UC San Diego
UC San Diego Electronic Theses and Dissertations

Title
Biochemical Characterization of Proteolytic Activities and Substrate Specificities

Permalink
https://escholarship.org/uc/item/4vm209n3

Author
Jiang, Zhenze

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Biochemical Characterization of Proteolytic Activities and Substrate Specificities

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in

Chemistry

by

Zhenze Jiang

Committee in charge:

      Professor Anthony J. O'Donoghue, Chair
      Professor Elizabeth A. Komives, Co-chair
      Professor Vivian Hook
      Professor Tadeusz F. Molinski
      Professor Wei Wang

2020

The dissertation of Zhenze Jiang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-chair

_____
Chair

University of California San Diego

2020

DEDICATION

I dedicate this work to my parents for their unwavering support, love, encouragement and

guidance. I wouldn't have accomplished this without them.

EPIGRAPH

Let life be beautiful like summer flowers and death like autumn leaves

~ Rabindranath Tagore

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ADC	Antibody–drug conjugates
AOMK	Acyloxymethyl ketone
AMC	7-amino-4-methylcoumarin
AMP	Antimicrobial peptides
ANOVA	Analysis of variance
Aspph	*Aspergillus phoenicis*
BLH	Bleomycin hydrolase
CG	Chromaffin granule
DMSO	Dimethyl sulfoxide
DNP	2,4-dinitrophenol
Ex/Em	Excitation/Emission
GuHCl	Guanidine hydrochloride
hBD2	Human β-defensin 2
HNP3	Human neutrophil peptide 3
IQ	Internally quenched
Mca	(7-methoxycoumarin-4-yl)acetyl
MSP-MS	Multiplex substrate profiling by mass spectrometry
PAP	Protease-activated prodrug
PENK	Proenkephalin
RMSD	Simply root-mean-square deviation
TFA	Trifluoroacetic acid
TMT	Tandem mass tags

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my Ph.D. advisor, Dr. Anthony O'Donoghue for his support, mentorship, encouragement and advice. I am very lucky to be his first graduate student to have the experience helping him building us his lab. I really appreciate him being so approachable to chat, open to ideas and different opinions, and granting me a lot of trust and freedom to work on my projects. I am grateful for all the research, collaborations and conferences opportunities that I had during my Ph.D. tenure, which well prepared me for my career. It has been an absolute pleasure and an honor working with him.

Next, I'd like to thank Chris Lietz, Steven Wang, Pavla Fajtova, Betsaida Bibo Verdugo, Michael Yoon, Chenxi Liu, Sonia Podvin, Brian Suzuki and all other present and past group members for a wonderful lab environment and teamwork. I would like to give special thanks to Chris Lietz, who is my mentor on mass spectrometry and proteomics. He patiently answered all my questions, taught me everything he knows in details and helped me troubleshooting LC-MS problems. I wouldn't have acquired so much knowledge and skills without the help of him. I would like to sincerely thank my colleague and friend Steven Wang. He's a brilliant person to work with, and I am grateful for all the helps he offered in maintaining LC-MS, the knowledge he shared and all the funs we had in and out of the lab.

I must thank all my collaborators, especially Vivian Hook, Dennis Wolan, Conor Caffrey and M. Joanne Lemieux. It was fantastic to work with them and their team members. I am glad to get involved in many very different research projects and have learnt a lot knowledge outside of my field of study. And I am glad that their groups and our group have established good relationships and long-term collaboration. I want to also thank my committee members, especially my co-advisor Elizabeth Komives. She was incredibly supportive, providing guidance for my PhD

study and my career development. I would like to also thank my friends for their support and the all times we spent together.

Last but not least, I would like to again thank my dearest parents for their unconditional love. They have been incredible supportive for all the choices I made, and never doubt my ability. I am so lucky to have them who understand and encourage me to pursue my goals.

Chapter II is a full reprint of the material as it appears in the article published in ACS chemical biology, 13(9), pp.2513-2521 by Janice H. Xu, Zhenze Jiang, Angelo Solania, Sandip Chatterjee, Brian Suzuki, Christopher B. Lietz, Vivian Y.H, Hook, Anthony J. O'Donoghue, and Dennis W. Wolan. The thesis author is the primary researcher and author of the article.

Chapter III is a modified reprint of the material as it appears in the manuscript prepared to be submitted, by Zhenze Jiang, Christopher B. Lietz, Sonia Podvin, Michael Yoon, Vivian Hook, Anthony J. O'Donoghue. The thesis author is the primary researcher and author of the article.

Chapter IV is a full reprint of the material as it appears in the article published in Molecular & Cellular Proteomics, 18(5), pp.968-981 by John D. Lapek, Zhenze Jiang, Jacob M. Wozniak, Elena Arutyunova, Steven C. Wang, M. Joanne Lemieux, David J. Gonzalez, and Anthony J. O'Donoghue. The thesis author is the primary researcher and author of the article.

VITA

2015   Bachelor of Science, Chemistry, Shaanxi Normal University, China

2017   Master of Science, Chemistry, University of California San Diego

2020   Doctorate of Philosophy, Chemistry, University of California San Diego

ABSTRACT OF THE DISSERTATION

Biochemical Characterization of Proteolytic Activities and Substrate Specificities

by

Zhenze Jiang

Doctor of Philosophy in Chemistry

University of California San Diego

Professor Anthony O'Donoghue, Chair

Proteases are crucial components of life. They regulate numerous biological pathways, such blood coagulation, neurotransmission, cell proliferation and apoptosis. They are also found to be involved in the development cancer, Alzheimer's and many infectious diseases. Proteases, as a post-translational modification enzyme, regulate biological processes through proteolysis.

And like many other enzymes, one of the key features of proteases is their substrate specificity, which is tightly controlled by specific interactions between the amino acid sequences of the substrates and their binding pockets. It's important to characterize proteases' specificities as it is a key to understand their biological functions. This work focused on the development and applications of a mass spectrometry-based technology, named Multiplex Substrate Profiling by Mass Spectrometry (MSP-MS), to characterize the specificities of proteases, and the design of fluorescent substrates and inhibitors based on protease specificity profiles.

Chapter II demonstrates the application of MSP-MS in uncovering the substrate specificity of a bacterial protease, Pd_dinase, which is a putative C1B-like cysteine protease commonly secreted by the human gut commensal *Parabacteroidetes distasonis*. We designed and synthesized a potent protease inhibitor glycine-arginine-*AOMK* based on its specificity profile. Furthermore, we revealed that Pd_dinase hydrolyzes several human antimicrobial peptides, such as β-defensin 2 and keratin-derived antimicrobial peptides, indicating that it may be secreted into the extracellular milieu to assist in gut colonization by inactivation of host antimicrobial peptides.

Chapter III presents the study to extend the application of the MSP-MS technology, together with proteomics and peptidomics, to establish a multi-omics platform to studying complex biological samples. In this study, we isolated bovine chromaffin granules (CGs) and quantified the endogenous proteins and peptides through proteomics and peptidomics. In addition, we performed degradation assays to profile proteolytic processing of proneuropeptides and MSP-MS assays to characterize proteas activities. With comprehensive profile of proteases, substrates and proteolytic specificities, we were able to discover six catalytically active proteases and assign their activities to some of the cleavages on proneuropeptides.

Chapter IV presents the study of developing quantitative Multiplex Substrate Profiling by

Mass Spectrometry (qMSP-MS) method by combining the quantitative power of tandem mass tags (TMTs) with our previously established peptide cleavage assay, MSP-MS. We validated the method with papain, a well-characterized cysteine peptidase and uncovered the substrate specificity of two minimally characterized intramembrane rhomboid proteases. We further showed that activity from multiple peptidases in complex biological samples can be deciphered, including secretions from lung cancer cell lines. Discovery of the protease specificity at the site of the disease highlights the potential for qMSP-MS to guide the development of protease-activated prodrugs for cancer and infectious disease.

Chapter I

Introduction

A. Proteases are pivotal components in biological processes

Proteolysis is one of the important post-translational modifications (PTMs) that is regulated by a class of enzymes called proteases. They comprise approximately 2% of total proteins in human and were categorized 5 classes, which are aspartic, cysteine, serine, metallo and threonine, based on the catalytic mechanisms at their active sites[1-3]. Not similar to most PTM enzymes, proteases generally irreversibly hydrolyze peptide bonds, and thus it is imperative that their activities are tightly controlled. They regulate numerous biological processes, including blood coagulation, digestion, apoptosis, neuropeptide maturation, cell proliferation and immune response[4-12]. The molecular mechanisms of proteolytic regulation can be generally divided into two aspects. On the one hand, many proteins, such as receptors, kinases, transcription factors and even protease themselves, are cleaved proteases at specific sites to get activated or alter their present functions. For example, in the last step of blood coagulation, thrombin cleaves soluble fibrinogen into insoluble strands of fibrin which polymerizes into clots[5]. On the other hand, some proteases are involved in cleaving proteins to inactivate their functions. For example, misfolded, damaged or unneeded proteins will be targeted by ubiquitination and sent to proteasome, which is a proteolytic complex consists three catalytic subunits, for degradation[13].

B. Proteolytic enzymes as emerging drug targets

As pivotal elements of life, proteases are also found to be associated with a variety of diseases[14-16]. Many caspases such as caspase 8 and 10 were often found to be inactivated or silenced through somatic mutations or DNA methylation[17]. On the contrary, cysteine cathepsins are frequently found with an increased expression and aberrant localization in various tumors[18, 19]. Down-regulation or inhibition of cathepsins, such as cathepsin B and cathepsin K, reduces

tumor growth, invasion and angiogenesis in several *in vitro* and *in vivo* studies[20, 21]. Extracellular proteases are also involved in cancer progression. Matrix metalloproteases were a family of 23 proteases that found to be over secreted into tumor microenvironment. They have been associated with tumor expansion and metastasis by degrading extracellular matrix (ECM)[22-24]. They can also activate several growth factors such as TGF-β to induce angiogenesis[25]. In addition to cancer, dysregulated protease activities are also linked to diseases, including Alzheimer's, Parkinson's, arthritis and vascular diseases[9, 14, 26]. Proteases are also great targets to combat infectious diseases. Many efforts have been made to inhibit HIV-1 protease for the treatment of AIDS[27, 28]. SARS-CoV-2 virus of the current on-going COVID-19 pandemic utilizes 3-chymotrypsin-like protease (3CL^{pro}, as known as main protease (M^{pro})) and papain-like protease (PLpro) to process its polyprotein and are essential for its life cycle[29, 30]. Proteasomes are also well-established targets not only for multiple myeloma and autoimmune diseases, but also for parasitic infections[31].

Many studies have been focusing on developing small molecule protease inhibitors. Most successful compounds include human constitutive proteasome inhibitors bortezomib and carfilzomib for the treatment multiple myeloma[32]; dipeptidyl peptidase-4 (DPP-4) inhibitors vildagliptin and sitagliptin for the treatment of diabetes[33]; angiotensin-converting enzyme (ACE) inhibitors enalapril and captopril for the treatment of hypertension, diabetic kidney disease, and heart failure[34]; and many HIV protease inhibitors, such as saquinavir, indinavir, ritonavir, nelfinavir, amprenavir, lopinavir, etc., which were often used together for the treatment of AIDS[28]. Despite of the success of above-mentioned small molecule protease inhibitors, many compounds failed in clinical trials due to their high cytotoxicity which commonly caused by off-

target inhibition or nonselective-inhibition of the target protease that involved in other biological pathways[35].

Another strategy is protease-activated prodrugs (PAPs) where the drug is conjugated with a protease-activatable linker. The prodrug conjugate remains inactive until the linker is cleaved by the target protease at the disease site[36, 37]. There have been several successful examples such as FDA-approved antibody–drug conjugates (ADCs) brentuximab vedotin for treating Hodgkin lymphoma (HL) and anaplastic large cell lymphoma (ALCL)[38], and polatuzumab vedotin for the treatment of diffuse large B-cell lymphoma[39]. They both have the sequence of valine-cit designed in the linker that can be hydrolyzed by cathepsin B which is a lysosomal protease commonly found to be overexpressed in cancer. The therapeutic moiety of the conjugate can also be swapped with an imaging agent for disease diagnosis and image-aided surgery. PAPs are generally less cytotoxic as they do not exhibit direct inhibitory effect. The design of the linker is often the most challenging part since it has to avoid the proteolytic degradation by off-target proteases.

C. Technologies for characterizing protease specificity

One of the keys to understand the functions of proteases is to characterize their substrate specificities. The knowledge of proteases' specificities facilitates the discovery of their endogenous substrates and their roles in biological pathways. It also provides us with structural insights in designing peptide-like protease inhibitors, PAPs linkers, as well as lead optimization. Proteases recognize their endogenous substrates mainly based on their amino acid sequences. This simple protease-substrate interaction mechanism makes peptide degradation assays ideal for revealing protease specificity. With the development of modern biochemical technologies, many

peptide libraries and methods have been constructed for large-scale profiling of protease specificity[40-42].

Positional scanning substrate combinatorial library (PS-SCL) consists a mixture of synthetic peptides with amino acids at one position fixed and other positions randomized[43]. These peptides are attached to a 7-amino-4-methylcoumarin (AMC) or 7-amino-4-carbamoylmethylcoumarin (ACC) fluorophore at P1′ position, which serves as the signal readout. A high fluorescent intensity indicates a strong preference of the amino acid at that position. Some other fluorescent substrate libraries, such as hybrid combinatorial substrate library (HyCoSuL) and Counter Selection Substrate Library (CoSeSuL), were constructed with include unnatural amino acids[44, 45]. These libraries extended the pool of amino acids that were commonly used for specificity profiling and showed the advantages especially in designing selective and potent protease inhibitor, substrates and protease-activating prodrugs. However, most fluorescent substrate libraries are limited by P1′ fluorophore, and thus, are only profiling non-prime side specificity. To overcome this problem, internally quenched fluorescent (IQF) peptides originating from FRET (Förster Resonance Energy Transfer) were also developed to uncover the specificities of endopeptidases[46]. They are intact peptides with fluorophore and quencher labeled on two termini and a cleavage in between the peptide sequences will produce fluorescent signal. Though they allow researchers to study protease specificity at both sides of the cleavage site, they clearly have several disadvantages. Firstly, the cleavage site can't be simply determined with fluorescent readout as the cleavage occurs anywhere in between the sequence will produce fluorescence. Mass spectrometry analysis is often required to determine the cleavage site. Secondly, multiple cleavages can occur within the same IQF peptide sequence, making it difficult to determine the cleavage site precisely. Thirdly, these substrates can't be

used to study exopeptidases as their two termini of were labeled. Phage, yeast and bacterial display are also widely used to determine protease specificity[47, 48]. The substrate libraries are usually labeled with affinity tags/support or color dyes for separation of sensitive and resistant clones, and cleaved substrates are identified through DNA sequencing. Though these substrate display methods allow researchers to build largest, most comprehensive and diverse library, they are not able to identify cleavage sites precisely, and downstream validation by bioinformatics analysis and biochemical assays are required.

With the development of modern liquid chromatography and mass spectrometry technologies, many mass spectrometry-based substrate profiling methods have been developed. By using mass spectrometry, the cleavage site of each substrate can be easily and precisely determined. Many technologies have been using human proteome derived proteins and peptides as the substrate library, such as Proteomic Identification of protease Cleavage Sites (PICS)[49], Terminal amine isotopic labeling of substrates (TAILS)[50], subtiligase N-terminal labeling[51, 52], Protein Topography and Migration Analysis Platform (PROTOMAP)[53] and Fast Profiling of Protease Specificity (FPPS)[54, 55]. These libraries were constructed based on human proteome which best mimic protease endogenous substrates. The results of these approaches are, therefore, more biological relevant. Though these methods do not require affinity tags or fluorophores as signal readout, enrichment steps were often needed to separate cleaved peptides from the library substrates. The enrichment is achieved with modified substrates' termini and termini-modified substrate libraries are not suitable for studying exopeptidases.

D. Multiplex Substrate Profiling by Mass Spectrometry

Different from proteome-derived substrate library, our group previously developed a substrate profiling method, which is termed Multiplex Substrate Profiling by Mass Spectrometry (MSP-MS), using a rationally designed 228 synthetic tetradecapeptide library[56]. These peptides were synthesized to include all possible neighbor and near-neighbor natural amino acid pairs. Additionally, substrates have unmodified amino and carboxyl termini, allowing us to characterize not only endopeptidase but also amino- and carboxy-peptidase substrate specificity. The amino acid sequences of the synthetic substrates are distinct from endogenous proteins and peptides; therefore, when complex biological samples are assayed with this library, cleaved products are easily discriminated from endogenous peptides. MSP-MS assay is performed by simply incubating the substrate library with the protease of interest, followed by LC-MS/MS to determine the cleaved peptides. The cleaved peptides' sequences are then matched back to the sequences library peptides to determine the cleavage sites. Label free quantification is also performed to quantify both library and cleaved peptides, allowing us to calculate and rank the efficiency of each cleavage. This assay has been used to uncover the substrate specificity profiles of every protease family, and the data generated from these assays have guided the development of fluorescent substrates, peptide-mimetic inhibitors, activity-based probes, therapeutic peptides and activity-based biomarkers[57-62]. We have further coupled MSP-MS assay with isobaric tandem mass tags (TMTs) to yield Quantitative Multiplex Substrate Profiling by Mass Spectrometry (qMSP-MS)[63]. TMT labeling minimizes experimental and instrumental variation, greatly improving the reproducibility and accuracy, while also allows for a reduction in instrument time by 10-fold. Progression curves of each cleavage events are generated, and proteolytic kinetics can be calculated.

My thesis mainly focuses on development and application of MSP-MS, along with proteomics and peptidomics, for characterizing proteolytic activities and designing protease fluorescent substrates and inhibitors. Chapter II demonstrates the application of the MSP-MS assay in uncovering the substrate specificity of a cysteine protease that secreted by commensal bacterial, designing protease inhibitors and fluorescent substrates and predicting protease functions[64]. Chapter III presents the study of characterizing proteolytic processing of proneuropeptides in a complex biological system using proteomics, peptidomics and MSP-MS. Chapter IV shows the development of above-mentioned qMSP-MS technology and its application in profiling proteolytic substrate specificities of two integral membrane peptidases and lung cancer cell secretions[65].

E. References

1. Rawlings, N. D.; Salvesen, G., Handbook of proteolytic enzymes. 3rd ed.; Academic Press,: London ; Boston, 2013; pp. 1 online resource (3 volumes (liv, 3,932 pages)). v.1

2. Walsh, C. T.; Garneau-Tsodikova, S.; Gatto, G. J., Jr., Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl* **2005,** *44* (45), 7342-72.

3. Ehrmann, M.; Clausen, T., Proteolysis as a regulatory mechanism. *Annu Rev Genet* **2004,** *38*, 709-24.

4. Furie, B.; Furie, B. C., The molecular basis of blood coagulation. *Cell* **1988,** *53* (4), 505-18.

5. Coughlin, S. R., Thrombin signalling and protease-activated receptors. *Nature* **2000,** *407* (6801), 258-64.

6. Antalis, T. M.; Shea-Donohue, T.; Vogel, S. N.; Sears, C.; Fasano, A., Mechanisms of disease: protease functions in intestinal mucosal pathobiology. *Nat Clin Pract Gastroenterol Hepatol* **2007,** *4* (7), 393-402.

7. Chang, H. Y.; Yang, X., Proteases for cell suicide: functions and regulation of caspases. *Microbiol Mol Biol Rev* **2000,** *64* (4), 821-46.

8. Sukharev, S. A.; Pleshakova, O. V.; Sadovnikov, V. B., Role of proteases in activation of apoptosis. *Cell Death Differ* **1997,** *4* (6), 457-62.

9. Bai, G.; Pfaff, S. L., Protease regulation: the Yin and Yang of neural development and disease. *Neuron* **2011,** *72* (1), 9-21.

10. Scher, W., The role of extracellular proteases in cell proliferation and differentiation. *Lab Invest* **1987,** *57* (6), 607-33.

11. Kammerl, I. E.; Meiners, S., Proteasome function shapes innate and adaptive immune responses. *Am J Physiol Lung Cell Mol Physiol* **2016,** *311* (2), L328-36.

12. Bond, J. S., Proteases: History, discovery, and roles in health and disease. *J Biol Chem* **2019,** *294* (5), 1643-1651.

13. Glickman, M. H.; Ciechanover, A., The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol Rev* **2002,** *82* (2), 373-428.

14. Turk, B., Targeting proteases: successes, failures and future prospects. *Nat Rev Drug Discov* **2006,** *5* (9), 785-99.

15.   Lopez-Otin, C.; Bond, J. S., Proteases: multifunctional enzymes in life and disease. *J Biol Chem* **2008,** *283* (45), 30433-7.

