a). In the unfractionated C8 sample, 9 non-UniProt ORFs were identified. With GELFrEE fractionation, 62 non-UniProt ORFs were identified, of which 44 were uniquely identified with this method. With Superdex 30 fractionation, 90 non-UniProt ORFs were identified, of which 72 were uniquely identified with this method. In the C18-extracted sample, 139 non-UniProt ORFs were identified (Figure 2.14b). In the unfractionated C18 sample, 6 non-UniProt ORFs were identified. With GELFrEE fractionation, 59 non-UniProt ORFs were identified, of which 41 were uniquely identified with this method. With Superdex 30 fractionation, 98 non-UniProt ORFs were identified, of which 80 were uniquely identified with this method. These findings indicate that our fractionation was able to increase the number of non-UniProt ORFs that were identified.

Features of non-UniProt ORFs (location, MW, start codon usage)

We assessed the confidence of the detection of the 242 non-UniProt ORFs identified in our dataset. While previous studies have applied visual inspection of the annotated mass spectra to evaluate non-UniProt identifications[29], our non-UniProt dataset contained over 550 individual spectra to inspect, this number of spectra proved to be too large to inspect in a timely and consistent basis. Visual inspection is also subject to individual bias, and annotations of mass spectra could be unfairly rejected.

Rather than perform a visual inspection of all matched mass spectra, we assigned each of the 242 non-UniProt ORFs to three categories: high confidence, medium confidence, and low confidence (Figure 2.15a). Higher confidence matches have more matched tryptic peptides and spectral counts, thereby reducing the likelihood of a match to a random transcript. The 11 non-UniProt ORFs in the high confidence category had more than one tryptic peptide matched or one peptide with greater than 10 spectral counts. The 41 non-UniProt ORFs in the medium confidence category had one tryptic peptide matching with between two and ten spectral counts. The 190 non-UniProt ORFs in the low confidence category had one tryptic peptide matched and one spectral count.

Next, we analyzed the length distribution of all non-UniProt ORFs (Figure 2.15b). We found that 222 had a length of 150 amino acids or less, which would be classified as microproteins. This category would also include peptides and small proteins that might function as neuropeptides or peptide hormones but have not yet been characterized or annotated in UniProt.

We analyzed the start codon usage of the 242 non-UniProt ORFs (Figure 2.15c). We found that 60 of these used an AUG start, 94 used a near-cognate start codon, and 88 used other codons. These other codons were usually the furthest upstream codon in the transcript, which

suggests that ORF might be translated from an alternative RNA transcript that is not adequately represented in the RefSeq database.

We also analyzed the relative positions of the non-UniProt ORFs on their RNA transcripts (Figure 2.16a). We assigned these non-UniProt ORFs to one of five categories: 26 were in the 5' untranslated region (UTR), 52 overlapped with the coding sequence (CDS), 63 were in the 3' UTR, 14 were from non-coding RNAs, and 3 were transcribed from the antisense strand. Examples of each category are given in Figure 2.16b. Further experiments, including acquiring reference spectra using synthetic peptides and generating recombinant constructs to verify expression, will be required to validate the list of non-UniProt ORFs identified in the mouse brain.

Pde1b Upstream Open Reading Frame (PDURF) reveals new potential modes of regulation of the *Pde1b* gene

We examined the annotation of the non-UniProt ORF with annotated transcript coordinates of 103503190-103503568 (+ strand) on chromosome 15, which encodes a putative protein of 120 amino acids (Figure 2.17a). This non-UniProt ORF is immediately upstream of the *Pde1b* gene. The *Pde1b* gene is expressed in the mouse brain and encodes a calcium/calmodulin-dependent 3'/5'-cyclic nucleotide phosphodiesterase involved in second messenger signaling. We will refer to this non-UniProt ORF as Pde1B Upstream Open Reading Frame (*PDURF*). The start codon used was the furthest upstream near-cognate AGG codon embedded in a Kozak consensus sequence. Three other near-cognate codons were also found downstream, which might also serve as translation initiation sites.

We detected three tryptic peptides that showed confident mass spectra matches. Each peptide had precursor ion mass errors of less than 1 p.p.m., extensive coverage of the *b*- and *y*-ion fragmentation series and had most major peaks matched in the mass spectrum. Ribosome profiling studies of the mouse brain[34] have shown that the presence of ribosome footprints in this region is consistent with the translation of PDURF (Figure 2.18). The genomic structure of *PDURF* and *Pde1b* raises the possibility of bicistronic gene expression and regulation in cis. *PDURF* might act as an upstream ORF (uORF) that regulates the expression of *Pde1b* by leaky scanning. A similar mechanism regulates the expression of human *FRAT2*[5], mammalian *ATF4*[35], and yeast *GCN4*[36] genes.

Alternatively, the PDURF microprotein might interact with and regulate the PDE1B protein in trans. A recent study has shown that the peptides from HAUS6, HMGA2, FBXO9, MIEF, and DDIT3 uORFs interact with and form a protein complex with their respective downstream ORFs in human cell lines[37]. The interplay between the *PDURF* and *Pde1b* ORFs can be determined by introducing recombinant versions that encode different epitope tags on the same plasmid. The expression of each protein can then be assessed by western blotting. Mutating the start codon or introducing a premature stop codon in either coding sequence can then be performed to determine the effect of protein expression from the other ORF. Immunoprecipitation followed by western blotting can be performed to check for protein complex formation.

2.5 Conclusion

In this chapter, we described the development and application of an integrated proteogenomic strategy to identify peptides and small proteins from mouse brain tissue. The

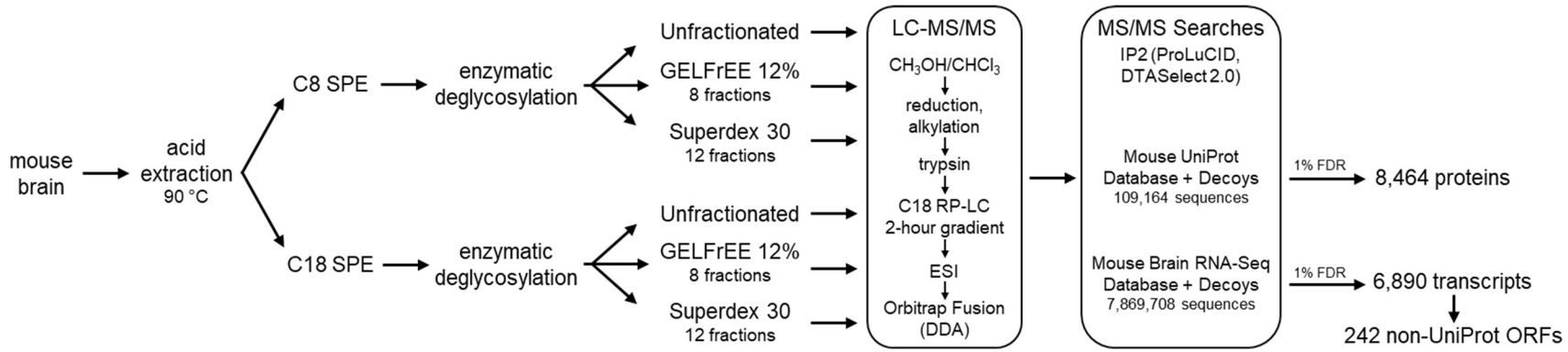
strategy included extraction, fractionation, and proteogenomic steps to target peptides and small proteins specifically. The use of GELFrEE and size-exclusion chromatography fractionation steps resulted in a nearly two-fold increase in the number of protein identifications. Our strategy uncovered known peptide hormones, neurotransmitters, and neuropeptides. Our proteogenomic strategy also revealed the existence of 242 unannotated ORFs, of which 222 are smORFs. These smORFs might act in cis to regulate the expression of neighboring ORFs or interact in trans with the translated protein products. Our results suggest that the mechanisms regulating gene expression and the repertoire of translated ORFs in the mouse brain are more diverse than previously thought.

2.6 Acknowledgements

Chapter 2, in part, is currently being prepared for submission for publication of the material. Mak, Raymond; Vaughan, Joan; Shokhirev, Max; Diedrich, Jolene; Saghatelian, Alan. "Proteogenomic discovery of open reading frames encoding peptides and small proteins in mouse brain". The dissertation author was the primary investigator and author of this material. This work was supported by the Mass Spectrometry Core of the Salk Institute with funding from NIH-NCI CCSG: P30 014195 and the Helmsley Center for Genomic Medicine. This work was supported by the The Razavi Newman Integrative Genomics and Bioinformatics Core Facility of the Salk Institute with funding from NIH-NCI CCSG: P30 014195, and the Helmsley Trust.

2.7 Figures

Figure 2.1: An integrated proteogenomic strategy for the discovery of translated polypeptides from the mouse brain. Mouse brain tissue was homogenized in strong acid at 90 °C. polypeptides were subjected to solid-phase extraction (SPE) using C8 or C18 silica. Following enzymatic deglycosylation, polypeptides were fractionated using GELFrEE or Superdex 30. All fractions containing proteins less than 25 kDa were analyzed by liquid chromatography coupled to electrospray ionization and tandem mass spectrometry (LC-MS/MS). Proteins were first precipitated using methanol/chloroform, reduced and alkylated, digested with trypsin, then subjected to reverse-phase liquid chromatography (RP-LC) using a C18 column and a 2-hour gradient. Eluted polypeptides were ionized using electrospray (ESI) and analyzed on an Orbitrap Fusion instrument operated in data-dependent acquisition (DDA) mode. MS/MS spectra were then searched using a target-decoy strategy against a database of mouse UniProt reference proteins or a database generated from the in silico translation of mouse brain RNA-Seq data. Protein and transcript identifications were reported with a false-discovery rate (FDR) of 1%.


mouse brain
acid extraction 90 °C
C8 SPE
C18 SPE
enzymatic deglycosylation
enzymatic deglycosylation
Unfractionated
GELFrEE 12% 8 fractions
Superdex 30 12 fractions
Unfractionated
GELFrEE 12% 8 fractions
Superdex 30 12 fractions
LC-MS/MS
CH3OH/CHCl3
reduction, alkylation
trypsin
C18 RP-LC 2-hour gradient
ESI
Orbitrap Fusion (DDA)
MS/MS Searches
IP2 (ProLuCID, DTASelect 2.0)
Mouse UniProt Database + Decoys 109,164 sequences
Mouse Brain RNA-Seq Database + Decoys 7,869,708 sequences
1% FDR
8,464 proteins
1% FDR
6,890 transcripts
242 non-UniProt ORFs

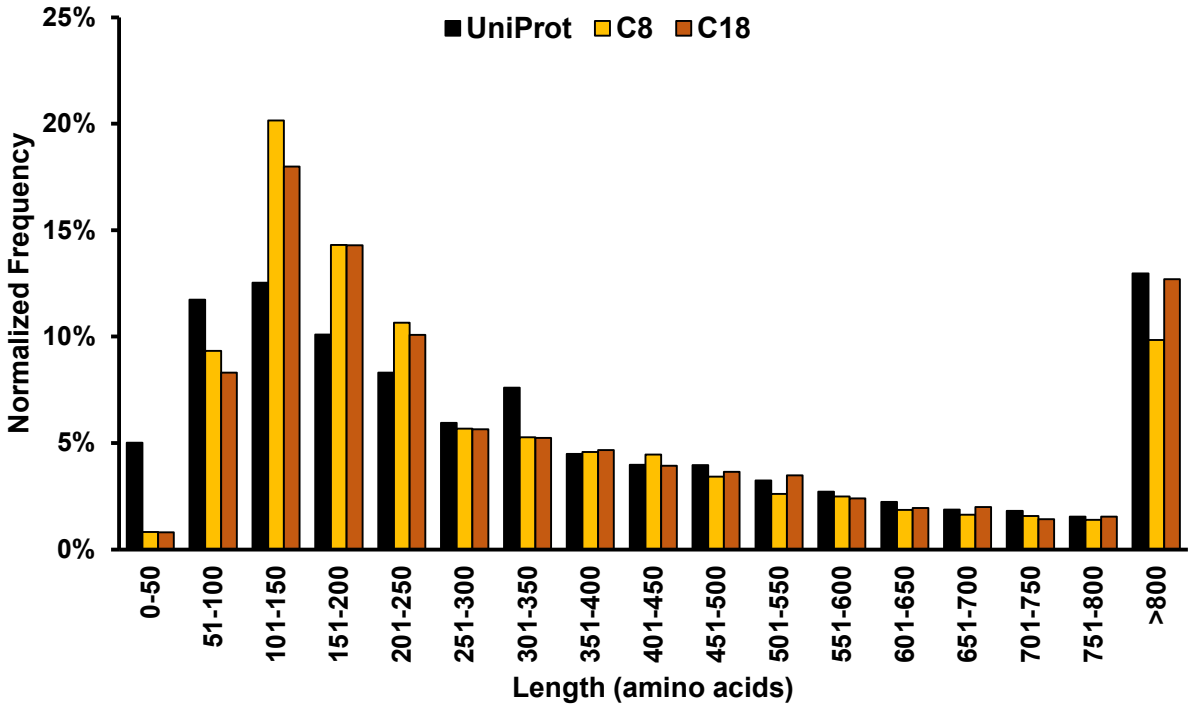
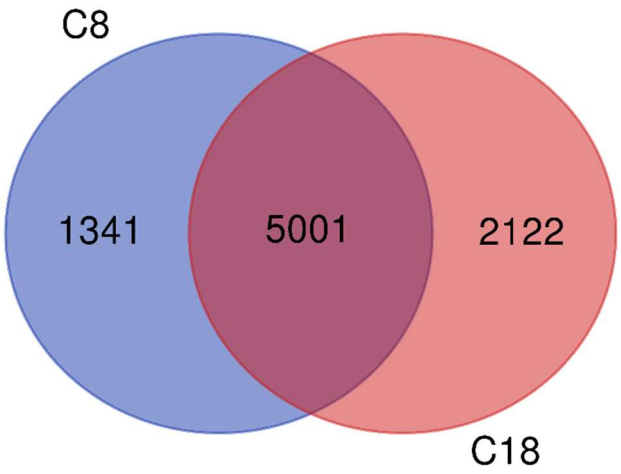


Figure 2.2: Length distribution of UniProt-annotated mouse brain protein identifications. Lengths of canonical sequences were retrieved from the UniProtKB database and plotted as a normalized frequency in 50 amino-acid length windows. All entries in UniProtKB mouse reference proteome, black bars. C8-extracted proteins, yellow bars. C18-extracted proteins, orange bars. UniProt contains multiple redundant entries for proteins less that were not removed; this appears to skew the distribution of proteins less than 150 amino acids in length.

(A)



(B)

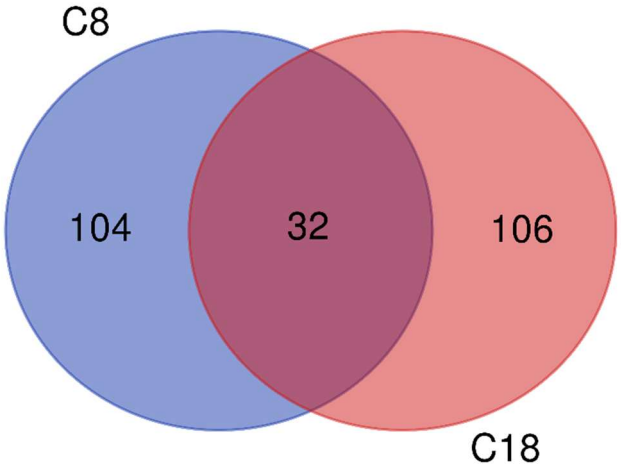


Figure 2.3: Overlap of protein identifications in mouse brain. (A) UniProt-annotated proteins. **(B)** Non-UniProt ORFs. C8-extracted proteins, blue circle. C18-extracted proteins, red circle.

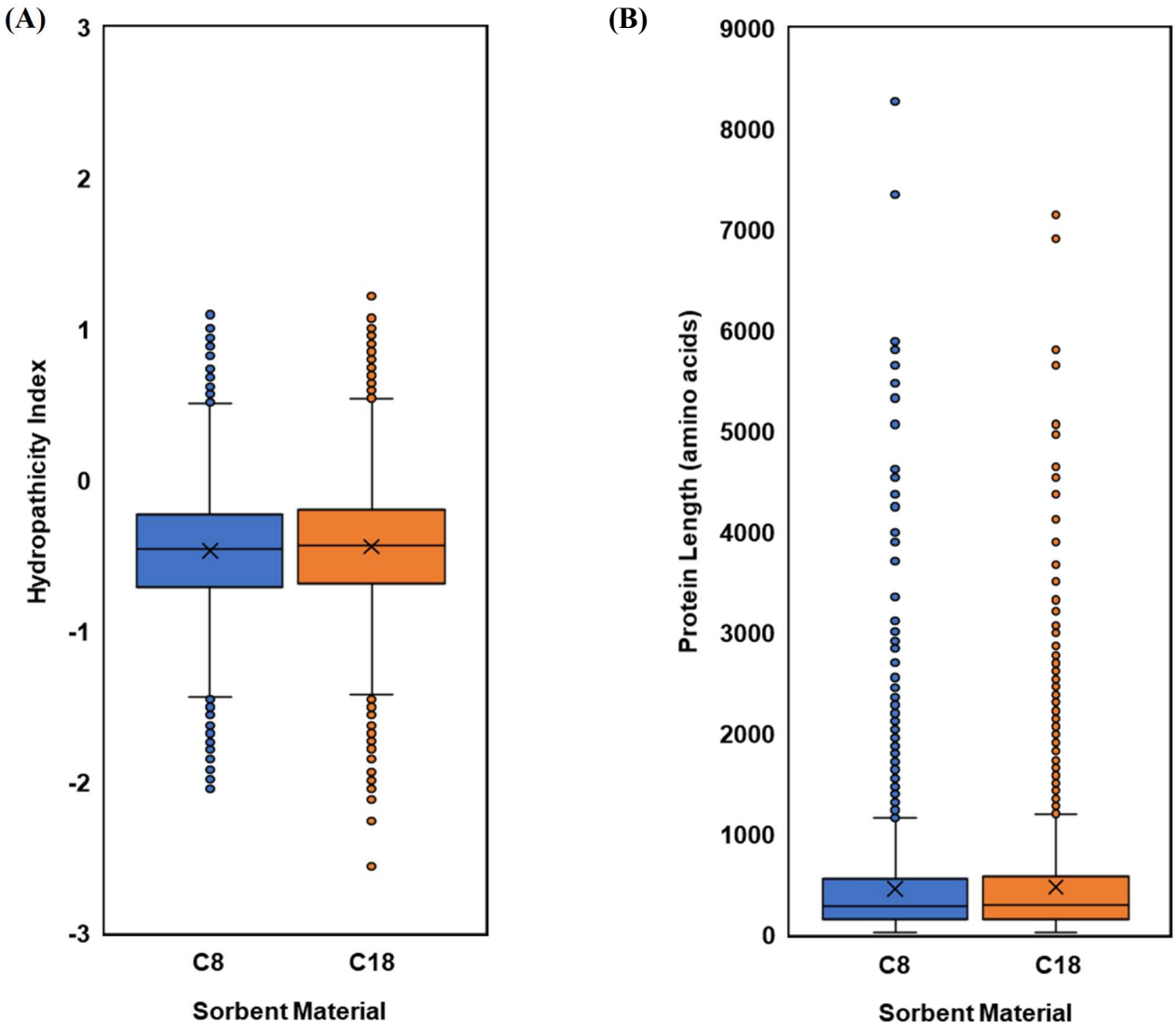


Figure 2.4: Comparison of biochemical properties of C8- and C18-extracted UniProt-annotated proteins. Box plots of **(A)** hydropathicity index or **(B)** protein length. Quartiles (25%, 50%, 75%) are represented by the box. One standard deviation from the mean is represented by the whiskers. Outliers, including minimum and maximum values, are plotted as points. The arithmetic mean is plotted as an "X." Hydropathicity index was calculated using Kyte & Doolittle values for each amino acid in the protein[33]. C8-extracted proteins, blue. C18-extracted proteins, orange.

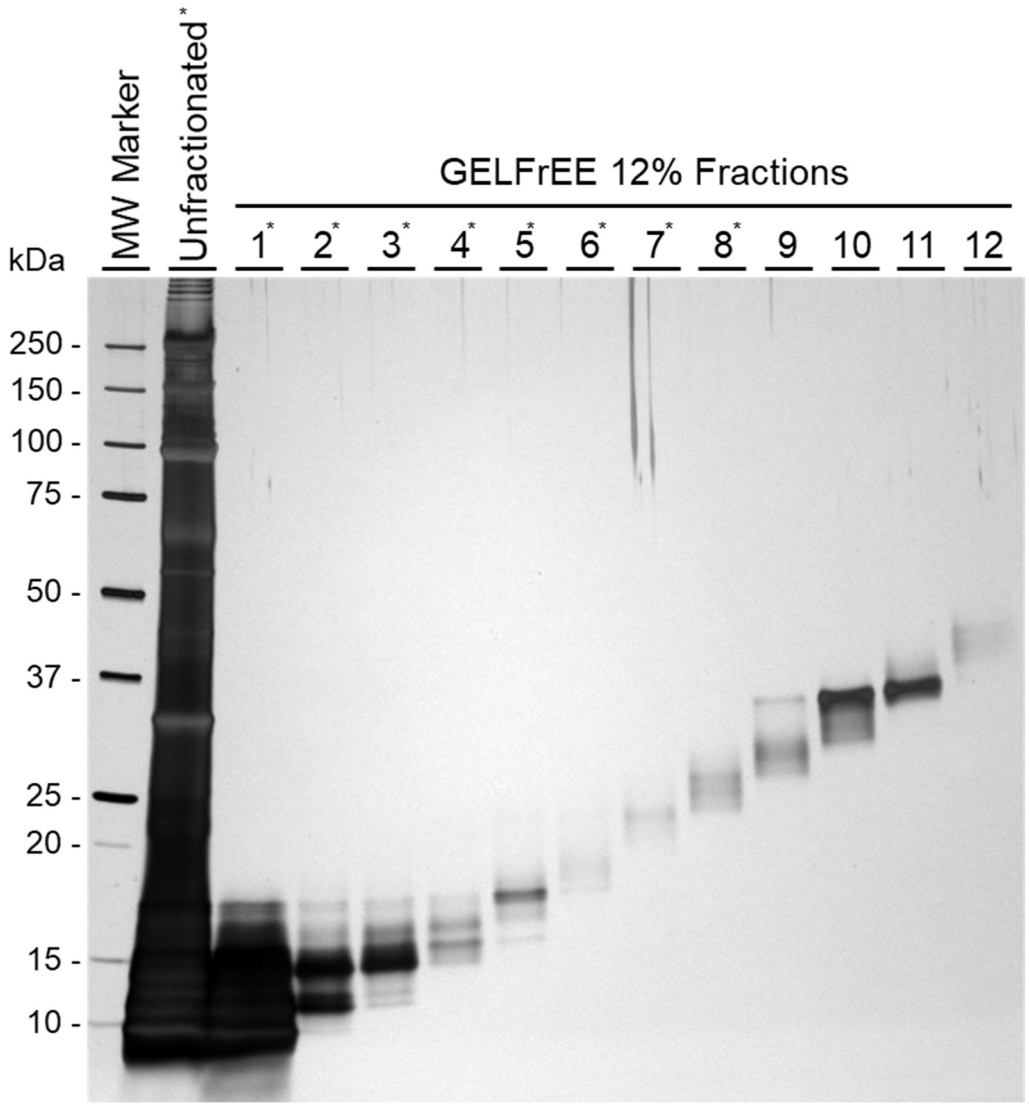
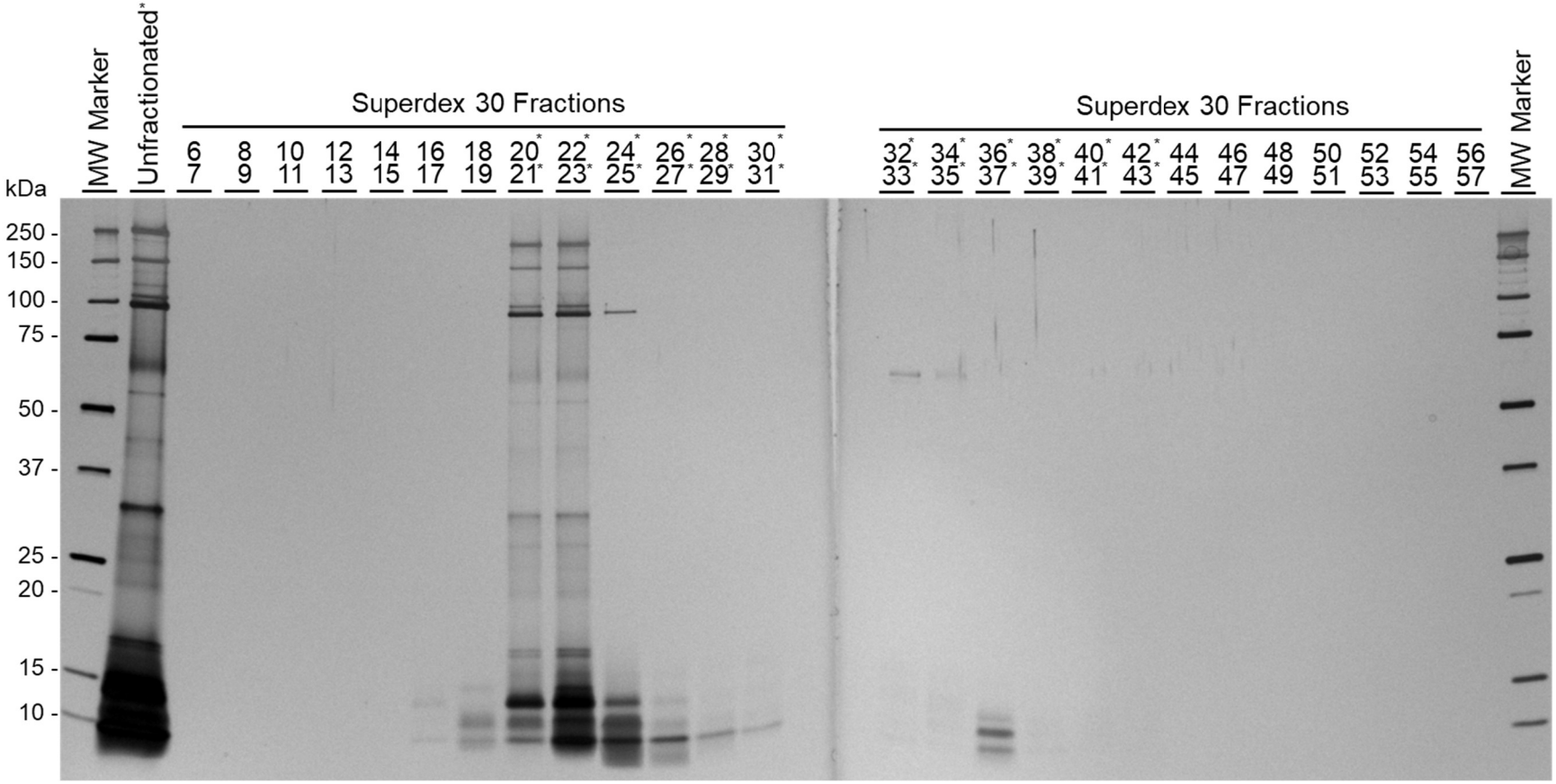


Figure 2.5: Silver staining of C8-extracted proteins separated by a GELFrEE device with a 12% cartridge. Proteins from each fraction were separated by SDS-PAGE and visualized by silver staining. Unfractionated material and fractions 1-8 (marked by *) containing proteins less than 25 kDa were analyzed by LC-MS/MS. MW, molecular weight. kDa, kilodalton.

Figure 2.6: Silver staining of C8-extracted proteins separated by size-exclusion chromatography using a Superdex 30 column. Fractions were combined pairwise. Proteins from each pair of fractions were separated by SDS-PAGE and visualized by silver staining. Unfractionated material and fractions 20-43 (marked by *) containing proteins less than 25 kDa were analyzed by LC-MS/MS. MW, molecular weight. kDa, kilodalton.


MW Marker
Unfractionated*
Superdex 30 Fractions
6 7
8 9
10 11
12 13
14 15
16 17
18 19
20* 21*
22* 23*
24* 25*
26* 27*
28* 29*
30* 31*
Superdex 30 Fractions
32* 33*
34* 35*
36* 37*
38* 39*
40* 41*
42* 43*
44 45
46 47
48 49
50 51
52 53
54 55
56 57
MW Marker
kDa
250 -
150 -
100 -
75 -
50 -
37 -
25 -
20 -
15 -
10 -

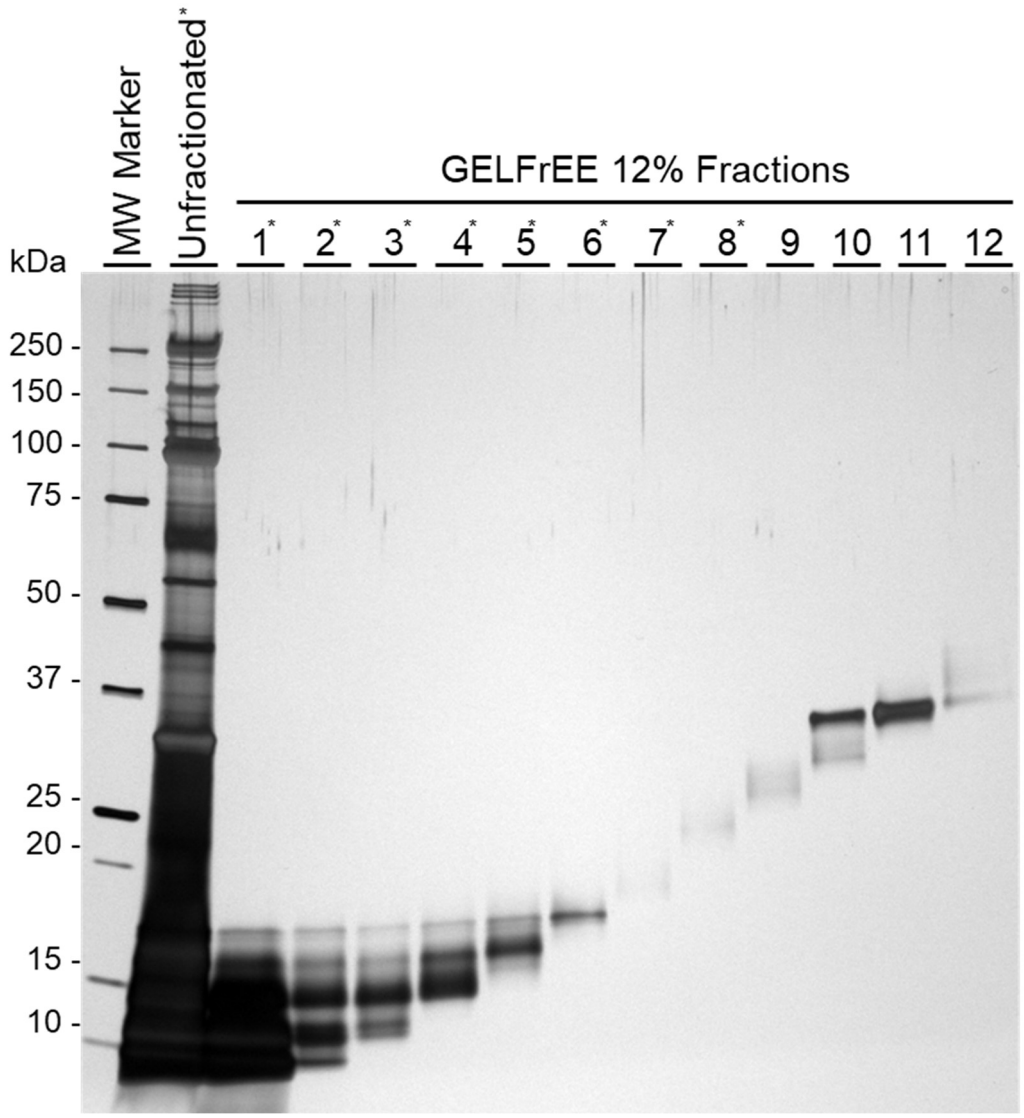


Figure 2.7: Silver staining of C18-extracted proteins separated by a GELFrEE device with a 12% cartridge. Proteins from each fraction were separated by SDS-PAGE and visualized by silver staining. Unfractionated material and fractions 1-8 (marked by *) containing proteins less than 25 kDa were analyzed by LC-MS/MS. MW, molecular weight. kDa, kilodalton.