16.   Lopez-Otin, C.; Matrisian, L. M., Emerging roles of proteases in tumour suppression. *Nat Rev Cancer* **2007,** *7* (10), 800-8.

17.   Jager, R.; Zwacka, R. M., The enigmatic roles of caspases in tumor development. *Cancers (Basel)* **2010,** *2* (4), 1952-79.

18.   Turk, V.; Kos, J.; Turk, B., Cysteine cathepsins (proteases)--on the main stage of cancer? *Cancer Cell* **2004,** *5* (5), 409-10.

19.   Olson, O. C.; Joyce, J. A., Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. *Nat Rev Cancer* **2015,** *15* (12), 712-29.

20.   Aggarwal, N.; Sloane, B. F., Cathepsin B: multiple roles in cancer. *Proteomics Clin Appl* **2014,** *8* (5-6), 427-37.

21.   Xia, L.; Kilb, J.; Wex, H.; Li, Z.; Lipyansky, A.; Breuil, V.; Stein, L.; Palmer, J. T.; Dempster, D. W.; Bromme, D., Localization of rat cathepsin K in osteoclasts and resorption pits: inhibition of bone resorption and cathepsin K-activity by peptidyl vinyl sulfones. *Biol Chem* **1999,** *380* (6), 679-87.

22.   Das, A.; Monteiro, M.; Barai, A.; Kumar, S.; Sen, S., MMP proteolytic activity regulates cancer invasiveness by modulating integrins. *Sci Rep* **2017,** *7* (1), 14219.

23.   Vartak, D. G.; Gemeinhart, R. A., Matrix metalloproteases: underutilized targets for drug delivery. *J Drug Target* **2007,** *15* (1), 1-20.

24.   Kessenbrock, K.; Plaks, V.; Werb, Z., Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell* **2010,** *141* (1), 52-67.

25.   Yu, Q.; Stamenkovic, I., Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-beta and promotes tumor invasion and angiogenesis. *Genes Dev* **2000,** *14* (2), 163-76.

26.   Slack, M. A.; Gordon, S. M., Protease Activity in Vascular Disease. *Arterioscler Thromb Vasc Biol* **2019,** *39* (10), e210-e218.

27.   Brik, A.; Wong, C. H., HIV-1 protease: mechanism and drug discovery. *Org Biomol Chem* **2003,** *1* (1), 5-14.

28.   Lv, Z.; Chu, Y.; Wang, Y., HIV protease inhibitors: a review of molecular selectivity and toxicity. *HIV AIDS (Auckl)* **2015,** *7*, 95-104.

29.   Jin, Z.; Zhao, Y.; Sun, Y.; Zhang, B.; Wang, H.; Wu, Y.; Zhu, Y.; Zhu, C.; Hu, T.; Du, X.; Duan, Y.; Yu, J.; Yang, X.; Yang, X.; Yang, K.; Liu, X.; Guddat, L. W.; Xiao, G.; Zhang, L.; Yang, H.; Rao, Z., Structural basis for the inhibition of SARS-CoV-2 main protease by antineoplastic drug carmofur. *Nat Struct Mol Biol* **2020,** *27* (6), 529-532.

30.   Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R., Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. *Science* **2020,** *368* (6489), 409-412.

31.   Bibo-Verdugo, B.; Jiang, Z.; Caffrey, C. R.; O'Donoghue, A. J., Targeting proteasomes in infectious organisms to combat disease. *FEBS J* **2017,** *284* (10), 1503-1517.

32.   Dimopoulos, M.; Siegel, D.; White, D. J.; Boccia, R.; Iskander, K. S.; Yang, Z.; Kimball, A. S.; Mezzi, K.; Ludwig, H.; Niesvizky, R., Carfilzomib vs bortezomib in patients with multiple myeloma and renal failure: a subgroup analysis of ENDEAVOR. *Blood* **2019,** *133* (2), 147-155.

33.   Kothny, W.; Lukashevich, V.; Foley, J. E.; Rendell, M. S.; Schweizer, A., Comparison of vildagliptin and sitagliptin in patients with type 2 diabetes and severe renal impairment: a randomised clinical trial. *Diabetologia* **2015,** *58* (9), 2020-6.

34.   Packer, M.; Lee, W. H.; Yushak, M.; Medina, N., Comparison of captopril and enalapril in patients with severe chronic heart failure. *N Engl J Med* **1986,** *315* (14), 847-53.

35.   Drag, M.; Salvesen, G. S., Emerging principles in protease-based drug discovery. *Nat Rev Drug Discov* **2010,** *9* (9), 690-701.

36.   Poreba, M., Protease-activated prodrugs: strategies, challenges, and future directions. *FEBS J* **2020,** *287* (10), 1936-1969.

37.   Choi, K. Y.; Swierczewska, M.; Lee, S.; Chen, X., Protease-activated drug development. *Theranostics* **2012,** *2* (2), 156-78.

38.   van de Donk, N. W.; Dhimolea, E., Brentuximab vedotin. *MAbs* **2012,** *4* (4), 458-65.

39.   Sehn, L. H.; Herrera, A. F.; Flowers, C. R.; Kamdar, M. K.; McMillan, A.; Hertzberg, M.; Assouline, S.; Kim, T. M.; Kim, W. S.; Ozcan, M.; Hirata, J.; Penuel, E.; Paulson, J. N.; Cheng, J.; Ku, G.; Matasar, M. J., Polatuzumab Vedotin in Relapsed or Refractory Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **2020,** *38* (2), 155-165.

40.   Chen, S.; Yim, J. J.; Bogyo, M., Synthetic and biological approaches to map substrate specificities of proteases. *Biol Chem* **2019,** *401* (1), 165-182.

41.   auf dem Keller, U.; Schilling, O., Proteomic techniques and activity-based probes for the system-wide study of proteolysis. *Biochimie* **2010,** *92* (11), 1705-14.

42.   Diamond, S. L., Methods for mapping protease specificity. *Curr Opin Chem Biol* **2007,** *11* (1), 46-51.

43.   Schneider, E. L.; Craik, C. S., Positional scanning synthetic combinatorial libraries for substrate profiling. *Methods Mol Biol* **2009,** *539*, 59-78.

44.   Poreba, M.; Salvesen, G. S.; Drag, M., Synthesis of a HyCoSuL peptide substrate library to dissect protease substrate specificity. *Nat Protoc* **2017,** *12* (10), 2189-2214.

45.   Poreba, M.; Solberg, R.; Rut, W.; Lunde, N. N.; Kasperkiewicz, P.; Snipas, S. J.; Mihelic, M.; Turk, D.; Turk, B.; Salvesen, G. S.; Drag, M., Counter Selection Substrate Library Strategy for Developing Specific Protease Substrates and Probes. *Cell Chem Biol* **2016,** *23* (8), 1023-35.

46.   Poreba, M.; Szalek, A.; Rut, W.; Kasperkiewicz, P.; Rutkowska-Wlodarczyk, I.; Snipas, S. J.; Itoh, Y.; Turk, D.; Turk, B.; Overall, C. M.; Kaczmarek, L.; Salvesen, G. S.; Drag, M., Highly sensitive and adaptable fluorescence-quenched pair discloses the substrate specificity profiles in diverse protease families. *Sci Rep* **2017,** *7*, 43135.

47.   Ratnikov, B.; Cieplak, P.; Smith, J. W., High throughput substrate phage display for protease profiling. *Methods Mol Biol* **2009,** *539*, 93-114.

48.   Boulware, K. T.; Daugherty, P. S., Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proc Natl Acad Sci U S A* **2006,** *103* (20), 7583-8.

49.   Schilling, O.; Huesgen, P. F.; Barre, O.; Auf dem Keller, U.; Overall, C. M., Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. *Nat Protoc* **2011,** *6* (1), 111-20.

50.   Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J. N.; Overall, C. M., Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat Protoc* **2011,** *6* (10), 1578-611.

51.   Mahrus, S.; Trinidad, J. C.; Barkan, D. T.; Sali, A.; Burlingame, A. L.; Wells, J. A., Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* **2008,** *134* (5), 866-76.

52.   Wiita, A. P.; Seaman, J. E.; Wells, J. A., Global analysis of cellular proteolysis by selective enzymatic labeling of protein N-termini. *Methods Enzymol* **2014,** *544*, 327-58.

53.   Fuhrman-Luck, R. A.; Silva, L. M.; Hastie, M. L.; Gorman, J. J.; Clements, J. A., Determining Protease Substrates Within a Complex Protein Background Using the PROtein TOpography and Migration Analysis Platform (PROTOMAP). *Methods Mol Biol* **2017,** *1574*, 145-170.

54.   Vizovisek, M.; Vidmar, R.; Van Quickelberghe, E.; Impens, F.; Andjelkovic, U.; Sobotic, B.; Stoka, V.; Gevaert, K.; Turk, B.; Fonovic, M., Fast profiling of protease specificity reveals similar substrate specificities for cathepsins K, L and S. *Proteomics* **2015,** *15* (14), 2479-90.

55.   Vidmar, R.; Vizovisek, M.; Turk, D.; Turk, B.; Fonovic, M., Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *EMBO J* **2017,** *36* (16), 2455-2465.

56.   O'Donoghue, A. J.; Eroy-Reveles, A. A.; Knudsen, G. M.; Ingram, J.; Zhou, M.; Statnekov, J. B.; Greninger, A. L.; Hostetter, D. R.; Qu, G.; Maltby, D. A.; Anderson, M. O.; Derisi, J. L.; McKerrow, J. H.; Burlingame, A. L.; Craik, C. S., Global identification of peptidase specificity by multiplex substrate profiling. *Nat Methods* **2012,** *9* (11), 1095-100.

57.   O'Donoghue, A. J.; Knudsen, G. M.; Beekman, C.; Perry, J. A.; Johnson, A. D.; DeRisi, J. L.; Craik, C. S.; Bennett, R. J., Destructin-1 is a collagen-degrading endopeptidase secreted by Pseudogymnoascus destructans, the causative agent of white-nose syndrome. *Proc Natl Acad Sci U S A* **2015,** *112* (24), 7478-83.

58.   Winter, M. B.; La Greca, F.; Arastu-Kapur, S.; Caiazza, F.; Cimermancic, P.; Buchholz, T. J.; Anderl, J. L.; Ravalin, M.; Bohn, M. F.; Sali, A.; O'Donoghue, A. J.; Craik, C. S., Immunoproteasome functions explained by divergence in cleavage specificity and regulation. *Elife* **2017,** *6*.

59.   Roncase, E. J.; Moon, C.; Chatterjee, S.; Gonzalez-Paez, G. E.; Craik, C. S.; O'Donoghue, A. J.; Wolan, D. W., Substrate Profiling and High Resolution Co-complex Crystal Structure of a Secreted C11 Protease Conserved across Commensal Bacteria. *ACS Chem Biol* **2017,** *12* (6), 1556-1565.

60.   Li, H.; O'Donoghue, A. J.; van der Linden, W. A.; Xie, S. C.; Yoo, E.; Foe, I. T.; Tilley, L.; Craik, C. S.; da Fonseca, P. C.; Bogyo, M., Structure- and function-based design of Plasmodium-selective proteasome inhibitors. *Nature* **2016,** *530* (7589), 233-6.

61.   Lentz, C. S.; Ordonez, A. A.; Kasperkiewicz, P.; La Greca, F.; O'Donoghue, A. J.; Schulze, C. J.; Powers, J. C.; Craik, C. S.; Drag, M.; Jain, S. K.; Bogyo, M., Design of Selective Substrates and Activity-Based Probes for Hydrolase Important for Pathogenesis 1 (HIP1) from Mycobacterium tuberculosis. *ACS Infect Dis* **2016,** *2* (11), 807-815.

62.   Joshi, S.; Chen, L.; Winter, M. B.; Lin, Y. L.; Yang, Y.; Shapovalova, M.; Smith, P. M.; Liu, C.; Li, F.; LeBeau, A. M., The Rational Design of Therapeutic Peptides for Aminopeptidase N using a Substrate-Based Approach. *Sci Rep* **2017,** *7* (1), 1424.

63.   Lapek, J. D.; Jiang, Z.; Wozniak, J. M.; Arutyunova, E.; Lemieux, M. J.; Gonzalez, D. J.; O'Donoghue, A. J., Quantitative Multiplex Substrate Profiling of Proteases by Mass Spectrometry. **2018**.

64.   Xu, J. H.; Jiang, Z.; Solania, A.; Chatterjee, S.; Suzuki, B.; Lietz, C. B.; Hook, V. Y. H.; O'Donoghue, A. J.; Wolan, D. W., A Commensal Dipeptidyl Aminopeptidase with Specificity for N-Terminal Glycine Degrades Human-Produced Antimicrobial Peptides in Vitro. *ACS Chem Biol* **2018,** *13* (9), 2513-2521.

65.   Lapek, J. D., Jr.; Jiang, Z.; Wozniak, J. M.; Arutyunova, E.; Wang, S. C.; Lemieux, M. J.; Gonzalez, D. J.; O'Donoghue, A. J., Quantitative Multiplex Substrate Profiling of Peptidases by Mass Spectrometry. *Mol Cell Proteomics* **2019**.

Chapter II

A Commensal Dipeptidyl Aminopeptidase with Specificity for N-terminal Glycine Degrades Human-produced Antimicrobial Peptides *In Vitro*

A. Introduction

Efforts to structurally define new protein folds and assign potential function to uncharacterized proteins were spearheaded by high-throughput Protein Structure Initiative Centers over the last 15 years and several of these centers focused on the elucidation of proteins from commensal microbiome-associated bacteria.[1-4] One such structure (PDB ID: 3PW3) determined by the Joint Center for Structural Genomics (JCSG) was of an uncharacterized protein (Uniprot ID: A6LE66) from the gut commensal organism *Parabacteroidetes distasonis*. This bacterial species has been implicated in inflammatory bowel diseases (IBD); however, the role of *P. distasonis* in gut inflammation remains to be elucidated.[5, 6] Primary sequence analysis revealed that this protease, which we have called Pd_dinase, has limited homology to papain-like cysteine proteases. Yet, the apo Pd_dinase structure is strongly conserved in tertiary scaffold features with members of the bleomycin hydrolase (BLH) C1B (PF03051) family (Figure 2.1A).
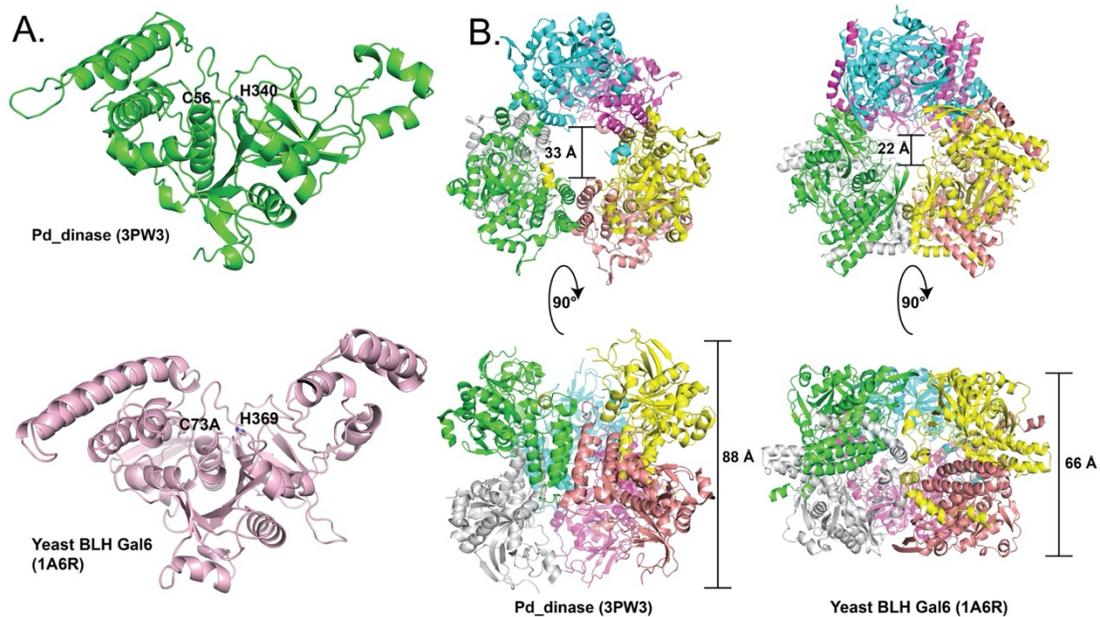
Figure 2.1. Pd_dinase is a member of the C1B cysteine protease family. A. Monomer subunits of apo Pd_dinase (green, PDB ID: 3PW3) and yeast BLH Gal6 (rose, PDB ID: 1A6R) demonstrate high structural conservation. The catalytic cysteine and histidine residues are shown as sticks and labeled accordingly. B. The biologically relevant homohexamer arrangements of Pd_dinase (left) and yeast BLH (right) diverge in similarity despite individual subunit conservation. The central channel of Pd_dinase and yeast BLH Gal6 are 33 Å and 22 Å, respectively, as viewed from above (top). A 90° rotation to a side view (bottom) shows Pd_dinase has a longer central channel of 88 Å in length as compared to 66 Å for yeast BLH.

BLHs are cysteine hydrolases originally identified in humans for their ability to deamidate the anticancer glycopeptide bleomycin.[7-9] Broadly, members of the BLH family are responsible for the hydrolysis of homocysteine thiolactone and a range of other biological processes.[10, 11] For example, murine BLH degrades filaggrin,[12] and yeast BLH (Gal6) is a negative regulator of the galactose regulon[13] and has aminopeptidase activity.[14] In humans, BLH (hBLH) represents a potential drug discovery target to combat Alzheimer's disease,[15] as the protease processes amyloid precursor protein.[16] While the structures of Gal6 and hBLH demonstrate a strong conservation with cysteine proteases of the broader C1 papain-like family, the BLH subfamily is unique as these proteases form homo-hexamers and possibly homo-tetramers as reported for hBLH.[17] In its hexameric state, Gal6 and hBLH form a barrel-like structure akin to the proteasome with the active sites located within a central cavity (Figure 2.1B). Limited accessibility to the cavity active sites are hypothesized to regulate proteolysis and control the spectrum of peptide/protein BLH substrates.[18, 19]

Proteases are essential for a wide variety of physiological reactions that are required for organismal survival. In the distal gut, proteases secreted by commensal bacteria are predicted to be involved in degradation of proteins and peptides for nutrient provision while also playing a role in evasion from host immune system recognition.[20-23] These proteases, along with other enzymatic functions can benefit the host via the breakdown of otherwise indigestible dietary constituents. Notwithstanding, secreted bacterial proteases are aberrantly increased in IBD and therefore, assigning their biological importance in both health and disease is of critical importance.[22-28]

Analysis of the primary sequence of A6LE66 produced by *P. distasonis*, suggested that Pd_dinase is highly conserved throughout the Bacterial Kingdom (Figure 2.2). In addition, the

protease contains a signal sequence that lacks a lipobox motif for post-translational lipid modification for insertion into bacterial membranes, and therefore is likely to be secreted into the distal gut. In this study, we sought to biochemically and structurally characterize the Pd_dinase protease. We confirmed that this protein is enzymatically active and found that it preferentially removes dipeptides from the amino terminus of oligopeptides. An irreversible dipeptide inhibitor was synthesized based on the substrate specificity and was used for producing a co-crystal structure with Pd_dinase. This structure revealed several active site features that clearly distinguished Pd_dinase from other C1B family members and demonstrated why the protease preferentially removes dipeptides from the amino terminus of oligopeptides including the secreted human gut antimicrobial peptides, β-defensins and keratin-derived antimicrobial peptides. This in-depth biochemical and structural analysis of a C1B family protease may explain how commensal organisms colonize the human distal gut in the presence of host-secreted antimicrobials.
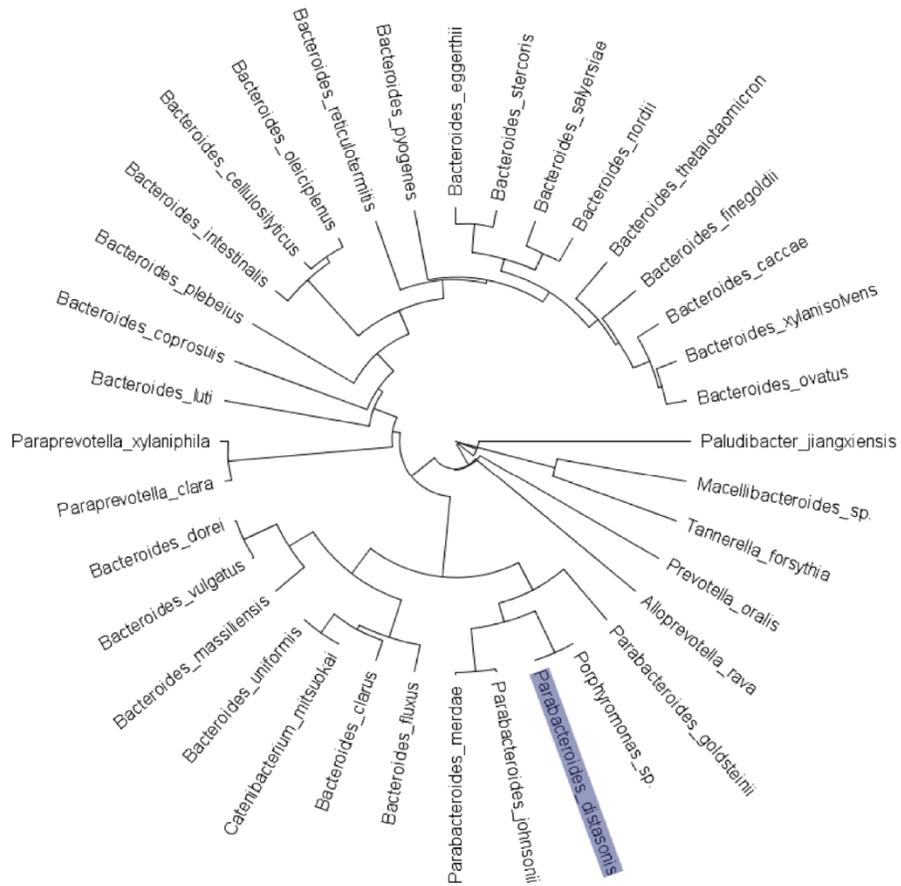
Figure 2.2. Conservation of *P. distasonis* Pd_dinase cysteine protease (highlighted) across a small sampling of bacterial strains that belong to the Bacteroidales order.

B. Materials and Methods

1. Pd_dinase expression and purification

The Pd_dinase WT clone from *Parabacteroidetes distasonis* ATCC 8503 (Protein Accession: WP_005854242, Uniprot ID: A6LE66) consisted of residues 24-405 without the N-terminal secretion leader sequence (residues 1-23) and was kindly provided by JCSG. Pd_dinase is over-expressed as a N-terminal His$_6$-tag fusion with a TEV cleavage site (additional amino acids for the affinity tag include MGSDKI-H6-ENLYFQG-) from E. coli BL21DE3pLysS (Strategene) in a pSpeedET vector. Cells were grown in 2xYT media supplemented with 50 µg/ml kanamycin at 37 °C to an OD600-nm of 0.6-0.8. Flasks were then transferred to 22 °C and protein expression was induced with 0.2% L-arabinose for 16 h. Cells were immediately harvested and resuspended in ice cold 100 mM Tris, pH 8.0 and 100 mM NaCl (buffer A) and subjected to 3 cycles of lysis by microfluidization (Microfluidics). The cell lysate was clarified by centrifugation at 14,000 x g for 8 min at 4 °C and soluble fractions were loaded onto a 5 mL HisTrap FF crude Ni-NTA affinity column (GE Amersham) pre-equilibrated with buffer A and eluted with buffer A containing 250 mM imidazole. The eluted protein was immediately diluted 5-fold with buffer B (20 mM Tris, pH 8.0) and purified by anion-exchange chromatography (HiTrap Q HP, GE Amersham) with a 20-column volume gradient to 50% of buffer B containing 1 M NaCl. Fractions corresponding to Pd_dinase were pooled and immediately stored at -80 °C.

2. Peptide digestion assays

2 nM of Pd_dinase was incubated in triplicate with a mixture of 228 synthetic tetradecapeptides (0.5 µM each) in D-PBS with 2 mM DTT, and 10% of the reaction mixture was removed after 5, 15, 60, 240 and 1200 min of incubation. The activity was quenched with

6.4 M guanidine hydrochloride (GuHCl). A control reaction consisted of Pd_dinase treated with GuHCl prior to peptide exposure. 10 nM of Pd_dinase was incubated in triplicate with 5 μM of each antimicrobial peptide for 5, 15, 60, 240 and 1200 min and 10% of the reaction volume quenched with 6.4 M urea. A control reaction consisted of Pd_dinase treated with urea prior to peptide exposure. hBD2 and HNP3 were incubated with 5 mM DTT for 45 min at 55 °C. Afterwards, free sulfhydryl groups were carbamidomethylated using 15 mM iodoacetamide for 30 min at room temperature in the dark. hBD2 and HNP3 were further digested with proteomic grade trypsin and GluC (Promega), respectively, for 20 h in 1:50 enzyme: protein ratio at 37 °C to generate detectable N-terminal fragments. All samples were stored at -80 °C prior to desalting and LC-MS/MS analysis.

3. Desalting, LC-MS/MS and data analysis

All peptide digestion samples were thawed, acidified to pH < 3.0 with 1% formic acid and desalted using a C18 LTS tips (Rainin). MSP-MS peptides were dissolved in 0.1% formic acid and 0.22 ng were separated by reverse-phase chromatography on an Ultimate 3000 HPLC equipped with a C18 column (1.7 um bead size, 75 um x 20 cm), heated to 65 °C. Separation occurred at a flow rate of 400 nl min^{-1} using a 55-min linear gradient from 5% B to 30% B, with solvent A: 0.1% formic acid in water and solvent B: 0.1% formic acid in acetonitrile. Survey scans were recorded over a 150–2000 m/z range at 70000 resolution at 200 m/z (AGC target 1×10^6, maximum IT 75 ms). MS/MS was performed in data-dependent acquisition mode with HCD fragmentation (NCE 30) on the 10 most intense precursor ions at 17500 resolution at 200 m/z (AGC target 5×10^4, maximum IT 120 ms, dynamic exclusion 15 s).

For antimicrobial peptide, 500 fmol of each peptide were separated by reverse-phase chromatography on a C18 column (1.7 um bead size, 75 um x 25 cm) heated to 60 °C. Separation occurred at a flow rate of 300 nl min-1 using a 30-min linear gradient from 5% B to 40% B. The Q Exactive was operated in the PRM mode (positive polarity, r = 17,500 at 200 m/z, AGC target 2e^5, maximum IT 120 ms, MSX count 1, isolation window 2.0 m/z, NCE 30), with acquisition of targeted peptides.

Peak integration and data analysis were performed using Peaks software (Bioinformatics Solutions Inc.). Label-free quantification results are normalized by LOWESS and filtered by 0.3 peptide significance. Imputation was performed when 2 or 3 replicates reported missing or as zero. Missing and zero values are imputed with normally distributed random numbers that fit smallest 5% of the data. IceLogo software was used for visualization of amino-acid frequency surrounding the cleavage sites using the sequences of all diamino-cleavages whose abundance changed greater than 5-fold in 5 min and passed statistical test (ANOVA q < 0.05). Skyline software[29] was used for quantification of target peptides (Table 2.1). All mass spectrometry data can be accessed here: ftp://massive.ucsd.edu/MSV000082228.

Table 2.1. List of peptides used in PRM analysis. Fragment ions marked in red were ones used for peptide quantification. All cysteines in hBD2 and HNP3 were modified with iodoacetamide (IAA, +57.02146).

Sample	Sequence	Precursor m/z	Ions for PRM
hBD2	GIGDPVTC[IAA]LK	530.278792 (+2)	y8, y6, y5, y4
hBD2	GDPVTC[IAA]LK	445.226028 (+2)	y6, y5, y4
HNP3	DC[IAA]YC[IAA]RIPAC[IAA]IAGE	792.833811 (+2)	b5, b6, b8, b9
HNP3	YC[IAA]RIPAC[IAA]IAGE	655.305015 (+2)	y7, b5, b6
KAMP-14	GGLSSVGGGSSTIK	603.819866 (+2)	y11, y10, y9, y8, y7
KAMP-14	LSSVGGGSSTIK	546.798403 (+2)	y10, y8, y7
KAMP-14	SVGGGSSTIK	446.740356 (+2)	y8, y7, y6
KAMP-13	AIGGGLSSVGGGS	509.76182 (+2)	y5, y4, y2
KAMP-13	GGGLSSVGGGS	834.395186 (+1)	y5, y4, y2
KAMP-13	GLSSVGGGS	720.3530 (+1)	y5, y4, y3, y2

4. Fluorescent protease assays

Substrates containing aminomethylcoumarin (AMC) were purchased from Bachem and all assays were performed in D-PBS, 2 mM DTT. Pd_dinase (8 nM), human aminopeptidase B (2 nM) and bovine trypsin (100 nM) were assayed with 50 µM NH_2-arginine-AMC (Arg-AMC), NH_2-Gly-Arg-AMC (Gly-Arg-AMC), and Glutaryl-Gly-Arg-AMC for 30 min and the change in fluorescence monitored with excitation 360 nm and emission 460 nm using a Synergy HTX microplate reader (BioTek). Kinetic values, K_M and v_{max} were calculated over a range of 50 nM to 100 µM of Gly-Arg-AMC using 2 nM of Pd_dinase. For inhibitor screening, the reaction was initiated by adding 4 nM of Pd_dinase to a mixture of substrate and inhibitor such that the final enzyme, substrate and inhibitor concentrations were, 2 nM, 50 µM and 12.5 µM, respectively. Activity was monitored for 4 h. Inhibition values were calculated using 0.2 nM of enzyme, 300 µM GR-AMC and 0.5 nM to 2.048 µM of GR-AOMK.

5. Synthesis of Fmoc-Arg(Pbf)-AOMK

To a stirred solution of amino acid 1 (6.49 g, 10.0 mmol) in 100 mL dry THF at -10 °C within a 200 mL flame polished round bottom flask, N-methylmorpholine (1.06 g, 10.5 mmol) and isobutyl chloroformate (1.434 g, 10.5 mmol) were added dropwise. After 15 min, ethereal diazomethane was generated and distilled from Diazald® (6.43 g, 30.0 mmol) in accordance to procedures outlined in Aldrich Technical Bulletin AL-180 into stirred solution over the course of 30 min. After distillation, the reaction was allowed to warm to room temperature and continued for an additional 1 h. Glacial acetic acid was added dropwise after being chilled to quench excess diazomethane, and 33% HBr in acetic acid added dropwise until a red tint persisted for more than 5 min. The solvent was removed *in vacuo*, re-dissolved in ethyl acetate and subsequently

washed with water, sat. aq. NaHCO₃ twice, sat. aq. NaCl and dried over MgSO4. A flame dried 20 mL scintillation vial charged with anhydrous potassium fluoride (5 g, 100 mmol) and 2,6-dimethylbenzoic acid (7.509 g. 50 mmol) in 10 mL anhydrous DMF was sonicated for 5 min. **2**, dissolved in a minimal amount of anhydrous DMF, was added dropwise to stirred solution of carboxylic acid and base. After 30 min the solution was diluted with 250 mL EA, washed with 200 mL sat. aq. NaCl 2x, quickly with 1 M NaOH, sat. aq. NaHCO₃, sat. aq. NaCl and dried over MgSO₄. The crude oil was purified by flash chromatography using 1:2 hexanes: ethyl acetate to yield an off white solid, **3,** in 40 % yield.
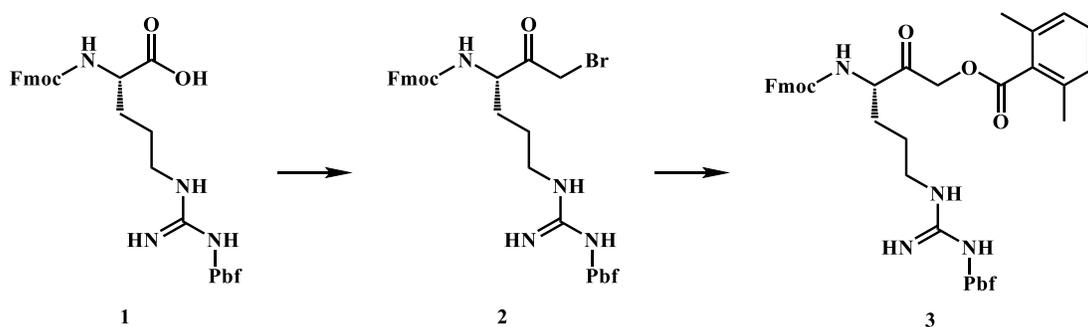
Figure 2.3. Synthesis of Fmoc-Arg(Pbf)-AOMK.

6. Semicarbazide aminomethyl polystyrene resin

A flame-dried 500 mL round bottom flask charged with a magnetic stir bar, aminomethyl polystyrene resin (25 g, 28.75 mmol), N, N'-carbonyldiimidazole (46.62 g, 287.5 mmol) in 250 mL anhydrous DCM was stirred under positive argon pressure for 3 h. The resin was washed once with anhydrous DCM, once with anhydrous DMF, transferred into a new flame dried vessel and resuspended in 250 mL of anhydrous DMF. To this stirred solution, anhydrous hydrazine (55.29 g, 54.15 mL, 1725 mmol) was added gradually over 5 min. The reaction was stirred at

room temperature for 1 h. The resin was filtered, washed with DCM 5x, MeOH 5x, dried thoroughly *in vacuo,* and stored at 4 °C.

A flame-dried 100 mL round bottom charged with amino acid **3** (1.59 g, 2.0 mmol) and **5** (1.00 g, 1.15 mmol/g) was dried *in vacuo* for 6 h and suspended in 20 mL anhydrous THF. This stirred solution was heated at 70 °C for 18 h. The excess amino acid derivative was recovered, and the resin washed with DMF 2x, DCM 2x, MeOH 2x, dried thoroughly and stored at -20 °C.
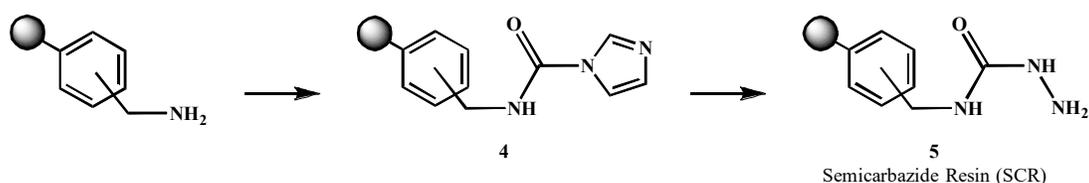
Figure 2.4. Synthesis of Semicarbazide aminomethyl polystyrene resin.

7. Crystallization and x-ray data collection

Inhibitor GR-AOMK was added in a 2-fold molar excess to Pd_dinase (10.5 mg/mL), incubated for 1.5 h at 25 °C, and immediately used for co-crystallization experiments as described in the Supplemental Information. The protein:inhibitor solution was diluted with 20 mM Tris, pH 8.0 and 150 mM NaCl to 6 mg/mL and crystals were grown by sitting drop-vapor diffusion by mixing equal volumes (1.5 μl) of the Pd_dinase:GR-AOMK complex and reservoir solution consisting of 0.29 M K/Na Tartrate, 22% PEG3350, and 1 mM ZnCl at 25 °C. Data was collected on single, flash-cooled crystals at 100 K in a cryoprotectant consisting of mother liquor without 1 mM ZnCl and 25% glycerol and were processed with HKL2000 in orthorhombic space group $P2_12_12_1$. The calculated Matthews' coefficient ($V_M = 3.14$ Å^3 Da^{-1}) suggested six monomers per asymmetric unit with a solvent content of 60%. X-ray data was collected to 2.63

Å resolution on beamline 12.2 at the Stanford Synchrotron Radiation Lightsource (SSRL) (Menlo Park, CA). Data collection and processing statistics are summarized in Table 2.2.

8. Structure solution and refinement

The Pd_dinase:GR-AOMK structure was determined by molecular replacement (MR) with Phaser [30] using the previously published apo structure (PDB ID: 3PW3) determined by the JCSG as the initial search model. The structure was manually built with WinCoot [31] and iteratively refined using Phenix [32] with cycles of conventional positional refinement with isotropic B-factor refinement. Non-crystallographic symmetry (NCS) constraints were applied. The electron density maps clearly identified that GR-AOMK was covalently attached to Cys56 within the active site. Water molecules were automatically positioned by Phenix using a 2.5σ cutoff in F_o-F_c maps and manually inspected. The final R_{cryst} and R_{free} are 17.3% and 21.2%, respectively (Table 2.1). The model was analyzed and validated with PROCHECK [33], WHATCHECK [34], and Molprobity [35] on the JCSG webserver. Analysis of backbone dihedral angles with the program PROCHECK indicated that all residues are located in the most favorable and additionally allowed regions in the Ramachandran plot. Coordinates and structure factors have been deposited in the PDB with accession entry 5WDL. Structure refinement statistics are shown Table 2.2.

Table 2.2. Data collection and refinement statistics. $^a R_{meas} = \{\Sigma_{hkl}[N/(N-1)]1/2\Sigma_i|I_{i(hkl)} - <I_{(hkl)}>|\}$ $/\Sigma_{hkl}\Sigma_i$ $I_{i(hkl)}$, where $I_{i(hkl)}$ are the observed intensities, $<I_{(hkl)}>$ are the average intensities and N is the multiplicity of reflection hkl. $^b R_{merge} = \Sigma_{hkl}\Sigma_i|I_{i(hkl)} - <I_{(hkl)}>|/ \Sigma_{hkl}\Sigma_i I_{i(hkl)}$ where $I_{i(hkl)}$ is the ith measurement of reflection h and $< I_{(hkl)}>$ is the average measurement value. $^c R_{p.i.m.}$ (precision-indicating R_{merge}) $= \Sigma_{hkl}[1/(N_{hkl} - 1)]^{1/2}\Sigma_i|I_{i(hkl)} - <I_{(hkl)}>|/\Sigma_{hkl}\Sigma_i I_{i(hkl)}$. ^{d}Reflections with I > 0 were used for refinement[36-38]. $^e R_{cryst} = \Sigma_h||F_{obs}|-|F_{calc}||/\Sigma|F_{obs}|$, where F_{obs} and F_{calc} are the calculated and observed structure factor amplitudes, respectively. R_{free} is R_{cryst} with 5.0% test set structure factors. ^{f}Cross-validated (CV) Luzzati coordinate errors.

Structure	5WDL
Space group	P $2_1 2_1 2_1$
Cell dimensions	
a, b, c; Å	104.8, 140.3, 221.1
α, β, γ; º	90, 90, 90
Data Processing	
Resolution, Å (outer shell)	50.0-2.65 (2.70-2.65)
Completeness, %	99.6 (98.5)
Unique reflections	96,577 (4,696)
Redundancy	5.9 (5.1)
R_{meas} (%)^a	17.6 (118.4)
R_{merge} (%)^b	15.0 (172.1)
$R_{p.i.m.}$ (%)^c	7.1 (49.8)
Average I / Average σ (I)	12.7 (1.7)
$CC_{1/2}$	91.2 (56.4)
Refinement	
Resolution, Å (outer shell)	50.0-2.65 (2.66-2.65)
No. reflections (test set)^d	96,495 (4,868)
R_{cryst} (%)^e	17.3 (21.2)
R_{free} (%)	21.1 (31.0)
Protein atoms / Waters	17,899 / 794
CV coordinate error (Å)^f	0.28
Rmsd bonds (Å) / angles º	0.002 / 0.48
B-values protein/waters/ligands (Å^2)	39 / 34 / 40
Ramachandran Statistics (%)	
Preferred	98.9
Allowed	1.1
Outliers	0

9. Zinc Inhibition Assay

IC_{50} was calculated over a range of 8 nM to 500 µM of $ZnCl_2$ using 18 nM of Pd_dinase and 50 µM Gly-Arg-AMC. The assay was performed in DPBS, 2 mM DTT. The change in fluorescence was monitored using EnVision 2104 plate reader with excitation 355 nm and emission 486 nm.

10. Synthesis of keratin-derived antimicrobial peptides

KAMP-14 (NH_2-GGLSSVGGGSSTIK-COOH) and KAMP-13 (NH_2-AIGGGLSSVGGGS-COOH) were synthesized using standard FMOC-protected solid-phase peptide synthesis. The peptides were purified by preparatory reverse phase HPLC (19x150 mm XBridge C18, CH_3CN/H_2O/0.1% TFA, 10:90 to 50:50, over 13 min; 20 mL/min) and lyophilization. hBD2 was purchased from Peptides International, Inc.

11. Data deposition

The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.wwpdb.org (PDB ID code 5WDL). All mass spectrometry data can be accessed here: ftp://massive.ucsd.edu/MSV000082228.

C. Results and Discussion

1. Pd_dinase is a homo-hexamer

Pd_dinase and members of the BLH family, we speculated that the biologically relevant quaternary structure of Pd_dinase would likely be a multi-oligomer. We subjected our purified Pd_dinase to size exclusion chromatography coupled with multi-angle static light scattering (SEC-MALS) to accurately determine the oligomeric state and determined the mass of the complex to be 254 kD -/+ 1.1%. This mass corresponds to a homo-hexameric complex, as observed for yeast and human BLH (Figure 2.5).