Figure 2.8: Silver staining of C18-extracted proteins separated by size-exclusion chromatography using a Superdex 30 column. Fractions were combined pairwise. Proteins from each pair of fractions were separated by SDS-PAGE and visualized by silver staining. Unfractionated material and fractions 20-43 (marked by *) containing proteins less than 25 kDa were analyzed by LC-MS/MS. MW, molecular weight. kDa, kilodalton.

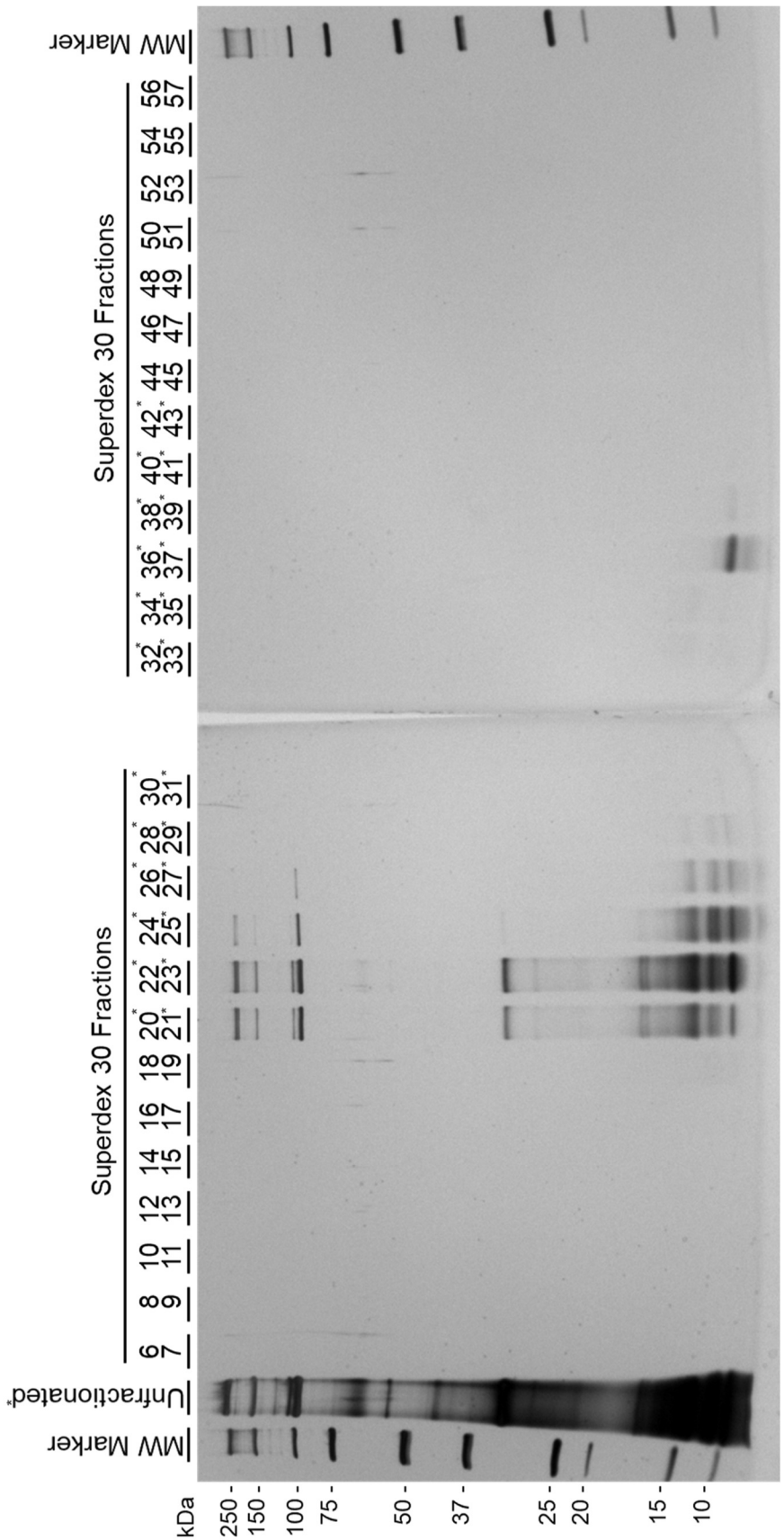

MW Marker
Unfractionated*
Superdex 30 Fractions
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20* 21* 22* 23* 24* 25* 26* 27* 28* 29* 30* 31*
Superdex 30 Fractions
32* 33* 34* 35* 36* 37* 38* 39* 40* 41* 42* 43* 44 45 46 47 48 49 50 51 52 53 54 55 56 57
MW Marker
kDa
250
150
100
75
50
37
25
20
15
10

Figure 2.9: Elution chromatogram of C8-extracted proteins separated by size-exclusion chromatography using a Superdex 30 column. Protein elution from the column was monitored by an in-line ultraviolet detector at 280 nm. Fractions (red) were combined in a pairwise manner. Void volume, V_0. Total volume, V_t.

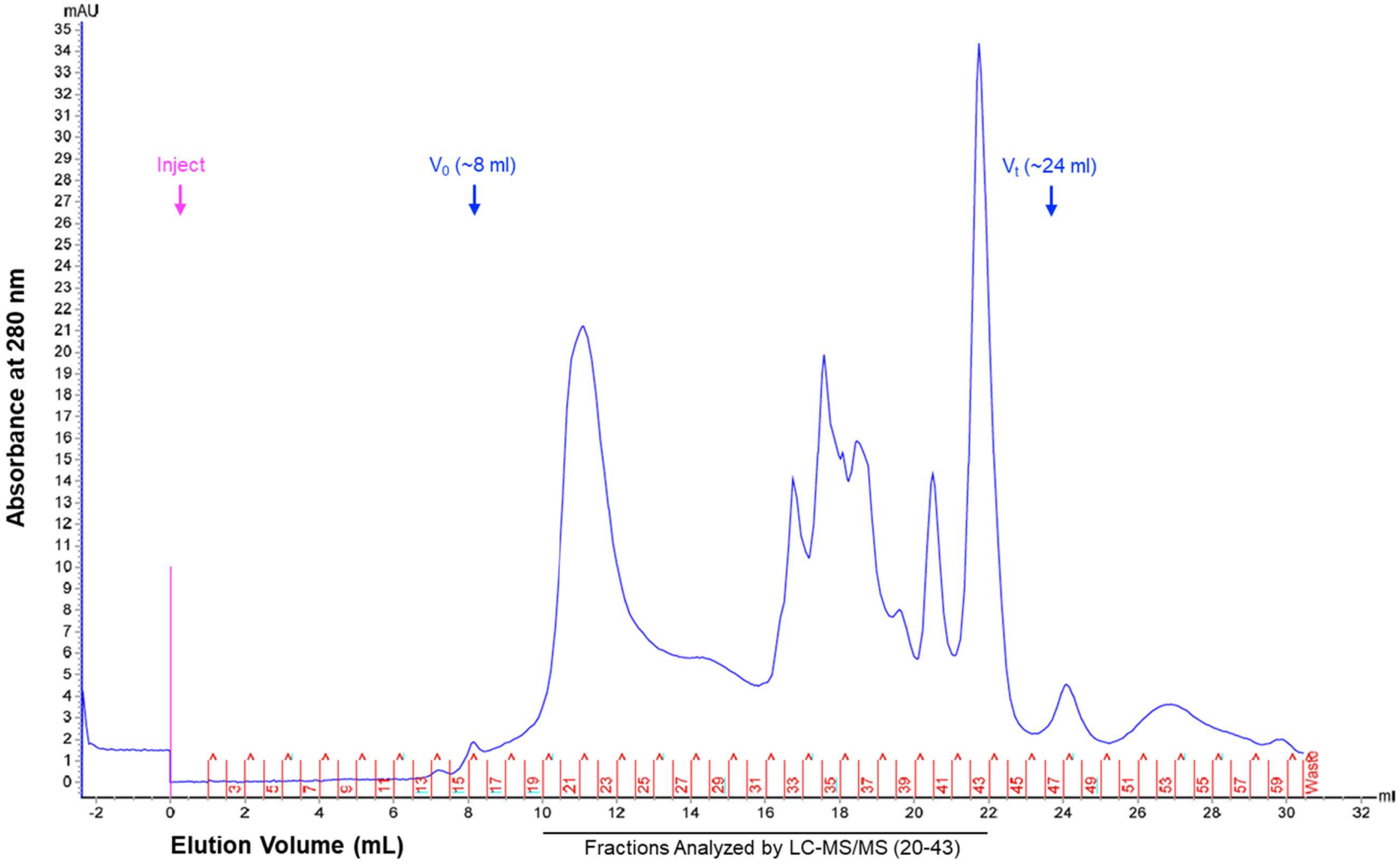
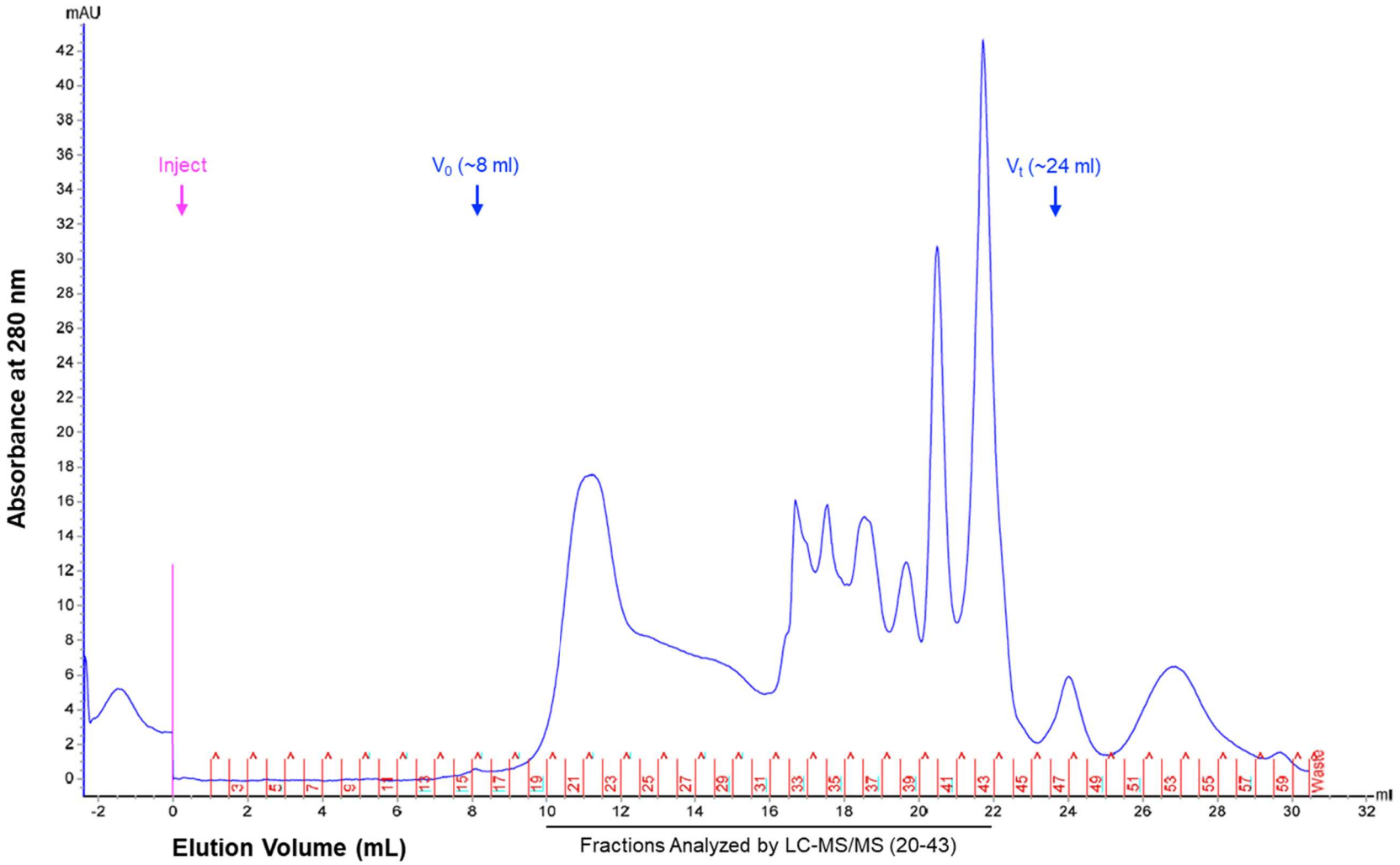
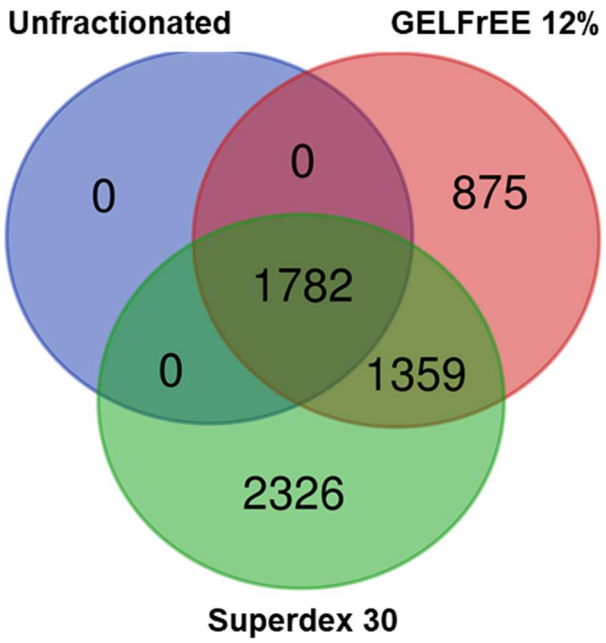

mAU
Absorbance at 280 nm
Inject
V_0 (~8 ml)
V_t (~24 ml)
Elution Volume (mL)
Fractions Analyzed by LC-MS/MS (20-43)
ml
Waste

Figure 2.10: Elution chromatogram of C18-extracted proteins separated by size-exclusion chromatography using a Superdex 30 column. Protein elution from the column was monitored by an in-line ultraviolet detector at 280 nm. Fractions (red) were combined in a pairwise manner. Void volume, V_0. Total volume, V_t.


mAU
Absorbance at 280 nm
Inject
V_0 (~8 ml)
V_t (~24 ml)
Elution Volume (mL)
Fractions Analyzed by LC-MS/MS (20-43)
ml
Waste

(A)



(B)

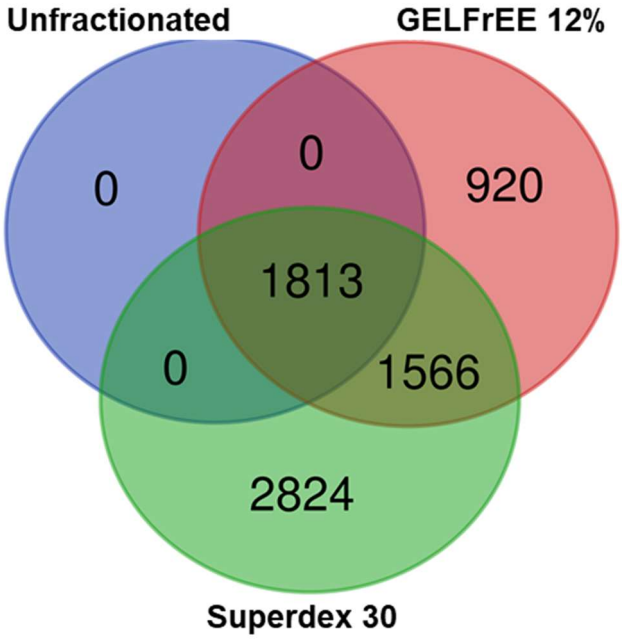


Figure 2.11: Comparison of fractionation methods on UniProt protein identifications in mouse brain. (A) C8-extracted proteins and **(B)** C18-extracted proteins were fractionated by GELFrEE or Superdex 30 and analyzed by LC-MS/MS.

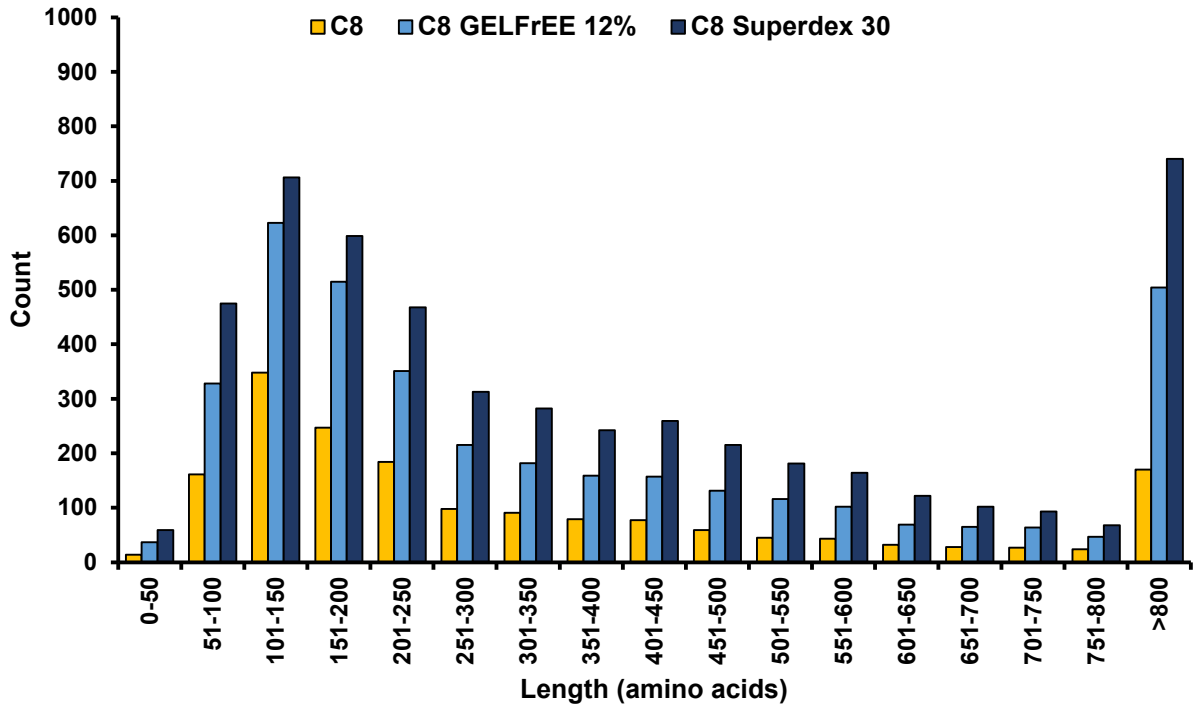


Figure 2.12: Length distribution of UniProt-annotated mouse brain protein identifications after C8 protein extraction and fractionation shows that length distribution does not skew input material. Lengths of canonical sequences were retrieved from the UniProtKB database and plotted in 50 amino-acid length windows. Unfractionated C8-extracted proteins, yellow bars and correspond to the yellow bars in Figure 2.2. C8-extracted proteins fractionated by GELFrEE 12%, light blue bars. C8-extracted proteins fractionated by Superdex 30, dark blue bars.

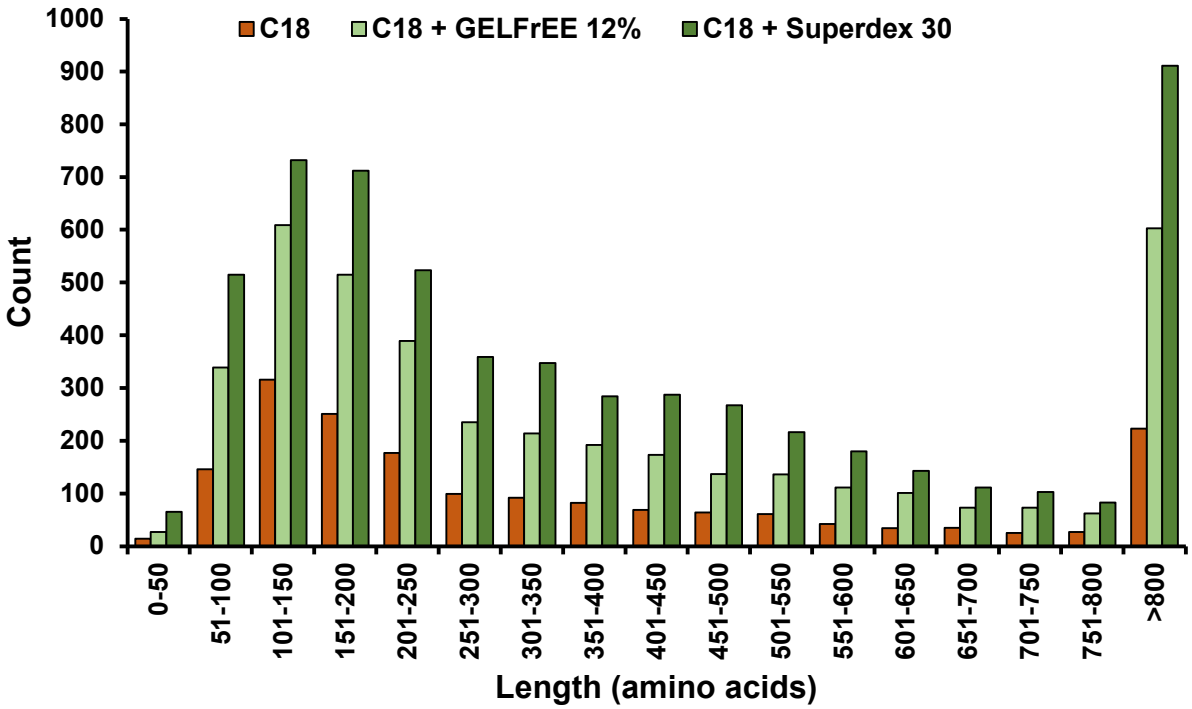
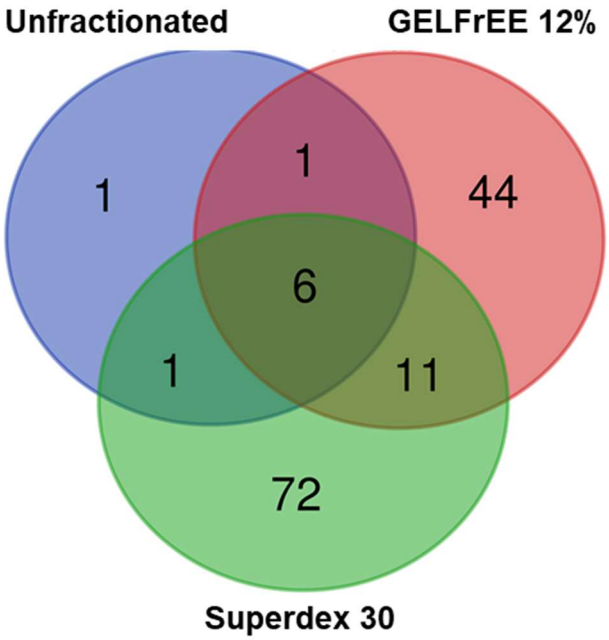


Figure 2.13: Length distribution of UniProt-annotated mouse brain protein identifications after C18 protein extraction and fractionation. Lengths of canonical sequences were retrieved from the UniProtKB database and plotted in 50 amino-acid length windows. Unfractionated C18-extracted proteins, orange bars and correspond to the orange bars in Figure 2.2. C18-extracted proteins fractionated by GELFrEE 12%, light green bars. C18-extracted proteins fractionated by Superdex 30, dark green bars.

(A)



(B)

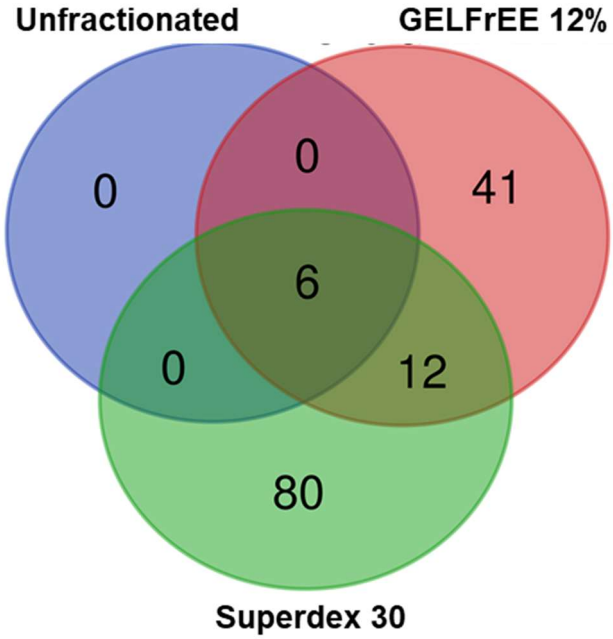


Figure 2.14: Comparison of fractionation methods on non-UniProt ORF identifications in the mouse brain. (A) C8-extracted proteins and **(B)** C18-extracted proteins were fractionated by GELFrEE or Superdex 30 and analyzed by LC-MS/MS. Tryptic peptides were mapped onto RNA transcripts, from which ORFs were annotated.

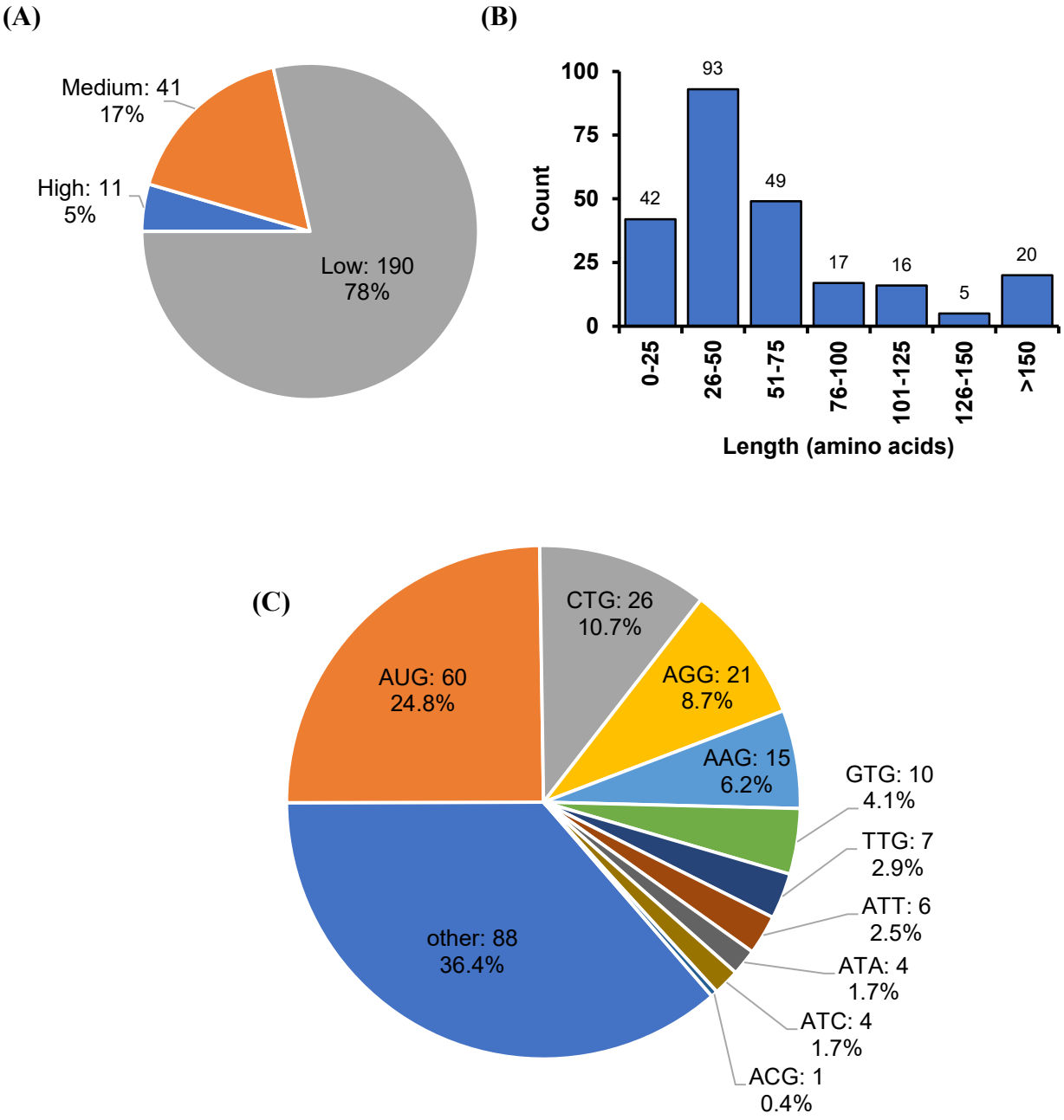
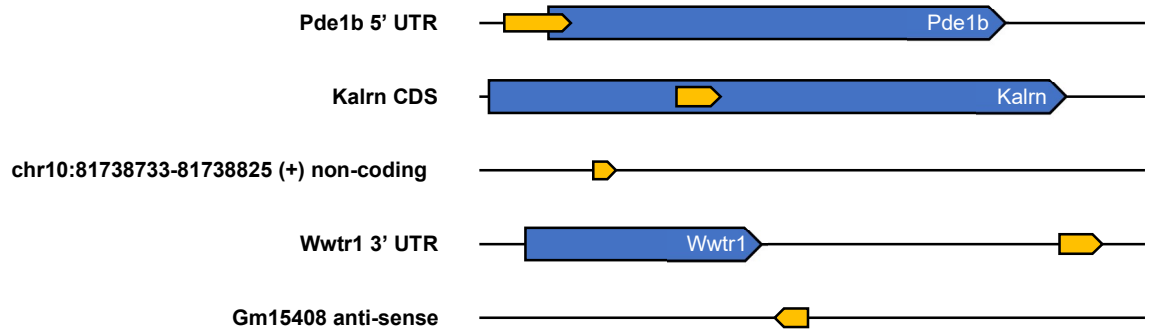


Figure 2.15: Features of 242 non-UniProt ORF identifications in the mouse brain. (A) Confidence of detection. Non-UniProt ORFs were assigned to one of three confidence categories: (1) High, >1 non-UniProt peptide detected or >10 peptide spectrum matches, (2) Medium, 1 non-UniProt peptide detected and between 2 and 10 peptide spectrum matches, (3) Low, 1 non-UniProt peptide detected and 1 peptide spectrum match. **(B)** Length distribution of non-UniProt ORFs annotated with the furthest upstream start codon in the transcript. **(C)** Distribution of possible start codons of non-UniProt ORFs.

(A)



(B)

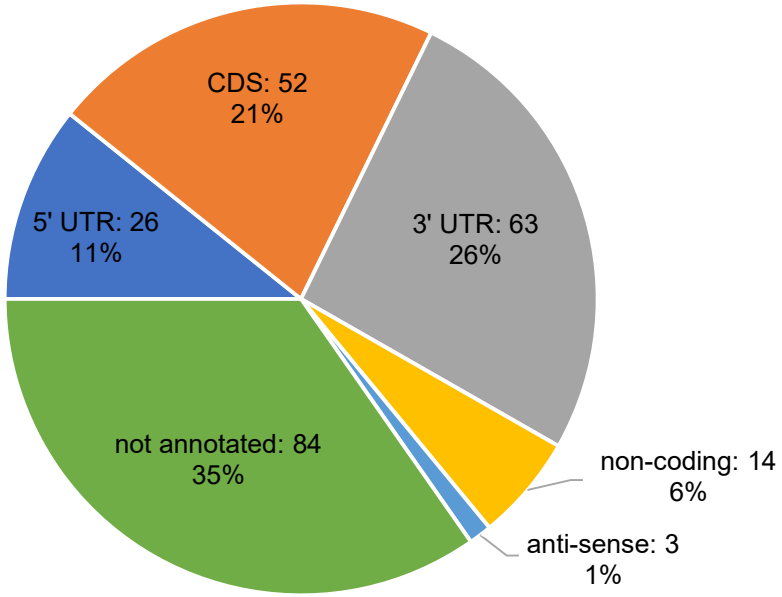
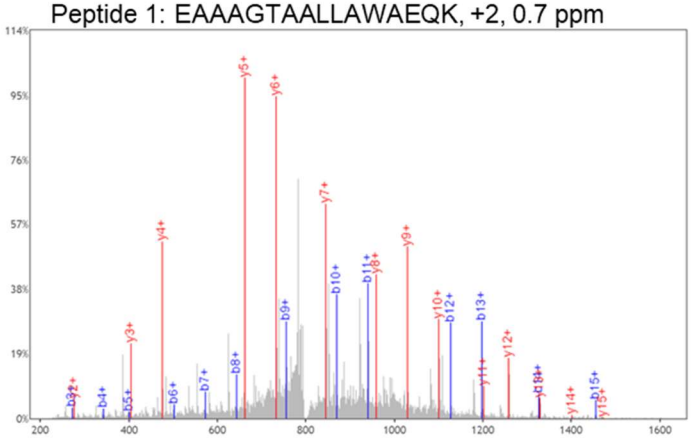


Figure 2.16: Transcript locations of 242 non-UniProt ORFs. (A) Examples of ORFs localized to the 5' untranslated region (UTR), coding sequence (CDS), non-coding RNA, 3' UTR, and anti-sense. RNA transcripts represented by thin black lines in the 5' to 3' direction. Non-UniProt ORFs, yellow arrows. Annotated CDS, blue arrows. Not drawn to scale. **(B)** Distribution of transcript locations of 242 non-UniProt ORFs.