2. *In vitro* Pd_dinase substrate specificity

In order to biochemically characterize Pd_dinase, it was essential to uncover the substrate specificity of this enzyme. Pd_dinase was incubated with an equimolar mixture of 228 tetradecapeptides and the resulting cleavage products were evaluated by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS). This approach, termed multiplex substrate profiling by mass spectrometry (MSP-MS), has previously been used by our group to generate substrate specificity information for cysteine,[39] serine,[40] and threonine[41] proteases. Pd_dinase preferentially removed dipeptides from the amino terminus resulting in the formation of dodecapeptides (Figure 2.6A). Many of these new cleavage products were subsequently degraded to decapeptides and then to octapeptides over time. For example, one such sequence among the 228 peptides, substrate GIQSTYFHDLNPYL, is hydrolyzed to FHDLNPYL by sequential removal of the dipeptides GI, QS and TY from the amino terminus (Figure 2.6B).

Figure 2.5. Size-exclusion chromatography with multi-angle light scattering (SEC-MALS) confirms Pd_dinase is a homohexamer.
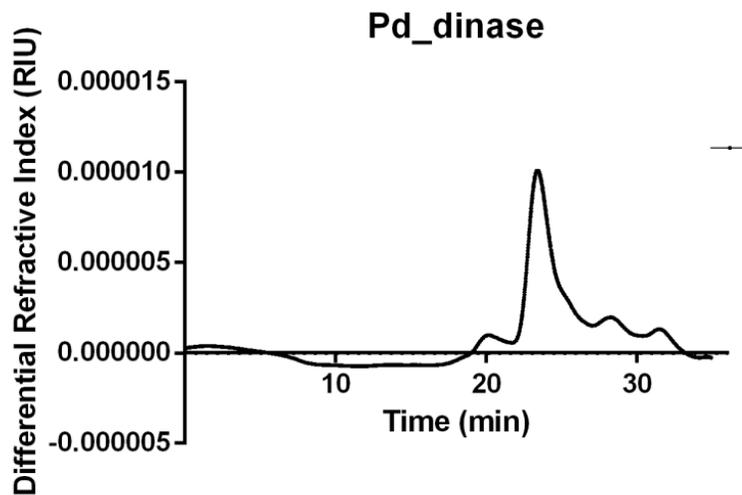
Protein	Calculated Molecular Weight (kDa)	Uncertainty (%)
Pd_dinase	260.9	1.1
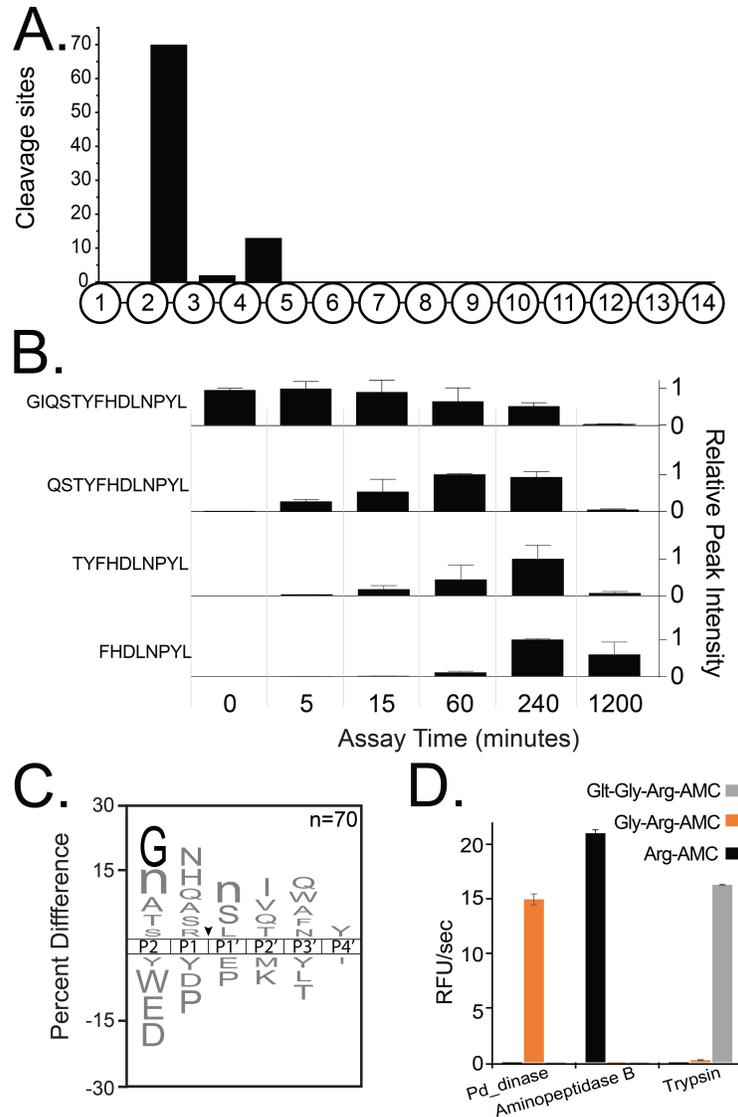
Figure 2.6. Pd_dinase is a sequential di-aminopeptidase. A. Analysis of cleavage site location within the tetradecapeptide substrate library shows that hydrolysis after 5 minutes incubation generally occurs between the 2nd - 3rd, and 4th - 5th amino acids. B. A sample peptide from the library showing time-dependent decrease in substrate with concomitant increase in degradation products. C. An iceLogo plot illustrating the amino acids that are most frequently observed above X-axis and least frequently observed (below X-axis) at the P2 to P4′ positions that surround the cleavage site (black arrow). Glycine at P2 is highlighted in black, as it is the only residue that is significantly enriched (p-value < 0.01), lowercase 'n' corresponds to the non-natural amino acid, norleucine. D. Evaluating Pd_dinase for mono-, di- and tri-aminopeptidase activity. Human aminopeptidase B and trypsin were used as control enzymes for the Arg-AMC and Glt-Gly-Arg-AMC substrates, respectively.

We next generated a substrate specificity profile of the 70 cleavage sites detected between the 2nd and 3rd amino acids of the oligopeptides and determined that Pd_dinase has a strong preference for amino terminal Gly (*p*-value = 0.01), which corresponds to the P2 position (Figure 2.6C). In addition, peptides containing small amino acids on the N-terminal, such as Ala, Ser and Thr, were frequently cleaved; however, peptides containing large hydrophobic or negatively charged amino acids at P2, such as Trp, Arg, Glu and Asp were unfavored (Figure 2.6, Figure 2.7).

Although this enzyme is clearly a di-aminopeptidase, two of the peptides in the library were hydrolyzed between the 3rd - 4th amino acids, indicating that Pd_dinase may have minor mono- or tri-aminopeptidase activity (Figure 2.6A). Therefore, we utilized a selection of fluorescent substrates of increasing length to quantitatively compare the specificity of different aminopeptidases. Each substrate consisted of Arg linked to a cleavable aminomethylcoumarin (AMC) molecule as this amino acid is commonly used to assay C1B hydrolases.[11, 42] We predicted that Pd_dinase would preferentially cleave Gly-Arg-AMC relative to Arg-AMC or the N-terminally capped, glutamyl-Gly-Arg-AMC (Glt-Gly-Arg-AMC). As controls, we assayed the same set of substrates with human aminopeptidase B and bovine trypsin. Both enzymes have a preference for Arg in the P1 position; however, trypsin is an endopeptidase and aminopeptidase B cleaves single amino acids from the N-terminus.[43] Our assays confirmed that Pd_dinase does not cleave Arg-AMC or Glt-Gly-Arg-AMC but rapidly hydrolyzes Gly-Arg-AMC with k_{cat} and K_M values of 1.25 ± 0.02 sec^{-1} and 25.6 ± 1.7 μM (Figure 2.6D, Figure 2.8). In contrast, aminopeptidase B only cleaved Arg-AMC while trypsin cleaved Glt-Gly-Arg-AMC 80-fold more rapidly than Gly-Arg-AMC and failed to hydrolyze Arg-AMC. Taken together, these

studies show that Pd_dinase is a sequential di-aminopeptidase with an unusual preference for N-terminal Gly residues.

In order to investigate the structural features that define the P2-Gly specificity, we synthesized a dipeptide inhibitor consisting of acyloxymethyl ketone (AOMK) reactive warhead. The AOMK warhead strongly reacts with nucleophilic cysteine residues and we showed that Gly-Arg-AOMK irreversibly inactivates Pd_dinase with a K_i of 683 ± 36 nM and k_{inact} of $3.6 \times 10^{-4} \pm 1.0 \times 10^{-5}$ sec^{-1} (Figure 2.9A-C). To confirm that inhibition was not due to the reactivity of the warhead, we synthesized AOMK inhibitors consisting of N-terminally acetylated (Ac) and propynylated (Prop) Arg residue in place of the Gly-Arg dipeptide. While 50 µM of Gly-Arg-AOMK completely inhibited 2 nM of Pd_dinase, the same concentration of Ac-Arg-AOMK and Prop-Arg-AOMK had no effect on activity (Figure 2.9D). These studies clearly show that potency of the Gly-Arg-AOMK was due to selectivity in the S2 pocket for Gly.
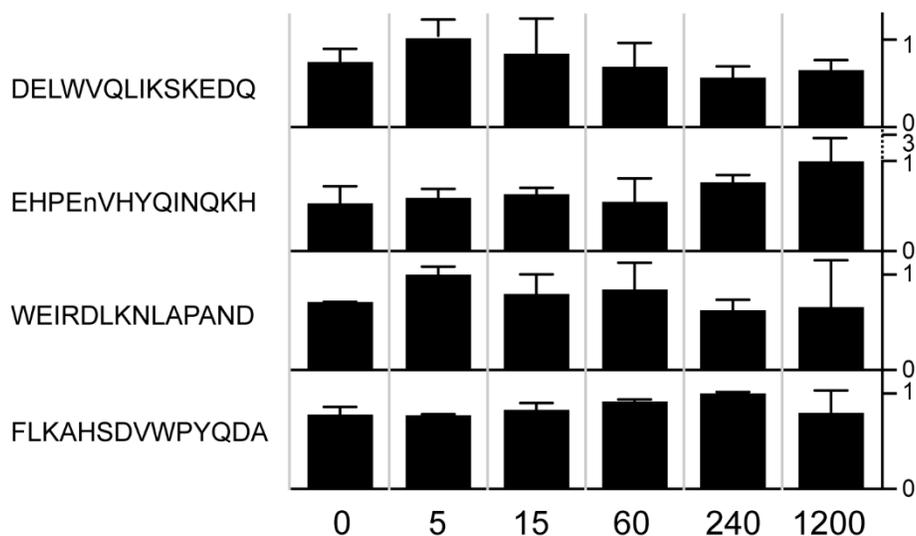
Figure 2.7. Large or negatively charged amino acids at P2 are unfavored by Pd_dinase. There were no significant reductions of library peptides with negatively charged and large hydrophobic amino acids at the N-termini. Cleavage products pertaining to these peptides were also not detected throughout the course of the assay. Relative peak area is calculated from triplicate assays and normalized to the time interval with the highest mean peak area. Data presented as mean ± sd.
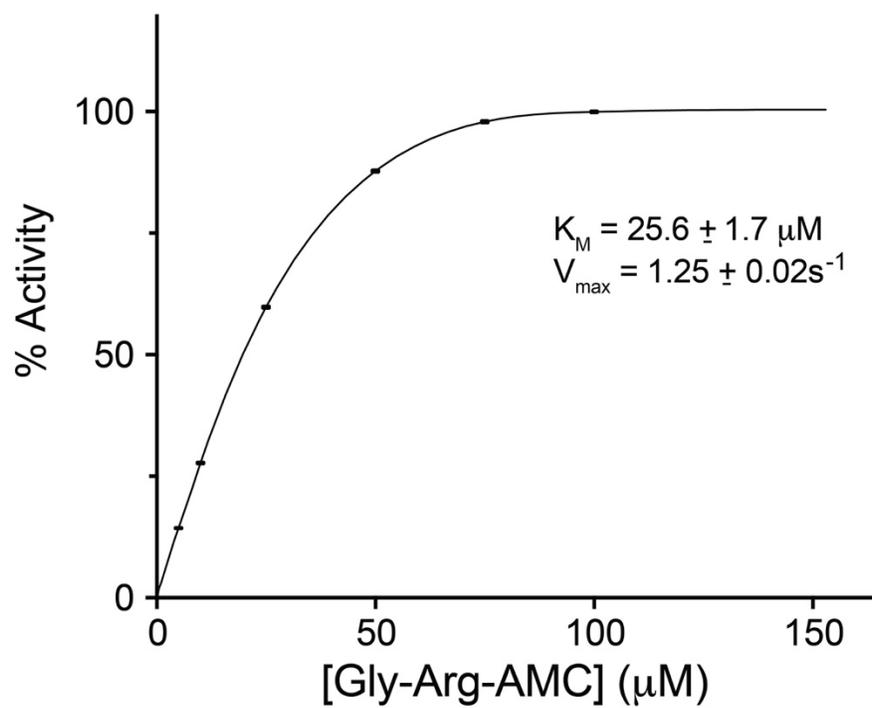
Figure 2.8. Kinetics of GR-AMC. Pd_dinase hydrolyzed Gly-Arg-AMC with $k_{cat} = 1.25 \pm 0.02$ s^{-1} and $K_M = 25.6 \pm 1.7$ μM. Mean $\pm$ SD values are shown (n = 4).
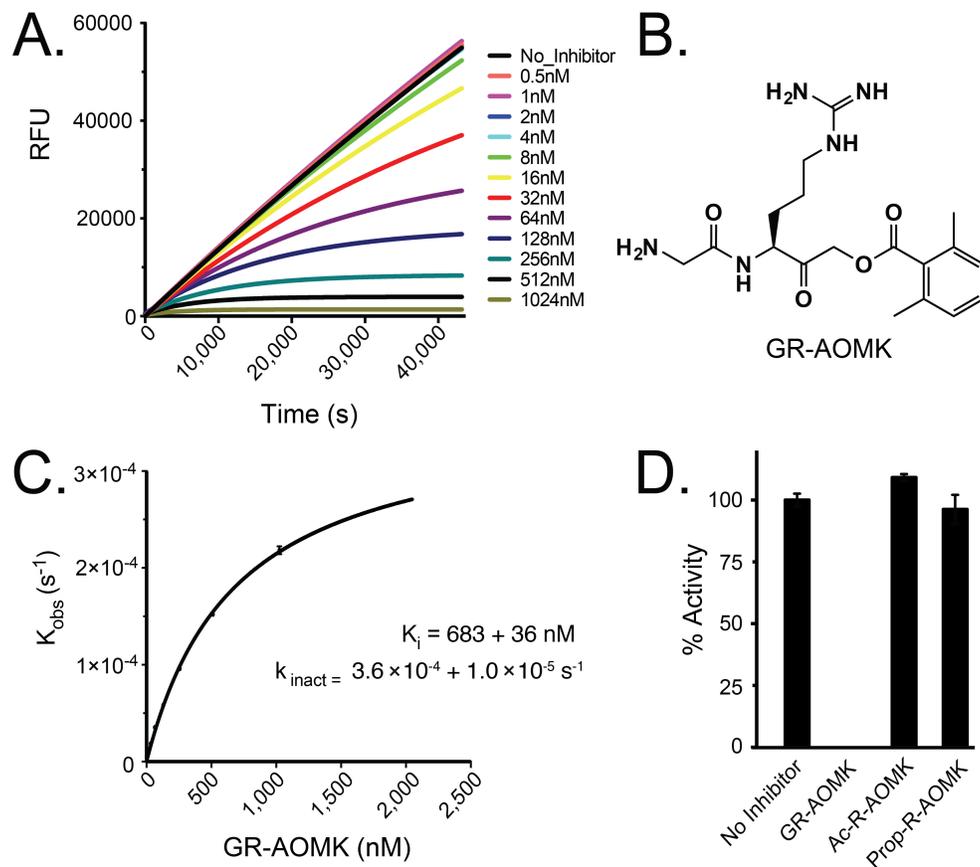
Figure 2.9. Pd_dinase is selectively inhibited by dipeptide inhibitors. A. Incubation of Pd_dinase with 0.5 to 1024 nM of Gly-Arg-AOMK. B. Chemical structure of Gly-Arg-AOMK. C. Calculation of Ki and kinact. D. Comparison of Pd_dinase activity in the presence of 50 μM of different peptide-AOMK inhibitors.

3. Co-crystal structure of Pd_dinase in complex with NH₂-glycine-arginine-AOMK

We determined the x-ray crystal structure of Pd_dinase (residues 24-405) in complex with GR-AOMK to 2.62 Å resolution (PDB ID: 5WDL) (Table 2.2). Each individual Pd_dinase monomer consists of 12 β-strands and 13 α-helices (Figure 2.10). The protease has a core antiparallel β-sheet comprised of 6 β-strands, extending through the center of the protease that is protected by 6 α-helices donated by residues from the N-terminus on one face and 5 α-helices on the opposing face (Figure 2.10). This canonical α/β/α sandwich is conserved among the cysteine peptidases. Interestingly, Pd_dinase has a structural feature unique to known C1 protease structures in that the longest β-strand (β9, residues 340-350) kinks at Gly346 and results in the contribution of 2 separate anti-parallel β-sheets (Figure 2.11). The six monomers form a barrel-like structure with a cavity of approximately 33 Å and 88Å in diameter and length, respectively. The final R_{cryst} and R_{free} values were 0.171 and 0.212, respectively, with 99% of the residues residing in the most favored region of the Ramachandran plot (Table 2.2). Excluding loop 286-310 (discussed below), there are few conformational changes between the GR-AOMK complex and 3PW3 apo structure with an average rmsd = 0.17 Å and a maximum displacement of 1.31 Å. Electron density was not visible for residues 290-310 for subunits A, B, C, and D, as they comprise a loop with direct exposure to a solvent channel. Only residues 305-310 in subunit E lacked density and is likely due to stabilization of the loop region via interactions with a crystal contact. Continuous main-chain density was observed for chain F throughout the extended loop of 290-310, as F was the only subunit that established direct contact with an adjacent subunit within the hexamer (residues 365-370 of B). There was no direct evidence as to why only subunit F adopts this specific 290-310 loop conformation relative to the other 5 subunits.
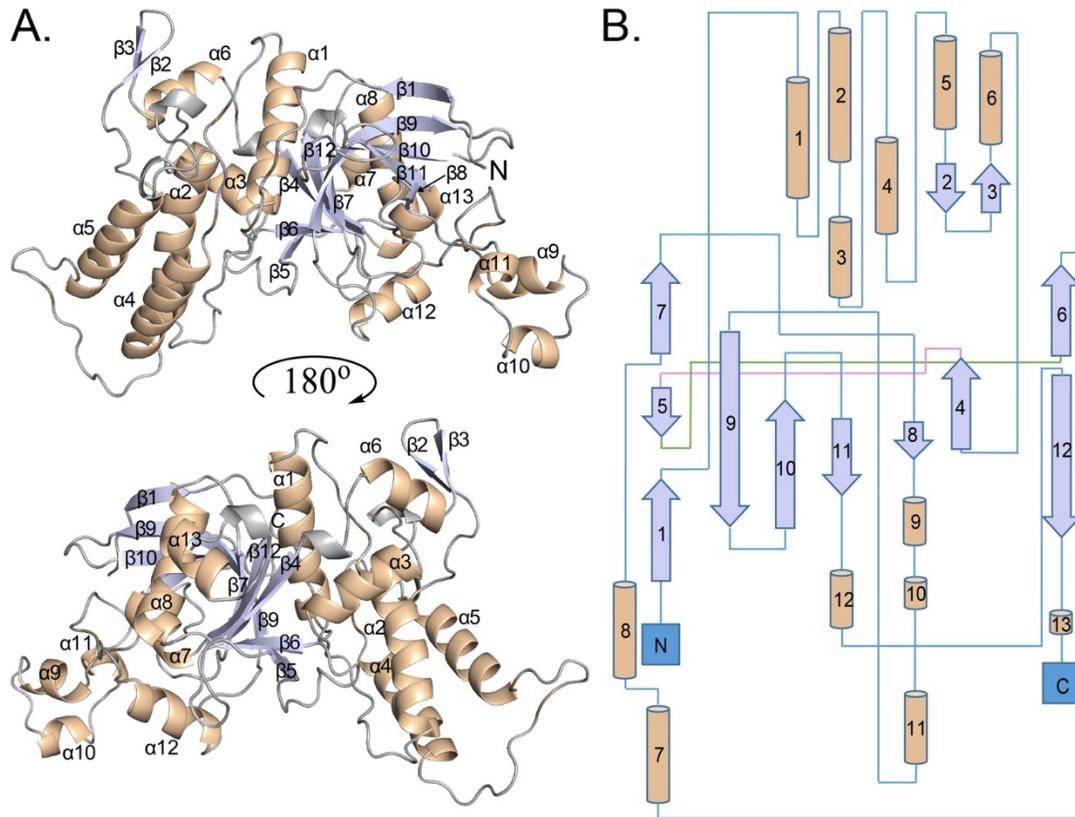
Figure 2.10. Secondary structure of a Pd_dinase monomer. A. The Pd_dinase monomer consists of 12 β-strands (purple) and 13 α-helices (beige). B. A topology diagram of the monomeric subunit depicts the arrangement and connectivity of the α-helices (beige cylinders) and β-strands (purple arrows) with conserved numbers to panel A.
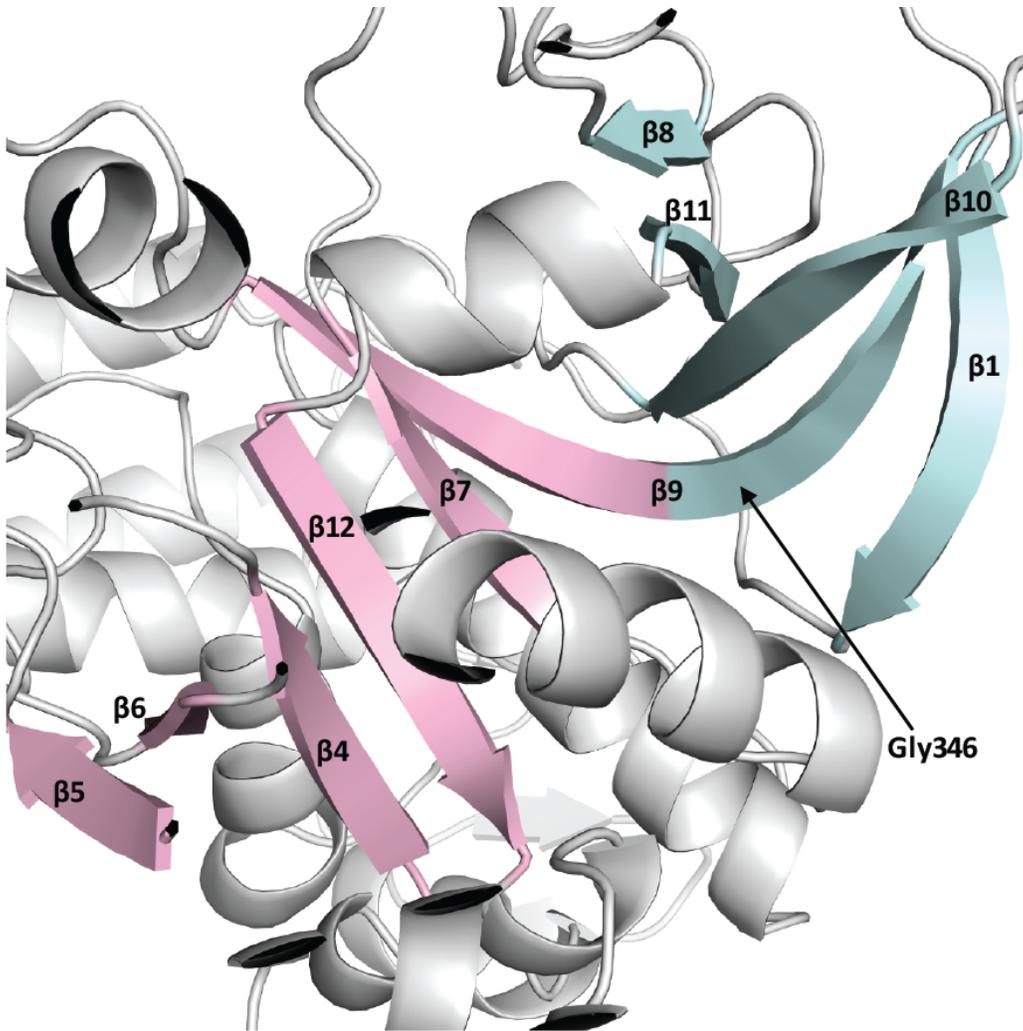
Figure 2.11. Pd_dinase has a core antiparallel β-sheet that is unique to C1 proteases. β-9 forms two separate anti-parallel β-sheets with β-7 and β-10 due to the flexibility of G346.

Four β-sheets (β-7,9,10,11) and three α-helices (α-1,3,7) form the active site and are highly conserved across the C1 papain-like peptidase family (Figure 2.10). Likewise, the Pd_dinase active site consists of the cysteine-histidine dyad common to cysteine proteases with the catalytic Cys56 positioned over the N-terminus of α-1 that promotes the pK_a perturbation of the side-chain thiol (Figure 2.12A). The naïve electron density maps clearly outlined the entire Gly-Arg peptide within the Pd_dinase active site covalently attached to Cys56 (Figure 2.12B). Aside from the covalent bond between Cys56 and the inhibitor, the most significant interactions between Pd_dinase active site residues and the dipeptide are primarily via hydrogen bonds to the P2 Gly residue. The P2 primary amine is highly coordinated with three potential hydrogen bond acceptors, including the carboxylate side chains of Asp237 and Asp338 and the main-chain carbonyl of Gly109 (Figure 2.12C). Presence of aspartic acids in the active site strongly support why substrates containing like charges (*i.e.*, Asp and Glu) at the N-terminus are disfavored (Figure 2.6C). Non-specific peptide backbone interactions make up the rest of the potential hydrogen bonds. The P2 carbonyl interacts with main-chain amide of Gly109 and the amide nitrogen of P1 hydrogen bonds with the main-chain carbonyl of Asp339. The ketone moiety of the AOMK inhibitor is within hydrogen bonding distance to the His340 side chain; however, the side chain is not optimally oriented to accommodate an interaction with the inhibitor ketone in any of the 6 active sites within the structure. The P1 Arg side chain does not directly contact the protein and is held in position via a water-mediated network to the Asp237 side-chain carboxylate and Ser106 side-chain hydroxyl (Figure 2.12C). Our structural studies strongly support the substrate specificity data, including: 1) the P1 side chain does not confer any substrate specificity (Figure 2.12C); 2) the active site can only bind two non-prime side amino acids from substrates; 3) the shallow S2 subsite that interacts with P2 preferentially

40

accommodates short aliphatic (Gly, Ala) or hydrogen bond donating (Ser, Thr) side chains that could provide hydrogen bonds with the main-chain carbonyls of Asp338 and Asp339 (Figure 2.12D); 4) the active site discriminates against substrates with N-terminal Asp and Glu due to electrostatic repulsion and Trp and Tyr because of steric clashes (Figure 2.12D); and 5) significant interactions between the N-terminal amine of the P2 residue substrate and the active site show that biologically relevant substrates likely require a free N-terminal amine with no post-translational modifications, including acetylation or propynylation (Figure 2.12C).

Analysis of the active site surface and corresponding electrostatic potential revealed some surprising structural features. In general, the Pd_dinase active site is highly electronegative and fairly non-descript in the region that indirectly positions the P1 residue (Figure 2.12D). Conversely, the P2 Gly of GR-AOMK is positioned directly over an entrance to a narrow channel that connects the interior active site to the solvent-exposed exterior (Figure 2.12D). The channel measures approximately 10 Å and 5 Å in length and width, respectively, and is also negatively charged. The structure depicted additional structured water molecules within the channel in comparison to the bound structures and may have a role in the hydrolysis mechanism of the enzyme. We posit that this channel strongly discriminates against most amino acids except residues with small and positively or uncharged side chains, such as Gly. Thus, this structural feature of Pd_dinase is the predominant restriction that confines substrate recognition to the small aliphatic P2 residues observed in the MSP-MS results (Figure 2.6). The P2 channel contains several ordered water molecules and likely supplies the solvent necessary for regenerating the free cysteine thiol upon release of the hydrolyzed products (Figure 2.12D).

4. Comparison of the apo and bound Pd_dinase structures

The unbound active site of the apo structure coordinates a Zn^{2+} ion via residues Cys56 and His340 introduced by $ZnCl_2$ from the crystallization buffer (Figure 2.13A). When added in excess, Zn^{2+} inhibits substrate turnover and is likely caused by coordination to the Cys56 and His340 as well as the side-chain carboxylate oxygens of D298 from an adjacent subunit (Figure 2.13A, Figure 2.14). The bound Zn^{2+} locks loop 290-310 into an orientation whereby the tip of the loop (residues 296-303) occludes the neighboring active site pocket. The loop residues 290-310 are missing electron density in the GR-AOMK co-complex structure and suggest that the Zn^{2+} and loop are expelled upon active site binding (Figure 2.13B). One monomer (F) in our GR-AOMK co-complex structure has density for residues 290-310 and superposition with the apo structure shows the loop moves approximately 30 Å upon active site binding and interacts with the same adjacent subunit within the hexamer (residues 365-370 of B) via a hydrogen bond between the main-chain carbonyl of Ala300 and the side-chain amide of Asn366 (Figure 2.13B). Interestingly, the loop 290-310 conformation when upon occupation of the active site by substrate or inhibitor forms a small pore approximately 26 Å in width and 15 Å in height with primarily electronegative surface (Figure 2.13C).
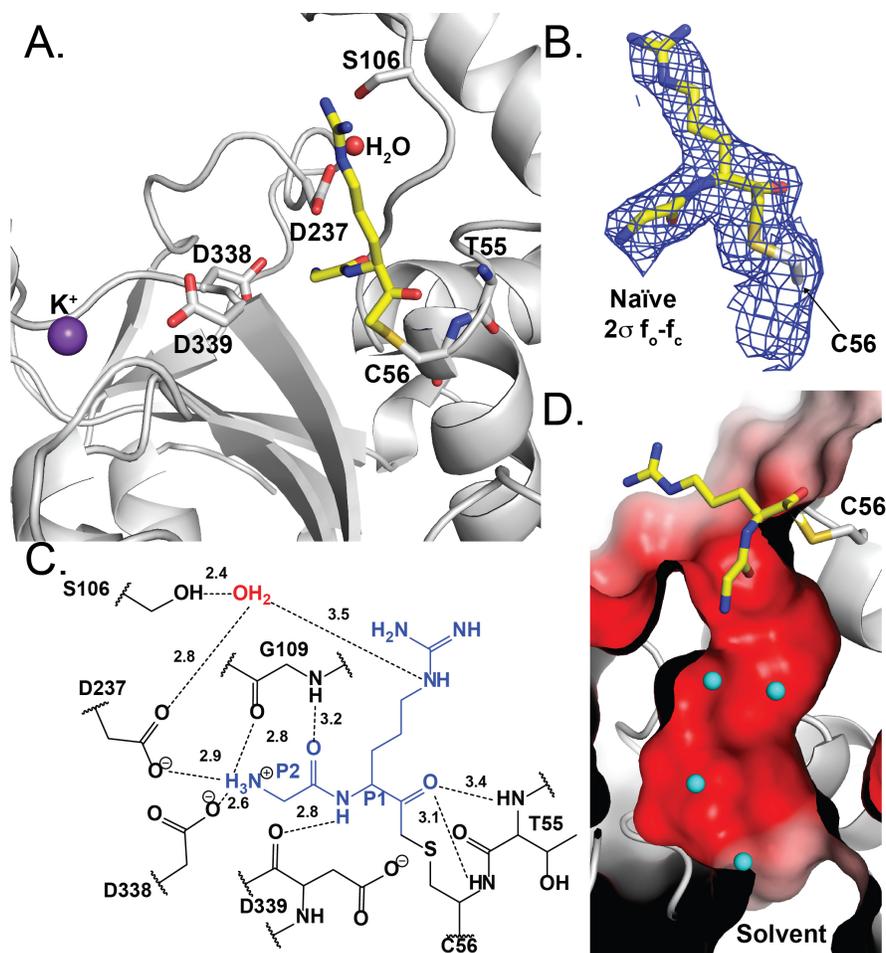
Figure 2.12. Pd_dinase in complex with the dipeptide inhibitor GR-AOMK determined by co-crystallization. A. Cartoon representation of the residues and ordered water (red sphere) within the Pd_dinase:GR-AOMK structure. GR-AOMK (yellow carbon), covalently bound to the active Cys56, and active site residues (grey carbon) are shown as sticks (red oxygen, blue nitrogen, mustard sulfur). B. The naïve f_o-f_c density map (blue) contoured at 2σ clearly indicated the orientation of the GR-AOMK dipeptide and covalent attachment to Cys56. C. The hydrogen bonding network between GR-AOMK (blue), ordered water (red), and Pd_dinase active site residues (black) show that the majority of interactions with the inhibitor are to the N-terminal P2 Gly residue. D. The surface electrostatic potential of the Pd_dinase active site consists of a highly electronegative channel that connects the solvent-exposed surface with the negatively charged active site, as shown with a cross-sectional representation. Ordered water molecules (cyan) are found throughout the channel. The electrostatic potential is colored: blue, positive potential (10 mV); white, neutral potential (0 mV); and red, negative potential (-10 mV).

43
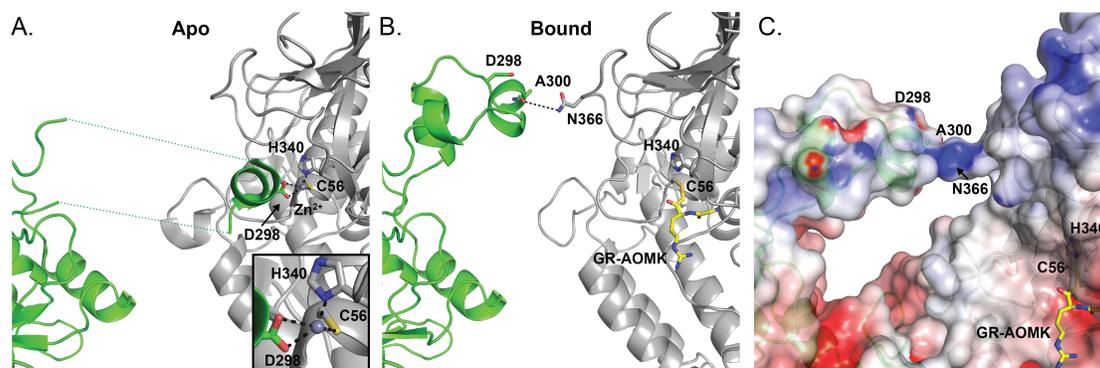
Figure 2.13. Loop 290-310 conformationally occludes neighboring active site with Zn^{2+} ion. Side-by-side comparison of A. the apo Pd_dinase (PDB ID: 3PW3) and B. Pd_dinase:GR-AOMK structures demonstrate the flexibility of loop 290-310 from an active-site inhibited (apo, Zn^{2+} ion bound) conformation (see inset in A) to an active-site open (GR-AOMK-bound, B) with colors as in Figure 2.12A. Potential hydrogen bonds are depicted as black dashes and residues 287-295 and 304-309 missing from the apo structure are represented by green dashed lines. C. The active-site open conformation (GR-AOMK-bound) shows that the Ala300 main-chain carbonyl and the Asn366 side chain from the neighboring subunit hydrogen bond to form a pore. The surface electrostatic potential is as in Figure 2.12D.

ZnCl$_2$ Inhibition in DPBS +/+
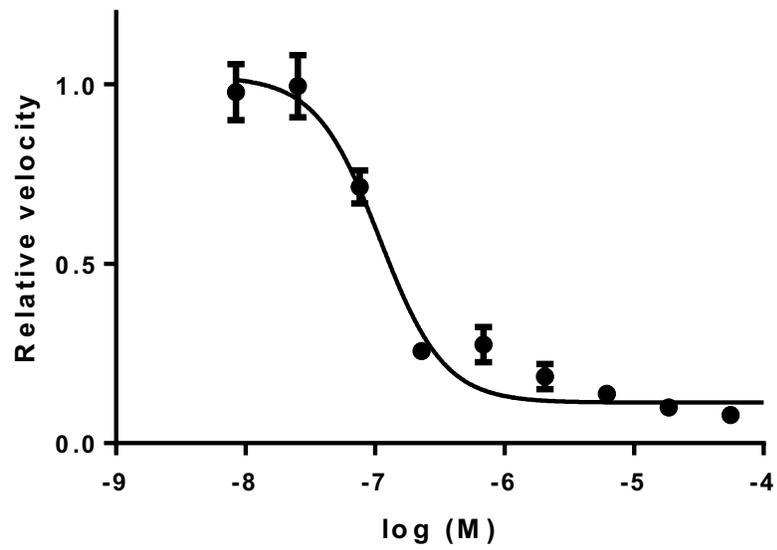
Figure 2.14. Zn^{2+} inhibits Pd_dinase activity with IC$_{50}$ = 109 ± 14 nM in DPBS buffer containing 0.9 mM Ca$^+$ and 0.5 m Mg^{2+}.

This unique conformational rearrangement revealed by the superposition of the apo and co-crystal structures likely has relevance to the biophysical mechanism of the protease. The flexible loop (residues 290-310) coordinated within an adjacent subunit's active site via a Zn^{2+} ion in the apo structure is expelled upon binding substrate or inhibitor and forms an electronegative pore (Figure 2.13C). Such loop rearrangements have been observed in HIV-1 protease[44] and Crk signaling protein[45] that serve as a funnel for substrate binding and as a self-inhibitory mechanism, respectively. The conformational changes observed in these proteins are conferred by either Gly flexibility or Pro isomerization at a specified hinge region. Consistent with the other proteins, the flexible loop of Pd_dinase is flanked by Pro286, Pro316, and Pro318 on either side of the loop. We hypothesize that isomerization of one or more of the proline residues are likely to the shift between the open and closed conformation and these rearrangements could serve as a regulatory mechanism or may influence substrate recognition. Importantly, the much wider central channel formed by the homo-hexamers lacks any specific structural features or electrostatic charges that would help guide substrates to the active sites.

5. Comparisons to C1 papain-like structures

Dali structure similarity searches with Pd_dinase:GR-AOMK identified a list of papain-like cysteine proteases and endopeptidases despite little to no sequence homology. As expected, proteins with similar structural features include the C1B proteases hBLH (1CB5, Z = 32.9, rmsd = 3.1 Å, 321/453 residues),[11] Gal6 (1A6R, Z = 31.9, rmsd = 3.1 Å, 324/459 residues),[14] as well as *Lactobacillus rhamnosus* PepW (4K7C, Z = 34.0, rmsd = 2.9 Å, 322/446 residues). The structural similarities extend to include other members of the broader C1 papain-like proteases, such as *Clostridioides difficile* Cwp84 (4D5A, Z = 18.2, rmsd = 2.9 Å, 207/406 residues),[46] *Dionaea muscipula* dionain-1 (5A24, Z = 17.8, rmsd = 2.8 Å, 197/222 residues),[47] and EP-B2 (2FO5, Z = 17.6, rmsd = 2.8 Å, 194/224 residues)[48] (Figure 2.15). Like most papain-like proteases, Pd_dinase prefers hydrophobic amino acids in the P2 position, but unlike the papain-like proteases, Pd_dinase strongly prefers Gly at the P2 position[49] and has no P1 specificity. In addition to different substrate specificity, Pd_dinase also forms an oligomer, which is unique to the C1B family. The most distinguishing feature that separates Pd_dinase from the rest of the aforementioned proteases is its unique substrate specificity, binding pocket, and channel by the active site.
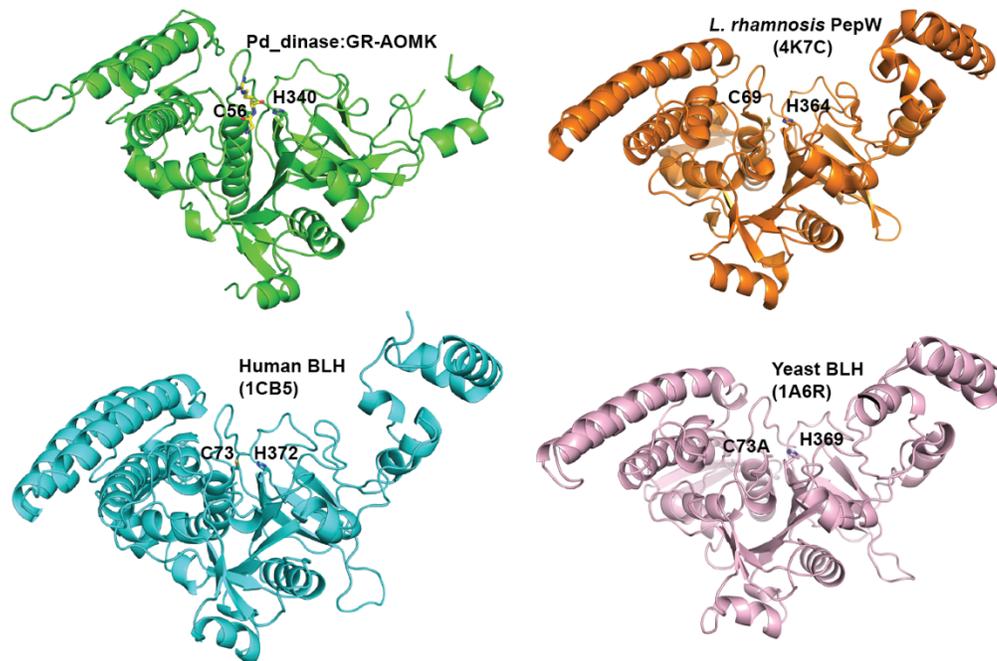
Figure 2.15. DALI search results demonstrate Pd_dinase has structural similarities with the C1B proteases, including *L. rhamnosis* PepW (PDB ID: 4K7C), human and yeast bleomycin hydrolase (PDB ID: 1CB5, 1A6R). All active site His and Cys residues are labelled.