Figure 2.17: Tryptic peptides identified in Pde1b Upstream Open Reading Frame (PDURF). (A) The location of three tryptic peptides were identified in PDURF. Annotated mass spectra and list of *b*- and *y*-ions identified in peptides **(B)** EAAAGTAALLAWAEQK, **(C)** QPQTVAEHGAVPPQSSR, **(D)** DAGVGLPVTPGAEVSPQQENVD. Precursor ion charge states and precursor ion mass tolerance for each peptide are also listed. Parts per million, ppm. Chromosome, chr. RNA transcript represented by a thin black line in the 5' to 3' direction. Peptides detected, black vertical lines. PDURF, yellow arrow. Pde1b, blue arrow. Not drawn to scale.

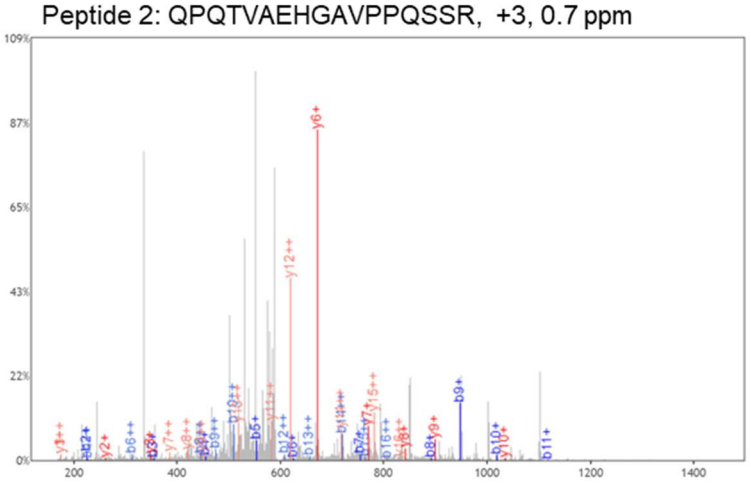
(A)

(B)



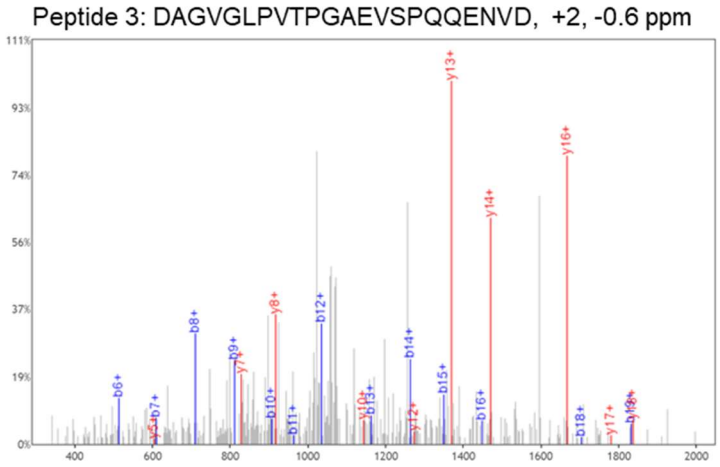
b+	#	Seq	#	y+
130.0499	1	E	16	
201.0870	2	A	15	1471.7904
272.1241	3	A	14	1400.7532
343.1612	4	A	13	1329.7161
400.1827	5	G	12	1258.6790
501.2304	6	T	11	1201.6575
572.2675	7	A	10	1100.6099
643.3046	8	A	9	1029.5728
756.3886	9	L	8	958.5356
869.4727	10	L	7	845.4516
940.5098	11	A	6	732.3675
1126.5891	12	W	5	661.3304
1197.6262	13	A	4	475.2511
1326.6688	14	E	3	404.2140
1454.7274	15	Q	2	275.1714
	16	K	1	147.1128

(C)



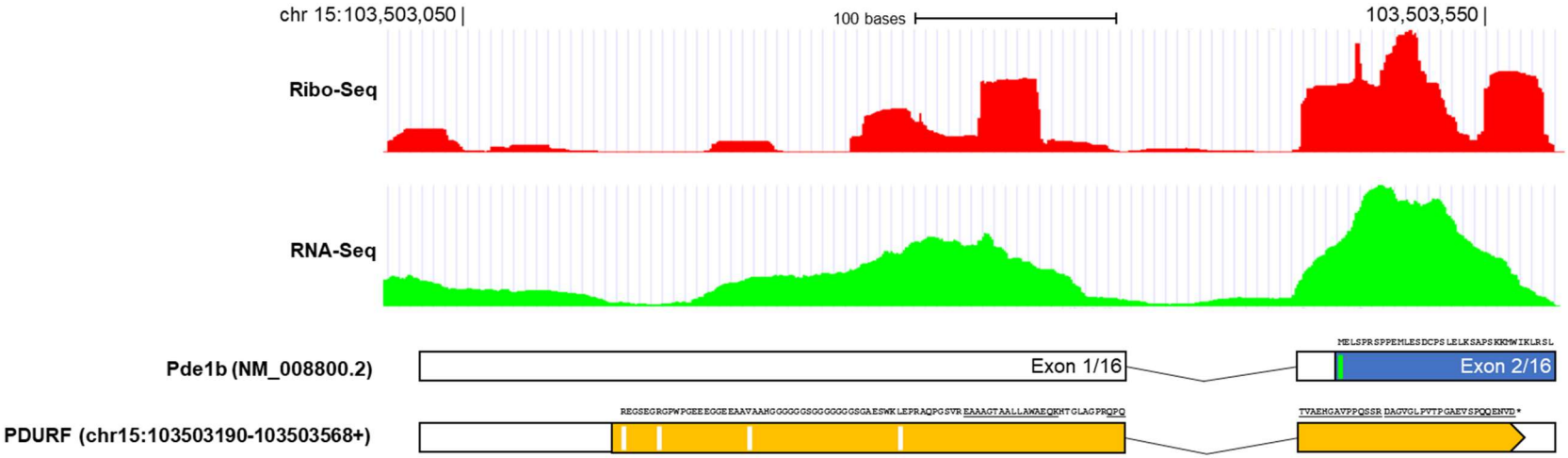
b+	b2+	#	Seq	#	y+	y2+
129.0659	65.0366	1	Q	17		
226.1186	113.5629	2	P	16	1660.8402	830.9237
354.1772	177.5922	3	Q	15	1563.7874	782.3973
455.2249	228.1161	4	T	14	1435.7288	718.3680
554.2933	277.6503	5	V	13	1334.6811	667.8442
625.3304	313.1688	6	A	12	1235.6127	618.3100
754.3730	377.6901	7	E	11	1164.5756	582.7914
891.4319	446.2196	8	H	10	1035.5330	518.2701
948.4534	474.7303	9	G	9	898.4741	449.7407
1019.4905	510.2489	10	A	8	841.4526	421.2300
1118.5589	559.7831	11	V	7	770.4155	385.7114
1215.6117	608.3095	12	P	6	671.3471	336.1772
1312.6644	656.8359	13	P	5	574.2944	287.6508
1440.7230	720.8651	14	Q	4	477.2416	239.1244
1527.7550	764.3812	15	S	3	349.1830	175.0951
1614.7871	807.8972	16	S	2	262.1510	131.5791
		17	R	1	175.1190	88.0631

(D)



b+	#	Seq	#	y+
116.0342	1	D	22	
187.0713	2	A	21	2064.0244
244.0928	3	G	20	1992.9873
343.1612	4	V	19	1935.9658
400.1827	5	G	18	1836.8974
513.2667	6	L	17	1779.8759
610.3195	7	P	16	1666.7919
709.3879	8	V	15	1569.7391
810.4356	9	T	14	1470.6707
907.4884	10	P	13	1369.6230
964.5098	11	G	12	1272.5703
1035.5469	12	A	11	1215.5488
1164.5895	13	E	10	1144.5117
1263.6579	14	V	9	1015.4691
1350.6900	15	S	8	916.4007
1447.7427	16	P	7	829.3686
1575.8013	17	Q	6	732.3159
1703.8599	18	Q	5	604.2573
1832.9025	19	E	4	476.1987
1946.9454	20	N	3	347.1561
2046.0138	21	V	2	233.1132
	22	D	1	134.0448

Figure 2.18: Ribosomal occupancy of PDURF in mouse brain. Screenshot taken from GWIPS-Viz Genome Browser[38] at the Pde1b locus on mouse chromosome 15. Approximate genomic coordinates are shown. Relative heights of bars represent sequencing reads from ribosome-profiling (Ribo-Seq, red) and RNA sequencing (RNA-Seq, red). Approximate location of Pde1b transcript (NM_008800.2) exons and introns are shown. The model of PDURF exon and intron structure is shown. Untranslated regions are depicted in a solid black outline. The coding sequence of Pde1b (blue) and PDURF (yellow) are shown. Peptide sequences are shown. Underline indicates the detected sequence. AUG translation start site, green vertical line. Near-cognate translation start site, vertical white line.


chr 15:103,503,050 |
100 bases
103,503,550 |
Ribo-Seq
RNA-Seq
Pde1b (NM_008800.2)
Exon 1/16
Exon 2/16
PDURF (chr15:103503190-103503568+)

2.8 References

1. Burge, C. B. & Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8,** 346–354 (1998).

2. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: Beautiful needles in the haystack. *Genome Res.* **7,** 768–771 (1997).

3. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. Life with 6000 Genes. *Science* **274,** 546–567 (1996).

4. Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L. & Grimmond, S. M. The abundance of short proteins in the mammalian proteome. *PLoS Genet.* **2,** 515–528 (2006).

5. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. & Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

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

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

8. Burbach, J. P. H. Neuropeptides and cerebrospinal fluid. *Ann. Clin. Biochem.* **19,** 269–277 (1982).

9. Liu, L. & Duff, K. A technique for serial collection of cerebrospinal fluid from the cisterna magna in mouse. *J. Vis. Exp.* 10–12 (2008). doi:10.3791/960

10. Vale, W., Vaughan, J., Yamamoto, G., Bruhn, T., Douglas, C., Dalton, D., Rivier, C. & Rivier, J. Assay of corticotropin releasing factor. *Methods Enzymol.* **103,** 565–577 (1983).

11. Vale, W., Vaughan, J., Jolley, D., Yamamoto, G., Bruhn, T., Seifert, H., Perrin, M., Thorner, M. & Rivier, J. Assay of growth hormone-releasing factor. *Methods Enzymol.* **124,** 389–401 (1986).

12. Herraiz, T. & Casal, V. Evaluation of solid-phase extraction procedures in peptide analysis. *J. Chromatogr. A* **708,** 209–221 (1995).

13. Navare, A., Zhou, M., McDonald, J., Noriega, F. G., Sullards, M. C. & Fernandez, F. M. Serum biomarker profiling by solid-phase extraction with particle-embedded micro tips and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass*

Spectrom. **22,** 997–1008 (2008).

14. Moremen, K. W., Tiemeyer, M. & Nairn, A. V. Vertebrate protein glycosylation: Diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.* **13,** 448–462 (2012).

15. Levery, S. B., Steentoft, C., Halim, A., Narimatsu, Y., Clausen, H. & Vakhrushev, S. Y. Advances in mass spectrometry driven O-glycoproteomics. *Biochim. Biophys. Acta - Gen. Subj.* **1850,** 33–42 (2015).

16. Jedrychowski, M. P., Wrann, C. D., Paulo, J. A., Gerber, K. K., Szpyt, J., Robinson, M. M., Nair, K. S., Gygi, S. P. & Spiegelman, B. M. Detection and quantitation of circulating human irisin by tandem mass spectrometry. *Cell Metab.* **22,** 734–740 (2015).

17. Wang, H., Chang-Wong, T., Tang, H. Y. & Speicher, D. W. Comparison of extensive protein fractionation and repetitive LC-MS/MS analyses on depth of analysis for complex proteomes. *J. Proteome Res.* **9,** 1032–1040 (2010).

18. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201 (2004).

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

20. Nesvizhskii, A. I. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11,** 1114–1125 (2014).

21. Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., Gingeras, T. R., Ecker, J. R. & Snyder, M. P. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 17224–17229 (2014).

22. Chu, Q., Martinez, T. F., Novak, S. W., Donaldson, C. J., Tan, D., Vaughan, J. M., Chang, T., Diedrich, J. K., Andrade, L., Kim, A., Zhang, T., Manor, U. & Saghatelian, A. Regulation of the ER stress response by a mitochondrial microprotein. *Nat. Commun.* **10,** 1–13 (2019).

23. He, L., Diedrich, J., Chu, Y. Y. & Yates, J. R. Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. *Anal. Chem.* **87,** 11361–11367 (2015).

24. Xu, T., Park, S. K., Venable, J. D., Wohlschlegel, J. A., Diedrich, J. K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X., Wong, C. C. L., Fonslow, B., Delahunty, C., Gao, Y., Shah, H. & Yates, J. R. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteomics* **129,** 16–24 (2015).

25. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).

26. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** 61–65 (2007).

27. Tagore, D. M., Nolte, W. M., Neveu, J. M., Rangel, R., Guzman-Rojas, L., Pasqualini, R., Arap, W., Lane, W. S. & Saghatelian, A. Peptidase substrates via global peptide profiling. *Nat. Chem. Biol.* **5,** 23–25 (2009).

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

29. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. & Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

30. Wang, Y., Wang, M., Yin, S., Jang, R., Wang, J., Xue, Z. & Xu, T. NeuroPep: A comprehensive resource of neuropeptides. *Database* **2015,** 1–9 (2015).

31. Kim, Y., Bark, S., Hook, V. & Bandeira, N. NeuroPedia: Neuropeptide database and spectral library. *Bioinformatics* **27,** 2772–2773 (2011).

32. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47,** D506–D515 (2019).

33. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).

34. Cho, J., Yu, N.-K., Choi, J.-H., Sim, S.-E., Kang, S. J., Kwak, C., Lee, S.-W., Kim, J., Choi, D. Il, Kim, V. N. & Kaang, B.-K. Multiple repressive mechanisms in the hippocampus during memory formation. *Science* **350,** 82–87 (2015).

35. Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11269–11274 (2004).

36. Abastado, J. P., Miller, P. F., Jackson, B. M. & Hinnebusch, A. G. Suppression of ribosomal reinitiation at upstream open reading frames in amino acid-starved cells forms the basis for GCN4 translational control. *Mol. Cell. Biol.* **11,** 486–496 (1991).

37. Chen, J., Brunner, A. D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D. & Weissman, J. S. Pervasive functional translation of noncanonical human open reading frames. *Science* **367,** 140–146 (2020).

38. Michel, A. M., Fox, G., M. Kiran, A., De Bo, C., O'Connor, P. B. F., Heaphy, S. M., Mullan, J. P. A., Donohue, C. A., Higgins, D. G. & Baranov, P. V. GWIPS-viz: Development of a ribo-seq genome browser. *Nucleic Acids Res.* **42,** 859–864 (2014).

Chapter 3

Integrated proteogenomics strategy for the identification of secreted peptides and small proteins

3.1 Abstract

Peptide hormones, neurotransmitters, and neuropeptides are examples of secreted peptides with important signaling roles in regulating growth, development, and function of the human body and brain. While several hundred peptide hormones and neuropeptides have been discovered using classical biochemical approaches, recent advances in next-generation sequencing, bioinformatics, and proteomics have revealed the existence of hundreds of small open reading frames (smORFs). These smORFs may encode secreted peptide hormones or neuropeptides that have not yet been discovered. In this chapter, I will describe the application of an integrated proteogenomics strategy to identify secreted peptides and small proteins in the conditioned media of human cell lines and cerebrospinal fluid (CSF). We identified 5,923 proteins in HEK293T cells, 3,613 proteins in HEK293T conditioned media, and 1,272 proteins in CSF. We also found evidence for translation products from an additional 31 unannotated smORFs. These findings show that proteogenomics strategies can also be applied to extracellular fluids to detect secreted peptides.

3.2 Introduction

The discovery and characterization of the Tarsal-less or polished rice (Tal/Pri) gene revealed an emerging class of protein-coding peptides and small proteins derived from small open reading frames (smORFs)[1,2]. Classical bioactive peptides (i.e., insulin, glucagon, etc.) are

synthesized as a longer prohormone sequence and then processed into a shorter peptides prior to secretion[3,4]. By contrast, Tal/Pri is translated as an 11-amino acid peptide that operates within the cell without any known processing[5,6]. Since Tal/Pri is not secreted, it does not require a signal peptide and therefore does not need to be processed from a longer protein sequence.

Intriguingly, Tal/Pri was previously unannotated before its discovery because it falls below the length cutoff of traditional algorithms that search the genome for protein-coding genes. Initial analysis of genomes revealed a large number of open reading frames (ORFs) that were less than 100 codons long (i.e., smORFs). For example, the yeast genome, which contains approximately 5,000 genes, revealed an additional 260,000 protein-coding genes when simply scanning the genome for in-frame start and stop codons. Few of these genes were conserved, and plotting the number of smORFs as a function of length led to a distribution that looked like the theoretical decay for a randomly assembled set of codons, which led to the conclusion that many of these "genes" may simply be noise. As a result, algorithms designed to identify protein-coding genes from set a lower length limit of 100 codons to reduce the numbers of false positives in the data, but still acknowledging that some critical and functional protein-coding smORFs will be missed.

With the dramatic improvements in mass spectrometry proteomics and next-generation sequencing, it is now possible to identify translated smORFs empirically. The expressed peptides and small proteins generated from smORFs have been referred to as smORF-encoded polypeptides (SEPs), micropeptides, and microproteins. We use the microprotein terminology because these smORFs are translated in the same manner as ORFs. As more smORFs and microproteins have been found, the number with defined biological roles has grown as well. Several smORFs that encode peptides have been identified with roles in muscle biology,

including the peptide minion, which is necessary for the proper fusion of muscle cells into multinucleated fibers[7]. CYREN is another newly characterized microprotein that regulates DNA repair pathway choice during the cell cycle by inhibiting non-homologous end joining repair to favor the higher fidelity homology-directed repair (HDR)[8]. Lastly, PIGBOS, an outer mitochondrial transmembrane microprotein, interacts with the endoplasmic reticulum to regulate the unfolded protein response[9].

As a newly discovered protein-coding genes, there are still many unanswered questions about smORFs. Of interest to us it to ask how many smORFs (known or unknown) that produced microproteins that are prepropeptides that are processed in the secretory pathway into smaller peptides before expulsion from the cell. And, if we could detect any such peptides, can we deduce the processed form of the peptide for subsequent biological studies. Detecting and characterizing the processed form of a microprotein, even a known microprotein, would be vital because it would indicate that smORFs can encode peptide hormones or neuropeptides, spurring a broader effort to find and characterize these smORFs.

3.3 Methods and Materials

Plasmids and cell culture

All plasmids were purchased from GenScript in a pcDNA3.1+ vector with a single FLAG-tag sequence (DYKDDDDK) inserted before the stop codon. A cDNA clone of MSMP (OHu16861) was used. MEGA and MWIA cDNA clones were synthesized. HEK293T (Dharmacon), HeLa, and U2OS cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM, Corning) supplemented with 10% (v/v) fetal bovine serum (FBS, Corning) in a humidified 37 °C incubator with 5% CO_2 atmosphere. Cell culture treated plastic dishes

(Corning) were used. Cells were sub-passaged every 3 days when 90% confluence was reached using 0.25% trypsin-EDTA (Life Technologies). Cells were transfected in a 6-well dish at 80% confluence. A transfection mixture of 2 µg plasmid in 50 µL Opti-MEM (Life Technologies) was combined with 5 µL Lipofectamine 2000 (Life Technologies) and 45 µL Opti-MEM. After a 10-minute incubation at room temperature, the transfection mixture was added dropwise to cells. Cells were incubated with transfection mixture at 37 °C for 4-6 hours and subsequently re-plated in 10 cm dishes. After two days, cell culture medium was replaced with serum-free DMEM without phenol red (Corning). Three days after replacing the media, conditioned medium was decanted, passed through a 0.22 µm syringe filter to remove residual cells, supplemented with 1× protease inhibitors (Roche), and stored on ice. Cells were collected by pipetting in ice-cold phosphate-buffered saline (PBS), pelleted by centrifugation at 200 × *g* for 3 minutes at 4 °C, washed once more with PBS, and stored on ice.

Confocal imaging

Confocal imaging was performed as previously described[9]. HeLa and U2OS cells were grown on glass coverslips (Fisher Scientific) placed in a 6-well dish. Cells were transfected as described above. After 48 hours, cells were fixed with 4% paraformaldehyde (Polysciences) and permeabilized with 1% saponin (Alfa Aesar). After blocking with 4% bovine serum albumin (Fisher Scientific) in PBS for 1 hour at room temperature, fixed and permeabilized cells were probed with mouse anti-FLAG M2 antibodies (1:1000, Sigma-Aldrich) overnight at 4 °C on an orbital platform shaker. Nuclei were counterstained with Hoescht 33258 (1:2000, Sigma-Aldrich). Following three washes with PBS, coverslips were incubated with anti-mouse Alexa-Fluor 488 antibody conjugates (1:5000, Thermo Fisher Scientific) for 1 hour at room

temperature. Following three washes with PBS, coverslips were mounted on glass microscope slides using ProLong Gold Antifade Mountant (Invitrogen). Samples were imaged on an inverted confocal microscope (Zeiss LSM 880 with AiryScan) using a 63×1.4NA oil-immersion objective using ZEN software.

Protein extraction and SDS-PAGE

Cells were lysed in IP lysis buffer (Pierce) supplemented with 1× Halt protease inhibitor cocktail (Thermo Fisher Scientific). The lysate was clarified by centrifugation at 20,000 × *g* for 10 minutes at 4 °C. Conditioned medim was filtered through 0.22 um PES syringe filters before being concentrated using 3000 molecular weight cutoff centrifugal filters (EMD Millipore). Protein concentration was measured using a BCA protein assay kit (Pierce) in a 96-well microassay plate using bovine serum albumin standards (Pierce). Samples were combined with 4× SDS-PAGE buffer (250 mM Tris-Cl pH 6.8, 8% (w/v) SDS, 0.2% (w/v) bromophenol blue, 40% (v/v) glycerol, and 10% (v/v) 2-mercaptoethanol) to a working concentration of 1×. Samples were subsequently incubated at 95 °C for 3 min and cooled to room temperature. Samples were loaded on a pre-cast Novex 4-12% bis-tris polyacrylamide gel in 1× MOPS running buffer (Novex) and electrophoresed at 200 V until the dye-front migrated to the bottom of the gel. For silver staining, proteins were visualized by staining using a Pierce silver stain kit according to the manufacturer's instructions and imaged on a Bio-Rad ChemiDoc XRS+ gel imaging system using a white light transilluminator. For immunoblotting, proteins were transferred to PVDF membranes (Life Technologies, 0.2 µm pore size) using an iBlot2 transfer device (Life Technologies) using the pre-programmed P0 setting (20 V for 1 min, 23 V for 4 min, 25 V for 2 min). Membranes were incubated at room temperature for 1 hour in PBS

blocking buffer (LI-COR) and probed with anti-DYKDDDDK rabbit antibodies (1:2000, Cell Signaling Technology) in blocking buffer at 4 °C overnight on an orbital platform shaker. Membranes were washed three times in TBST (20 mM Tris-Cl pH 7.6, 150 mM NaCl, 0.1% (v/v) Tween-20) and probed with goat-anti-rabbit fluorescent conjugates (1:10,000, LI-COR 800CW) in blocking buffer at room temperature for 45 minutes in the dark. Membranes were washed four times with TBST, and proteins were visualized on an Odyssey imaging system (LI-COR).

Polypeptide enrichment for LC-MS/MS analysis

Polypeptides from human cerebrospinal fluid (CSF) were enriched from the flow-through by solid-phase extraction using Agilent Bond Elut 1 gram C8 or C18 silica cartridges on a vacuum manifold. Cartridges were pre-wet with one column volume of methanol and equilibrated with one column volume of triethylammonium formate (TEAF) buffer, pH 3.0. Samples were applied to the pre-equilibrated column, washed with one column volume of TEAF, and eluted with 2 ml of a mixture of 75% (v/v) acetonitrile and 25% (v/v) TEAF. Samples were evaporated to dryness in a vacuum centrifuge overnight and stored at -20 °C until further processing.

LC-MS/MS analysis

LC-MS/MS analysis was carried out as previously described[10]. Each sample was analyzed in duplicate. Samples were precipitated by methanol/chloroform and re-dissolved in 8 M urea/100 mM tetraethylammonium tetrahydroborate (TEAB), pH 8.5. Polypeptides were reduced with 5 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP, Sigma-Aldrich) and

alkylated with 10 mM chloroacetamide (Sigma-Aldrich). Polypeptides were digested overnight at 37 °C in 2 M urea/100 mM TEAB, pH 8.5, with trypsin (Promega). For disulfide linkage determination, chloroacetamide was omitted and replaced with 2 mM N-ethylmaleimide (Sigma-Aldrich). Digestion was quenched with formic acid, 5 % final concentration. The digested samples were analyzed on a Fusion Orbitrap tribrid mass spectrometer (Thermo Fisher Scientific). The digest was injected directly onto a 30 cm, 75 μm inner diameter column packed with BEH 1.7 μm C18 resin (Waters). Samples were separated at a flow rate of 300 nL/min on a nLC 1000 chromatography system (Thermo Fisher Scientific). Buffer A and B were 0.1% formic acid in water and 0.1% formic acid in 90% acetonitrile, respectively. A gradient of 1-35% B over 110 min, an increase to 50% B over 10 min, an increase to 90% B over 10 min and held at 90% B for a final 10 min was used for a 140 min total run time. The column was re-equilibrated with 20 μL of buffer A prior to the injection of the sample. Peptides were eluted directly from the tip of the column and nanosprayed directly into the mass spectrometer by application of 2.5 kV voltage at the back of the column. The Orbitrap Fusion was operated in a data-dependent mode. Full MS scans were collected in the Orbitrap at 120,000 resolution with a mass range of 400 to 1600 m/z and an automatic gain control (AGC) target of 5×10^5. The cycle time was set to 3 sec, and within this 3 sec, the most abundant ions per scan were selected for collision-induced dissociation MS/MS in the ion trap with an AGC target of 10^4 and a minimum intensity of 5000. Maximum fill times were set to 50 ms and 100 ms for MS and MS/MS scans, respectively. Quadrupole isolation at 1.6 m/z was used, monoisotopic precursor selection was enabled and dynamic exclusion was used with a duration of 5 sec.

Data processing of UniProt-annotated proteins

Data analysis was performed as previously described[10] with minor modifications. Tandem mass spectra data were extracted and deconvolved from .raw files using RawConverter[11] version 1.1.0.23 and searched using a target-decoy strategy with ProLuCID[12] and the Integrated Proteomics Pipeline version 6.0.5 (IP2, Integrated Proteomics Applications) data analysis platform. Samples from human CSF and HEK293T CM were searched against the UniProt human reference proteome (downloaded April 23, 2018). Common contaminant proteins and reverse decoy sequences were generated and appended to each database. The following search parameters were used: CID/HCD fragmentation mode, monoisotopic mass, 50 parts per million (ppm) precursor ion mass tolerance, 600 ppm fragment ion mass tolerance, 600-6000 mass range, trypsin was set as the enzyme, requiring at least one tryptic end, up to two missed cleavages allowed, and no differential modifications. Mass spectra matches were filtered to 10 ppm precursor ion mass tolerance and evaluated with DTASelect 2.0[13] using XCorr and Zscore as the primary and secondary score types, respectively. Protein identifications required at least one matched peptide per protein and were reported at a false discovery rate of 1%.

Data processing of unannotated proteins

Data analysis was performed as described previously[14] with minor modifications. The analysis procedure is the same as in the previous section, except that a custom protein database generated from the in silico translation in 3-frames of strand-specific RNA-Seq data from the spinal cord of male and female human embryos (ENCSR000AFH) and adult female brain cells (ENCSR274JRR) was used. The in-silico-translated database was generated as previously described[14]. Reverse decoy sequences were generated and appended to the database. The same search parameters were used, except that no differential modifications were specified. Peptide

sequences matching to reviewed and non-reviewed entries in the UniProt human reference proteome (downloaded on March 18, 2020), common laboratory contaminant proteins, and all reverse decoy sequences were filtered out using custom string-searching scripts. For the remaining peptide sequences, an RNA transcript was identified in the NCBI Mouse Reference Sequence (RefSeq) Database[15] using tBlastn. Where possible, known RefSeq transcript sequences ("NM_" prefix) were used for annotation. If no reviewed transcript was found, then transcripts from model transcript sequences ("XM_" prefix) were used. Open reading frames were annotated using the nearest in-frame start (AUG) and stop codons (UAG, UGA, UAA). For RNA sequences lacking an upstream AUG start codon, the furthest upstream near-cognate codon (ACG, AAG, CUG, etc.) in a Kozak sequence in the RNA transcript was assigned as the start codon. If a near-cognate start in a Kozak sequence was not identified, then one of three codons was assigned as the translation initiation site in the transcript: (1) the furthest upstream near-cognate codon, (2) the codon after the nearest upstream stop codon, or (3) the furthest upstream in-frame codon in the transcript. Open reading frames of less than 150 amino acids were designated as microproteins.

3.4 Results and Discussion

Characterization of MEGA and MWIA microproteins

To detect secreted small peptides and proteins that might function as hormones and neuropeptides, we wanted to develop an experimental workflow that could be applied to extracellular fluids. Our goal was to combine peptidomics techniques that can be used to enrich for peptides and small proteins from a sample and identify them with mass spectrometry-based proteomics techniques. To develop and optimize the workflow, we needed a secreted peptide or

small protein that could function as a positive control. To find such a secreted candidate, we considered two microproteins that were identified in a recent study from our research group that used ribosome profiling in human cell lines[16]. Both microproteins showed high phyloCSF scores, which indicates a high likelihood of conservation between mouse and human species. MEGA is a 48-amino acid microprotein (Figure 3.1), and MWIA is a 113-amino acid microprotein (Figure 3.2). The microprotein names are derived from the first four residues of their amino acid sequences in one-letter codes.

We analyzed both microprotein sequences using SignalP 4.1 and found that they had predicted signal peptides, which suggests that these microproteins are expressed in HEK293T cells and secreted. To confirm if these microproteins are expressed and secreted, we transfected HEK293T cells with plasmids expressing cDNA constructs of each microprotein with a single FLAG-tag immediately before the stop codon (Figure 3.3a). FLAG-tagged constructs have been used previously to validate microprotein expression[17] as antibodies to most newly discovered microproteins have not yet been developed. We then performed immunoprecipitation (IP) using anti-FLAG resin and performed an immunoblot (IB) to detect the presence of the FLAG-tagged microproteins in both cell lysates and conditioned medium at both 24 and 48 hours post-transfection. We could not detect the MEGA-FLAG microprotein in either the lysate or conditioned media samples, which suggests that this microprotein might not be expressed or is expressed at levels that are below the detection limit of our immunoblot assay (Figure 3.3b). We detected the MWIA-FLAG microprotein in cell lysates, but not in the conditioned media. These findings suggest that MEGA-FLAG is expressed but is not secreted or is secreted at a level that is below the detection limit of our immunoblot assay (Figure 3.3b).

To further confirm these findings, we performed confocal imaging of HeLa and U2OS cells that had been transfected with plasmids encoding MEGA-FLAG and MWIA-FLAG (Figure 3.4). MEGA-FLAG did not appear to be expressed in or secreted from either cell type. MWIA-FLAG appeared to form foci, which confirms that this microprotein is expressed. The foci might indicate that this microprotein is localized to secretory vesicles, but not secreted under basal conditions. Further experiments would be required to determine the subcellular localization and stimulus for secretion. These results confirm that neither MEGA-FLAG or MWIA-FLAG could be detected as secreted microproteins and that other candidates would have to be considered.

Characterization of MSMP microprotein

We searched for secreted microproteins candidates in the literature. Prostate-associated microseminoprotein (MSMP) is a 139-amino acid microprotein that highly expressed in the prostate cancer PC3 cell line[18] (Figure 3.5). We also found that MSMP has a 38-amino acid signal peptide sequence predicted by SignalP 4.1. MSMP functions as a chemokine in the immune response by binding to the CCR2 cell surface receptor on monocytes and stimulating chemotaxis[19]. Recombinant MSMP with C-terminal myc and his-tag epitopes was expressed in HEK293T cells and detected in both the cell lysate and conditioned medium[19]. We reasoned that MSMP with a C-terminal FLAG-tag might be a suitable secreted microprotein with which to develop a workflow to detect other secreted peptides and small proteins.

To test this, we transfected HEK293T cells with a plasmid carrying an MSMP-FLAG construct. We then performed an immunoprecipitation (IP) using anti-FLAG resin and performed an immunoblot (IB) to detect the presence of the FLAG-tagged microproteins in both cell lysates and conditioned media at both 24 and 48 hours post-transfection. MSMP-FLAG was detected in

both cell lysates and conditioned media. Secreted MSMP showed a slightly up-shifted band, which suggests that MSMP-FLAG is modified post-translationally. To confirm if MSMP-FLAG was secreted through the secretory pathway, rather than being released passively by cell lysis, we treated HEK293T cells transfected with an MSMP-FLAG construct with modulators of protein secretion. We found a dose-dependent increase of MSMP-FLAG in the conditioned medium with ionomycin stimulation[20], while treatment with Brefeldin A[21] reduced MSMP-FLAG secretion to below detectable levels (Figure 3.7). These findings confirm that MSMP-FLAG is targeted through the secretory pathway.

Development of a workflow to detect peptides and small proteins in HEK293T cells and conditioned media

In designing a workflow to detect secreted peptides and small proteins in extracellular fluids, we considered two technical challenges that usually obscure the number of identifications possible in a proteomics experiment: sample complexity and the limit of detection. To circumvent these issues, we designed a workflow that integrated peptide extraction techniques to enrich for peptides and small proteins and a fractionation step to reduce sample complexity. A previous study from our research group has shown that the use of molecular weight cutoff (MWCO) filters to enrich peptides and electrostatic repulsion hydrophilic interaction chromatography (ERLIC) increased the number of microproteins identified in K562 cells more effectively than using MWCO filters alone[14]. The complexity of peptide samples generated from biological tissues likely results in undersampling by the mass spectrometer. Indeed, the study went on to show that replicate measurements of the same biological sample showed a large proportion of microproteins detected in one of the replicates only. This finding supports the

observation that sample complexity is a major barrier to the identification of peptides and small proteins and that including a fractionation step is beneficial to maximize the number of peptides and small proteins identified in a sample.

Rather than use ERLIC fractionation, we decided to use gel-eluted liquid fraction entrapment electrophoresis (GELFrEE)[22] as a fractionation method. We used a cartridge with a fractionation range of 3.5 to 50 kDa range to better separate peptides and small proteins. With GELFrEE fractionation, smaller polypeptides elute in earlier fractions, and larger polypeptides elute in later ones, thereby also having the advantage of excluding fractions that contain larger polypeptides from downstream analysis. To test how well GELFrEE fractionation could be applied to microproteins, we transfected HEK293T cells with a plasmid encoding PSMP-FLAG as a positive control and marker for microprotein-containing fractions. We concentrated the conditioned medium using 3K MWCO filters and prepared lysates from cells and fractionated both cell lysates and conditioned media using GELFrEE (Figure 3.8). To determine the extent of GELFrEE fractionation, we performed SDS-PAGE followed by silver staining and detected MSMP-FLAG by immunoblotting (Figure 3.9). In both cell lysates and conditioned medium, we observed a good degree of separation between low and high molecular weight species across 12 fractions. In the conditioned media, a higher molecular weight species around 55 kDa is likely bovine serum albumin that was not completely removed. MSMP-FLAG appeared predominantly in the earlier fractions, suggesting that the microproteins could be fractionated away from larger proteins in the proteome. We then digested each fraction with trypsin and analyzed the digested peptides by liquid chromatography coupled to electrospray ionization and tandem mass spectrometry (LC-MS/MS).

To determine if GELFrEE fractionation was effective, we compared the number of proteins identified to an unfractionated sample. GELFrEE increased the number of proteins identified by nearly 1.5-fold (Figure 3.10). In cell lysates, 3,688 proteins were identified in the unfractionated sample. With GELFrEE fractionation, the number of proteins increased to 5,255. In conditioned media, 1,760 proteins were identified in the unfractionated sample. With GELFrEE fractionation, 3,457 proteins were identified. In the unfractionated cell lysates, 668 proteins were uniquely identified without fractionation, which suggests that proteins outside of the 3.5 to 50 kDa fractionation range might have been excluded from the fractions. Similarly, 156 proteins from conditioned medium were only identified in the unfractionated sample. Finally, our sequence database that was used to identify proteins also included unannotated microprotein sequences that were identified using ribosome profiling[16].

In all, we identified seven microproteins in HEK293T cells (Figure 3.10). There are several possibilities to explain this limited number of identifications: (1) the microproteins are at levels that are below the detection level of the mass spectrometers used in our experiments, (2) microproteins may be modified post-translationally in an unpredictable manner, (3) microproteins might have been degraded by proteases during sample concentration with MWCO filters or (4) the sequence database used did not contain the correct microprotein sequences. MSMP-FLAG was identified, and the number of spectral counts reflected the band intensity we observed during immunoblotting (Table 3). As MSMP-FLAG was produced from an overexpression vector, this finding indicates that other microproteins might exist at levels that are below the limit of detection in our experiments, thereby escaping detection. As we had already reached the loading capacity on the GELFrEE system, we had to consider alternative means to increasing the sample quantity analyzed in the workflow. These modifications included

using a sequence database generated from the in silico translation of mRNA expression data and using solid-phase extraction as a method to enrich peptides and proteins that had been used in a previously published study from our research group[10].

Table 3.1 Spectral counts of PSMP-FLAG and C4ORF48 detected in HEK293T conditioned media

		GELFrEE fraction (spectral counts)											
Protein	**Unf.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
PSMP-FLAG	18	79	64	81	101	48	40	25	12	5	7	5	11
C4ORF48	0	5	4	1	0	0	0	0	0	0	0	0	0

Unf., unfractionated

Development of a proteogenomic workflow to detect peptides and small proteins in human cerebrospinal fluid

To identify peptides and small proteins that might be functioning as neuropeptides or peptide hormones, we wanted to develop a peptidomics-based workflow and apply this workflow to human cerebrospinal fluid (Figure 3.11). Neuropeptides and peptide hormones are traditionally difficult to detect and characterize. We reasoned that advances in next-generation sequencing and mass spectrometry could be combined with traditional peptide enrichment techniques to identify new neuropeptides and peptide hormones. Based on our microprotein experiments with HEK293T cells, we decided to continue using GELFrEE fractionation because MSMP-FLAG was fractionated away from larger proteins and was identified by mass spectrometry.

To address the issue that peptides and small proteins might have degraded during sample preparation, we used solid-phase extraction (SPE) using C8 silica. SPE has the advantage of being carried out under acidic conditions, thereby reducing the activity of proteases that might be present in the sample. SPE can also serve to concentrate peptides from large volumes, thereby

further enriching for peptides and small proteins and excluding larger proteins. We then fractionated the C8-extracted material using GELFrEE, digested each fraction with trypsin, and analyzed the digested peptides using LC-MS/MS.

When we searched our mass spectrometry data using a protein sequence database containing the UniProt human reference proteome, we identified a total of 1,272 proteins (Supplementary Table 4). Of these, 585 were identified in the unfractionated sample, and this number increased to 1,236 with GELFrEE fractionation (Figure 3.12a). Silver staining of the GELFrEE fractions showed an effective separation of peptides and smaller proteins from larger ones (Figure 3.12b). Of the proteins identified, 39 had a known neuropeptide function, showing that our workflow was able to identify biologically relevant peptides in CSF (Supplementary Table 5).

Detection of non-UniProt ORFs

To identify peptides and small proteins that might function as neuropeptides in human cerebrospinal fluid, but that might not be annotated in the human proteome, we decided to use a proteogenomic approach (Figure 3.11). We searched the acquired mass spectrometry data against a sequence database from the in silico translation of publicly available RNA-Seq datasets. The RNA-Seq data was obtained from the spinal cord from male and female human embryos (ENCSR000AFH) and adult female brain cells (ENCSR274JRR) that are publicly available from the encyclopedia of DNA elements (ENCODE) consortium[23]. The strand-specific RNA-Seq data was translated in all three reading frames from stop codon to stop codon, thereby generating all theoretically possible translation products. In this manner, the translation of protein sequences that initiate at near-cognate start codons is also represented alongside those that use the cognate

ATG translation start codon. A similar type of protein sequence database has been used to discover unannotated microproteins in K562 cells[14,17].

In our dataset, we mapped tryptic peptides to 1,999 RNA transcripts found in the RefSeq database. We filtered out sequences that matched those found in the UniProt human reference proteome. For the remaining sequences, open reading frames (ORFs) were annotated using in-frame start (AUG) and stop (UAG, UGA, UAA) codons. If an upstream AUG was not found, then the furthest upstream near-cognate start codon was used instead. If an upstream near-cognate codon was not present, then the furthest upstream codon after a stop codon or the furthest upstream in-frame codon was designated as the translation start site. In all cases, we assigned the translation start site to the furthest upstream codon to avoid introducing biases towards shorter lengths. As a result, we identified 33 ORFs that were not annotated in UniProt (Supplementary Table 6). We identified two transcripts for which no downstream stop was present, and an ORF could not be annotated. These transcripts might have harbor downstream frameshift mutations that place the stop codon out-of-frame. Another three transcripts were excluded from further consideration as their peptides were in-frame with an existing ORF, as in the case of a 5' or 3' extension product from a splice isoform.

Features of non-UniProt ORFs (location, MW, start codon usage)

We assessed the confidence of the detection of the 33 non-UniProt ORFs identified in our dataset. We assigned each of the 33 non-UniProt ORFs into three categories: high confidence, medium confidence, and low confidence (Figure 3.13a). Higher confidence matches have more matched tryptic peptides and spectral counts, thereby reducing the likelihood of a match to a random transcript. The two non-UniProt ORFs in the high confidence category had more than

one tryptic peptide matched. The four non-UniProt ORFs in the medium confidence category had one tryptic peptide matching with between two and five spectral counts. The 27 non-UniProt ORFs in the low confidence category had one tryptic peptide matched and one spectral count.

Next, we analyzed the length distribution of all non-UniProt ORFs (Figure 3.13b). We found that 31 had a length of 150 amino acids or less, which would be classified as microproteins. We analyzed the start codon usage of the 33 non-UniProt ORFs (Figure 2.13c). We found that nine of these used an AUG start, 17 used a near-cognate start codon, and 7 used other codons. These other codons were usually the furthest upstream codon in the transcript, which suggests that ORF might be translated from an alternative RNA transcript form that is not adequately represented in the RefSeq database.

We also analyzed the relative positions of the non-UniProt ORFs on their RNA transcripts (Figure 3.14a). We assigned these non-UniProt ORFs to one of four categories: three were in the 5' untranslated region (UTR), three overlapped with the coding sequence (CDS) in a different reading frame, three were in the 3' UTR, one was from a non-coding RNA, none were translated from the antisense strand, and the remaining 23 translation products were from RNA transcripts that were not annotated and did not appear to contain known ORFs. Examples of each category are given in Figure 3.14b.

A large number of sequences in the translated RNA-Seq database used to search the mass spectrometry data might result in a higher number of false peptide-spectrum matches. The combined brain and spinal cord dataset contained approximately 32,000,000 sequences, including reverse decoys. By comparison, the mouse brain RNA-Seq database used in Chapter 2 contained approximately 8,000,000 sequences. This suggests that the transcript diversity of the human brain and spinal cord RNA-Seq dataset was very high, resulted in multiple entries for

nearly identical sequences. As the size of the sequence database increases, so does the likelihood that random matches occur. Further experiments, such as acquiring reference spectra using synthetic peptides and generating recombinant constructs to verify expression, will be required to validate the list of non-UniProt ORFs identified in human CSF.

3.5 Conclusions

In this chapter, we described the development and application of an integrated proteogenomic strategy to identify intracellular and secreted peptides and small proteins from HEK293T cells and human CSF. The use of GELFrEE resulted in a nearly two-fold increase in the number of protein identifications. Our strategy uncovered known peptide hormones, neurotransmitters, and neuropeptides. Our proteogenomic strategy also revealed the existence of 33 unannotated ORFs, of which 31 are smORFs. These smORFs might act in cis to regulate the expression of neighboring ORFs or interact in trans with the translated protein products. Our results suggest that the mechanisms regulating gene expression and the repertoire of translated ORFs in human brain are more diverse than previously thought.

3.6 Acknowledgements

This work was supported by the Mass Spectrometry Core of the Salk Institute with funding from NIH-NCI CCSG: P30 014195 and the Helmsley Center for Genomic Medicine. This work was supported by the The Razavi Newman Integrative Genomics and Bioinformatics Core Facility of the Salk Institute with funding from NIH-NCI CCSG: P30 014195, and the Helmsley Trust.

3.7 Figures

(A)

(B)

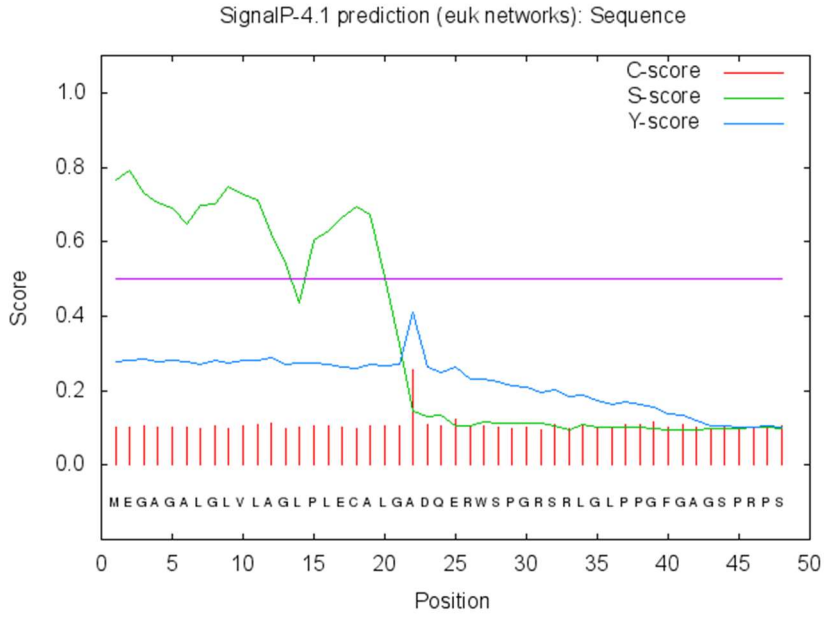
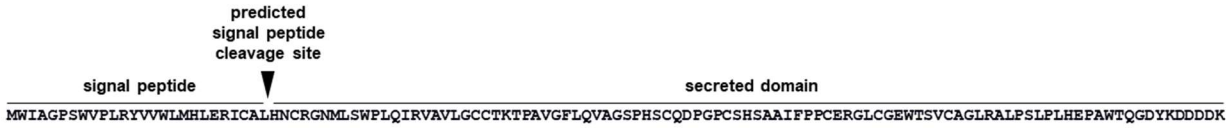


Figure 3.1: MEGA microprotein. (A) The sequence of human MEGA microprotein with a c-terminal FLAG-tag shown in one-letter amino acid abbreviations. **(B)** The output of SignalP 4.1 prediction of signal peptide cleavage sites. Raw cleavage site score, C-score (red). Signal peptide score, S-score (green). Combined cleavage site score, Y-score (blue). Score cutoff used in determining signal peptide cleavage sites shown as a dashed horizontal magenta line.

(A)



(B)

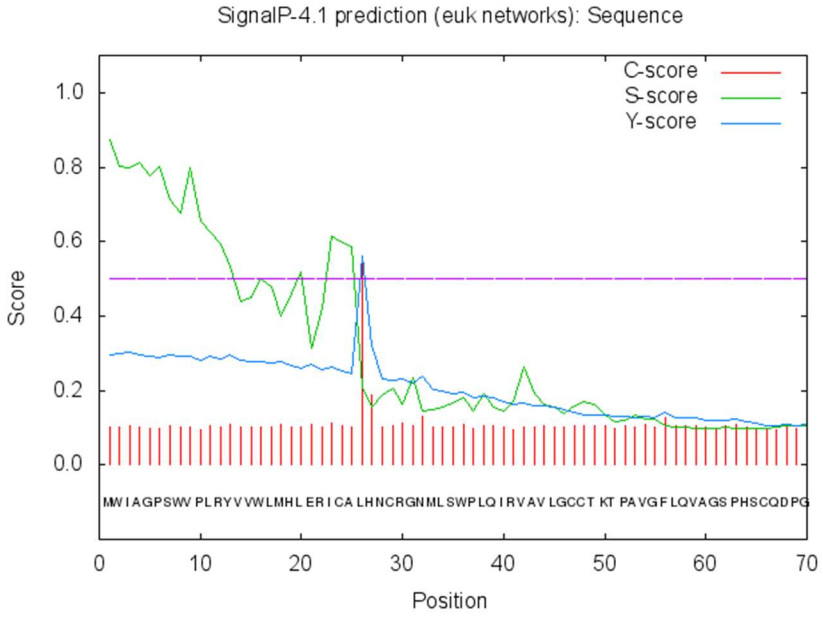


Figure 3.2: MWIA microprotein. (A) The sequence of human MWIA microprotein with a c-terminal FLAG-tag shown in one-letter amino acid abbreviations. **(B)** The output of SignalP 4.1 prediction of signal peptide cleavage sites. Raw cleavage site score, C-score (red). Signal peptide score, S-score (green). Combined cleavage site score, Y-score (blue). Score cutoff used in determining signal peptide cleavage sites shown as a dashed horizontal magenta line.

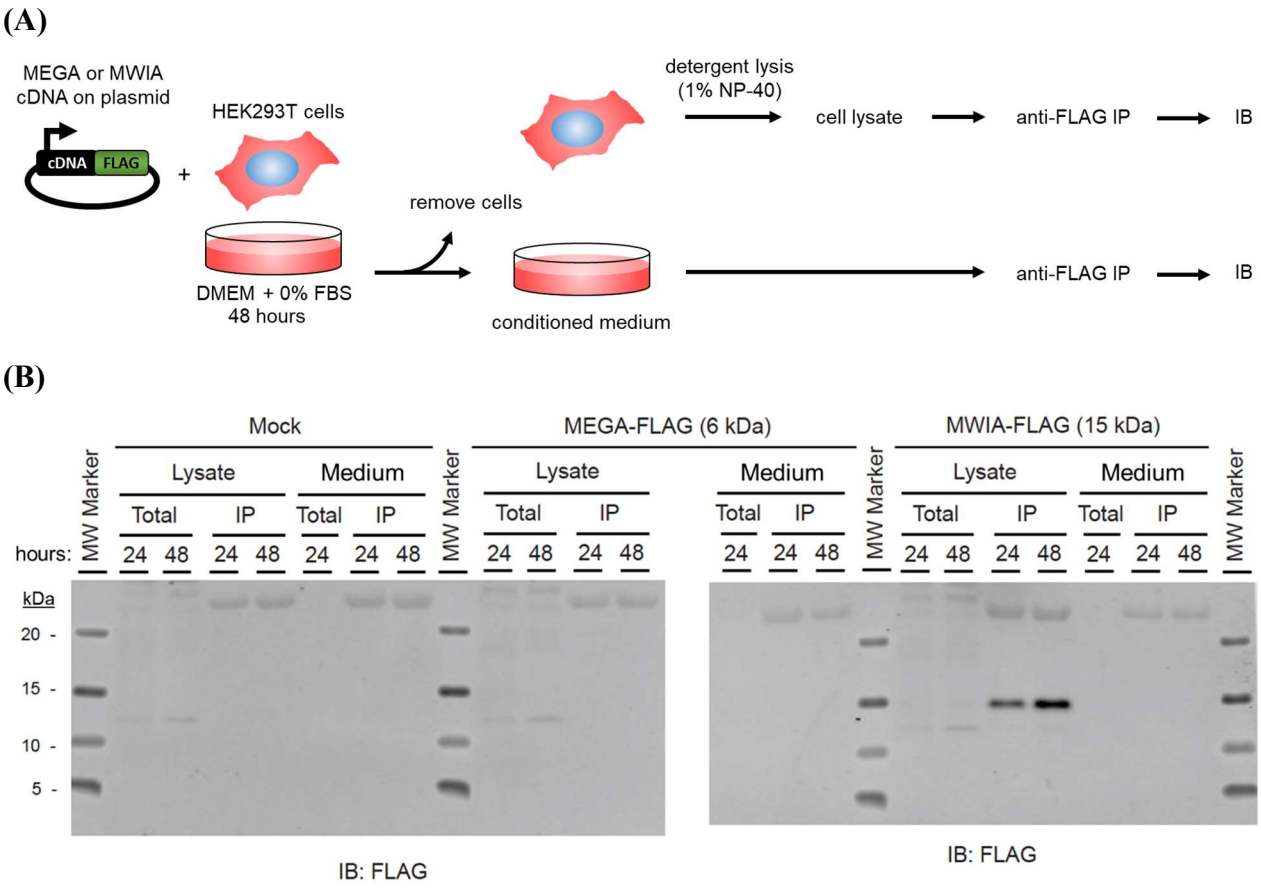


Figure 3.3: Expression of MEGA-FLAG and MWIA-FLAG in HEK293T cells. (A) HEK293T cells were transfected with plasmids encoding MEGA-FLAG or MWIA-FLAG. Cells were grown in serum-free medium for 48 hours. Cells were lysed, and FLAG-tagged proteins were immunoprecipitated (IP). **(B)** Immunoblots, IB. Fetal bovine serum, FBS.

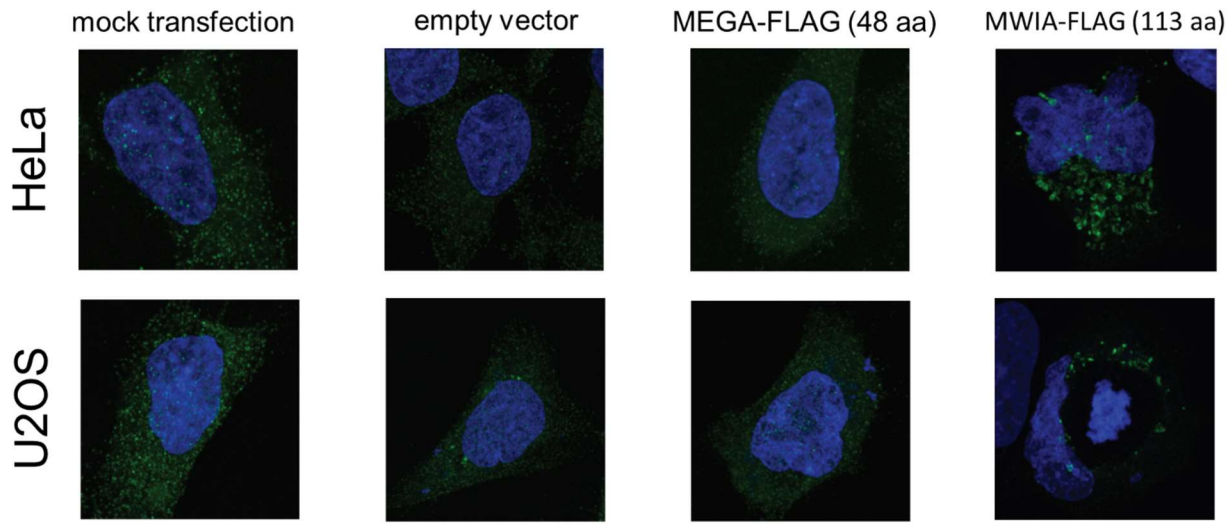
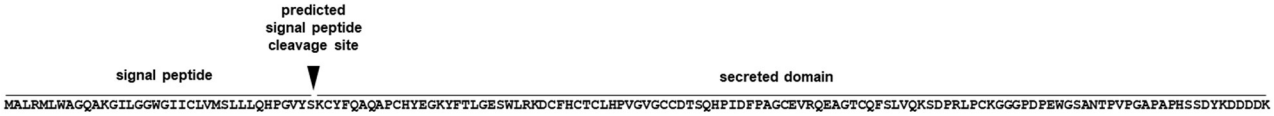


Figure 3.4: Confocal images of human cells expressing MEGA-FLAG or MWIA-FLAG. Human cell lines were transfected with the indicated plasmids were imaged by confocal microscopy. HeLa cells, top row. U2OS, bottom row. Anti-FLAG, green. Nuclei are counterstained with Hoechst (blue).

(A)



(B)

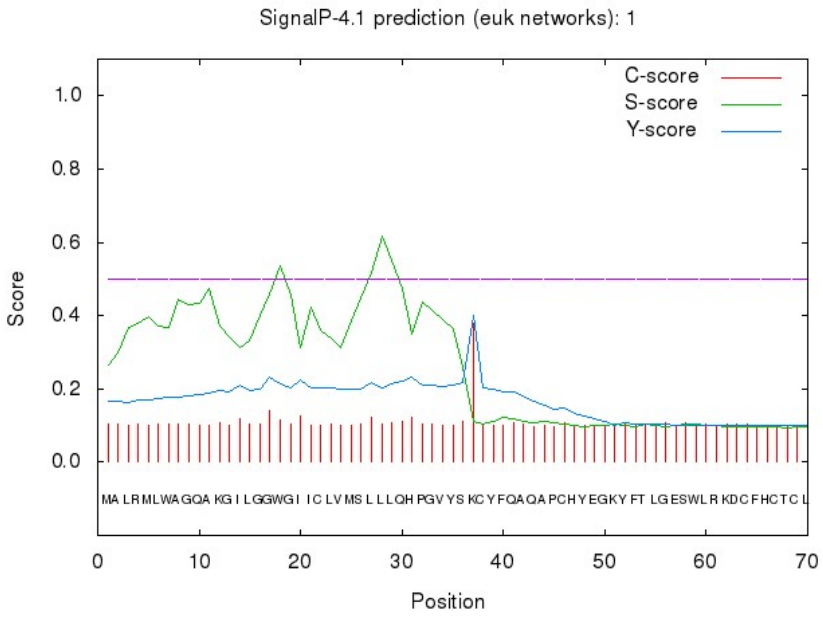
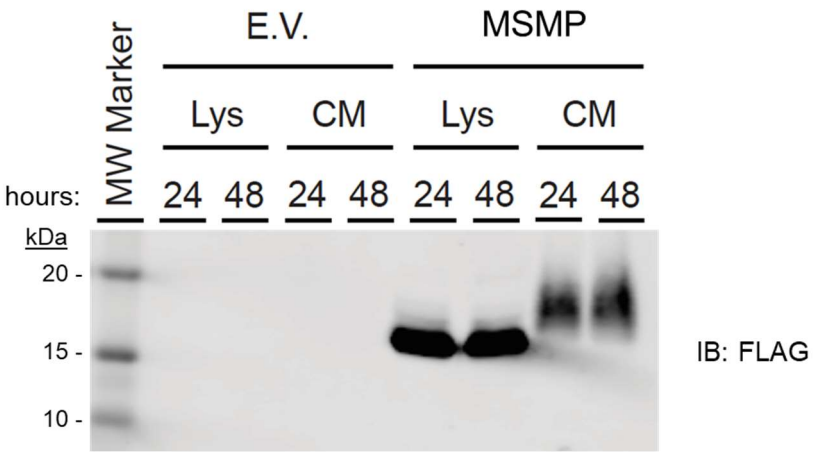


Figure 3.5: Human prostate-associated microseminoprotein. (A) The sequence of human prostate-associated microsemino protein (UniProt accession Q1L6U9) with a c-terminal FLAG-tag shown in one-letter amino acid abbreviations. **(B)** The output of SignalP 4.1 prediction of signal peptide cleavage sites. Raw cleavage site score, C-score (red). Signal peptide score, S-score (green). Combined cleavage site score, Y-score (blue). Score cutoff used in determining signal peptide cleavage sites shown as a dashed horizontal magenta line.

(A)



(B)

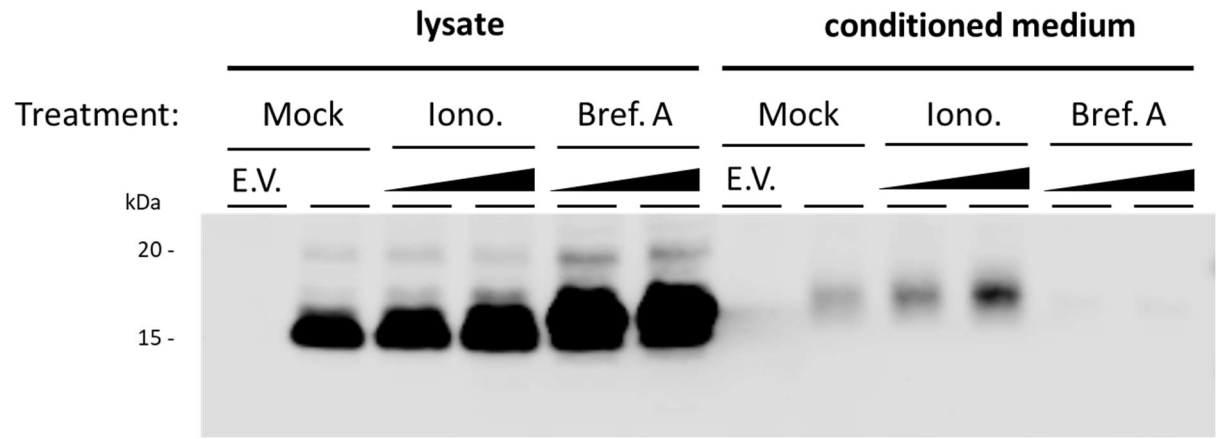


Figure 3.6: MSMP-FLAG is a secreted microprotein. (A) HEK293T cells were transfected with plasmids encoding MSMP-FLAG. An anti-FLAG immunoprecipitation was performed from lysate (Lys) or conditioned medium (CM) after 24 or 48 hours. **(B)** HEK293T cells were treated with ionomycin (Iono.) or Brefeldin A (Bref. A) or mock-treated, and an anti-FLAG immunoprecipitation was carried out from lysate or conditioned media. Triangle indicates concentration. Empty vector, E.V.

Figure 3.7: Workflow to detect secreted peptides and small proteins in HEK293T cells.