Of the papain superfamily, only members of the C1B BLH-like family are currently known to form biologically relevant oligomers. Despite the high structural conservation between the monomeric units of Pd_dinase and the BLHs, the oligomerization states vary between tetrameric and hexameric forms, as both Pd_dinase and Gal6 are hexameric (Figure 2.1B). Human BLH appears to be tetrameric *in vitro*; however, crystal structures suggest the protein can exist as both a tetramer and hexamer.[17] Despite Gal6 and Pd_dinase existing as hexamers, significant differences are observed in both the shape and diameter of the central channel as well as the electrostatic surface potentials (Figure 2.16A). The central channel of Gal6 is shorter and has a narrower opening with a length and diameter of 66 Å and 22 Å that expands to 45 Å in the protein interior, respectively (Figure 2.16A, 2.1B). This narrow opening has been predicted to confer selectivity for partially unfolded protein substrates that are funneled into the active sites. The central channel of Gal6 is lined with 60 lysine residues and as a result has a strong positive electrostatic potential (Figure 2.16B). This charge distribution suggested that the central channel was the site of nucleic acid binding, as mutation of 3 lysine residues to alanine at the entrance to the channel abolished the nucleic acid binding activity.[11, 13] The surface-exposed residues within the central cavity and the outer region of Pd_dinase are primarily negative (Figure 2.16B), as seen with the central channel of hexameric hBLH, which correlates with hBLH inability to bind nucleic acid. Conversely, Gal6 is more positive, which is consistent with its ability to bind single stranded DNA and RNA.
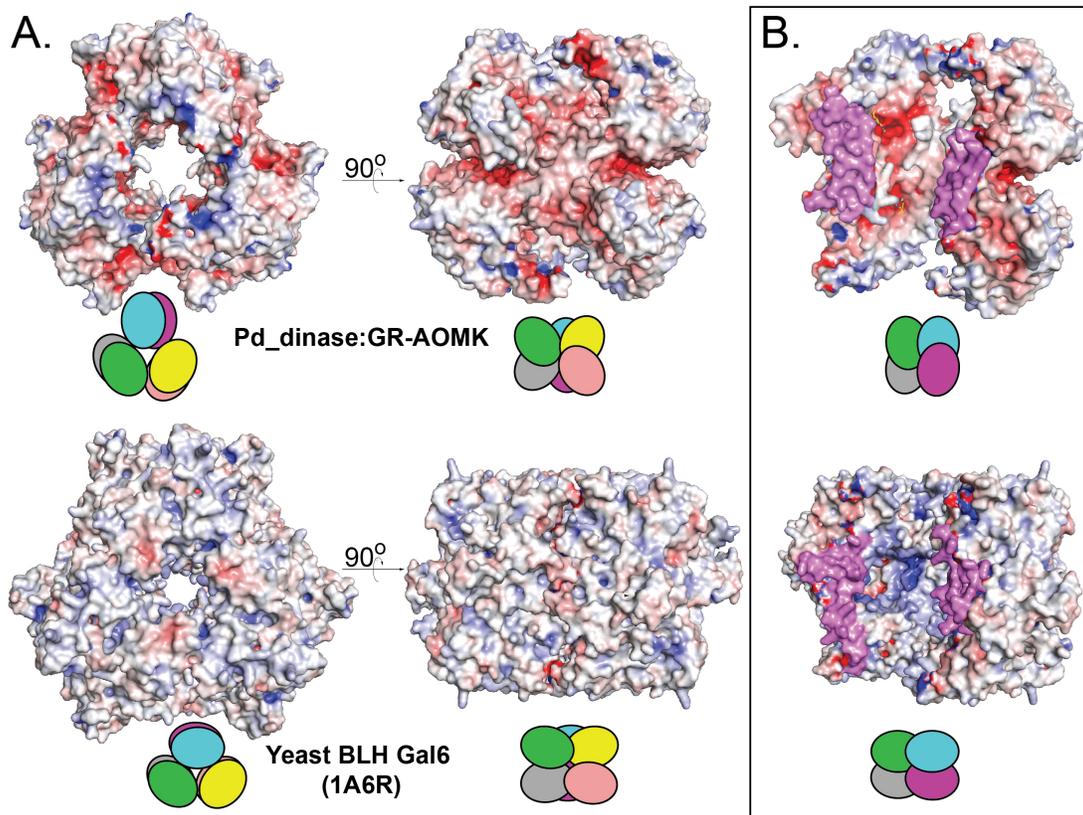
Figure 2.16. Comparison between Pd_dinase and yeast BLH homohexamers. A. Surface representation and electrostatic potential of Pd_dinase (top) and yeast BLH Gal 6 (bottom) homohexamers accompanied by a schematic to illustrate the position of each chain. B. Two chains are removed from the homohexamers to exhibit the electrostatic potential of the central channel. The interior of yeast BLH (bottom) is strongly positive while Pd_dinase (top) is slightly negative. The oligomer interface is colored magenta. The electrostatic surface potentials are calculated and colored according to Figure 2.12.

Localization of the proteases are also different. Native Pd_dinase is expressed with a signal sequence and is thus predicted to be secreted into the host environment. The BLH family, instead, are largely localized in the cytoplasm. This further supports that despite similar structural features, the two proteases serve different biological functions.

6. Antimicrobial peptide cleavage by Pd_dinase

We posited that potential biologically relevant Pd_dinase substrates would have an increased frequency of Gly residues. Certain classes of antimicrobial peptides (AMPs) are rich in Gly and represent a common theme for host defense across different organisms, including bacteria,[50] plants,[51] and humans.[52] These AMPs have been found to mainly target Gram⁻ bacteria through outer membrane pore formations and DNA interactions. One set of human-derived Gly-rich AMPs are the keratin-derived antimicrobial peptides (KAMPs). KAMPs represent peptide fragments, ranging from 10 to 36 residues that are shed from the C-terminal region of the constitutively expressed human keratin 6A of stratified epithelial cells.[53] KAMP-14 (a 14-mer fragment) is bactericidal against *Pseudomonas aeruginosa*, *E. coli*, *Streptococcus pyogenes*, and *Staphylococcus aureus.*[52, 54] We synthesized KAMP-14 (GGLSSVGGGSSTIK) and KAMP-13 (AIGGGLSSVGGGS) using standard solid-phase peptide synthesis and incubated these peptides with Pd_dinase. Using the same LC-MS/MS-based approach to detect cleavage of the peptide library (Figure 2.6A), we found that KAMP-14 is degraded into a 12-mer peptide due to removal of the amino terminal Gly-Gly dipeptide. This product is further processed into a 10-mer peptide after extended incubation with Pd_dinase (Figure 2.17A). KAMP-13 was degraded into an 11-mer peptide following removal of Ala-Ile; however, this processing was slow with only 15.5% (p = 0.013) of the substrate being degraded by 1200 min. The 11-mer cleavage product consisted of

an amino terminal Gly residue and was therefore rapidly hydrolyzed to a 9-mer peptide (Figure 2.17B). We next incubated Pd_dinase with human β-defensin 2 (hBD2), a potent AMP produced in human epithelial cells that is upregulated in patients with ulcerative colitis.[55] The amino acids in position 2 and 3 are identical to KAMP-13; however, this substrate has the preferred amino terminal Gly residue. Therefore, hBD2 was hydrolyzed faster than KAMP-13 with 51.2% at the substrate degraded by 60 min (Figure 2.17C). Further cleavage of hBD2 by Pd_dinase between Asp and Pro could not be evaluated in this assay because partial hydrolysis of this bond occurred even in the absence of enzyme. The Asp-Pro bond is acid-labile[56] and was likely hydrolyzed during the sample acidification step prior to LC-MS/MS. As a control, we also incubated Pd_dinase with human neutrophil peptide 3 (HNP3), a well-studied human AMP that participates in systemic innate immunity.[57] We predicted that this would not be cleaved by Pd_dinase because in contains disfavored Asp residue at the amino terminus. After a 4-h incubation with Pd_dinase, we failed to find a cleavage product of HNP3, as well as no significant reduction of parent peptide (Figure 2.18).
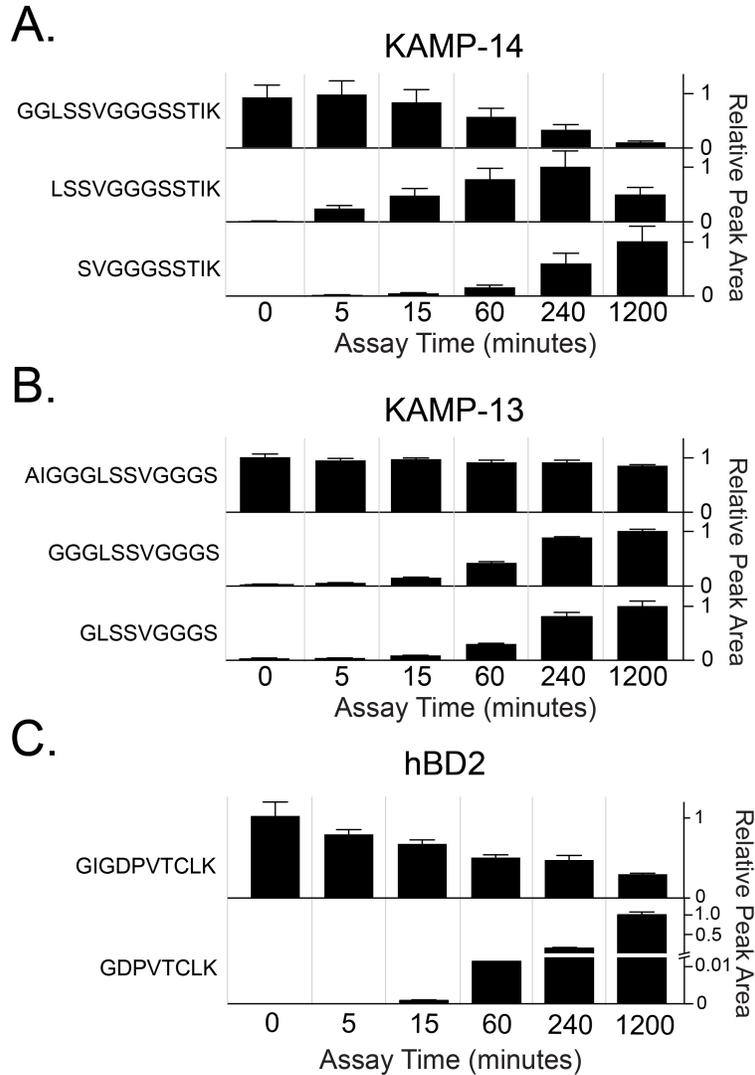
Figure 2.17. Hydrolysis of human antimicrobial peptides by Pd_dinase. A. KAMP-14 is sequentially cleaved by Pd_dinase to a 12-mer and 10-mer. B. Degradation of KAMP-13 to a 11-mer product is formed at a slower rate than KAMP-14 hydrolysis. The 11-mer product is efficiently cleaved further into a 9-mer peptide. C. 51.2% of hBD2 is cleaved after 60 min incubation. Relative peak area is calculated from triplicate assays and normalized to the time interval with the highest mean peak area.
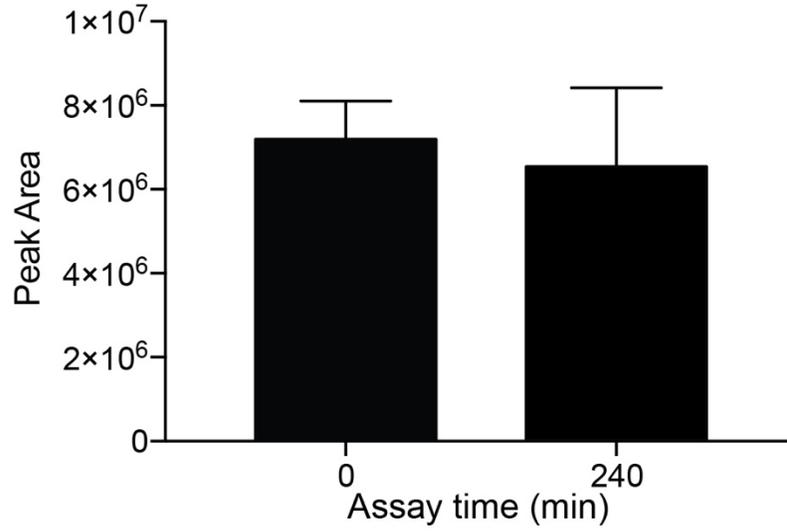
Figure 2.18. HNP3 wasn't be cleaved by Pd_dinase. There was no significant reduction of HNP3 after a 4-h incubation with Pd_dinase. Similarly, the Pd_dinase cleavage product of HNP3 was not found. Data presented as mean ± sd and the assays were performed in triplicate.

The cleavage of the three human AMPs, with important roles as antibacterial, antiviral, and antifungal agents, clearly demonstrate the potential of these peptides as biological substrates of secreted Pd_dinase as well as related homologs from other commensal bacteria. While human AMP LL-37 has been shown to be hydrolyzed by bacterial proteases from pathogenic bacteria, including *S. aureus*[58] and *Bacillus anthracis,*[59] we show that AMP degradation may aid in the colonization of the human gut by commensal organisms. The breadth of AMPs susceptible to Pd_dinase hydrolysis is a subject for further study, as most characterized secreted AMPs are 10-50 amino acids in length[60] and Pd_dinase prefers unmodified N-terminal Gly residues.

Pd_dinase may have additional biological substrates that are rich in Gly residues, including plant material and peptidoglycan layers of other bacteria. For example, many plants introduced through diet produce Gly-rich proteins often localized in cell walls and vacuoles and represent potential Pd_dinase substrates as a nutritional source for both the microbiome constituents and host. Similarly, the peptidoglycan layer of certain bacteria may be a source of Pd_dinase substrates, such as *S. aureus*. The *S. aureus* peptidoglycan is comprised of poly-Gly chains employed by sortase A (SpA) to covalently adhere proteins with a LPXT/G motif. Additional studies will determine the scope of Pd_dinase targets and role in survival of *P. distasonis.*

D. Conclusions

We report the biochemical and structural characterization of a cysteine protease containing a canonical N-terminal secretion signal derived from the gut commensal bacteria *P. distasonis*. This protease is highly conserved across commensal gut bacteria and arranges in a homohexamer with all six active sites directed along a central cavity akin to bleomycin

hydrolase. Using MSP-MS, kinetics, and x-ray crystallography, we have determined that this protease acts as a di-aminopeptidase with preference for N-terminal glycine. We demonstrate that potential biologically relevant substrates include several glycine-rich human antimicrobial peptides that are readily degraded by the protease. Our studies provide new mechanistic insights into how commensal bacteria may colonize the distal gut and overcome the presence of host antimicrobial agents.

E. Acknowledgement

F. References

1.   Sheydina, A.; Eberhardt, R. Y.; Rigden, D. J.; Chang, Y.; Li, Z.; Zmasek, C. C.; Axelrod, H. L.; Godzik, A., Structural genomics analysis of uncharacterized protein families overrepresented in human gut bacteria identifies a novel glycoside hydrolase. *BMC Bioinformatics* **2014,** *15*, 112.

2.   Fleischman, N. M.; Das, D.; Kumar, A.; Xu, Q.; Chiu, H. J.; Jaroszewski, L.; Knuth, M. W.; Klock, H. E.; Miller, M. D.; Elsliger, M. A.; Godzik, A.; Lesley, S. A.; Deacon, A. M.; Wilson, I. A.; Toney, M. D., Molecular characterization of novel pyridoxal-5'-phosphate-dependent enzymes from the human microbiome. *Protein Sci.* **2014,** *23* (8), 1060-76.

3.   Bhowmik, S.; Jones, D. H.; Chiu, H. P.; Park, I. H.; Chiu, H. J.; Axelrod, H. L.; Farr, C. L.; Tien, H. J.; Agarwalla, S.; Lesley, S. A., Structural and functional characterization of BaiA, an enzyme involved in secondary bile acid synthesis in human gut microbe. *Proteins* **2014,** *82* (2), 216-29.

4.   Michalska, K.; Tan, K.; Li, H.; Hatzos-Skintges, C.; Bearden, J.; Babnigg, G.; Joachimiak, A., GH1-family 6-P-beta-glucosidases from human microbiome lactic acid bacteria. *Acta Crystallogr. D Biol. Crystallogr.* **2013,** *69* (Pt 3), 451-63.

5.   Kverka, M.; Zakostelska, Z.; Klimesova, K.; Sokol, D.; Hudcovic, T.; Hrncir, T.; Rossmann, P.; Mrazek, J.; Kopecny, J.; Verdu, E. F.; Tlaskalova-Hogenova, H., Oral administration of Parabacteroides distasonis antigens attenuates experimental murine colitis through modulation of immunity and microbiota composition. *Clin. Exp. Immunol.* **2011,** *163* (2), 250-259.

6.   Dziarski, R.; Park, S. Y.; Kashyap, D. R.; Dowd, S. E.; Gupta, D., Pglyrp-Regulated Gut Microflora Prevotella falsenii, Parabacteroides distasonis and Bacteroides eggerthii Enhance and Alistipes finegoldii Attenuates Colitis in Mice. *PLoS One* **2016,** *11* (1).

7.   Schwartz, D. R.; Homanics, G. E.; Hoyt, D. G.; Klein, E.; Abernethy, J.; Lazo, J. S., The neutral cysteine protease bleomycin hydrolase is essential for epidermal integrity and bleomycin resistance. *Proc. Natl. Acad. Sci. U. S. A.* **1999,** *96* (8), 4680-5.

8.   Umezawa, H.; Takeuchi, T.; Hori, S.; Sawa, T.; Ishizuka, M., Studies on the mechanism of antitumor effect of bleomycin of squamous cell carcinoma. *J. Antibiot. (Tokyo)* **1972,** *25* (7), 409-20.

9.   Lazo, J. S.; Humphreys, C. J., Lack of metabolism as the biochemical basis of bleomycin-induced pulmonary toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **1983,** *80* (10), 3064-8.

10.   Zimny, J.; Sikora, M.; Guranowski, A.; Jakubowski, H., Protective mechanisms against homocysteine toxicity: the role of bleomycin hydrolase. *J. Biol. Chem.* **2006,** *281* (32), 22485-92.

11.   O'Farrell, P. A.; Gonzalez, F.; Zheng, W.; Johnston, S. A.; Joshua-Tor, L., Crystal structure of human bleomycin hydrolase, a self-compartmentalizing cysteine protease. *Structure* **1999,** *7* (6), 619-27.

12.   Kamata, Y.; Taniguchi, A.; Yamamoto, M.; Nomura, J.; Ishihara, K.; Takahara, H.; Hibino, T.; Takeda, A., Neutral cysteine protease bleomycin hydrolase is essential for the breakdown of deiminated filaggrin into amino acids. *J. Biol. Chem.* **2009,** *284* (19), 12829-36.

13.   Zheng, W.; Xu, H. E.; Johnston, S. A., The cysteine-peptidase bleomycin hydrolase is a member of the galactose regulon in yeast. *J. Biol. Chem.* **1997,** *272* (48), 30350-5.

14.   Zheng, W.; Johnston, S. A.; Joshua-Tor, L., The unusual active site of Gal6/bleomycin hydrolase can act as a carboxypeptidase, aminopeptidase, and peptide ligase. *Cell* **1998,** *93* (1), 103-9.

15.   Papassotiropoulos, A.; Bagli, M.; Jessen, F.; Frahnert, C.; Rao, M. L.; Maier, W.; Heun, R., Confirmation of the association between bleomycin hydrolase genotype and Alzheimer's disease. *Mol. Psychiatry* **2000,** *5* (2), 213-5.

16.   Lefterov, I. M.; Koldamova, R. P.; Lazo, J. S., Human bleomycin hydrolase regulates the secretion of amyloid precursor protein. *FASEB J.* **2000,** *14* (12), 1837-47.

17.   Koldamova, R. P.; Lefterov, I. M.; Gadjeva, V. G.; Lazo, J. S., Essential binding and functional domains of human bleomycin hydrolase. *Biochemistry* **1998,** *37* (8), 2282-90.