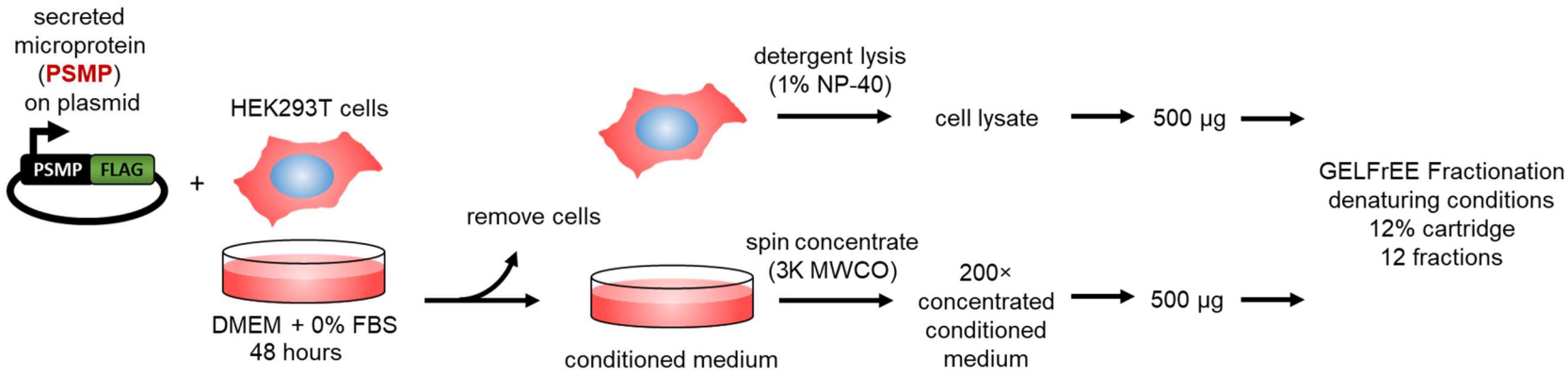
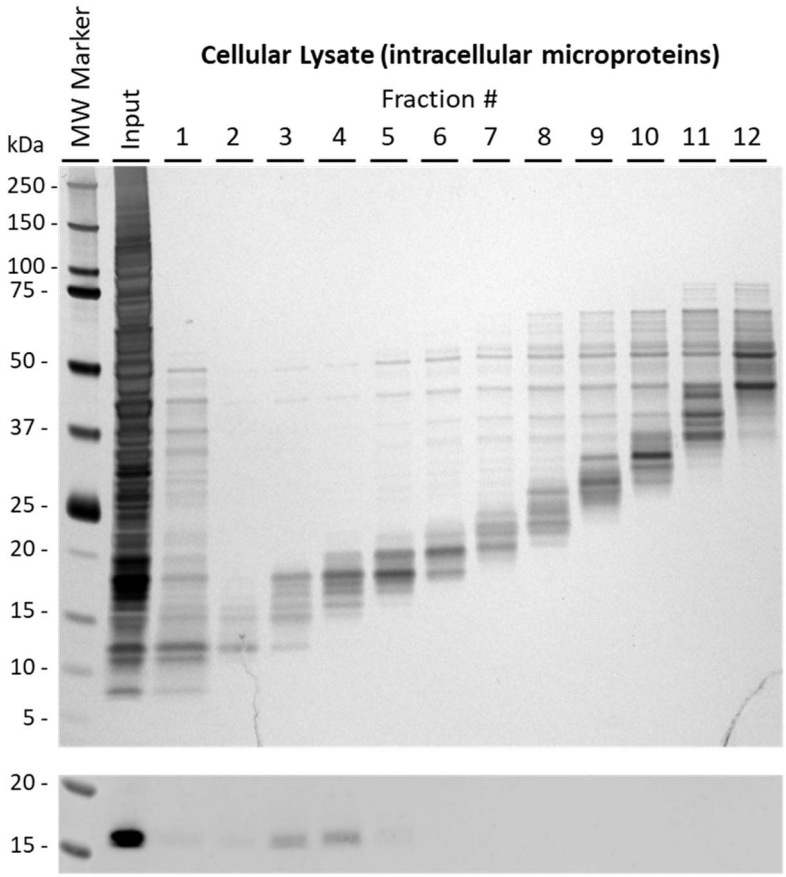
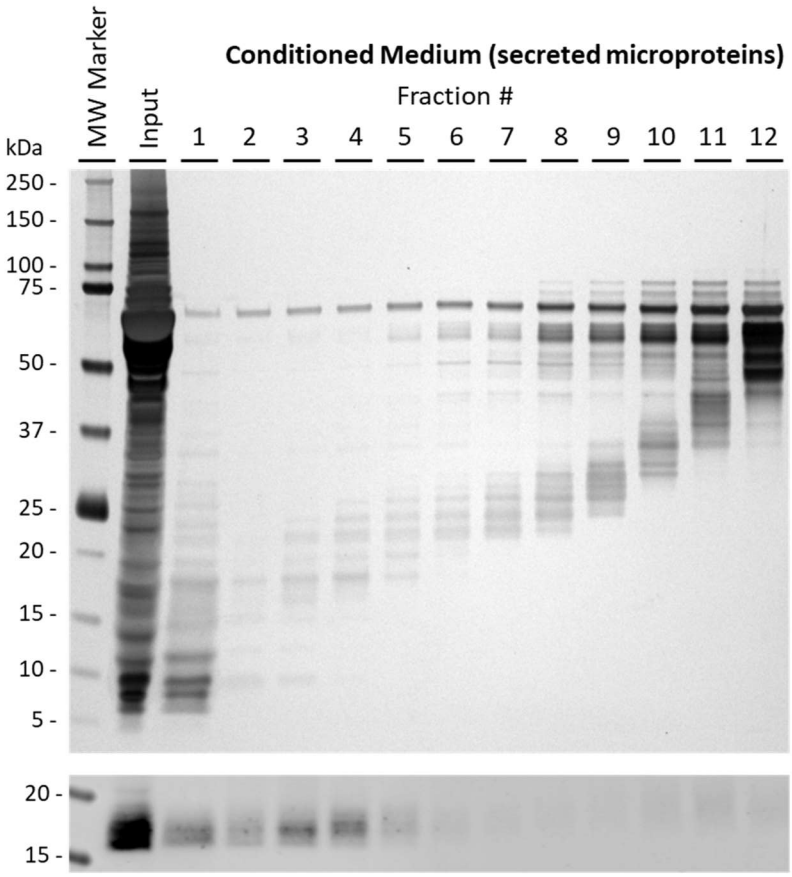
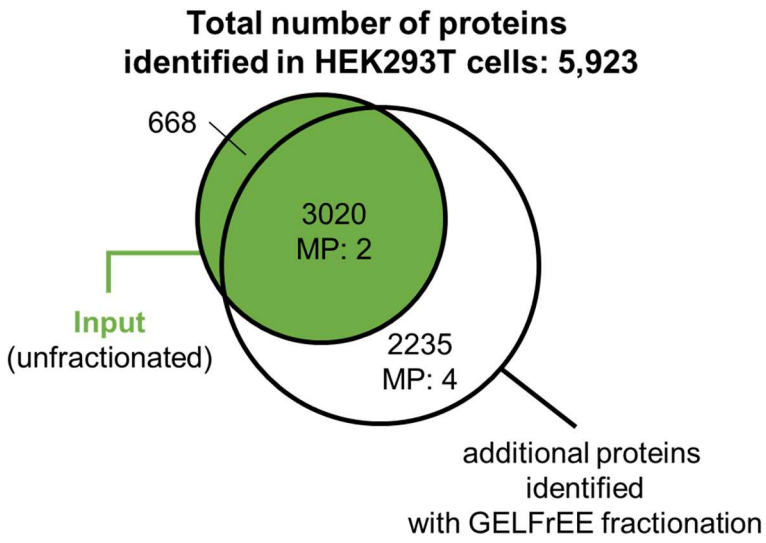

secreted
microprotein
(PSMP)
on plasmid
PSMP
FLAG
+
HEK293T cells
DMEM + 0% FBS
48 hours
remove cells
detergent lysis
(1% NP-40)
cell lysate
500 µg
spin concentrate
(3K MWCO)
conditioned medium
200×
concentrated
conditioned
medium
500 µg
GELFrEE Fractionation
denaturing conditions
12% cartridge
12 fractions

Figure 3.8: Silver staining and immunoblotting of C8-extracted proteins separated by a GELFrEE device with a 12% cartridge. Proteins from each fraction were separated by SDS-PAGE and visualized by silver staining. Unfractionated material and fractions 1-12 containing were analyzed by LC-MS/MS. MW, molecular weight. kDa, kilodalton.


Cellular Lysate (intracellular microproteins)
Fraction #
MW Marker
Input
1
2
3
4
5
6
7
8
9
10
11
12
kDa
250 -
150 -
100 -
75 -
50 -
37 -
25 -
20 -
15 -
10 -
5 -
20 -
15 -


Conditioned Medium (secreted microproteins)
Fraction #
MW Marker
Input
1
2
3
4
5
6
7
8
9
10
11
12
kDa
250 -
150 -
100 -
75 -
50 -
37 -
25 -
20 -
15 -
10 -
5 -
20 -
15 -

(A)



(B)

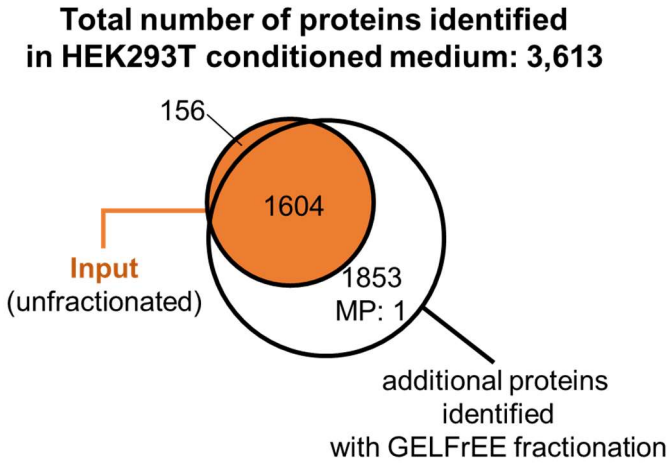


Figure 3.9: Overlap of protein identifications in HEK293T cells. Uniprot-annotated proteins identified in **(A)** HEK293T cells and **(B)** conditioned medium. Microprotein, MP

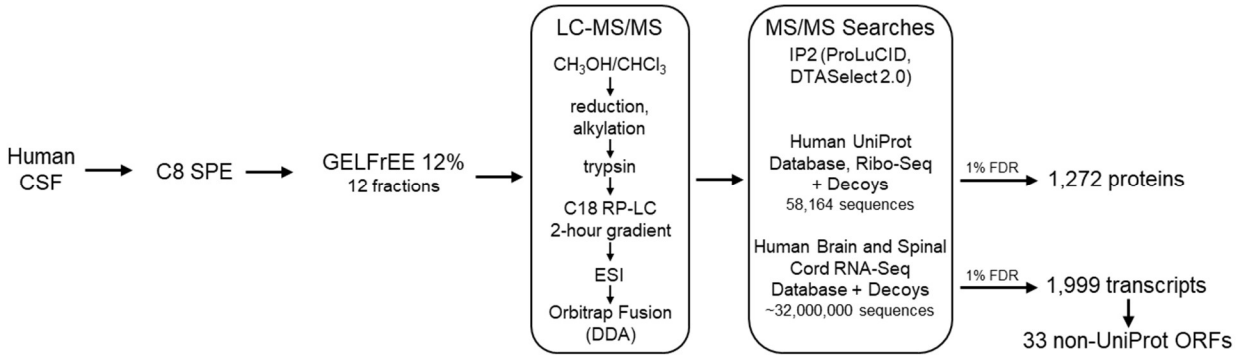
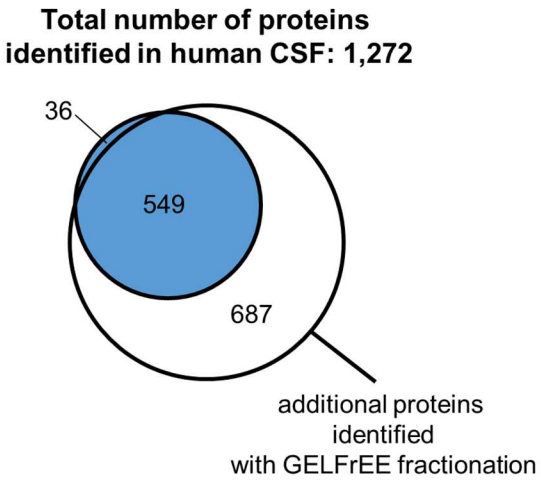


Figure 3.10 Integrated proteogenomic strategy for the discovery of translated polypeptides in human cerebrospinal fluid. Human cerebrospinal fluid (CSF) was subjected to solid-phase extraction (SPE) using C8 silica. Extracted polypeptides were fractionated using a GELFrEE system. Fractions were analyzed by liquid chromatography coupled to electrospray ionization and tandem mass spectrometry (LC-MS/MS). Proteins were first precipitated using methanol/chloroform, reduced and alkylated, digested with trypsin, then subjected to reverse-phase liquid chromatography (RP-LC) using a C18 column and a 2-hour gradient. Eluted polypeptides were ionized using electrospray (ESI) and analyzed on an Orbitrap Fusion instrument operated in data-dependent acquisition (DDA) mode. The MS/MS spectra were then searched using a target-decoy strategy against a database of mouse UniProt reference proteins or a database generated from the in silico translation of human spinal cord and RNA-Seq data. Protein and transcript identifications were reported with a false-discovery rate (FDR) of 1%.

(A)



(B)

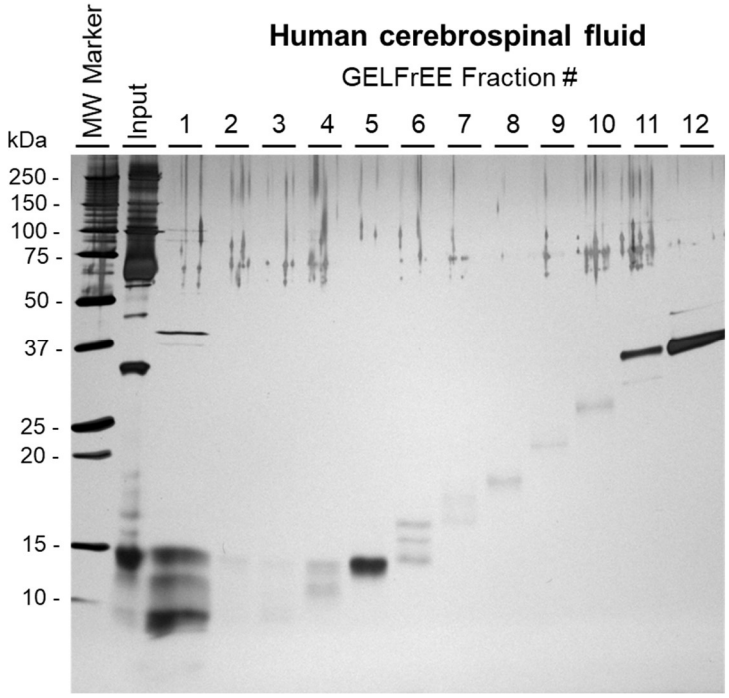


Figure 3.11: Identification of proteins in human cerebrospinal fluid. (A) Overlap of UniProt-annotated protein identifications. **(B)** Silver staining and immunoblotting of C8-extracted proteins separated by a GELFrEE device with a 12% cartridge.

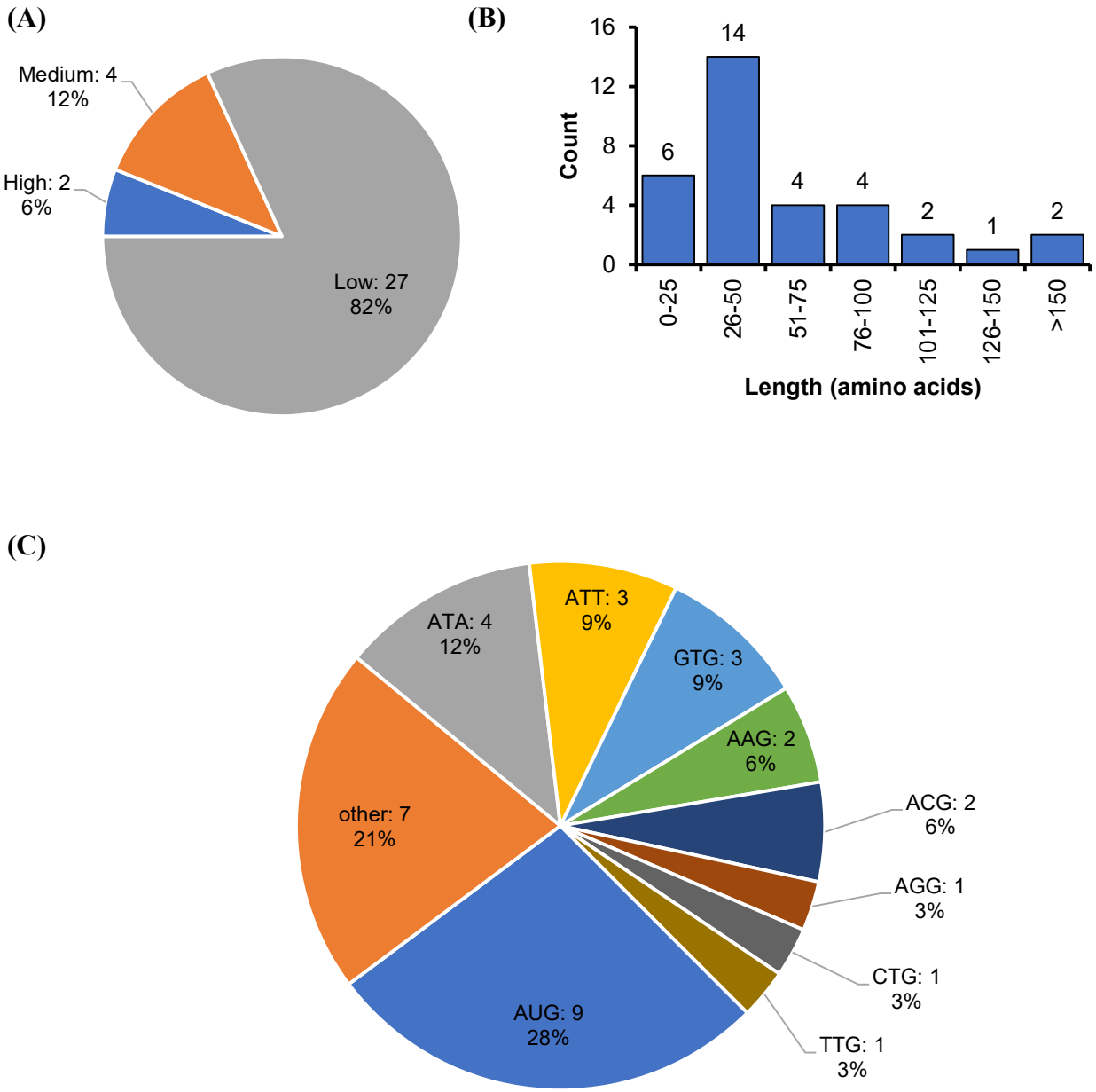
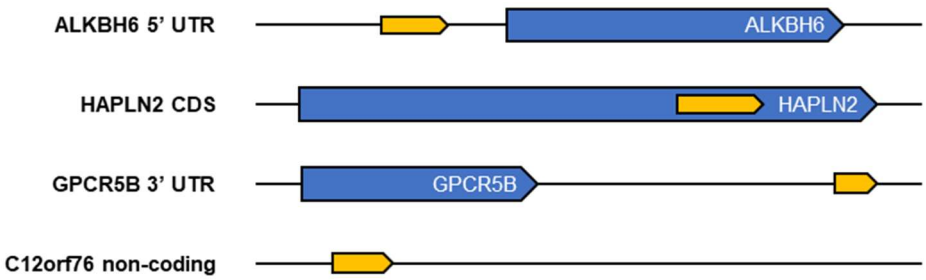


Figure 3.12: Features of 33 non-UniProt ORF identifications in human cerebrospinal fluid. **(A)** Confidence of detection. Non-UniProt ORFs were assigned to one of three confidence categories: (1) High, >1 non-UniProt peptide detected or >10 peptide spectrum matches, (2) Medium, 1 non-UniProt peptide detected and between 2 and 10 peptide spectrum matches, (3) Low, 1 non-UniProt peptide detected and 1 peptide spectrum match. **(B)** Length distribution of non-UniProt ORFs annotated with the furthest upstream start codon in the transcript. **(C)** Distribution of possible start codons of non-UniProt ORFs.

(A)



(B)

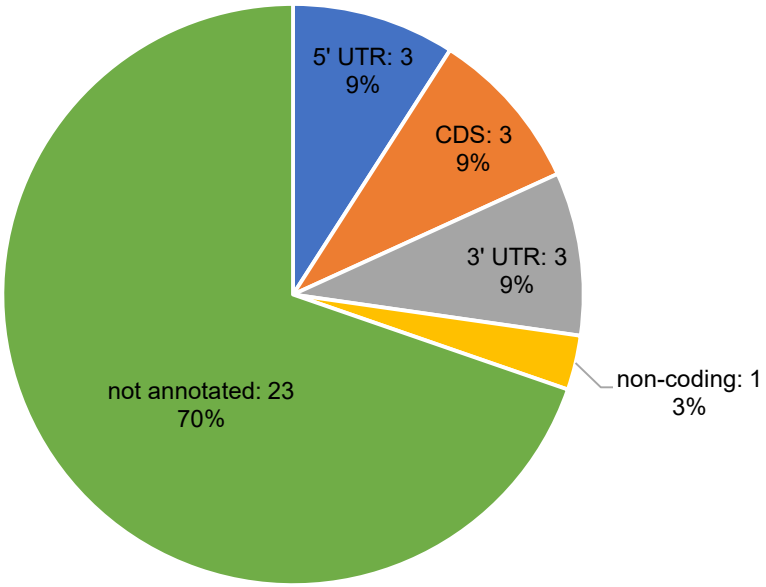


Figure 3.13: Transcript locations of 33 non-UniProt ORFs. (A) Examples of ORFs localized to the 5' untranslated region (UTR), coding sequence (CDS), non-coding RNA, 3' UTR, and anti-sense. RNA transcripts represented by thin black lines in the 5' to 3' direction. Non-UniProt ORFs, yellow arrows. Annotated CDS, blue arrows. Not drawn to scale. **(B)** Distribution of transcript locations of 33 non-UniProt ORFs.

3.8 References

1. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* **11,** 909–916 (2015).

2. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **18,** 575–589 (2017).

3. Behrens, O. K. & Grinnan, E. L. Polypeptide hormones. *Annu. Rev. Biochem.* **38,** 38–112 (1969).

4. Tager, H. S. & Steiner, D. F. Peptide hormones. *Annu. Rev. Anal. Chem.* **43,** 509–538 (1974).

5. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. & Couso, J. P. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.* **5,** 1052–1062 (2007).

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

7. Zhang, Q., Vashisht, A. A., O'Rourke, J., Corbel, S. Y., Moran, R., Romero, A., Miraglia, L., Zhang, J., Durrant, E., Schmedt, C., Sampath, S. C. & Sampath, S. C. The microprotein Minion controls cell fusion and muscle formation. *Nat. Commun.* **8,** (2017).

8. Arnoult, N., Correia, A., Ma, J., Merlo, A., Garcia-Gomez, S., Maric, M., Tognetti, M., Benner, C. W., Boulton, S. J., Saghatelian, A. & Karlseder, J. Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. *Nature* **549,** 548–552 (2017).

9. Chu, Q., Martinez, T. F., Novak, S. W., Donaldson, C. J., Tan, D., Vaughan, J. M., Chang, T., Diedrich, J. K., Andrade, L., Kim, A., Zhang, T., Manor, U. & Saghatelian, A. Regulation of the ER stress response by a mitochondrial microprotein. *Nat. Commun.* **10,** 1–13 (2019).

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

11. He, L., Diedrich, J., Chu, Y. Y. & Yates, J. R. Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. *Anal. Chem.* **87,** 11361–11367 (2015).

12. Xu, T., Park, S. K., Venable, J. D., Wohlschlegel, J. A., Diedrich, J. K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X., Wong, C. C. L., Fonslow, B., Delahunty, C., Gao, Y., Shah, H. & Yates, J. R. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteomics* **129,** 16–24 (2015).

13. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).

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

15. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** 61–65 (2007).

16. Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N. & Saghatelian, A. Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16,** 458–468 (2019).

17. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. & Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

18. Valtonen-André, C., Bjartell, A., Hellsten, R., Lilja, H., Härkönen, P. & Lundwall, Å. A highly conserved protein secreted by the prostate cancer cell line PC-3 is expressed in benign and malignant prostate tissue. *Biol. Chem.* **388,** 289–295 (2007).

19. Pei, X., Sun, Q., Zhang, Y., Wang, P., Peng, X., Guo, C., Xu, E., Zheng, Y., Mo, X., Ma, J., Chen, D., Zhang, Y., Zhang, Y., Song, Q., Guo, S., Shi, T., Zhang, Z., Ma, D. & Wang, Y. PC3-Secreted Microprotein Is a Novel Chemoattractant Protein and Functions as a High-Affinity Ligand for CC Chemokine Receptor 2. *J. Immunol.* **192,** 1878–1886 (2014).

20. Bennett, J. P., Cockcroft, S. & Gomperts, B. D. Ionomycin stimulates mast cell histamine secretion by forming a lipid-soluble calcium complex. *Nature* **282,** 851–853 (1979).

21. Orcl, L., Tagaya, M., Amherdt, M., Perrelet, A., Donaldson, J. G., Lippincott-Schwartz, J., Klausner, R. D. & Rothman, J. E. Brefeldin A, a drug that blocks secretion, prevents the assembly of non-clathrin-coated buds on Golgi cisternae. *Cell* **64,** 1183–1195 (1991).

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

23. 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., 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., Kheradpour, P., 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., 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., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., 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., 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., 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., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., 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., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., 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., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., Van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K. K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S.,

Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J. & Lochovsky, L. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

Chapter 4

Biochemical characterization of C4ORF48 and Gm1673 neuropeptides

4.1 Abstract

Neuropeptides are a diverse class of signaling peptides that are used in cell-cell communication in the brain. In our proteogenomic analysis of the conditioned media of HEK293T cells and human cerebrospinal fluid (Chapter 3), we identified a putative neuropeptide candidate, C4ORF48. The molecular function of this neuropeptide is unknown. In this chapter, I will first review evidence that C4ORF48 and its mouse homolog Gm1673 are secreted neuropeptides encoded by small open reading frames. I will then discuss C4ORF48 and Gm1673 as possible biomarkers for human disease. I will finally describe experimental evidence that C4ORF48 and Gm1673 are secreted peptides and conclude with biochemical characterization of both peptides processed forms, including proteolytic, glycosylation, and disulfide modifications.

4.2 Introduction

In this chapter, I will describe the biochemical characterization of a peptide hormone and neuropeptide encoded by the *C4ORF48* gene in humans and the orthologous *Gm1673* gene in mice. This neuropeptide was cloned by Endele *et al.* in a 2011 study[1] to identify novel genes associated with mental retardation or intellectual disability in humans. The authors show that C4ORF48 is expressed in human brain tissue by western blotting. They further establish that *Gm1673* is expressed in the mouse brain by northern blotting and RNA in situ hybridization. The Endele *et al.* study was limited to tissue samples only. This chapter builds on their previously

published work by studying the secreted version of C4ORF48 and Gm1673. The molecular function of this neuropeptide remains unknown.

Identification of C4ORF48 as a neuropeptide

We initially sought to detect secreted peptides and small proteins in both human cerebrospinal fluid and HEK293T conditioned media. We reasoned that peptides and small proteins identified in these extracellular fluids could have peptide hormone or neuropeptide functions. We further deduced that peptides and small proteins identified in both extracellular fluids, rather than just one, would be ideal candidates as peptide hormones and neuropeptides for the following reasons: (1) it minimizes the likelihood that a peptide or small protein is released passively by cell lysis and is assumed to be secreted, (2) candidate peptides and small proteins can be cloned and readily expressed in HEK293T cell lines for further study, and (3) these two extracellular fluids have not been as extensively studied as intracellular proteomes. The results of these experiments were described in Chapter 3.

After analyzing the HEK293T conditioned media peptidomics dataset, we did not detect any novel microproteins. There are several possible explanations for this result: secreted microproteins were at abundance levels below the limit of detection, secreted microproteins degraded rapidly once secreted from the cell, the sequence database containing smORFs from ribosome profiling experiments lacked the relevant sequences of secreted microproteins, or all secreted peptides and small proteins from HEK293T are already known. The lack of novel microprotein identifications in our proteomics dataset of HEK293T conditioned media prompted us to consider peptides and small proteins with existing UniProt entries, but with uncharacterized molecular function. C4ORF48 emerged from this list as a candidate smORF-derived secreted

microprotein that warranted further study. We identified C4ORF48 in both human cerebrospinal fluid and HEK293T conditioned media. This initial finding establishes that C4ORF48 is a secreted microprotein. Our work with C4ORF48 is used to develop a general approach for characterizing secreted microproteins, which we hope to apply to novel smORFs in the future.

C4ORF48 as a biomarker of human disease

The *C4ORF48* gene was first cloned and characterized by Endele *et al.* while searching for genes linked to mental retardation and intellectual disability[1]. The *C4ORF48 gene* is encoded in a microdeletion region of the short arm of chromosome 4 in a patient with a milder form of Wolf-Hirschhorn syndrome (WHS)[2]. WHS patients typically present with craniofacial and skeletal deformities, convulsions, and severe intellectual disability[3]. These symptoms suggest that the deleted region on chromosome 4 encodes genes that are important in the development and/or functioning of the skeletal and central nervous systems. *C4ORF48* was likely discounted in the initial characterization of this microdeletion region as a protein-coding gene because of the small size of its translated protein product.

C4ORF48 encodes a 95-amino acid microprotein that is predicted to be processed into a 61-amino acid peptide that is secreted. The sequence of the secreted domain is highly conserved in *Gnathostomata* (jawed vertebrate) species. The mouse ortholog is *Gm1673*, which encodes a 90-amino acid microprotein that is predicted to be processed into a 62-amino acid peptide that is secreted. The high degree of conservation in the secreted domain indicates that C4ORF48 and GM1673 likely play essential roles in the functioning and development of the central nervous systems in other organisms. We found further evidence for this in the temporal gene expression patterns in human and mouse brain. In both species, gene expression levels in the brain are high

in the fetal stages, but low or undetectable in the adult stages[4,5]. Consistent with our detection of C4ORF48 in human cerebrospinal fluid, C4ORF48 likely encodes a neuropeptide and functions to promote the proper development of the brain.

As *C4ORF48* has already been linked to Wolf-Hirschhorn syndrome, we wondered if other disease phenotypes or disease models showed abnormal levels of *C4ORF48* or *Gm1673* expression. We searched the Expression Atlas database[6] for instances of differential gene expression that were associated with disease states compared with baseline states. In humans, *C4ORF48* expression was upregulated in several cancers, including squamous cell carcinoma, esophageal adenocarcinoma, thyroid carcinoma, and breast carcinoma. *C4ORF48* expression was downregulated in pancreatic adenocarcinoma, hepatobiliary carcinoma, Down syndrome, and Crohn's disease. In mouse disease models, *Gm1673* expression was upregulated in clear cell sarcoma, osteosarcoma, and synovial sarcoma. *Gm1673* expression was downregulated in Duchenne muscular dystrophy. As *C4ORF48* and *Gm1673* are both expressed in a variety of tissues throughout the body, this neuropeptide might also function more generally as a peptide hormone in tissues outside of the brain.

The rationale for studying C4ORF48 and GM1673

In wanting to study this peptide further, we realized that fundamental experiments, such as validation of the cleavage site and structure of the disulfide bonds between cysteines have not yet been carried out. The application of modern proteomics tools may rapidly answer these questions, and we sought to use C4ORF48 (and the mouse homolog Gm1673) as a test case to determine how well the processed forms of smORF-derived peptides can be characterized, which can spur the search and characterization of additional smORFs with secreted peptides.

4.3 Methods and Materials

Plasmids and mutagenesis

All plasmids were purchased from GenScript in a pcDNA3.1+ vector with a single FLAG-tag sequence (DYKDDDDK) inserted before the stop codon. cDNA clones used were MSMP (OHu16861), C4orf48 (OHu108340), Gm1673 (OMu75071). O-linked glycosylation sites were predicted using NetOGlyc 4.0[7]. Mutagenic primers for all predicted glycosylation sites were designed to introduce an alanine substitution using NEBaseChanger v1.2.9 (New England Biolabs, accessed at https://nebasechanger.neb.com/). Primer sequences used were C4orf48 S39A (GCCCGCCGGGGCTGCCGTCCCCGCGCAGAGC and TCGGCGCGGGCCCGGGCG), Gm1673 T32A (CGAGCCCGCCGCCGGGAGCGC and GCGCGGGCGCCCAGCAGC), GM1673 S34A (CGCCACCGGGGCCGCTGTCCCCGCTCAGAGC and GGCTCGGCGCGGGCGCCC), Gm1673 S40A (CCCCGCTCAGGCCCGCCCGTGCG and ACAGCGCTCCCGGTGGCG). Non-glycosylation acceptor mutant clones were generated using a Q5 site-directed mutagenesis kit (New England Biolabs) following the manufacturer's instructions. Gm1673 3xA (T32A, S34A, T40A) was generated by GenScript. All mutation sites were verified by DNA sequencing.

Cell culture

HEK293T (Dharmacon) and CHO-S cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM, Corning) supplemented with 10% (v/v) fetal bovine serum (Corning) in a humidified 37 °C incubator with 5% CO_2 atmosphere. Cell culture treated plastic dishes (Corning) were used. Cells were sub-passaged every 3 days when 90% confluence was reached using 0.25% trypsin-EDTA (Life Technologies). Cells were transfected in a 6-well dish at 80%

confluence. A transfection mixture of 2 µg plasmid in 50 µL Opti-MEM (Life Technologies) was combined with 5 µL Lipofectamine 2000 (Life Technologies) and 45 µL Opti-MEM. After a 10-minute incubation at room temperature, the transfection mixture was added dropwise to cells. Cells were incubated with transfection mixture at 37 °C for 4-6 hours and subsequently re-plated in 10 cm dishes. After two days, cell culture medium was replaced with serum-free DMEM without phenol red (Corning). Three days after replacing the media, conditioned medium was decanted, passed through a 0.22 µm syringe filter to remove residual cells, supplemented with 1× protease inhibitors (Roche), and stored on ice. Cells were collected by pipetting in ice-cold phosphate-buffered saline (PBS), pelleted by centrifugation at 200 × *g* for 3 minutes at 4 °C, washed once more with PBS, and stored on ice.

Protein extraction and immunoprecipitation

Cells were lysed in IP lysis buffer (Pierce) supplemented with 1× Halt protease inhibitor cocktail (Thermo Fisher Scientific). The lysate was clarified by centrifugation at 20,000 × *g* for 10 min at 4 °C. Protein concentration was measured using a BCA protein assay kit (Pierce) in a 96-well microtiter plate using bovine serum albumin standards (Pierce). The lysate and filtered conditioned media were applied to 25 µL anti-FLAG beads (Sigma-Aldrich) pre-equilibrated in IP lysis buffer, incubated overnight (>12 h) at 4 °C with end-over-end rotation, and washed four times with 1 mL TBSG (25 mM Tris-Cl pH 7.5, 150 mM NaCl, 5% (v/v) glycerol). Beads were pelleted by centrifugation at 500 × *g* for 2 mins at room temperature during the equilibration and wash steps. Bound proteins were eluted with TBSG + 0.1 mg/mL 3×FLAG peptide (Sigma-Aldrich) overnight at 4 °C with end-over-end rotation. Supernatants were stored at -20 °C until further processing.

SDS-PAGE and western blotting

Samples were combined with 4× SDS-PAGE buffer (250 mM Tris-Cl pH 6.8, 8% (w/v) SDS, 0.2% (w/v) bromophenol blue, 40% (v/v) glycerol, and 10% (v/v) 2-mercaptoethanol) to a working concentration of 1×. For non-reducing conditions, 2-mercaptoethanol was omitted from the buffer and replaced with water. Samples were subsequently incubated at 95 °C for 3 mins and cooled to room temperature. Samples were loaded on a pre-cast Novex 10-20% tris-glycine polyacrylamide gel in a running buffer at pH 8.3 (25 mM Tris, 192 mM glycine, 0.1% SDS) and electrophoresed at 200 V until the dye-front migrated to the bottom of the gel. Proteins were transferred to polyvinylidene difluoride membranes (Life Technologies, 0.2 µm pore size) using an iBlot2 transfer device (Life Technologies) using the pre-programmed P4 setting (15 volts, 7 minutes). Membranes were incubated at room temperature for 1 hour in PBS blocking buffer (LI-COR) and probed with anti-DYKDDDDK rabbit antibodies (1:2000, Cell Signaling Technology) in blocking buffer at 4 °C overnight on an orbital platform shaker. Membranes were washed three times in TBST (20 mM Tris-Cl pH 7.6, 150 mM NaCl, 0.1% (v/v) Tween-20) and probed with goat-anti-rabbit fluorescent conjugates (1:10,000, LI-COR 800CW) in blocking buffer at room temperature for 45 minutes in the dark. Membranes were washed four times with TBST, and proteins were visualized on an Odyssey imaging system (LI-COR).

Polypeptide enrichment for LC-MS/MS analysis

Polypeptides from human cerebrospinal fluid (CSF) and HEK293T conditioned medium were enriched from the flow-through by solid-phase extraction using Agilent Bond Elut 1 gram C8 or C18 silica cartridges on a vacuum manifold. Cartridges were pre-wet with one column volume of methanol and equilibrated with one column volume of triethylammonium formate

(TEAF) buffer, pH 3.0. Samples were applied to the pre-equilibrated column, washed with one column volume of TEAF, and eluted with 2 ml of a mixture of 75% (v/v) acetonitrile and 25% (v/v) TEAF. Samples were evaporated to dryness in a vacuum centrifuge overnight and stored at -20 °C until further processing. Polypeptides from mouse primary astrocytes were concentrated using 3000 molecular weight cutoff centrifugal filters (EMD Millipore) prior to analysis by LC-MS/MS.

LC-MS/MS analysis

LC-MS/MS analysis was carried out as previously described[8]. Each sample was analyzed in duplicate. Samples were precipitated by methanol/chloroform and re-dissolved in 8 M urea/100 mM tetraethylammonium tetrahydroborate (TEAB), pH 8.5. Polypeptides were reduced with 5 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP, Sigma-Aldrich) and alkylated with 10 mM chloroacetamide (Sigma-Aldrich). Polypeptides were digested overnight at 37 °C in 2 M urea/100 mM TEAB, pH 8.5, with trypsin (Promega). For disulfide linkage determination, chloroacetamide was omitted and replaced with 2 mM N-ethylmaleimide (Sigma-Aldrich). Digestion was quenched with formic acid, 5 % final concentration. The digested samples were analyzed on a Fusion Orbitrap tribrid mass spectrometer (Thermo Fisher Scientific). The digest was injected directly onto a 30 cm, 75 μm inner diameter column packed with BEH 1.7 μm C18 resin (Waters). Samples were separated at a flow rate of 300 nL/min on an nLC 1000 chromatography system (Thermo Fisher Scientific). Buffer A and B were 0.1% formic acid in water and 0.1% formic acid in 90% acetonitrile, respectively. A gradient of 1-35% B over 110 min, an increase to 50% B over 10 min, an increase to 90% B over 10 min and held at 90% B for a final 10 min was used for a 140 min total run time. The column was re-equilibrated with 20 μL

of buffer A prior to the injection of the sample. Peptides were eluted directly from the tip of the column and nanosprayed directly into the mass spectrometer by application of 2.5 kV voltage at the back of the column. The Orbitrap Fusion was operated in a data-dependent mode. Full MS scans were collected in the Orbitrap at 120,000 resolution with a mass range of 400 to 1600 m/z and an automatic gain control (AGC) target of 5×10^5. The cycle time was set to 3 sec, and within this 3 seconds, the most abundant ions per scan were selected for collision-induced dissociation MS/MS in the ion trap with an AGC target of 10^4 and a minimum intensity of 5000. Maximum fill times were set to 50 ms and 100 ms for MS and MS/MS scans, respectively. Quadrupole isolation at 1.6 m/z was used, monoisotopic precursor selection was enabled and dynamic exclusion was used with a duration of 5 sec.

Data processing of UniProt-annotated proteins

Data analysis was performed as previously described[8] with minor modifications. Tandem mass spectra data were extracted and deconvolved from .raw files using RawConverter[9] version 1.1.0.23 and searched using a target-decoy strategy with ProLuCID[10] and the Integrated Proteomics Pipeline version 6.0.5 (IP2, Integrated Proteomics Applications) data analysis platform. Samples from human CSF and HEK293T CM were searched against the UniProt human reference proteome (downloaded April 23, 2018). The mouse primary astrocyte CM samples were searched against the UniProt mouse reference proteome (downloaded on February 15, 2019). Common contaminant proteins and reverse decoy sequences were generated and appended to each database. The following search parameters were used: CID/HCD fragmentation mode, monoisotopic mass, 50 parts per million (ppm) precursor ion mass tolerance, 600 ppm fragment ion mass tolerance, 600-6000 mass range, trypsin was set as the

enzyme, requiring at least one tryptic end, up to two missed cleavages allowed, and no differential modifications. Mass spectra matches were filtered to 10 ppm precursor ion mass tolerance and evaluated with DTASelect 2.0[11] using XCorr and Zscore as the primary and secondary score types, respectively. Protein identifications required at least one matched peptide per protein and were reported at a false discovery rate of 1%.

Disulfide linkage determination

Disulfide linkages were analyzed using pLink2[12], as previously described[13]. Mass spectrometry .RAW files were searched against a protein database containing proteins identified in the IP and reverse decoy sequences using the following parameters: HCD-SS flow type, SS linkers, trypsin as the enzyme with up to 3 missed cleavages, peptide mass between 400 and 4000, peptide length between 4 and 60, 50 ppm precursor mass tolerance, 500 ppm fragment mass tolerance, and N-ethylmaleimide on cysteine as a variable mass modification. Results were filtered to 10 ppm mass tolerance and 1% at the spectrum level. Compute E-value option was enabled. Mass spectra from peptides with disulfide linkages were visually inspected with pLabel 2.4, and peptide spectrum matches with an E-value of less than 10^{-7} were reported.

4.4 Results and Discussion

C4ORF48 and GM1673 are secreted neuropeptides

To identify secreted microprotein candidates, we applied a peptidomic approach to profile extracellular fluids (Figure 4.1). We used C8 solid-phase extraction to enrich peptides and small proteins from human cerebrospinal fluid (CSF) and HEK293T conditioned medium (CM). The enriched samples were digested with trypsin and analyzed by liquid chromatography

coupled to electrospray ionization and tandem mass spectrometry (LC-MS/MS). The mass spectra were searched against a database containing the UniProt human reference proteome. We focused on peptides and small proteins that were less than 150 amino acids in length and with unknown molecular function. Among the proteins identified was C4ORF48, a putative neuropeptide of 95 amino acids[1]. We detected four tryptic peptides across both CSF and CM datasets, covering 47% of the protein sequence (Figure 4.2). To further determine the confidence with which this neuropeptide was identified, we examined the annotated mass spectra of one unique tryptic peptide, TETLLLQAER. We applied a stringent set of quality control criteria that have been used to detect novel microproteins[14]. The annotated spectra of this peptide in both CSF (Figure 4.3) and CM (Figure 4.4) datasets showed that these peptides were high-confidence matches, having precursor mass errors of less than 5 parts per million and peak matches to nearly all *b*- and *y*- ions in the mass spectra, thereby reducing the likelihood that this spectrum was matched randomly to this peptide sequence. Furthermore, C4ORF48 has been detected in a previous proteomic study of CSF and arachnoid cyst fluid from human patients[15], confirming that C4ORF48 is a secreted polypeptide and likely functions as a neuropeptide. This finding prompted us to ask if C4ORF48 might function more generally in the development or function of the central nervous system. When we searched for orthologous genes in the NCBI Entrez Gene database[16], we found 241 orthologs in *Gnathostomata* (jawed vertebrates), including mouse and zebrafish. Of these two model organisms, we reasoned that the mouse ortholog GM1673 would likely be more similar in function to human C4ORF48 than zebrafish. While pools of human CSF can be obtained commercially in quantities sufficient to detect C4ORF48, obtaining similar quantities of CSF from mice is much more technically challenging[17]. As such, we reasoned that if we could detect GM1673 in the conditioned media of a mouse brain cell line, this would be a

suitable alternative to show that GM1673 is a secreted polypeptide. Nicola Allen (Salk Institute) has performed peptidomic profiling experiments from the conditioned media of a primary culture of fetal mouse astrocytes (personal correspondence). Five tryptic peptides from GM1673 were detected in the conditioned media that had been concentrated using molecular weight cutoff filters, resulting in a sequence coverage of 60% (unpublished data, Figure 4.5). The annotated mass spectrum of the unique mouse tryptic peptide TETLLLQAER also showed characteristics of a high-confidence match (Figure 4.6). These findings confirm that GM1673, the mouse ortholog of human C4ORF48, is also a secreted neuropeptide.

RNA Transcript variants of C4ORF48

Two reviewed RNA transcript variants of C4ORF48 are annotated in the NCBI GenBank database[18] and are likely produced from alternative splicing events. Transcript variant 1 (accession NM_001168243.1) is a 433 bp linear mRNA that encodes a 128 amino acid polypeptide. Transcript variant 2 (accession NM_001168243.1) is a 532 bp linear mRNA that encodes a 95 amino acid polypeptide. To determine which transcript of the two variants is predominantly expressed, we ordered cDNA clones of both variants in a mammalian expression vector with a FLAG-tag inserted before the stop codon. We transfected these vectors into HEK293T cells and performed western blots with an anti-FLAG antibody (Figure 4.9).

The western blot revealed a prominent band at ~10 kDa for both variants, which corresponds more closely to the calculated molecular weight of 11.2 kDa of the 95 amino acid polypeptide with a FLAG-tag, rather than the 14.7 kDa calculated molecular weight of the 128 amino acid polypeptide with a FLAG-tag, indicating that variant 2 is the predominantly expressed form in human cells.

The longer version of transcript 1 might be processed at the RNA level to the length of version 2. This finding is consistent with northern blots from mouse brain tissue confirming that variant 2 of GM1673 is the predominant transcript version[1]. Alternatively, the expression of transcript 1 might be regulated at the level of the ribosome, where the downstream AUG start codon of variant 2 was used. Even though the mechanism that produces a protein product that matches to transcript 2 predominantly is not yet known, we focused solely on transcript variant 2 of C4ORF48 and GM1673 in our studies.

The signal peptide of C4ORF48 and GM1673

As we detected both C4ORF48 and GM1673 in extracellular fluids, this prompted us to ask if a signal peptide sequence was encoded in both genes. We analyzed both sequences using SignalP 4.1[19] and found that C4ORF48 has a predicted signal peptide sequence of 34 amino acids (Figure 10a), and GM1673 has a predicted signal peptide sequence of 28 amino acids (Figure 4.10b). We then mapped the predicted human and mouse signal peptide sequences onto a multiple sequence alignment of *Gnathostomata* orthologs and found that the secreted domain is highly conserved (Figure 4.11). The percent of identical residues in the secreted domain is 56%, including four cysteine residues. The percent of identical residues across the full-length sequence is 37%. The length of the signal peptide sequence ranged from 22 amino acids in zebrafish to 35 amino acids in cows.

Given that the signal peptide is usually cleaved co-translationally following ribosomal docking with the translocon channel at the endoplasmic reticulum[20], we revisited our tryptic peptide maps of C4ORF48 in Figure 4.2 and GM1673 in Figure 4.5. We noticed that tryptic peptides were detected in the secreted domain only, suggesting that the signal peptide sequence

was correctly predicted by SignalP 4.1. Furthermore, by only considering the secreted domain, this increased the sequence coverage to 74% of C4ORF48 and 87% of GM1673.

To further confirm if the signal peptide cleavage site was correctly predicted, we used a top-down proteomics approach. We first expressed C4ORF48 with a C-terminal FLAG-tag in HEK293T cells and performed immunoprecipitation (IP) using anti-FLAG resin from conditioned medium. We then performed an intact mass spectrometry experiment by omitting the usual trypsin digestion step and employing a multi-stage activation LC-MS/MS/MS method instead. From a fragment C4ORF48 peptide ion scan, we identified a precursor ion that corresponded to the correctly predicted secreted domain of C4ORF48 (Figure 4.12). The measured m/z of the precursor ion with +6 charge was 1329.9515, which matches the theoretical m/z of 1329.9644 (with carbamidomethylated cysteine residues) of the secreted C4ORF48 domain (Figure 4.13). These findings confirm that the signal peptide cleavage site of C4ORF48 by SignalP 4.1 is likely correct.

To confirm if the signal peptide sequence was targeting C4ORF48 through the secretory pathway, we treated HEK293T cells transfected with a C4ORF48-FLAG construct with modulators of protein secretion. We found a dose-dependent increase of C4ORF48-FLAG in the conditioned medium with ionomycin stimulation[21], while treatment with Brefeldin A[22] reduced C4ORF48-FLAG secretion to below detectable levels (Figure 4.14). These findings confirm that a 34-amino acid signal peptide targets C4ORF48 through the secretory pathway and is co-translationally cleaved after ribosomal docking at the translocon channel.

Glycosylation sites of C4ORF48 and GM1673

After we had performed an IP from conditioned media of both C4ORF48-FLAG and GM1673-FLAG, we noticed the presence of several upshifted bands on western blots. Secreted C4ORF48-FLAG from HEK293T cells had one upshifted band and GM1673-FLAG from CHO-S cells had two upshifted bands (Figure 4.15). To determine if these bands represented glycosylated forms, we treated the proteins from conditioned medium captured on anti-FLAG beads to a mixture of deglycosylases prior to western blot analysis. The upshifted band of C4ORF48 collapsed upon deglycosylase treatment (Figure 4.15a) and one of the two upshifted bands of GM1673 collapsed (Figure 4.15b). These results indicated that C4ORF48 and GM1673 are O-linked glycosylated on serine or threonine residues. Both neuropeptide sequences lack asparagine resides, thereby precluding the possibility of N-linked glycosylation.

To determine potential sites of O-linked glycosylated of C4ORF48 and GM1673, we analyzed the amino acid sequences using NetOGlyc[7]. C4ORF48 had one predicted site of glycosylation (S39) and GM1673 had three sites (T32, S34, and S40) (Figure 4.15c). We generated plasmids that encoded glycosylation acceptor-deficient versions of each neuropeptide as single alanine point mutations and a triple alanine mutant of the mouse neuropeptide with C-terminal FLAG epitopes. After the transfection of these plasmids into cells and western blot analysis of IPs using anti-FLAG resin from conditioned media, the same banding pattern was observed following deglycosylase treatment (Figure 4.15d, e). The mutation of all three glycosylation sites of GM1673 were required to abrogate the glycosylation modification. These results confirm that a sub-population of secreted C4ORF48-FLAG and GM1673-FLAG are modified by O-linked glycosylation. Further analysis will be needed to determine the identity of the sugar moiety.

The role of this glycosylation is not yet known and has not yet been reported in the literature. The remaining upshifted band on GM1673-FLAG that does not collapse after deglycosylase treatment and mutation of T32, S34, and S40 to alanine residues likely represents another modification that is not hydrolyzed by the deglycosylases used in this particular study. This modification occurs on residues other than T32, S34, and S40. We excluded phosphorylation as a possible modification by treating the IPs with lambda phosphatases as well (data not shown). Lastly, the extent of post-translational modification might be specific to the cell lines, over-expression vector, and recombinant cDNA construct that were used here and might differ from the form that is found in CSF. Some of the modified forms might be below the detection of the western blot used in this study. Multiple modified forms of C4ORF48 and GM1673 might also be produced during transit through the secretory pathway. Further examination of these forms will be required, especially to identify any modifications that are found on the bioactive forms of C4ORF48 and GM1673.

Disulfides of C4ORF48 and GM1673

The presence of four conserved cysteine residues in the sequence alignment of *Gnathostomata* orthologs prompted us to ask if disulfides were formed in the C4ORF48 and GM1673 structures (Figure 4.11). Consistent with this idea, our IPs of C4ORF48-FLAG and GM1673-FLAG from the conditioned media of cell lines showed a smear pattern by both silver staining and western blotting (Figure 4.16). This smear pattern appeared when reducing agents were omitted but collapsed into distinct bands when reducing agents were added. The formation of disulfides is characteristic of many secreted proteins as they transit through the endoplasmic reticulum[23]. The cysteine residues are grouped in pairs. The first pair of cysteine residues (C-V-

D-C) is found in the conserved region and is proximal to the signal peptide cleavage site. The second pair of cysteine residues (C-A-C) is found towards the C-terminus. The positioning of these cysteine residue pairs raised the possibility that two disulfide loops were being formed. The hominoid species also include a fifth cysteine residue, embedded in a conserved T-A-C-S-L sequence. The presence of this unpaired cysteine suggested that an inter-protein cysteine linkage might be formed between two C4ORF48 molecules.

To characterize the disulfide structure of each secreted form, we sought to determine the sites of disulfide formation by using a non-reducing proteomics workflow that preserves disulfides between cysteine residues[13]. We performed an IP of C4ORF48-FLAG and GM1673-FLAG from conditioned medium using anti-FLAG resin. Bound proteins were eluted and divided into two pools. The first pool was subjected to reduction, alkylation, and trypsin digestion to identify proteins associated with C4ORF48-FLAG or GM1673-FLAG during the IP. The second pool of the IP was subjected to trypsin digestion only, without the addition of reduction or alkylating agents, which preserves disulfides and generates cross-linked peptides. We used pLink 2[12] to identify disulfide-linked cysteines with a sequence database containing only the proteins identified in the first pool of the anti-FLAG IP. Based on these results, we generated structural models that combined all disulfides identified in the sample.

In C4ORF48-FLAG, two intra-molecular disulfide loops (C48-C51 and C88-90) and one inter-molecular disulfide linkage (C69-C69) were idenfied (Figure 4.17). In GM1673-FLAG, two intra-molecular disulfide loops (C43-C46 and C83-C85) were identified (Figure 4.18). Multiple inter-molecular disulfides between C83 and C85 were also identified, suggesting a mechanism by which GM1673 oligomers might form. Examples of annotated mass spectra for an intramolecular disulfide loop (Figure 4.19) and an intermolecular disulfide linkage (Figure 4.20)

are shown. The number of peptide spectrum matches for the disulfides can be found in Table 4.1 (C4ORF48-FLAG) and Table 4.2 (GM1673-FLAG). Based on the presence of multiple bands in the western blots in Figure 4.16b, multiple disulfides structures might exist for both C4ORF48-FLAG and GM1673-FLAG.

As with the glycosylation post-translational modifications, a limitation of our study is that the disulfide structures identified here might be specific to the cell lines, over-expression vector, recombinant cDNA construct, and purification strategy that was used here and might differ from the form that is found in CSF. We found evidence from structural studies of hepcidin that might mitigate these concerns[24]. Hepcidin is a 25-amino acid secreted peptide containing four disulfides that are prone to misfolding artifacts. In this study, Jordan *et al.* compared the disulfide structure of hepcidin that was produced recombinantly in E. coli, another version produced recombinantly from CHO-S cells, a chemically synthesized version, and the endogenous version purified from human urine. All four versions had the same disulfide structure.

Based on the findings from the Jordan *et al.* study, we chemically synthesized the 62-amino acid secreted domain of GM1673 without any epitope tags and subjected this peptide to an oxidative refolding procedure. We analyzed the disulfide structure by mass spectrometry as described in the preceding section and found the same pattern of disulfides as in GM1673-FLAG, confirming that our approach using recombinant GM1673-FLAG from conditioned medium had a similar disulfide structure. While we cannot completely exclude the possibility that our disulfide structures described in this study are not representative of the ones found in CSF, our efforts have established that only nanogram quantities of protein are required for the identification of disulfide linkages using mass spectrometry. We further plan to validate this

disulfide structure by synthesizing tryptic disulfide-linked peptides and acquire reference mass spectra with these peptides. Once an antibody to C4ORF48 has been developed, this opens the possibility of determining the disulfide structure of C4ORF48 after an IP from CSF and to compare this disulfide structure to the one obtained for C4ORF48-FLAG.

4.5 Conclusion

In this chapter, we characterized the biochemical structure of secreted human C4ORF48 and its mouse ortholog Gm1673. We detected C4ORF48 in cerebrospinal fluid and Gm1673 in the conditioned medium of fetal mouse astrocytes, confirming that this smORF-derived microprotein is a secreted neuropeptide. We further characterize the sites of signal peptide cleavage, glycosylation modification, and disulfide formation that are characteristic of other secreted peptides.

4.6 Acknowledgements

Chapter 4, in part, is currently being prepared for submission for publication of the material. Mak, Raymond; Vaughan, Joan; Diedrich, Jolene; Saghatelian, Alan. “Biochemical characterization of C4ORF48 and GM1673 neuropeptides”.The dissertation author was the primary investigator and author of this material. This work was supported by the Mass Spectrometry Core of the Salk Institute with funding from NIH-NCI CCSG: P30 014195 and the Helmsley Center for Genomic Medicine.

4.7 Figures

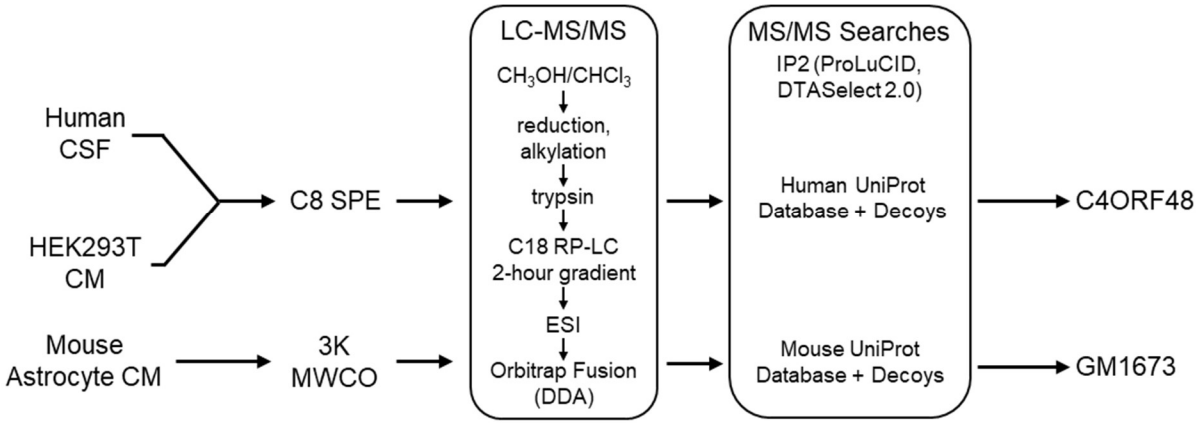
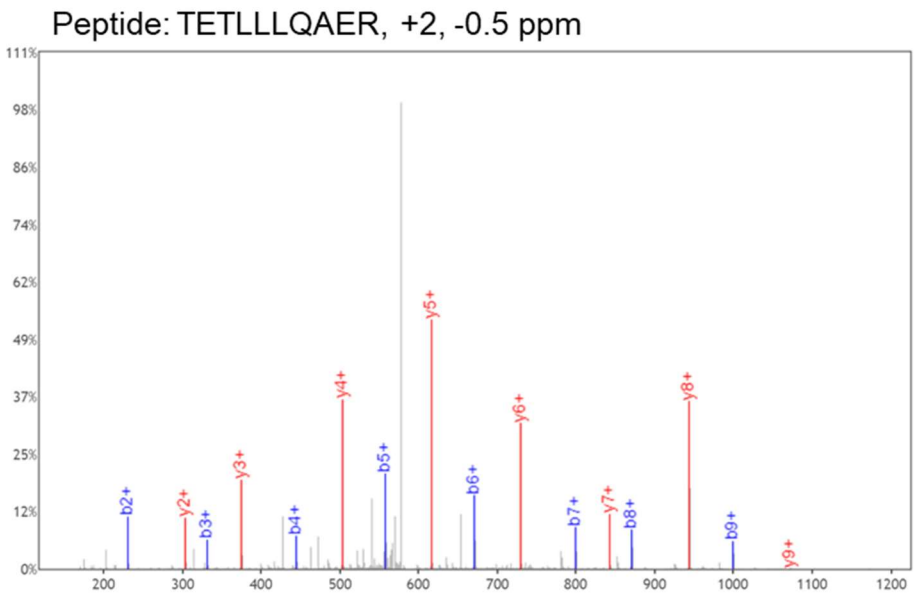


Figure 4.1: Integrated peptidomic strategy used to identify C4ORF48 and GM1673 in extracellular fluids. Human cerebrospinal fluid (CSF) and HEK293T conditioned medium (CM) subjected to solid-phase extraction (SPE) using C8 silica. CM from primary mouse astrocytes was concentrated using 3000 molecular weight cutoff (3K MWCO) filters. Peptides and small proteins were first precipitated using methanol/chloroform, reduced and alkylated, digested with trypsin, then subjected to reverse-phase liquid chromatography (RP-LC) using a C18 column and a 2-hour gradient. Eluted polypeptides were ionized using electrospray (ESI) and analyzed on an Orbitrap Fusion mass spectrometer operated in data-dependent acquisition (DDA) mode. Tandem mass spectra (MS/MS) were then searched using a target-decoy strategy against a database of either human or mouse UniProt reference proteins using the Integrated Proteomics Pipeline (IP2) data analysis platform. LC-MS/MS, liquid chromatography coupled to electrospray ionization and tandem mass spectrometry.

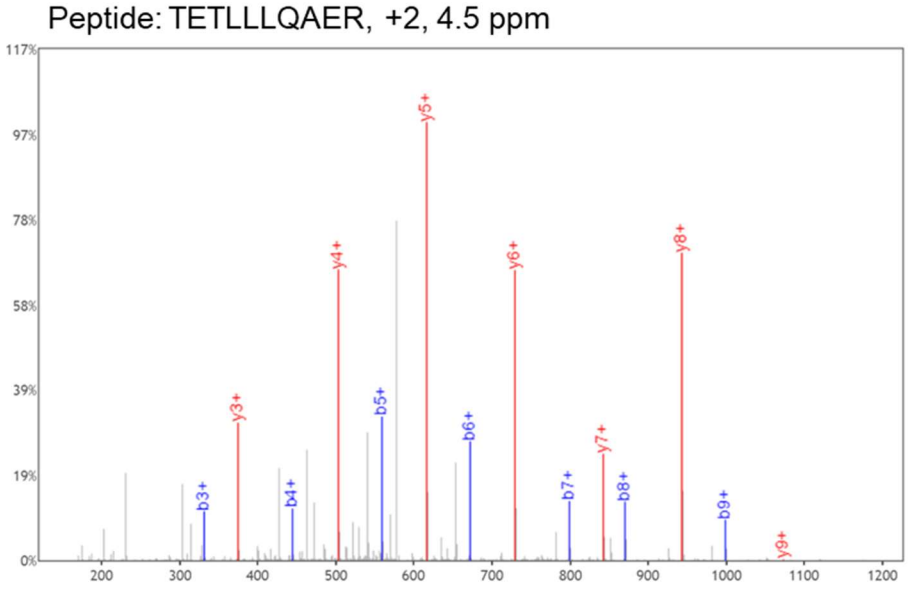
MAPPPACRSPMSPPPPPLLLLLLLSLALLGARA
RAEPAGSAVPAQSRPCVDCHAFEFMQRALQDL
RKTACSLDARTETLLLQAERRALCACWPAGH

Figure 4.2: Tryptic peptides identified in neuropeptide-like protein C4ORF48 by LC-MS/MS. The sequence of human neuropeptide-like protein C4ORF48 (UniProt accession Q5BLP8) shown in one-letter amino acid abbreviations. Yellow highlighting and underlined text indicate a detected tryptic peptide sequence.



b+	#	Seq	#	y+
102.0550	1	T	10	
231.0975	2	E	9	1072.5997
332.1452	3	T	8	943.5571
445.2293	4	L	7	842.5094
558.3134	5	L	6	729.4254
671.3974	6	L	5	616.3413
799.4560	7	Q	4	503.2572
870.4931	8	A	3	375.1987
999.5357	9	E	2	304.1615
	10	R	1	175.1190

Figure 4.3: Annotated tandem mass spectrum and list of b- and y-ions identified in the unique tryptic peptide TETLLLQAER in human cerebrospinal fluid. Precursor ion charge state and precursor ion mass error for each peptide are also listed. Parts per million, ppm. Mass-to-charge ratio, m/z.