18.   Lupas, A.; Flanagan, J. M.; Tamura, T.; Baumeister, W., Self-compartmentalizing proteases. *Trends Biochem. Sci.* **1997,** *22* (10), 399-404.

19.   Pickart, C. M.; Cohen, R. E., Proteasomes and their kin: proteases in the machine age. *Nat. Rev. Mol. Cell Biol.* **2004,** *5* (3), 177-87.

20.   Gibson, S. A.; McFarlan, C.; Hay, S.; MacFarlane, G. T., Significance of microflora in proteolysis in the colon. *Appl. Environ. Microbiol.* **1989,** *55* (3), 679-83.

21.   Macfarlane, G. T.; Allison, C.; Gibson, S. A.; Cummings, J. H., Contribution of the microflora to proteolysis in the human large intestine. *J. Appl. Bacteriol.* **1988,** *64* (1), 37-46.

22.   Carroll, I. M.; Maharshak, N., Enteric bacterial proteases in inflammatory bowel disease-pathophysiology and clinical implications. *World J. Gastroenterol.* **2013,** *19* (43), 7531-43.

23.   Roka, R.; Rosztoczy, A.; Leveque, M.; Izbeki, F.; Nagy, F.; Molnar, T.; Lonovics, J.; Garcia-Villar, R.; Fioramonti, J.; Wittmann, T.; Bueno, L., A pilot study of fecal serine-protease activity: a pathophysiologic factor in diarrhea-predominant irritable bowel syndrome. *Clin. Gastroenterol. Hepatol.* **2007,** *5* (5), 550-5.

24.   Carroll, I. M.; Ringel-Kulka, T.; Ferrier, L.; Wu, M. C.; Siddle, J. P.; Bueno, L.; Ringel, Y., Fecal protease activity is associated with compositional alterations in the intestinal microbiota. *PLoS One* **2013,** *8* (10), e78017.

25.   Cenac, N.; Andrews, C. N.; Holzhausen, M.; Chapman, K.; Cottrell, G.; Andrade-Gordon, P.; Steinhoff, M.; Barbara, G.; Beck, P.; Bunnett, N. W.; Sharkey, K. A.; Ferraz, J. G.; Shaffer, E.; Vergnolle, N., Role for protease activity in visceral pain in irritable bowel syndrome. *J. Clin. Invest.* **2007,** *117* (3), 636-47.

26.   Bustos, D.; Negri, G.; De Paula, J. A.; Di Carlo, M.; Yapur, V.; Facente, A.; De Paula, A., Colonic proteinases: increased activity in patients with ulcerative colitis. *Medicina (B. Aires)* **1998,** *58* (3), 262-4.

27.   Steck, N.; Mueller, K.; Schemann, M.; Haller, D., Bacterial proteases in IBD and IBS. *Gut* **2012,** *61* (11), 1610-8.

28.   Vergnolle, N., Protease inhibition as new therapeutic strategy for GI diseases. *Gut* **2016,** *65* (7), 1215-24.

29.   MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-968.

30.   Mccoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J., Phaser crystallographic software. *J. Appl. Crystallogr.* **2007,** *40*, 658-674.

31.   Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K., Features and development of Coot. *Acta Crystallographica Section D-Biological Crystallography* **2010,** *66*, 486-501.

32.   Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H., PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D-Biological Crystallography* **2010,** *66*, 213-221.

33.   Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M., Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* **1993,** *26*, 283-291.

34.   Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E., Errors in protein structures. *Nature* **1996,** *381* (6580), 272-272.

35.   Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C.,

MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007,** *35*, W375-W383.

36.   Weiss, M. S., Global indicators of X-ray data quality. *J. Appl. Crystallogr.* **2001,** *34*, 130-135.

37.   Weiss, M. S.; Hilgenfeld, R., On the use of the merging R factor as a quality indicator for X-ray data. *J. Appl. Crystallogr.* **1997,** *30*, 203-205.

38.   Karplus, P. A.; Diederichs, K., Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* **2015,** *34*, 60-68.

39.   Roncase, E. J.; Moon, C.; Chatterjee, S.; Gonzalez-Paez, G. E.; Craik, C. S.; O'Donoghue, A. J.; Wolan, D. W., Substrate Profiling and High Resolution Co-complex Crystal Structure of a Secreted C11 Protease Conserved across Commensal Bacteria. *ACS Chem. Biol.* **2017,** *12* (6), 1556-1565.

40.   Lentz, C. S.; Ordonez, A. A.; Kasperkiewicz, P.; La Greca, F.; O'Donoghue, A. J.; Schulze, C. J.; Powers, J. C.; Craik, C. S.; Drag, M.; Jain, S. K.; Bogyo, M., Design of Selective Substrates and Activity-Based Probes for Hydrolase Important for Pathogenesis 1 (HIP1) from Mycobacterium tuberculosis. *ACS Infect Dis* **2016,** *2* (11), 807-815.

41.   Li, H.; O'Donoghue, A. J.; van der Linden, W. A.; Xie, S. C.; Yoo, E.; Foe, I. T.; Tilley, L.; Craik, C. S.; da Fonseca, P. C.; Bogyo, M., Structure- and function-based design of Plasmodium-selective proteasome inhibitors. *Nature* **2016,** *530* (7589), 233-6.

42.   Bromme, D.; Rossi, A. B.; Smeekens, S. P.; Anderson, D. C.; Payan, D. G., Human bleomycin hydrolase: Molecular cloning, sequencing, functional expression, and enzymatic characterization. *Biochemistry* **1996,** *35* (21), 6706-6714.

43.   Gainer, H.; Russell, J. T.; Loh, Y. P., An Aminopeptidase Activity in Bovine Pituitary Secretory Vesicles That Cleaves the N-Terminal Arginine from Beta-Lipotropin60-65. *FEBS Lett.* **1984,** *175* (1), 135-139.

44.   Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C., HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (4), 915-920.

45.   Sarkar, P.; Saleh, T.; Tzeng, S. R.; Birge, R. B.; Kalodimos, C. G., Structural basis for regulation of the Crk signaling protein by a proline switch. *Nat. Chem. Biol.* **2011,** *7* (1), 51-57.

46.   Bradshaw, W. J.; Roberts, A. K.; Shone, C. C.; Acharya, K. R., Cwp84, a Clostridium difficile cysteine protease, exhibits conformational flexibility in the absence of its propeptide. *Acta Crystallographica Section F-Structural Biology Communications* **2015,** *71*, 295-303.

47.   Risor, M. W.; Thomsen, L. R.; Sanggaard, K. W.; Nielsen, T. A.; Thogersen, I. B.; Lukassen, M. V.; Rossen, L.; Garcia-Ferrer, I.; Guevara, T.; Scavenius, C.; Meinjohanns, E.; Gomis-Ruth, F. X.; Enghild, J. J., Enzymatic and Structural Characterization of the Major Endopeptidase in the Venus Flytrap Digestion Fluid. *J. Biol. Chem.* **2016,** *291* (5), 2271-2287.

48.   Bethune, M. T.; Strop, P.; Tang, Y. Y.; Sollid, L. M.; Khosla, C., Heterologous expression, purification, refolding, and structural-functional characterization of EP-B2, a self-activating barley cysteine endoprotease. *Chem. Biol.* **2006,** *13* (6), 637-647.

49.   Choe, Y.; Leonetti, F.; Greenbaum, D. C.; Lecaille, F.; Bogyo, M.; Bromme, D.; Ellman, J. A.; Craik, C. S., Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.* **2006,** *281* (18), 12824-12832.

50.   Bayer, A.; Freund, S.; Jung, G., Post-translational heterocyclic backbone modifications in the 43-peptide antibiotic microcin B17. Structure elucidation and NMR study of a 13C,15N-labelled gyrase inhibitor. *Eur. J. Biochem.* **1995,** *234* (2), 414-26.

51.   Park, C. J.; Park, C. B.; Hong, S. S.; Lee, H. S.; Lee, S. Y.; Kim, S. C., Characterization and cDNA cloning of two glycine- and histidine-rich antimicrobial peptides from the roots of shepherd's purse, Capsella bursa-pastoris. *Plant Mol. Biol.* **2000,** *44* (2), 187-97.

52.   Tam, C.; Mun, J. J.; Evans, D. J.; Fleiszig, S. M., Cytokeratins mediate epithelial innate defense through their antimicrobial properties. *J. Clin. Invest.* **2012,** *122* (10), 3665-77.

53.   Moll, R.; Divo, M.; Langbein, L., The human keratins: biology and pathology. *Histochem. Cell Biol.* **2008,** *129* (6), 705-733.

54.   Lee, J. T.; Wang, G.; Tam, Y. T.; Tam, C., Membrane-Active Epithelial Keratin 6A Fragments (KAMPs) Are Unique Human Antimicrobial Peptides with a Non-alphabeta Structure. *Front. Microbiol.* **2016,** *7*, 1799.

55.   Wehkamp, J.; Harder, J.; Weichenthal, M.; Mueller, O.; Herrlinger, K. R.; Fellermann, K.; Schroeder, J. M.; Stange, E. F., Inducible and constitutive beta-defensins are differentially expressed in Crohn's disease and ulcerative colitis. *Inflamm. Bowel Dis.* **2003,** *9* (4), 215-223.

56.   Fung, E. N.; Zambito, F.; Haulenbeek, J.; Piccoli, S. P.; Zhang, Y.; DeSilva, B.; Arnold, M.; Kozhich, A., Targeting an acid labile aspartyl-prolyl amide bond as a viable alternative to trypsin digestion to generate a surrogate peptide for LC-MS/MS analysis. *Bioanalysis* **2014,** *6* (22), 2985-2998.

57.   Cunliffe, R. N., alpha-Defensins in the gastrointestinal tract. *Mol. Immunol.* **2003,** *40* (7), 463-467.

58.   Sieprawska-Lupa, M.; Mydel, P.; Krawczyk, K.; Wojcik, K.; Puklo, M.; Lupa, B.; Suder, P.; Silberring, J.; Silberring, J.; Reed, M.; Pohl, J.; Shafer, W.; McAleese, F.; Foster, T.;   Travis, J.; Potempa, J., Degradation of human antimicrobial peptide LL-37 by Staphylococcus aureus-derived proteinases. *Antimicrob. Agents Chemother.* **2004,** *48* (12), 4673-4679.

59.   Thwaite, J. E.; Hibbs, S.; Titball, R. W.; Atkins, T. P., Proteolytic degradation of human antimicrobial peptide LL-37 by Bacillus anthracis may contribute to virulence. *Antimicrob. Agents Chemother.* **2006,** *50* (7), 2316-2322.

60.   Guilhelmelli, F.; Vilela, N.; Albuquerque, P.; Derengowski, L. D.; Silva-Pereira, I.; Kyaw, C. M., Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. *Front. Microbiol.* **2013,** *4*.

Chapter III

Detection and Characterization of Proteolytic Enzyme Activity and Neuropeptide Processing in Bovine Dense Core Secretory Vesicles

A. Introduction

Neuropeptides comprise hundreds to thousands of unique peptides that function as the major group of neurotransmitters and hormones, displaying diverse biological functions in neurotransmission, hormone signaling, as well as antimicrobial defense in human neuroendocrine systems, mammals, and vertebrate and invertebrate organisms[1-6]. The smaller group of classical small molecule neurotransmitters function with neuropeptides in neurotransmission and endocrine regulation. The diverse biological activities exerted by these peptides is defined by their primary amino acid sequence and post-translational modifications (PTMs).

Neuropeptides are generated from protein precursors, known as proneuropeptides, by the actions of several proteolytic enzymes[6-8]. The major site of neuropeptide biosynthesis occurs in the dense core secretory vesicle (DCSV) organelles of neuroendocrine and neuronal cells. Proneuropeptide precursors, proteases, and associated proteins are packaged from the Golgi apparatus into DCSV for processing, storage, and regulated release of neuropeptides into the external environment. In neurons, proneuropeptide processing occurs during axonal transport of DCSVs to synaptic nerve terminals where bioactive neuropeptides are ready to be released upon electrical stimulation for neurotransmission[9, 10]. The DCSVs from sympathoadrenal chromaffin cells of the adrenal medulla, also known as chromaffin granules (CGs), is an ideal model system for studying protease mechanisms of proneuropeptide processing and bioactive neuropeptide production. These cells are isolated from the peripheral sympathetic nervous system and contain a complex mixture of neuropeptides, catecholamines, and related signaling components which will be released in response to splanchnic nerve stimulation in stress responses[11, 12]. Previous studies have shown that chromaffin granules are rich in neuropeptides and their precursors,

including chromogranins, secretogranins, enkephalins, neuropeptide Y, Vasoactive intestinal polypeptide, and galanin[11, 13-20], as well as numerous proteases for biosynthesis of the rich spectrum of neuropeptides present in these DCSVs[21-24].

There have been studies focusing on examining selected proteases, one at a time, present in CG and DCSVs of neuroendocrine cells. These works showed that the cysteine proteases cathepsin L and cathepsin V, the serine proteases PC1 and PC2 (PC, proprotein convertase), and the exopeptidases CPE (carboxypeptidase E) and cathepsin H participate in neuropeptide production[6, 7]. Immune-depletion, inhibition, RNAi and gene knockout are traditional technologies to uncover the involvement of individual protease in neuropeptide biosynthesis analysis, but are usually labor-intensive, time-consuming, less accurate and often require prior knowledge of the target proteases and neuropeptides. And the complete repertoire of proteolytic activities for neuropeptide processing within the DCSV organelle has not yet been achieved in the field. Nowadays, LC-MS/MS-based technologies have been emerging as powerful tools in studying protein and peptide PTMs as they are more quantitative, efficient and able to directly measure the sequences and modifications of proteins and peptides in context of complex biological samples[12, 25, 26].

In this study, we sought to establish a multi-omics platform to comprehensively investigate protease-proneuropeptide interactions in model chromaffin cell DCSVs to uncover the role of proteases in proneuropeptide processing. We acquired and integrated data from proteomics, peptidomics, and multiplex substrate profiling by mass spectrometry (MSP-MS)[27] to identify all proteins, peptides, and to characterize protease activities in CGs. Additionally, we performed *in vitro* incubation assays to reveal the dynamic changes of neuropeptidome before and after proteolysis and incorporate class-specific inhibitors to deconvolute the complexity of

proteolysis. Altogether, we uncovered the proteolytic maps of CG proneuropeptides and discovered six catalytically active proteases that have may be involved in regulating the synthesis of active neuropeptides.

B. Materials and Methods

1. Purification of bovine chromaffin granules

Dense core secretory vesicles (DCSV), represented by CG present in adrenal medullary chromaffin cells, were isolated from fresh bovine adrenal medulla by differential sucrose density centrifugation as we have described previously[11]. Fresh adrenal medulla (bovine) tissue (from 20 adrenal glands) was gently homogenized in 0.32 M sucrose using a polytron and centrifuged at 365 × g (20 minutes at 4 °C) to remove nuclei (pellet). The CG-containing supernatant was centrifuged at 12,000 × g (20 minutes at 4 °C) to resulted in the CG pellet. The CG pellet was washed three times in 0.32 M sucrose (12,000 × g, 20 min., 4 °C). The washed CG pellet was resuspended in 0.32 M sucrose, layered on a 1.6 M sucrose step gradient with a cushion of 2.2 M sucrose, and centrifuged at 120,000 x g (120 min., 4 °C) in a Beckman SW28 rotor to result in purified CG at the interface of the 1.6/2.2 M sucrose. The CG were removed, resuspended in 0.32 M sucrose, layered on 1.6 M sucrose, and subjected to a second ultracentrifugation step (SW28 rotor, 120,000 × g). The resultant purified CG was resuspended in 15 mM KCl and stored at -70 °C for analysis.

We have documented the high purity of this preparation of isolated CG by electron microscopy and biochemical markers[11, 28, 29]. The purified CG lack markers for the subcellular organelles of lysosomes (acid phosphatase marker)[29], cytoplasm (lactate dehydrogenase marker)[30], mitochondria (fumarase and glutamate dehydrogenase markers)[28, 31], and endoplasmic

reticulum (glucose-6-phosphatase marker)[31]. Thus, the purity of the isolated CG has been documented in the literature[11, 28, 31]. Proteins were quantified using BCA Protein Assay (Thermo).

2. Identification of proteins in chromaffin granules

Proteins and proteases of CG were identified by proteomics analysis. All CG samples were prepared in triplicate. Proteins of the purified CGs (200 µg, in triplicate) were precipitated by incubating in ice-cold 90% MeOH (Thermo) for 15 min, followed by centrifugation at 14000 × g, 4 °C for 30 minutes. Protein precipitates were resuspended in the reduction buffer containing 8 M urea (MP Biomedicals), 50 mM Tris-HCl (MP Biomedicals), pH 8.0, 5 mM DTT (Sigma), and incubated at 55°C for 45 min for protein denaturation, following cysteine alkylation with 15 mM iodoacetamide (Sigma). Alkylation was quenched by addition of DTT and 50 mM Tris-HCl buffer, pH8.0, was added to the reduced and alkylated protein samples to dilute urea to a concentration < 1M. Proteins were then digested with trypsin (Promega V5113) at a 1:50 of trypsin:protein ratio for 20 hours at 37 °C. Reactions were quenched by adding 10% TFA (Thermo) to adjust the pH down to ~2 and desalted using homemade C18 Spin Tips[32] and dried with SpeedVac.

For each sample, 2 µg of the peptides were analyzed on a Q Exactive Mass Spectrometer (Thermo) equipped with an Ultimate 3000 HPLC (Thermo) for each injection. Each triplicate sample was injected into the LC-MS/MS 3 times as technical replicates. Peptides were separated by reverse phase chromatography on a C18 column (1.7 µm bead size, 75 µm × 25 cm, heated to 65 °C) at a flow rate of 300 nL/min using a 145-minute linear gradient from 5% B to 25% B, with solvent A: 0.1% formic acid/H_2O (Thermo) and solvent B: 0.1% formic acid/acetonitrile

(Thermo). Survey scans were recorded over a 310–1250 m/z range (35,000 resolutions at 200m/z, AGC target 3×10^6, 100 ms maximum IT). MS/MS was performed in data-dependent acquisition mode with HCD fragmentation (28 normalized collision energy) on the 20 most intense precursor ions (17,500 resolutions at 200m/z, AGC target 1×10^5, 50 ms maximum IT, dynamic exclusion 20 s).

Data was processed using PEAKS 8.5 (Bioinformatics Solutions Inc.). MS^2 data were searched against Bos taurus proteome (08/28/2018) with decoy sequences in reverse order. Fixed modifications of carbamidomethylation of cysteines (+57.02146 Da), variable modification of acetylation of protein N-termini (+42.0106) and oxidation of methionine (+15.99492 Da) were specified. A maximum of two missed cleavages of trypsin was allowed. A precursor tolerance of 20 ppm and 0.01 Da for MS^2 fragments was defined. Data were filtered to 1% peptide and protein level false discovery rates with the target-decoy strategy. Proteins that were identified in at least 2 out of 3 replicate injections were considered as identified in each replicate sample and proteins identified in at least 2 out of 3 replicate samples were considered as identified in CG. Proteases and proneuropeptides in the CG proteome were compiled according to the MEROPS database[33] and NeuroPedia[34], respectively.

3. Peptidomics study of chromaffin granules

Purified CG samples (1.25 mg per replicate sample, triplicate samples prepared) were diluted into 10 mg/mL and incubated at 37 °C for 0, 30, and 90 minutes at pH 5.5 (20 mM Citrate-Phosphate, pH5.5, 100 mM NaCl (Thermo), 1 mM $CaCl_2$ (Thermo), 1mM $MgCl_2$ (Thermo), 2 mM DTT (Sigma)), and at pH 7.4 (1× PBS (Thermo) buffer, pH7.4, 1 mM $CaCl_2$, 1mM $MgCl_2$, 2 mM DTT). Incubations were conducted in the absence of protease inhibitor

(0.1% DMSO, Sigma), and in the presence of the protease inhibitors of 4 μM Pepstatin (MP Biomedicals), 40 uM E64-c (Sigma), 4 mM AEBSF (Tocris) and 10 mM EDTA (MP Biomedicals) respectively. Endogenous peptides were then extracted by addition of ice cold HCl to 20 mM HCl (pH < 3), incubation on ice for 15 min, centrifugation for 30 min (14,000 × g, 4 °C), and the supernatant peptide fraction was collected. The acid extract was then brought 20% ACN and 10 mM HCl, filtered through a 10 kDa MW cut-off filter (Millipore MRCPRT010) (centrifuge for 45-60 min at 14,000 × g, 4 °C), followed by addition of 100 μl 0.5 M NaCl and 10 mM HCl to the filter and centrifugation. The low molecular weight filtrate was neutralized by addition of 1 M ammonium bicarbonate to 30 mM, dried in a Speedvac, and stored at -70 °C for the next step. Samples were then resuspended in 100 μl urea buffer (6 M urea, 60 mM Tris-HCl, pH 8), DTT was added to a final concentration of 5 mM and incubated at room temperature for 1 hour. Iodoacetamide (IAA) stock solution (200 mM IAA in urea buffer) was added to each sample to obtain 15 mM IAA and samples were incubated to 3 min at RT in the dark. The alkylation was quenched by addition of DTT to a final concentration of 10 mM, and samples were acidified by addition of TFA. Peptides were then collected by C18 stage-tip and eluted samples were dried in a Speedvac and stored at -70 °C. Samples were resuspended in water and were briefly vortexed and sonicated, and total peptide content was measured as described for CG proteomics. Samples were dried in a SpeedVac and resuspended in 2% ACN/0.1% TFA to a peptide concentration of 0.11 μg/μl.