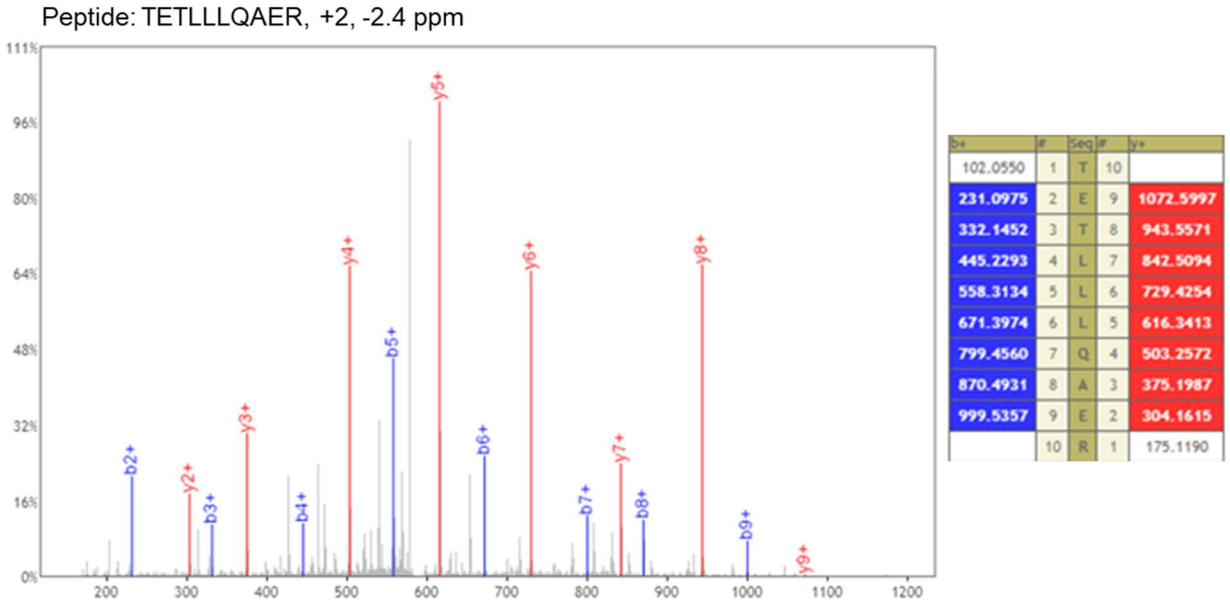


b+	#	Seq	#	y+
102.0550	1	T	10	
231.0975	2	E	9	1072.5997
332.1452	3	T	8	943.5571
445.2293	4	L	7	842.5094
558.3134	5	L	6	729.4254
671.3974	6	L	5	616.3413
799.4560	7	Q	4	503.2572
870.4931	8	A	3	375.1987
999.5357	9	E	2	304.1615
	10	R	1	175.1190

Figure 4.4: Annotated tandem mass spectrum and list of b- and y-ions identified in the unique tryptic peptide TETLLLQAER in HEK293T conditioned medium. Precursor ion charge state and precursor ion mass error for each peptide are also listed. Parts per million, ppm. Mass-to-charge ratio, m/z.

MAPALRSLLSPRTLLLLLLSLALLGARAEPAT
GSAVPAQSRPCVDCHAFEFMQRALQDLRKTAY
SLDARTETLLLQAERRALCACWPAGR

Figure 4.5: Tryptic peptides identified in neuropeptide-like protein homolog C4ORF48 by LC-MS/MS. The sequence of mouse neuropeptide-like protein C4ORF48 homolog (GM1673, UniProt accession Q3UR78) shown in one-letter amino acid abbreviations. Yellow highlighting and underlined text indicate a detected tryptic peptide sequence. Data provided by Nicola Allen (Salk Institute) and used with permission.



b+	#	Seq	#	y+
102.0550	1	T	10	
231.0975	2	E	9	1072.5997
332.1452	3	T	8	943.5571
445.2293	4	L	7	842.5094
558.3134	5	L	6	729.4254
671.3974	6	L	5	616.3413
799.4560	7	Q	4	503.2572
870.4931	8	A	3	375.1987
999.5357	9	E	2	304.1615
	10	R	1	175.1190

Figure 4.6: Annotated tandem mass spectrum and list of b- and y-ions identified in the unique tryptic peptide TETLLLQAER in mouse primary astrocyte conditioned media. Precursor ion charge state and precursor ion mass error for each peptide are also listed. Parts per million, ppm. Data provided by Nicola Allen (Salk Institute) and used with permission. Mass-to-charge ratio, m/z.

Figure 4.7: Ribosome occupancy and gene expression of C4ORF48 in HEK293T cells. The screenshot is taken from GWIPS-Viz Genome Browser at *C4ORF48* locus on human chromosome 4. Relative heights of bars represent sequencing reads from ribosome-profiling (Ribo-Seq, red) and mRNA sequencing (RNA-Seq, red). Ribo-Seq data of control-treated HEK293T cells are from Park *et al.* study[25]. Structure of exons (dark blue rectangles) and introns (dark blue dashed lines) are shown for *C4ORF48* transcript variant 1 (top) and variant 2 (bottom). Scale bar indicates 2-kilo base pairs (kb).

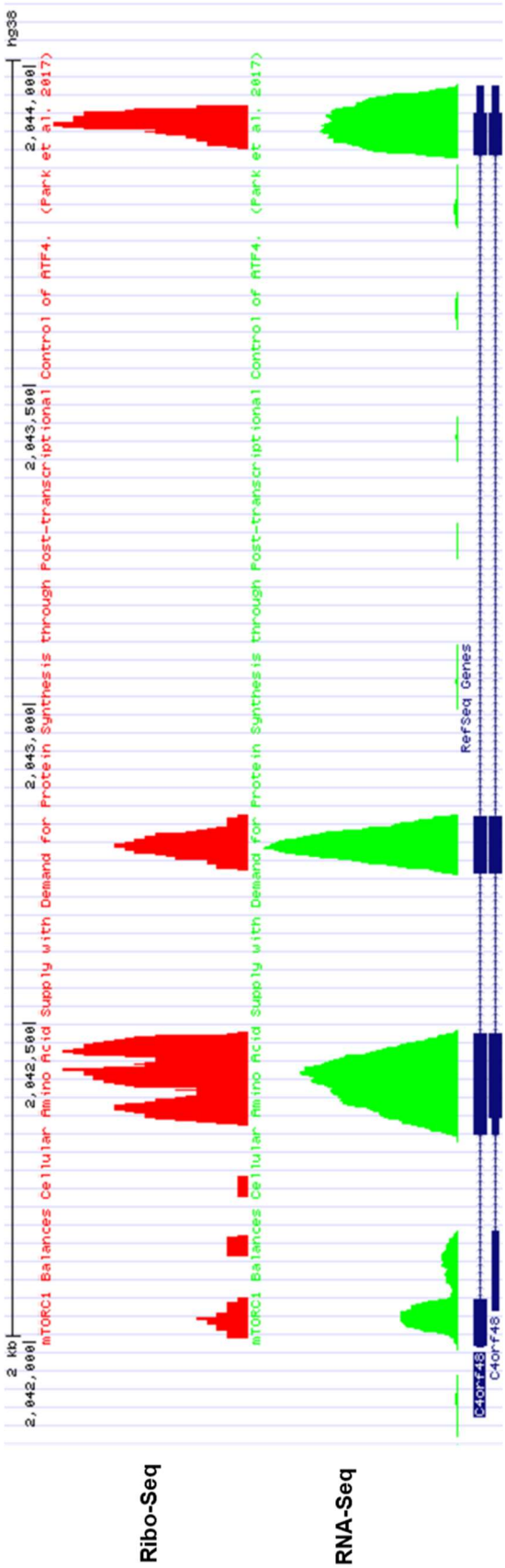
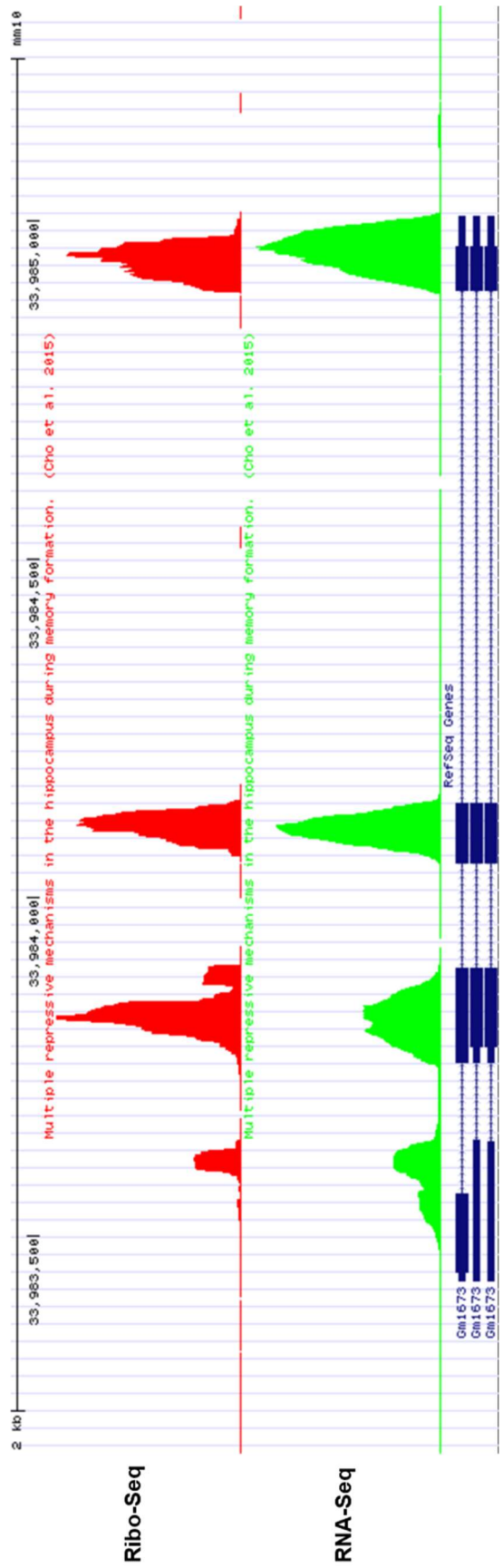

hg38
2 kb
2,042,000
2,042,500
2,043,000
2,043,500
2,044,000
mTORC1 Balances Cellular Amino Acid Supply with Demand for Protein Synthesis through Post-transcriptional Control of ATF4. (Park et al. 2017)
RefSeq Genes
C4orf48
C4orf48
Ribo-Seq
RNA-Seq

Figure 4.8: Ribosome occupancy and gene expression of Gm1673 in mouse primary hippocampal cells. The screenshot is taken from GWIPS-Viz Genome Browser at *Gm1673* locus on mouse chromosome 5. Relative heights of bars represent sequencing reads from ribosome-profiling (Ribo-Seq, red) and mRNA sequencing (RNA-Seq, red). Ribo-Seq data of control-treated mouse hippocampal primary cells are from Cho *et al.* study[26]. Structure of exons (dark blue rectangles) and introns (dark blue dashed lines) are shown for *C4ORF48* transcript variant 1 (top), variant 2 (middle), and variant 3 (bottom). Scale bar indicates 2-kilo base pairs (kb).


mm10
33,983,500
33,984,000
33,984,500
33,985,000
2 kb
Multiple repressive mechanisms in the hippocampus during memory formation. (Cho et al. 2015)
Multiple repressive mechanisms in the hippocampus during memory formation. (Cho et al. 2015)
RefSeq Genes
Gm1673
Gm1673
Gm1673
Ribo-Seq
RNA-Seq

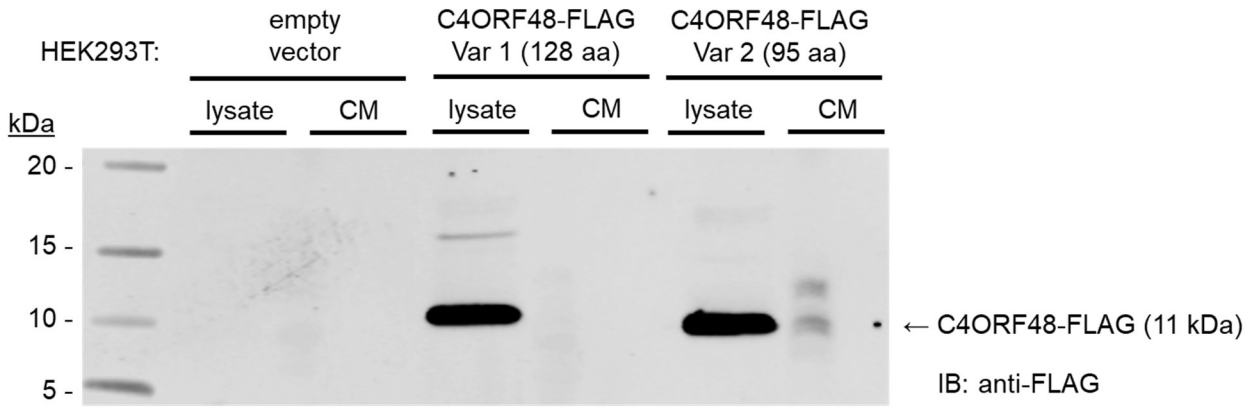
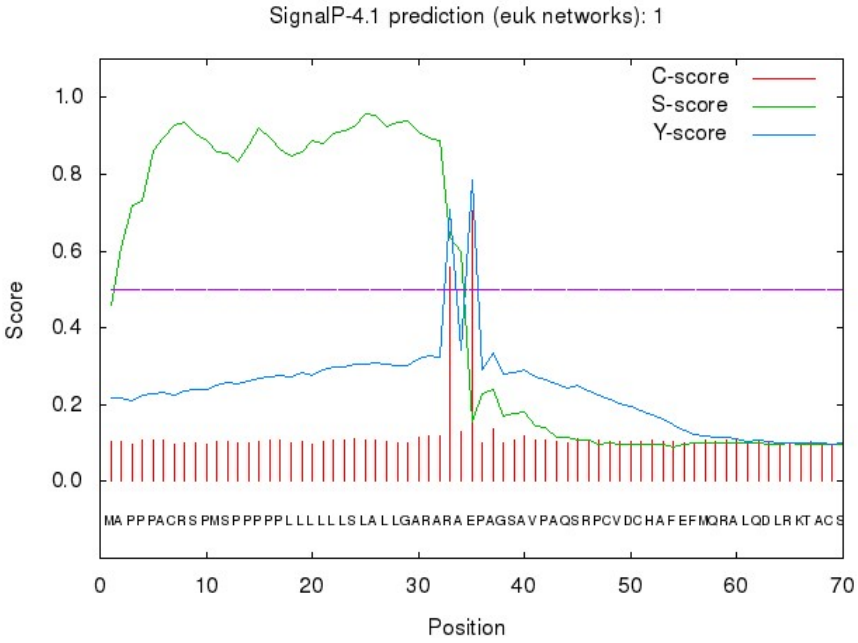


Figure 4.9: Western blot of C4ORF48 transcript variants in HEK293T cell lysates and conditioned media. HEK293T cells were transfected with plasmids containing cDNA constructs of C4ORF48 transcript variant (Var) 1 and variant 2 with a C-terminal FLAG-tag. An empty vector control plasmid containing the FLAG-tag only was also transfected. Conditioned media (CM) samples were also analyzed, but insufficient sample was loaded for detection. Kilodalton, kDa. Amino acid, aa.

(A)



(B)

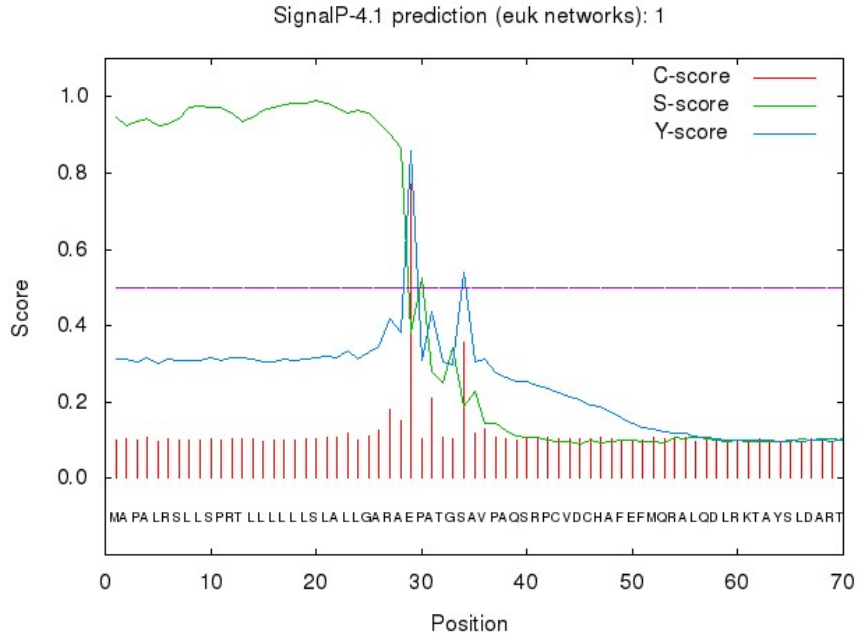


Figure 4.10: Output of SignalP 4.1 prediction of signal peptide cleavage sites. (A) C4ORF48 sequence. **(B)** GM1673 sequence. Raw cleavage site score, C-score (red). Signal peptide score, S-score (green). Combined cleavage site score, Y-score (blue). Score cutoff used in determining signal peptide cleavage site shown as a dashed horizontal magenta line.

Figure 4.11: Multiple sequence alignment of C4ORF48 and selected Gnathostomata orthologs. Protein sequences were aligned with T-COFFEE using default parameters. Identical amino acids in the alignment are indicated with a black background and an asterisk (*). Amino acids with similar chemical properties are indicated with a gray background and a colon (:). Amino acids with partially similar chemical properties are indicated with a white background and a period (.). Gaps in the alignment are indicated with a dash (-).

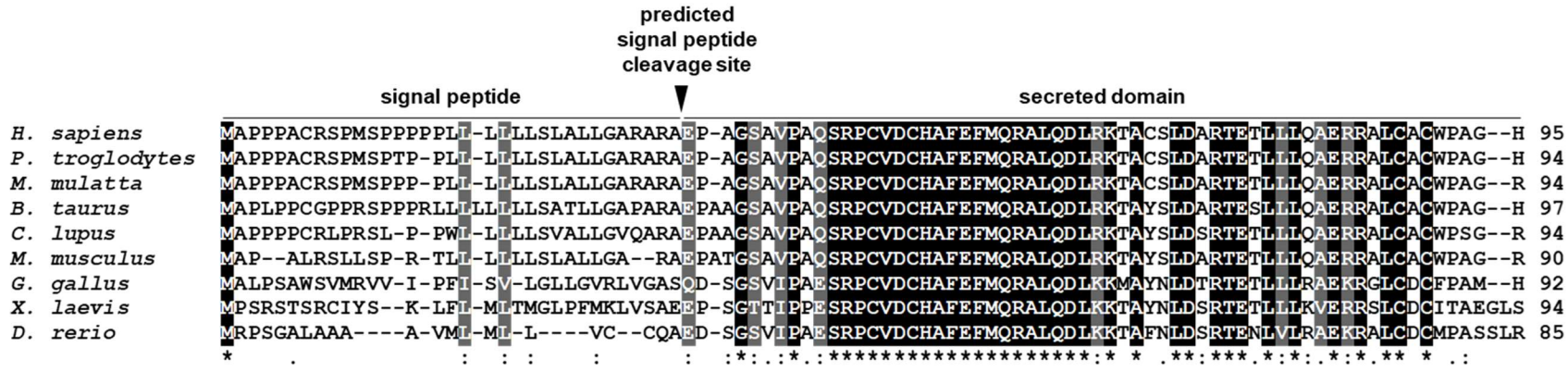
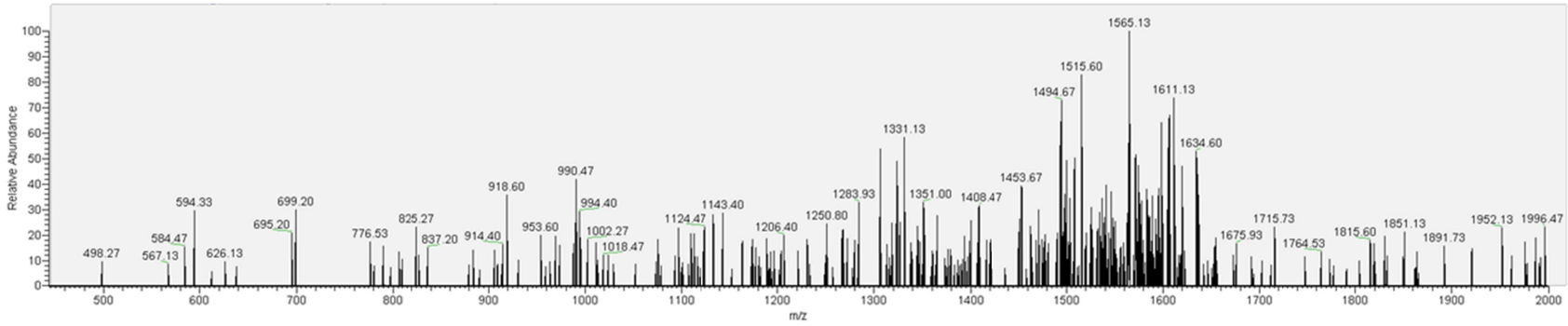
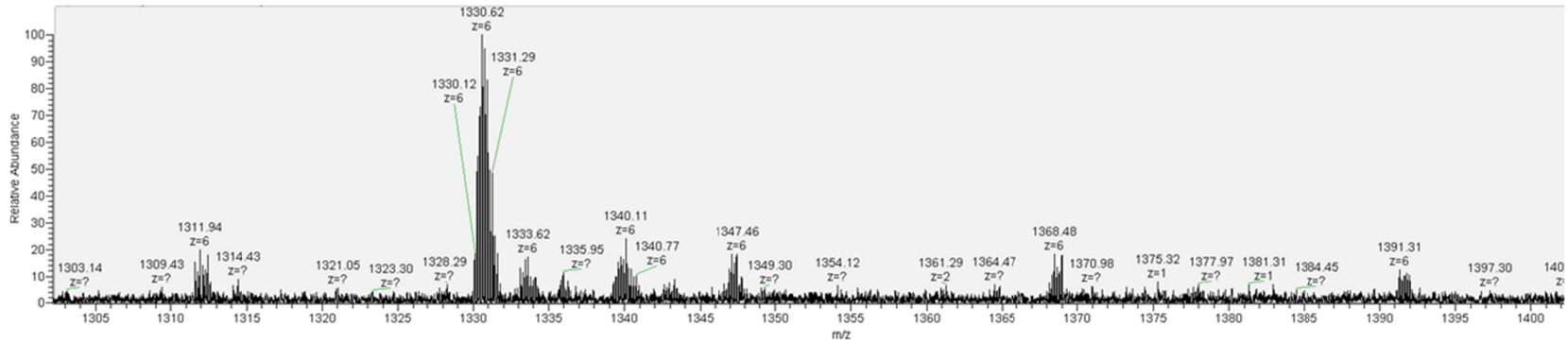

predicted
signal peptide
cleavage site
signal peptide
secreted domain
H. sapiens MAPPPACRSPMSPPPPPLL-LLLLSLALLGARARAEP-AGSAVPAQSRPCVDCHAFEFMQRALQDLRKTACSLDARTETLLLQAERRALCACWPAG--H 95
P. troglodytes MAPPPACRSPMSPTP-PLL-LLLLSLALLGARARAEP-AGSAVPAQSRPCVDCHAFEFMQRALQDLRKTACSLDARTETLLLQAERRALCACWPAG--H 94
M. mulatta MAPPPACRSPMSPPP-PLL-LLLLSLALLGARARAEP-AGSAVPAQSRPCVDCHAFEFMQRALQDLRKTACSLDARTETLLLQAERRALCACWPAG--R 94
B. taurus MAPLPPCGPPRSPPPRLLLLLLLLSATLLGAPARAEPAAGSAVPAQSRPCVDCHAFEFMQRALQDLRKTAYSLDARTESLLLQAERRALCACWPAG--H 97
C. lupus MAPPPPCRLPRSL-P-PWL-LLLLSVALLGVQARAEPAAGSAVPAQSRPCVDCHAFEFMQRALQDLRKTAYSLDSRTETLLLQAERRALCACWPSG--R 94
M. musculus MAP--ALRSLLSP-R-TLL-LLLLSLALLGA--RAEPATGSAVPAQSRPCVDCHAFEFMQRALQDLRKTAYSLDARTETLLLQAERRALCACWPAG--R 90
G. gallus MALPSAWSVMRVV-I-PFI-SV-LGLLGVRLVGASQD-SGSVIPAESRPCVDCHAFEFMQRALQDLKKMAYNLDTRTETLLLRAEKRGLCDCFPAM--H 92
X. laevis MPSRSTSRCIYS--K-LFL-MLTMGLPFMKLVSAEEP-SGTTIPPESRPCVDCHAFEFMQRALQDLKKTAYNLDSRTETLLLKVERRSLCDCITAEGLS 94
D. rerio MRPSGALAAA----A-VML-ML-L----VC--CQAED-SGSVIPAESRPCVDCHAFEFMQRALQDLKKTAFNLDSRTENLVIRAEKRALCDCMPASSLR 85

Figure 4.12: Acquired ion scans used to identify C4ORF48-FLAG from top-down proteomics. (A) Fragment ion and **(B)** precursor ion scans of C4ORF48-FLAG LC-MS/MS/MS without trypsin digestion.

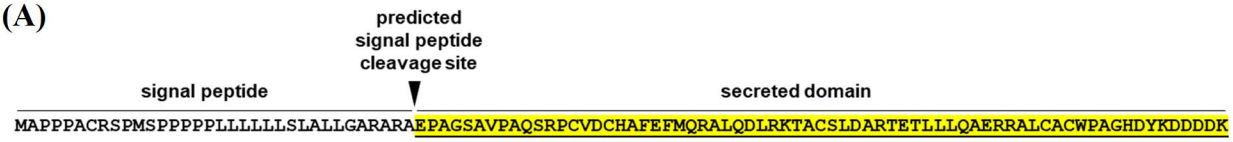
(A)


Relative Abundance
m/z
498.27
567.13
584.47
594.33
626.13
695.20
699.20
776.53
825.27
837.20
914.40
918.60
953.60
990.47
994.40
1002.27
1018.47
1124.47
1143.40
1206.40
1250.80
1283.93
1331.13
1351.00
1408.47
1453.67
1494.67
1515.60
1565.13
1611.13
1634.60
1675.93
1715.73
1764.53
1815.60
1851.13
1891.73
1952.13
1996.47

(B)


Relative Abundance
m/z
1303.14 z=?
1309.43 z=?
1311.94 z=6
1314.43 z=?
1321.05 z=?
1323.30 z=?
1328.29 z=?
1330.12 z=6
1330.62 z=6
1331.29 z=6
1333.62 z=6
1335.95 z=?
1340.11 z=6
1340.77 z=6
1347.46 z=6
1349.30 z=?
1354.12 z=?
1361.29 z=2
1364.47 z=?
1368.48 z=6
1370.98 z=?
1375.32 z=1
1377.97 z=?
1381.31 z=1
1384.45 z=?
1391.31 z=6
1397.30 z=?

(A)



(B)

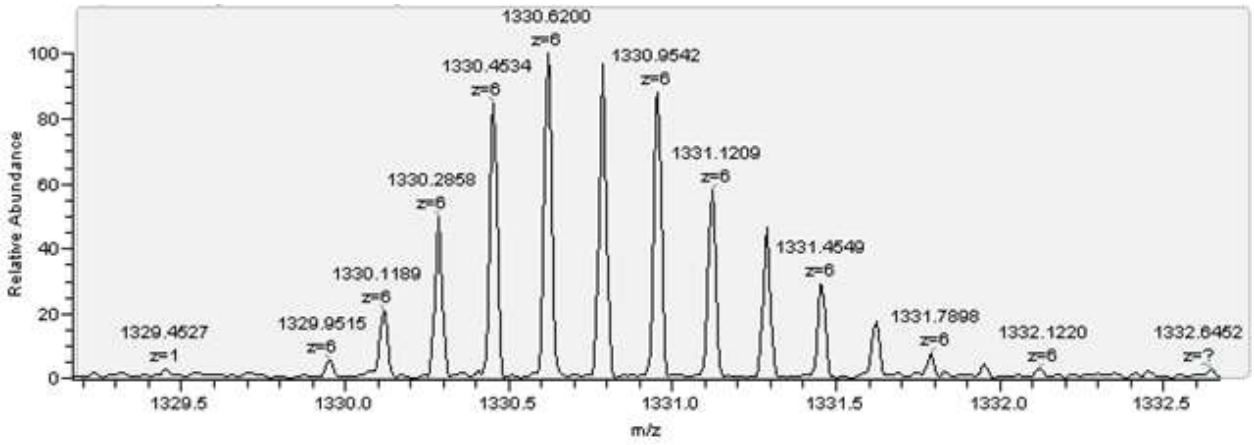


Figure 4.13: Detection of secreted C4ORF48-FLAG from HEK293T conditioned medium using top-down proteomics. (A) The sequence of human neuropeptide-like protein C4ORF48 (UniProt accession Q5BLP8) with a C-terminal FLAG-tag shown in one-letter amino acid abbreviations. A fragment ion of C4ORF48-FLAG was detected in an LC-MS/MS/MS experiment. Yellow highlighting and underlined text indicate the sequence corresponding to the precursor ion that resulted in the C4ORF48-FLAG fragment ion. (B) Enlargement of the mass envelope of the precursor ion. Measured parent mass of precursor ion is 1329.9515 m/z (z = +6). Calculated mass of precursor ion is 1329.9644 m/z (z = +6). Mass-to-charge ratio, m/z.

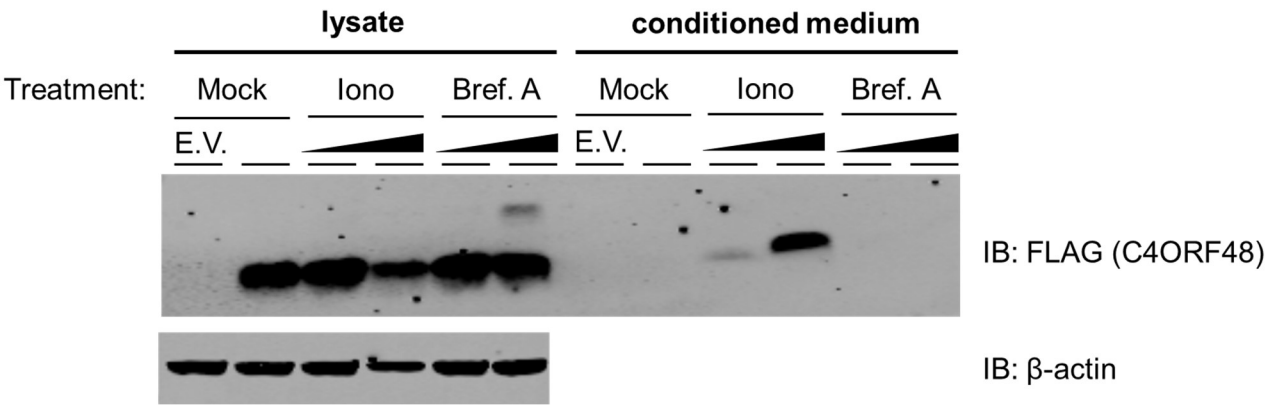


Figure 4.14: C4ORF48-FLAG is processed through the secretory pathway. HEK293T cells were transfected with an empty vector (E.V.) or a vector encoding C4ORF48-FLAG. Cells were treated with ionomycin (Iono), Brefeldin A (Bref. A), or mock-treated. Cell lysates and conditioned medium were prepared and analyzed by immunoblotting (IB).

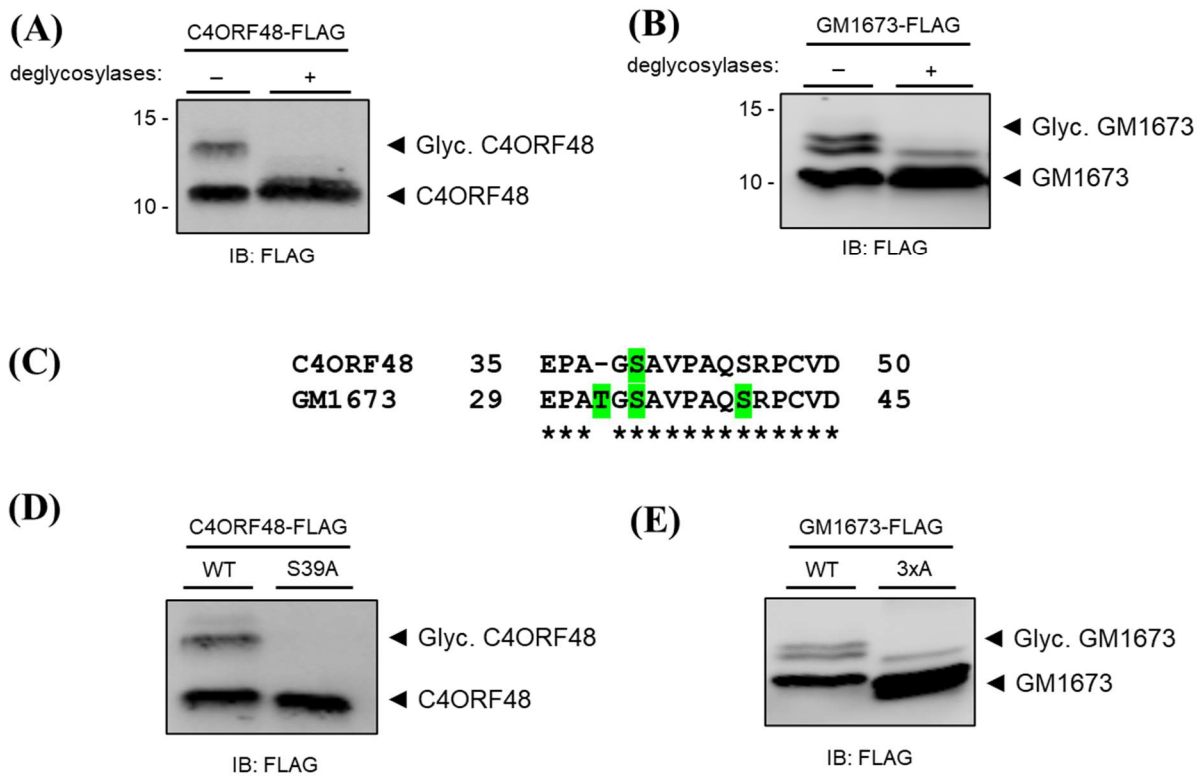


Figure 4.15. O-linked glycosylation of C4ORF48-FLAG and GM1673-FLAG. (A) C4ORF48-FLAG and **(B)** GM1673-FLAG from conditioned media were treated with (+) or without (-) deglycosylases and analyzed by anti-FLAG immunoblotting (IB). **(C)** Sites of O-linked glycosylation predicted by NetGlycO. Glycosylation-deficient mutants **(D)** C4ORF48-FLAG S39A and **(E)** GM1673-FLAG 3xA (T32A, S34A, S40A) were isolated from conditioned media and analyzed by anti-FLAG immunoblotting. Glycosylated (Glyc.) protein bands migrate more slowly relative to non-glycosylated proteins.

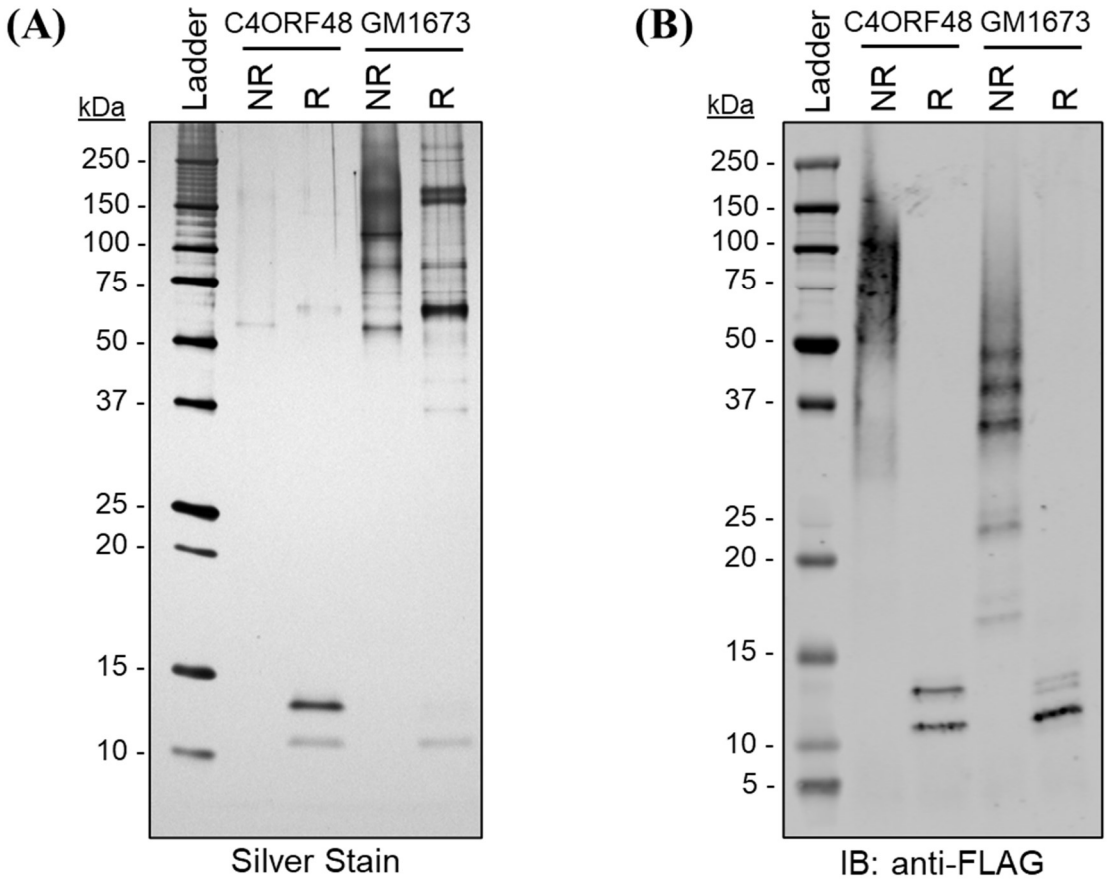
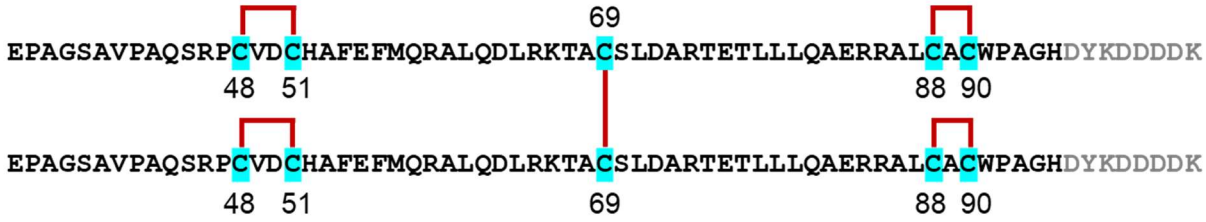


Figure 4.16: Disulfides of C4ORF48-FLAG and GM1673-FLAG from conditioned media. C4ORF48-FLAG and GM1673-FLAG were immunoprecipitated from conditioned media. **(A)** Silver staining and **(B)** anti-FLAG immunoblot (IB) following SDS-PAGE under non-reducing (NR) and reducing (R) conditions. Kilodalton, kDa.

(A)



(B)

Figure 4.17. Disulfide structural models. (A) C4ORF48-FLAG and **(B)** GM1673-FLAG. Disulfides indicated with red lines and residue numbers of cysteine residues are shown. FLAG-tag residues shown in gray lettering.

Figure 4.18: Example of an annotated mass spectrum showing an intra-molecular disulfide loop link. Mass spectra were searched for disulfides using pLink 2 and visualized in pLabel 2.4. Peptide PCVDCHAFEFMQR (z = +3) shows a loop link between C2 and C5. Matched *b*-ions peaks are shown in green. Matched *y*-ions are shown in orange.

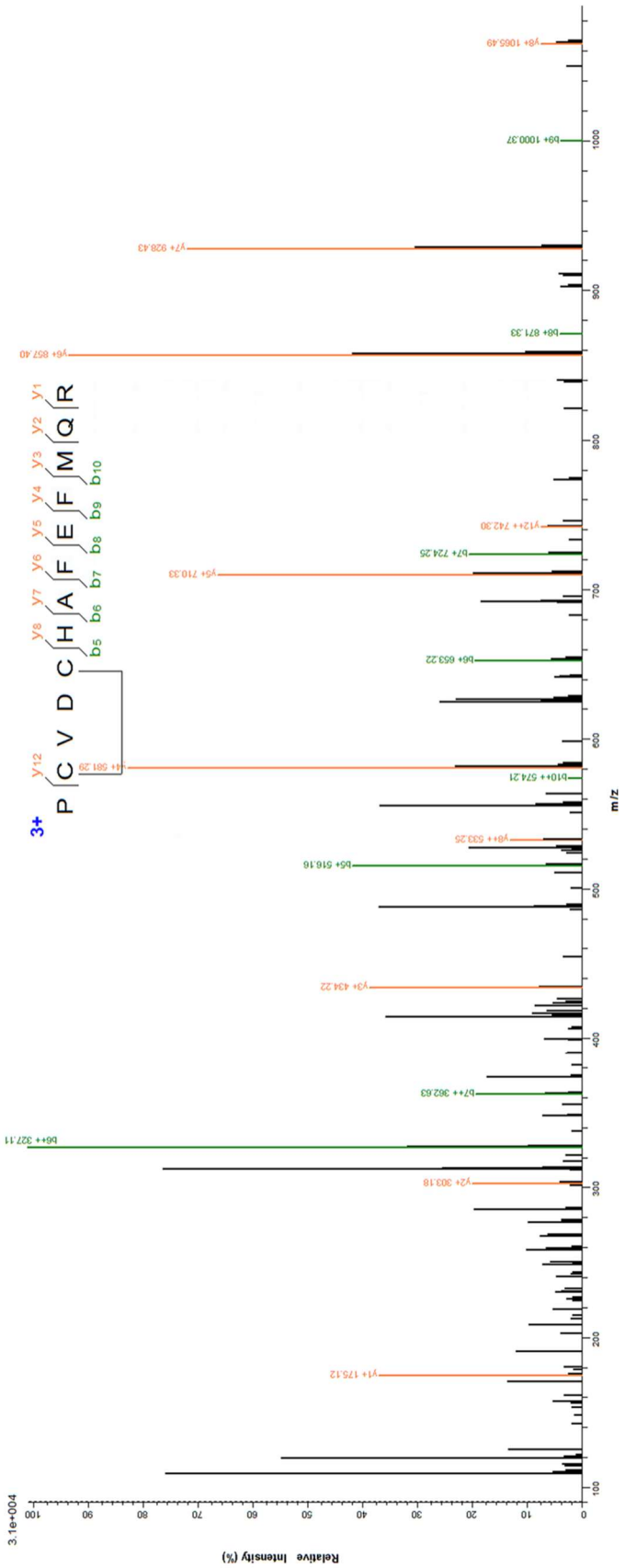
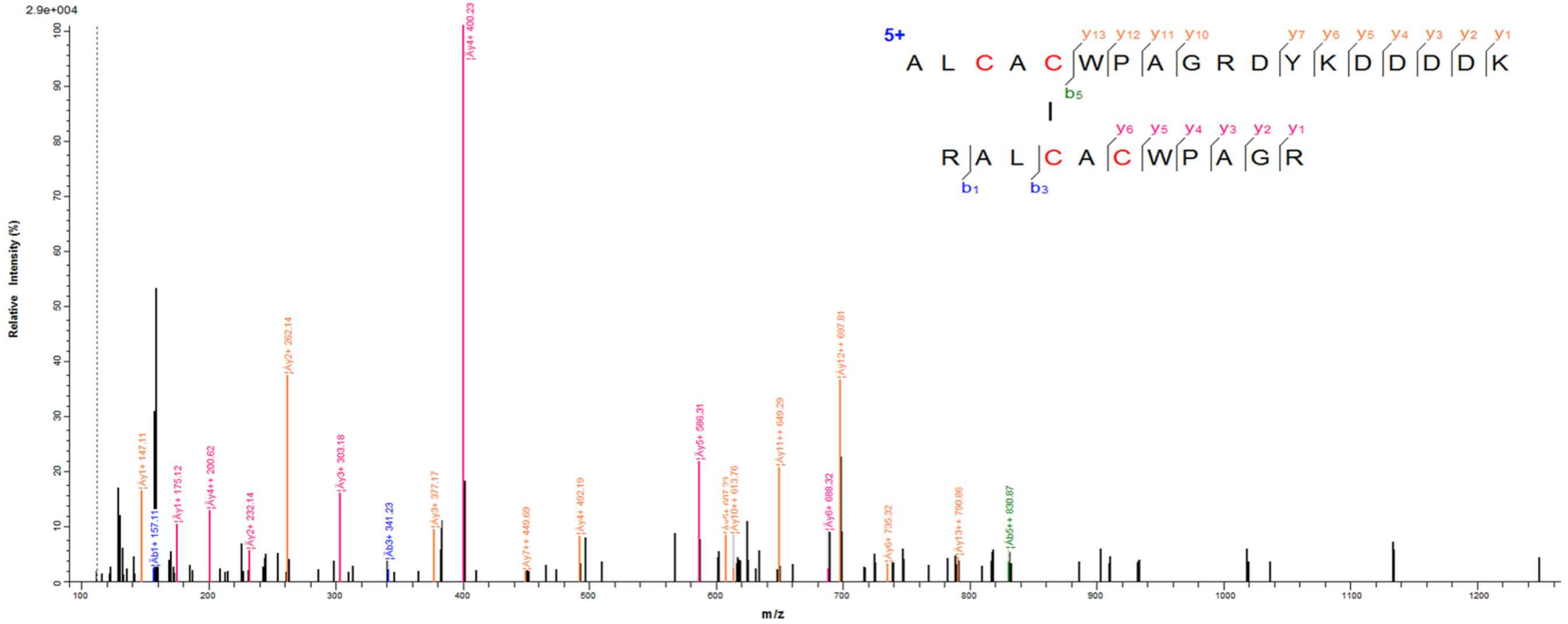

3+
P C V D C H A F E F M Q R
y12 y8 y7 y6 y5 y4 y3 y2 y1
b5 b6 b7 b8 b9 b10
Relative Intensity (%)
3.1e+004
m/z
y8+ 1065.49
b9+ 1000.37
y7+ 928.43
b8+ 871.33
y6+ 857.40
y12++ 742.30
b7+ 724.25
y5+ 710.33
b6+ 653.22
b10++ 574.21
y4+ 581.29
y9++ 533.25
b5+ 516.16
y3+ 434.22
b7++ 362.63
b6++ 327.11
y2+ 303.18
y1+ 175.12

Figure 4.19: Example of an annotated mass spectrum showing an inter-molecular disulfide cross-link. Mass spectra were searched for disulfides using pLink 2 and visualized in pLabel 2.4. Cross-linked peptides ALCACWPAGRDYKDDDDK and ALCACWPAGR (z = +5) shows a disulfide cross-link between C5 of peptide 1 and C4 of peptide 2. Matched *b*-ions peaks are shown in green and matched *y*-ions are shown in orange for peptide 1. Matched *b*-ions peaks are shown in blue and matched *y*-ions are shown in green for peptide 2.


2.9e+004
Relative Intensity (%)
m/z
5+
A L C A C W P A G R D Y K D D D D K
b5
R A L C A C W P A G R
b1
b3
y13 y12 y11 y10 y7 y6 y5 y4 y3 y2 y1
y6 y5 y4 y3 y2 y1
¡Ay4+ 400.23
¡Ay2+ 262.14
¡Ay1+ 147.11
¡Ab1+ 157.11
¡Ay1+ 175.12
¡Ay4++ 200.62
¡Ay2+ 232.14
¡Ay3+ 303.18
¡Ab3+ 341.23
¡Ay3+ 377.17
¡Ay7++ 449.69
¡Ay4+ 492.19
¡Ay5+ 588.31
¡Ay5+ 597.73
¡Ay10+ 613.76
¡Ay11++ 649.29
¡Ay12++ 697.81
¡Ay6+ 688.32
¡Ay6+ 726.32
¡Ay13++ 760.86
¡Ab5++ 830.87

4.8 References

1. Endele, S., Nelkenbrecher, C., Bördlein, A., Schlickum, S. & Winterpacht, A. C4ORF48, a gene from the Wolf-Hirschhorn syndrome critical region, encodes a putative neuropeptide and is expressed during neocortex and cerebellar development. *Neurogenetics* **12,** 155–163 (2011).

2. Rauch, A., Schellmoser, S., Kraus, C., Dörr, H. G., Trautmann, U., Altherr, M. R., Pfeiffer, R. A. & Reis, A. First known microdeletion within the Wolf-Hirschhorn-syndrome critical region refines genotype-phenotype correlation. *Am. J. Med. Genet.* **99,** 338–342 (2001).

3. Wilson, M. G., Towner, J. W., Coffin, G. S., Ebbin, A. J., Siris, E. & Brager, P. Genetic and clinical studies in 13 patients with the Wolf-Hirschhorn syndrome [del(4p)]. *Hum. Genet.* **59,** 297–307 (1981).

4. Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., Phatnani, H. P., Guarnieri, P., Caneda, C., Ruderisch, N., Deng, S., Liddelow, S. A., Zhang, C., Daneman, R., Maniatis, T., Barres, B. A. & Wu, J. Q. An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34,** 11929–11947 (2014).

5. Zhang, Y., Sloan, S. A., Clarke, L. E., Caneda, C., Plaza, C. A., Blumenthal, P. D., Vogel, H., Steinberg, G. K., Edwards, M. S. B., Duncan, J. A., Cheshier, S. H., Shuer, L. M., Chang, E. F., Grant, G. A., Gephart, M. G. H. & Barres, B. A. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89,** 1–17 (2016).

6. Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., Petryszak, R., Prada Medina, C. A., Talavera-López, C., Teichmann, S., Vizcaino, J. A. & Brazma, A. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48,** D77–D83 (2020).

7. Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T. B. G., Lavrsen, K., Dabelsteen, S., Pedersen, N. B., Marcos-Silva, L., Gupta, R., Paul Bennett, E., Mandel, U., Brunak, S., Wandall, H. H., Levery, S. B. & Clausen, H. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32,** 1478–1488 (2013).

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

9. He, L., Diedrich, J., Chu, Y. Y. & Yates, J. R. Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. *Anal. Chem.* **87,** 11361–11367 (2015).

10. Xu, T., Park, S. K., Venable, J. D., Wohlschlegel, J. A., Diedrich, J. K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X., Wong, C. C. L., Fonslow, B., Delahunty, C., Gao, Y., Shah, H. & Yates, J. R. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteomics* **129,** 16–24 (2015).

11. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).

12. Chen, Z.-L., Meng, J.-M., Cao, Y., Yin, J.-L., Fang, R.-Q., Fan, S.-B., Liu, C., Zeng, W.-F., Ding, Y.-H., Tan, D., Wu, L., Zhou, W.-J., Chi, H., Sun, R.-X., Dong, M.-Q. & He, S.-M. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10,** (2019).

13. Lu, S., Cao, Y., Fan, S.-B., Chen, Z.-L., Fang, R.-Q., He, S.-M. & Dong, M.-Q. Mapping disulfide bonds from sub-micrograms of purified proteins or micrograms of complex protein mixtures. *Biophys. Reports* **4,** 68–81 (2018).

14. Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. & Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

15. Berle, M., Kroksveen, A. C., Garberg, H., Aarhus, M., Haaland, Ø. A., Wester, K., Ulvik, R. J., Helland, C. & Berven, F. Quantitative proteomics comparison of arachnoid cyst fluid and cerebrospinal fluid collected perioperatively from arachnoid cyst patients. *Fluids Barriers CNS* **10,** 1–10 (2013).

16. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **33,** 54–58 (2005).

17. Lim, N. K. H., Moestrup, V., Zhang, X., Wang, W. A., Møller, A. & Huang, F. De. An improved method for collection of cerebrospinal fluid from anesthetized mice. *J. Vis. Exp.* **2018,** 1–7 (2018).

18. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **41,** 36–42 (2013).

19. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8,** 785–786 (2011).

20. Osborne, A. R., Rapoport, T. A. & van den Berg, B. Protein Translocation By the Sec61/Secy Channel. *Annu. Rev. Cell Dev. Biol.* **21,** 529–550 (2005).

21. Bennett, J. P., Cockcroft, S. & Gomperts, B. D. Ionomycin stimulates mast cell histamine secretion by forming a lipid-soluble calcium complex. *Nature* **282,** 851–853 (1979).

22. Orcl, L., Tagaya, M., Amherdt, M., Perrelet, A., Donaldson, J. G., Lippincott-Schwartz, J., Klausner, R. D. & Rothman, J. E. Brefeldin A, a drug that blocks secretion, prevents the

assembly of non-clathrin-coated buds on Golgi cisternae. *Cell* **64,** 1183–1195 (1991).

23. Ellgaard, L., Sevier, C. S. & Bulleid, N. J. How Are Proteins Reduced in the Endoplasmic Reticulum? *Trends Biochem. Sci.* **43,** 32–43 (2018).

24. Jordan, J. B., Poppe, L., Haniu, M., Arvedson, T., Syed, R., Li, V., Kohno, H., Kim, H., Schnier, P. D., Harvey, T. S., Miranda, L. P., Cheetham, J. & Sasu, B. J. Hepcidin revisited, disulfide connectivity, dynamics, and structure. *J. Biol. Chem.* **284,** 24155–24167 (2009).

25. Park, Y., Reyna-Neyra, A., Philippe, L. & Thoreen, C. C. mTORC1 Balances Cellular Amino Acid Supply with Demand for Protein Synthesis through Post-transcriptional Control of ATF4. *Cell Rep.* **19,** 1083–1090 (2017).

26. Cho, J., Yu, N. K., Choi, J. H., Sim, S. E., Kang, S. J., Kwak, C., Lee, S. W., Kim, J. Il, Choi, D. Il, Kim, V. N. & Kaang, B. K. Multiple repressive mechanisms in the hippocampus during memory formation. *Science (80-.).* **350,** 82–87 (2015).

Chapter 5

Concluding remarks: a personal reflection

When I was young, my parents purchased a book for me entitled *The Way Things Work: from levers to lasers, cars to computers – a visual guide to the world of machines* by David MacAuley (Houghton Mifflin, 1988). I spent hours poring over this book, becoming nauseated from the peculiarly scented ink. The pages were filled with insightful and detailed illustrations that explained how simple machines, such as levers and pulleys, and more complex ones, such as automobile transmissions and nuclear reactors, worked. I came to realize that one machine was nowhere to be found in the pages of this book: a biological cell. A cell can be thought of as a tiny and complex machine. It is composed of many parts, can store and convert energy, and—perhaps its most distinguishing feature—self-replicate. I have always been fascinated by how cells work, and primarily how they can function in a self-regulating and self-contained manner. I went to university to find out. Now, after having acquired not one, but three, advanced degrees in biology, I have more questions than answers.

Research in the biological sciences today is a truly global endeavor. I have been fortunate to share in this experience by studying and conducting research in four countries so far. I have benefited from encountering scientists from all over the world who, with their unique perspectives and approaches, are driven to answer some of the same questions that I continually pose. Some of these questions include: What are all the parts that are needed to make a biological cell? How are these parts assembled? How does a cell make an exact copy of itself? How does an organism composed of trillions of cells grow from a single one? How does a cell communicate with other cells around it? What happens when certain parts of a cell stop working correctly?

How can we repair these parts? These are complex and multifaceted problems. Solutions to these problems will only come from the continued collaborative efforts of the global community of scientists, not from working in isolation in fear of competition.

Let's briefly discuss the first question: What are all the parts that are needed to make a biological cell? We still don't know. Through decades of research, scientists found that each cell carries an instruction manual, a blueprint. The instructions are written in a chemical language, one that has just four letters—A, T, G, and C—that scientists have yet to fully understand. Yet, when read and interpreted by the cell's machinery, specifies all the parts that are needed to make a living cell. These "parts" are what we now know to be proteins, which we think of as carrying out nearly all the cell's processes. The four letters are represented by chemical bases in DNA strands. A DNA sequence that codes for a protein are called a gene. There have been extensive efforts to determine the sequence of the DNA. From knowing the DNA sequence, and applying the Central Dogma of molecular biology, we can accordingly predict what genes are present and what proteins might be produced.

We still haven't fully understood how information is encoded in the chemical language and which sequences of the DNA are translated into protein. Some of the results presented in this dissertation (Chapters 2 and 3) are part of a more recent wave of discoveries that reveal the presence of small protein-coding genes. These small genes were overlooked mainly because of their size, much like it is difficult to find a "needle in a haystack." The needle, in this case, is a small gene, and the haystack is the entire DNA sequence of an organism. Yet, as research continues to grow on these small genes, the results unambiguously suggest that they encode functional proteins and regulate other aspects of a cell, including how other proteins are made and function. These new "parts" are fundamental to a cell and await further study.

Eventually, there will need to be some reassessment of what to call these small genes–short open reading frames, small open reading frames, upstream open reading frames, noncanonical open reading frames are currently in use–and also what to call the protein products encoded within–peptides, small proteins, microproteins, micropeptides, miniproteins–to unify research in this field. The current terminology seems to suggest that these small genes and small proteins are in distinct classes when I think that they belong to the same class as any other gene or protein in a biological cell. From a chemical perspective, they are just polymers of nucleotides or amino acids. From a biological perspective, their newly uncovered functions should be viewed as expanding the repertoire of gene and protein function, rather than to be set apart. Much as the definition of genes and proteins have evolved over the years, and I think that these small genes and small proteins will challenge the current paradigm of gene and protein function.

Now, let's touch on another question: How does a cell communicate with other cells around it? Historically, this has been a challenging and time-intensive question to answer. We now know that cells use multiple types of signals to do so. Some of these signals are chemical in nature, such as peptide hormones. Peptide hormones act as chemical messengers that are produced and secreted by one cell, travel, and bind to a receptor on the surface on another cell. Once the hormone binds to a receptor, a signal is transmitted through the receptor to the inside of that cell, and the activity of an intracellular pathway is changed. The first discoveries concerning peptide hormones were focused on isolating the "active principle" or "active material" in tissue extracts that were responsible for an observable effect. In the case of diabetes in the early 1920s, insulin was isolated as the active principle from pancreatic extracts that had the effect of countering elevated blood sugar levels[1]. It took many years of research to determine that insulin

was a small protein, isolate it in pure form[2], and more than three decades to determine its protein sequence[3] and structure[4].

Hundreds of peptide hormone molecules, including neuropeptides that operate in the brain, have since been discovered. Have we discovered all peptide hormone-like molecules? Probably not, because recent studies that combine next-generation sequencing, proteomics, and bioinformatics have revealed that there are many more genes that encode small proteins than previously thought[5].

Discoveries in biology have always been driven by technological innovation. Some examples include the introduction of spectrophotometry to quantify and detect biomolecules, crystallography to determine molecular structure, and the polymerase chain reaction to amplify sequences of nucleic acid. I would argue that the introduction of mass spectrometers should be viewed in the same light. Rather than sequencing proteins one-at-a-time, now thousands of proteins can be sequenced from a sample.

I wondered if the application of modern proteomics techniques could be used to accelerate the discovery of peptide hormones and similar molecules. Indeed, this has been the foundation for much of this dissertation. I first started by developing a strategy to detect peptides and small proteins using mass spectrometry-based proteomics techniques. The main challenge was to find a way to reduce the complexity of the sample such that peptides and small proteins could be more readily detected.

As advanced as our current generation of mass spectrometers are, there are still too many peptides and proteins in a biological sample to be detected in one experiment only. To overcome this hurdle, I took a page out of the “classic” biochemist’s book: use fractionation. I fractionated tissues (brain, Chapter 2) and biological fluids (cerebrospinal fluid, Chapter 3) that would likely

contain undiscovered peptide hormones, neurotransmitters, and neuropeptides and subjected the fractions to an analysis by mass spectrometry. I uncovered some new genes that might encode small proteins with these functions. I then characterized the sequence of one of these small proteins that is found in the human brain (C4ORF48) and mouse brain (Gm1673) in Chapter 4. The molecular function of this small protein, or neuropeptide, isn't known yet, but it was essential to determine its structure before proceeding with functional studies. Part of what makes peptide hormones challenging to study is that they have highly modified structures. The structure is difficult, perhaps even impossible, to predict from its sequence alone. It is this specific structure of the active peptide that is the basis for binding to a specific cell-surface receptor and signal transduction. I used C4ORF48 and Gm1673 as a model of a small, secreted peptide that could be studied using mass spectrometry-based techniques.

The field of peptide biology, much like any biological field, will continue to benefit from the application of mass spectrometers. However, its application to biological samples required extensive adaptation and optimization at first. For example, ionization techniques had to be developed before mass spectrometers could be used to impart a charge on peptides without inducing fragmentation[6–8]. Computer algorithms that could interpret the complex fragment ion mass spectra generated from peptides also had to be developed[9]. As in the case with peptides, I see a few areas where the continued development of current technologies will be beneficial. The first is lowering the detection limit and increasing the scan speed of the mass analyzer. This will allow the more ready detection of peptide species found at low abundance. Another area includes fractionation methods that are optimized at separating peptides from larger proteins. The use of de novo peptide sequencing algorithms to interpret mass spectra will be invaluable in detecting peptides with extensive post-translational modification. Finally, the continued development of

algorithms to detect cross-linked peptides will be needed to detect disulfide (or other) structural modifications in bioactive peptides. With these tools in hand, the "active principle" or "active material" from an extract could be readily identified over the course of a few months, rather than decades.

At this point, I would like to unequivocally state that I am not interested in using mass spectrometry to compete in "the numbers game." I get increasingly worried that the competition to report increasing numbers of peptide and protein identifications has in itself become a justification for publishing scholarly papers. Unfortunately, in Chapters 2 and 3 of this dissertation, I have done just this. I would like to clarify that the numbers of protein identifications reported in those chapters and the lists of proteins included in the supplementary tables are not meant to showcase technological prowess. Instead, I wanted to be transparent in fully reporting what data were obtained and how. I also want to emphasize that adapting existing technologies to a biological problem is an extensive process, one that has consumed much of my graduate studies. I am sure that decades from now, these numbers will seem minuscule in comparison to studies that will be using even more advanced technology. I leave it up to the reader to determine which of the data to keep, which data to discard, and which data are worthy of further study.

Finally, what do I hope will become of this work? I realize that this dissertation and any chapter published as a research article will have a limited readership. As of this writing, no diseases have been cured, no groundbreaking drugs have been developed, and no lucrative patents have been filed as a result of my work. I take solace that I have uncovered new protein-coding gene sequences and have been successful at adapting mass spectrometry-based technologies to studying protein structure. These are small steps, but I hope that it will enable

other researchers to find new genes to study, uncover new bioactive peptides, define new avenues to pursue as therapeutic targets, and accelerate the rate at which other discoveries are made. I remain hopeful that I have contributed to advancing the field of biological research in a small and, eventually, meaningful way.

–Raymond H. Mak
San Diego, California
May 1, 2020

References

1. Banting, F. G., Best, C. H., Collip, J. B. & Macleod, J. J. R. The preparation of pancreatic extracts containing insulin. *Trans. R. Soc. Canada* **16,** 1–3 (1922).

2. Abel, J. J. & Geiling, E. M. K. Rechearches on insulin: Is insulin an unstable sulphur compound? *J. Pharmacol. Exp. Ther.* **25,** 423–448 (1925).

3. Sanger, F. Chemistry of Insulin: Determination of the structure of insulin opens the way to greater understanding of life processes. *Science* **129,** 1340–1344 (1959).

4. Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., Mercola, D. A. & Vijayan, M. Atomic positions in rhombohedral 2-zinc insulin crystals. *Nature* **231,** 506–511 (1971).

5. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* **11,** 909–916 (2015).

6. Fenn, J., Mann, M., Meng, C., Wong, S. & Whitehouse, C. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).

7. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. & Matsuo, T. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2,** 151–153 (1988).

8. Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Anal. Chem.* **60,** 2299–2301 (1988).

9. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).