4.6 μl (~500 ng) of the peptides of each sample were analyzed on a Q Exactive Mass Spectrometer (Thermo) equipped with an Ultimate 3000 HPLC (Thermo) for each injection. Peptides were separated by reverse phase chromatography on a C18 column (1.7 μm bead size, 75 μm × 25 cm, heated to 65 °C) at a flow rate of 300 nL/min using a 90-minute linear gradient

from 5% B to 25% B, with solvent A: 0.1% formic acid/H_2O (Thermo) and solvent B: 0.1% formic acid/acetonitrile (Thermo). Survey scans were recorded over a 310–1250 m/z range (70,000 resolutions at 200m/z, AGC target 3×10^6, 100 ms maximum IT). MS/MS was performed in data-dependent acquisition mode with HCD fragmentation (27 normalized collision energy) on the 15 most intense precursor ions (17,500 resolutions at 200m/z, AGC target 1×10^5, 70 ms maximum IT, dynamic exclusion 30 s).

Data were processed using PEAKS 8.5 (Bioinformatics Solutions Inc.). MS^2 data were searched against the Bos taurus proteome (04/22/2018) with decoy sequences in reverse order. Fixed modifications of carbamidomethylation of cysteines (+57.02146 Da), variable modification of acetylation of protein N-termini (+42.0106), oxidation of methionine (+15.99492 Da), Pyro-glu from Q (-17.03) and Phosphorylation (STY) (+79.97) were specified. A precursor tolerance of 20 ppm and 0.01 Da for MS^2 fragments was defined. No protease digestion was specified. Data were filtered to 1% peptide and protein level false discovery rates with the target-decoy strategy. Peptides were quantified with label free quantification and data were normalized by median and filtered by 0.3 peptide quality. Missing and zero values are imputed with random normally distributed numbers in the range of the average of smallest 5% of the data $\pm$ SD.

4. Multiplex substrate profiling of proteases in chromaffin granules by mass spectrometry

All MSP-MS assays were conducted in quadruplicate by incubating purified CG or recombinant proteases with the 228-peptide library at a final concentration of 0.5 µM for each peptide[27, 35] at pH 5.5 or pH 7.4 (using the same buffers as those used in the peptidomics methods) at 37 °C. 200 µg/mL of CG sample, 18.4 nM Cathepsin A (R&D Systems), 2.64 nM

Cathepsin B (R&D Systems), 19.6 nM Cathepsin C (R&D Systems), 100 nM Cathepsin D (R&D Systems), 3.84 nM Cathepsin L (R&D Systems), 18.8 nM Carboxypeptidase E (R&D Systems) was pre-incubated with 0.1% DMSO (Sigma), 4 uM Pepstatin (MP Biomedicals), 40 uM E64-c (Sigma), 4 mM AEBSF (Tocris) and 10 mM EDTA (MP Biomedicals) respectively for 30 min at room temperature. For each assay, 20 µL of the reaction mixture was removed after 30 and 90 minutes of incubation. Protease activity was quenched by addition of GuHCl (MP Biomedicals) to a final concentration of 6.4 M and samples were immediately stored at -80 °C. Control samples (Time 0) consisted of CG or recombinant protease samples pre-incubated with 6.4 M GuHCl prior to the addition of the peptide library. All samples were desalted with C18 and dried with SpeedVac.

~0.4 µg of peptides of each MSP-MS sample was injected into a Q-Exactive Mass Spectrometer (Thermo) equipped with an Ultimate 3000 HPLC. Peptides were separated by reverse phase chromatography on a C18 column (1.7 um bead size, 75 um x 25 cm, 65°C) at a flow rate of 300 nl/min using a 60-minute linear gradient from 5% to 30% B, with solvent A: 0.1% formic acid in water and solvent B: 0.1% formic acid in acetonitrile. Survey scans were recorded over a 150–2000 m/z range (70,000 resolutions at 200m/z, AGC target 3×10^6, 100 ms maximum). MS/MS was performed in data-dependent acquisition mode with HCD fragmentation (28 normalized collision energy) on the 12 most intense precursor ions (17,500 resolutions at 200m/z, AGC target 1×10^5, 50 ms maximum, dynamic exclusion 20 s).

Data were processed using PEAKS 8.5 (Bioinformatics Solutions Inc.). MS^2 data were searched against the tetradecapeptide library sequences with decoy sequences in reverse order. A precursor tolerance of 20 ppm and 0.01 Da for MS^2 fragments was defined. No protease digestion was specified. Data were filtered to 1% peptide level false discovery rates with the

target-decoy strategy. Peptides were quantified with label free quantification and data were normalized by median and filtered by 0.3 peptide quality. Missing and zero values are imputed with random normally distributed numbers in the range of the average of smallest 5% of the data ± SD.

5. Data analysis and visualization

Proteases were annotated based on Merops[33] and proneuropeptides were annotated based on NeuroPedia[34]. For peptidomics, peptides whose abundance increased by 2-fold with q value < 0.05 comparing the assays of 30- or 90-minutes incubation to Time 0 controls were considered to be significantly increased and whose abundance decreased by 2-fold with q value < 0.05 were considered to be significantly decreased. For MSP-MS assays, peptides whose abundance pi value[36] > 2.6138 (significance level < 0.005) comparing 90-minutes incubation to Time 0 controls were considered to be increased (cleavage products). Of those cleavage products whose pi value < -2.6138 comparing the assays of 90 min incubation with the inhibitor treated CG to DMSO controls were considered to be inhibited.

Proteolytic specificity profiles were generated using IceLogo software to visualize amino-acid frequency of surrounding the cleavage sites. Amino acids that were most frequently observed (above axis) and least frequently observed (below axis) from P4 to P4′ positions were illustrated (p < 0.3). Norleucine (Nle) was represented as 'n' in the reported profiles. Amino acids in opaque text were statistically significant (p < 0.05). Heatmaps were generated using peptides that were cleaved after 90-minute incubation without inhibitor at either pHs. Hierarchical clustering was performed using Euclidean distance and complete linkage. Pearson correlations of proteolytic specificity profiles were calculated based on z-scores of amino acids

from P4 to P4′ positions. Peptidomics profiles of each proneuropeptide and peptide alignment maps were generated using Peptigram[37].

6. Data deposition

All mass spectrometry data have been deposited in MassIVE and proteomeXchange with accession numbers: MassIVE: MSV000085957 (Proteomics), MSV000085956 (Peptidomics), MSV00008595 (MSP-MS); ProteomeXchange: PXD020926 (Proteomics), PXD020925 (Peptidomics).

C. Results and Discussion

1. Proteomic analysis reveals neuropeptides and proteases in bovine chromaffin granules

In this study, we isolated chromaffin granules (CGs) from bovine adrenal medulla and identified 2110 unique proteins through proteomics. Proteins were grouped into 24 categories by gene ontology analysis. As expected, many proteins were categorized as hydrolases, enzyme modulators and signaling molecules since the nervous system utilizes dense core secretory vesicles for regulated secretion of neuropeptides whose maturation requires hydrolytic processing (Figure 3.1A). Structural proteins, as well as proteins that regulate calcium binding, vesicular trafficking, and exocytosis are also abundant components of chromaffin granules and are likely to maintain a stable internal vesicle environment and induce the secretion of neurotransmitters and hormones. In the proteome, we identified 64 putative proteases, corresponding to 2.53% in abundance calculated by normalized spectral abundance factor (NSAF)[36] (Figure 3.1B). These enzymes were further categorized by catalytic classes, revealing

73

the presence of 2 aspartyl proteases, 17 cysteine proteases, 21 serine proteases and 24 metalloproteases and their relative abundance (Figure 3.2A, B). In addition, 15 well-characterized proneuropeptides were identified including chromogranin A and B (CHGA and CHGB), proenkephalin (PENK), Neuropeptide Y (NPY), and secretogranins (Figure 3.2C). The proneuropeptides accounted for 8.79% of the total protein abundance in the chromaffin granules (Figure 3.1B), with chromogranin A alone being the single most abundant protein (Figure 3.2C). While these proteomics studies provide us with a list of proteases and potential substrates, it does not provide a functional read-out of how these protein substrates are processed, what neuropeptides are produced and what active proteases are involved. Therefore, we performed peptidomics analysis to uncover the peptides in these samples as these are the endogenous products of the proteolytic enzymes.
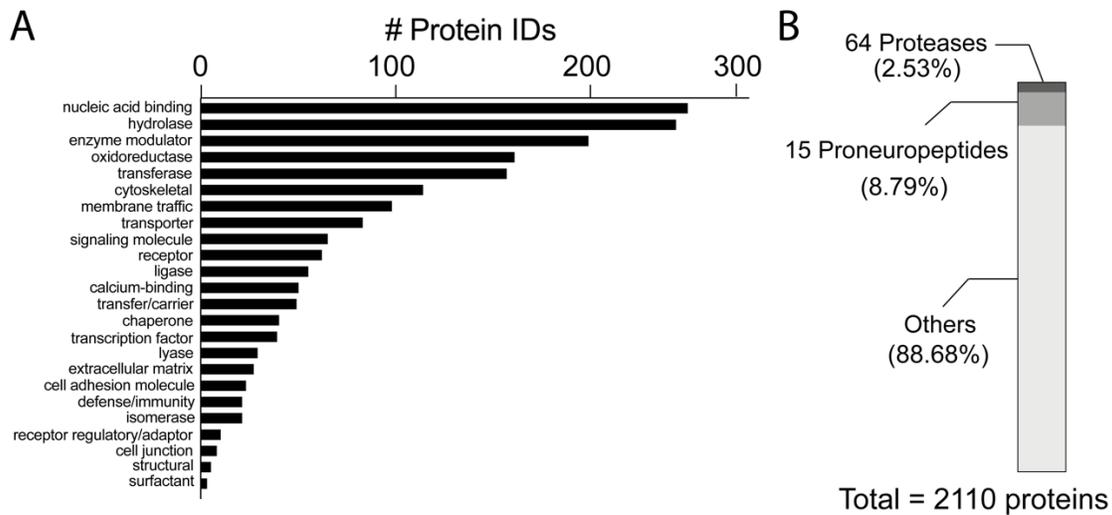
Figure 3.1. Proteins identified in CGs. A. Gene ontology analysis classified CG proteins into 24 categories. B. Relative abundance of proteases and proneuropeptides by spectral counting.
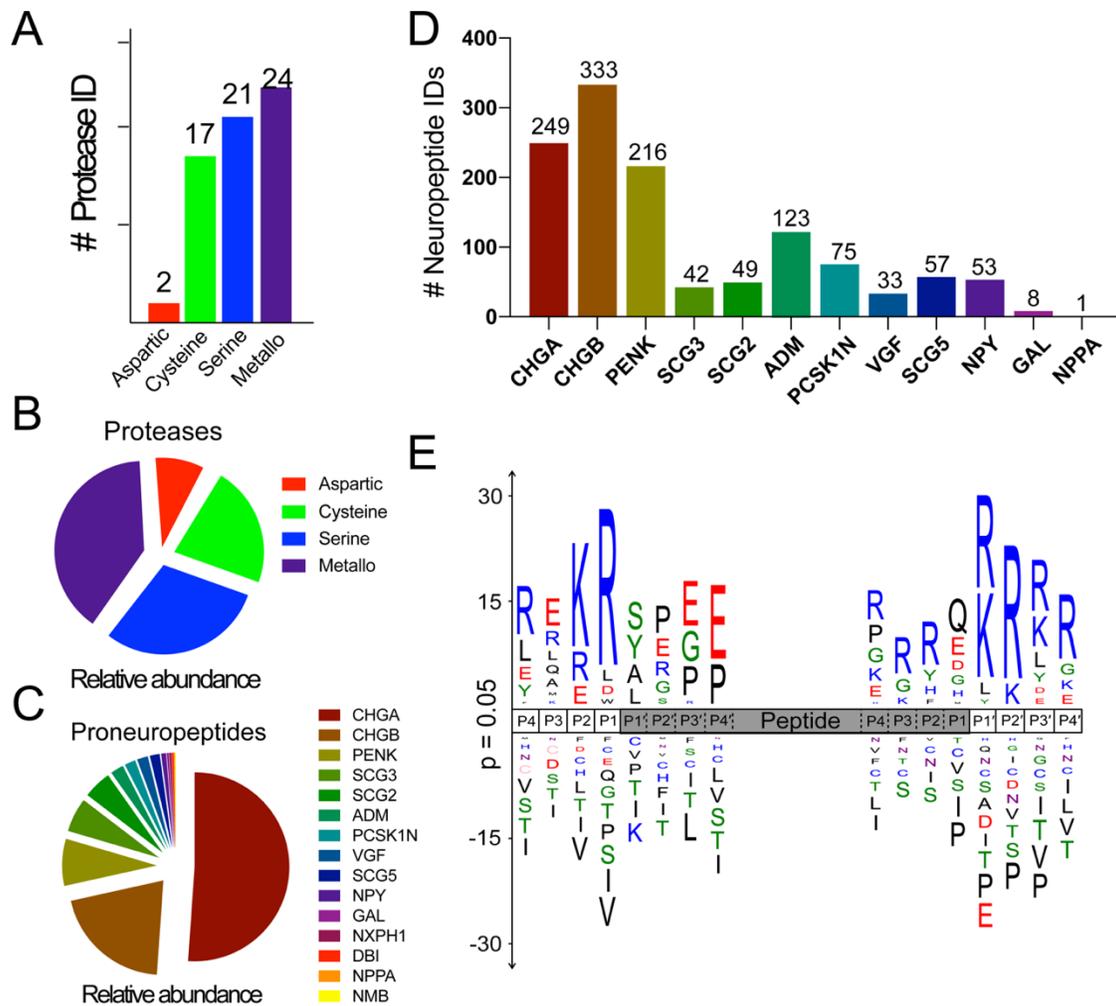
Figure 3.2. Diverse proteases and neuropeptides identified in CGs. A. 64 proteases were identified which were categorized to 4 different classes. B. Proteins were quantified by normalized spectral abundance factor (NSAF) method. Charts illustrate the relative abundance of proteases and C. proneuropeptides identified. D. Peptidomics identified 1239 out of 1981 total unique peptides, which are derived from 12 different proneuropeptides. E. A sequence logo showed amino acid sequences occurring at two termini of endogenous proneuropeptide-derived peptides. The identified peptide region is shaded in grey.

2. Peptidomic analysis reveals proneuropeptides processing in bovine chromaffin granules

The endogenous peptides in chromaffin granules were separated from the larger molecular weight proteins using molecular weight cut off filters and analyzed by LC-MS/MS. A total of 1981 unique peptides that ranged from 5 to 50 amino acids in length were identified and quantified. Of these, 1239 peptides were identified to be the proteolytic fragments of the 12 proneuropeptides and most of them were derived from chromogranins, PENK, adrenomedullin (ADM) and secretogranins (Figure 3.2D). A proteolytic specificity profile was generated by matching the proteolytic fragments to the substrate and calculating the frequency of amino acids occurring at both N and C termini of the peptides (Figure 3.2E). In protease nomenclature, cleavage occurs between the P1 and P1' amino acids and the more distal residues are defined as P2, P3, P4, etc., on the amino terminal side of the cleaved bond, and P2', P3', P4', etc., on the carboxy terminal side. We evaluated the frequency of amino acids from P4 to P4' positions on each end of the cleavage products and discovered that the cleavage specificity was remarkably different. In particular, at the N terminus, Pro and Glu were frequently found in the P2' to P4' positions while Arg and Lys were most abundant residues at the P1 to P4 positions. At the C-terminus, Arg and Lys were most frequently found from P4 to P4' with the notable exception being P1 where Gln, Glu and Asp were preferred. The frequent occurrence of basic amino acids near the peptide termini is, in general, consistent with previous studies showing that removal of di-basic amino acids is often the final step of proneuropeptides processing[6, 7]. However, it is clear that a distinct set of proteases are cleaving at each terminus as there is little similarity between the substrate profiles.

The peptidomics studies provided us with a static view of the peptidome but did not show how proneuropeptides were cleaved and what proteases were involved during these processes. When CGs were collected and purified, CG proteases remain mostly active. CGs also consist of proneuropeptides at different stages, which are fully processed, half processed and unprocessed. This allowed us to perform proneuropeptide degradation assays by incubating CGs *in vitro* to evaluate the dynamic changes of the peptidome after proteolysis. Additionally, to deconvolute the complexity of proteolytic activities in CGs, we added class-specific inhibitors to knockout the activity of a certain protease class to reveal their roles in processing proneuropeptides. In addition, we also performed the peptide degradation assays under pH 5.5 and pH 7.2 to mimic the acidic environment inside the granules and the neutral pH environment of the synaptic cleft, respectively.

CG samples (lacking exogenous protease inhibitors) were incubated for 30 and 90 minutes to quantify protease-mediated formation and degradation of peptides. At pH5.5, 646 neuropeptides were found to have increased in abundance, 486 of them were not previously found in the non-incubated peptidomics dataset. In addition, 225 neuropeptides found in the initial peptidome were significantly reduced following incubation at pH 5.5 (Figure 3.3A). At pH 7.4, 281 neuropeptides were identified to be increased and 209 of them decreased. Results indicated that proteolytic activities at pH 5.5 is much stronger than pH 7.2 as we identified more cleavage products at pH 5.5. It's interesting that the abundance of the majority of the peptides identified before incubation (T0) remained unchanged. This suggested that these peptides either reached specific sequences or structures that prevented proteases from processing them or their proteolytic formation and degradation reached an equilibrium. Both scenarios help these peptides to maintain relative stable forms and concentrations, which could be essential for their functions

when they are released from the secretory vesicles. Next, we compared the cleavage products formed at two different pHs. Results showed that there is only a small portion of overlapped peptides (Figure 3.3B). The frequency of amino acid sequences at two termini of the cleavage products of proneuropeptides at two pHs were also distinct (Figure 3.3C, D). Though Arg and Lys were still frequently identified at almost all positions, Leu and Glu occurred more frequently at P4 to P1' on N-termini and C-termini, suggesting different proteases are involved in generating these new cleavage products.

Figure 3.3. Neuropeptide production in CG after being incubated at pH 5.5 and pH 7.2. A. A Venn diagram showing the number of neuropeptides that are significantly increased or decreased after 90 min incubation compared to non-incubated control samples at pH 5.5 and pH 7.2. B. Comparison of neuropeptides whose abundance has significantly increased after 90 min incubation at two pHs C. Sequence logos showed amino acid sequences flanking at two termini of the cleavage products of proneuropeptides after 90 min incubation at pH 5.5 and D. at pH 7.2.

Proneuropeptide processing is usually driven by multiple CG proteases. To deconvolute the complexity of proteolytic processing, we added 4 exogenous class-specific inhibitors in peptide degradation assays. These inhibitors include pepstatin, E64-c, AEBSF and EDTA that inhibit aspartic, cysteine, serine and metalloproteases, respectively. By comparing inhibitor treated samples to control, we were able to determine which classes of proteases were responsible for the formation of neuropeptides. Results (Figure 3.4) showed some neuropeptides were formed at pH 5.5 but not pH 7.2 and were sensitive to pepstatin (group 2), indicating strong aspartic protease activity at pH 5.5 but not pH 7.2. On the contrary, group 6 neuropeptides were mostly formed at pH 7.2 but not pH 5.5. Group 5 neuropeptides displayed very similar profiles between two pHs, suggesting that there may be same proteases active at both two pHs and were responsible for generating these peptides. The formation of most neuropeptides was sensitive to multiple inhibitors, suggesting different proteases were involved in the cleavages of proneuropeptides. To further investigate proneuropeptide processing, we generate heatmap inhibition profiles for each proneuropeptides. The heatmap Figure 3.5 is an example summarized the processed regions of proneuropeptides and what classes of proteases were involved. For example, one of most frequently processed regions of PENK was 160-220 and this region is known to contain the sequences of functional PENKs and met-enkephalins. By further aligning the sequences of neuropeptides to the sequences of their precursor proneuropeptides, we were able to identify cleavage sites, which will be discussed later.
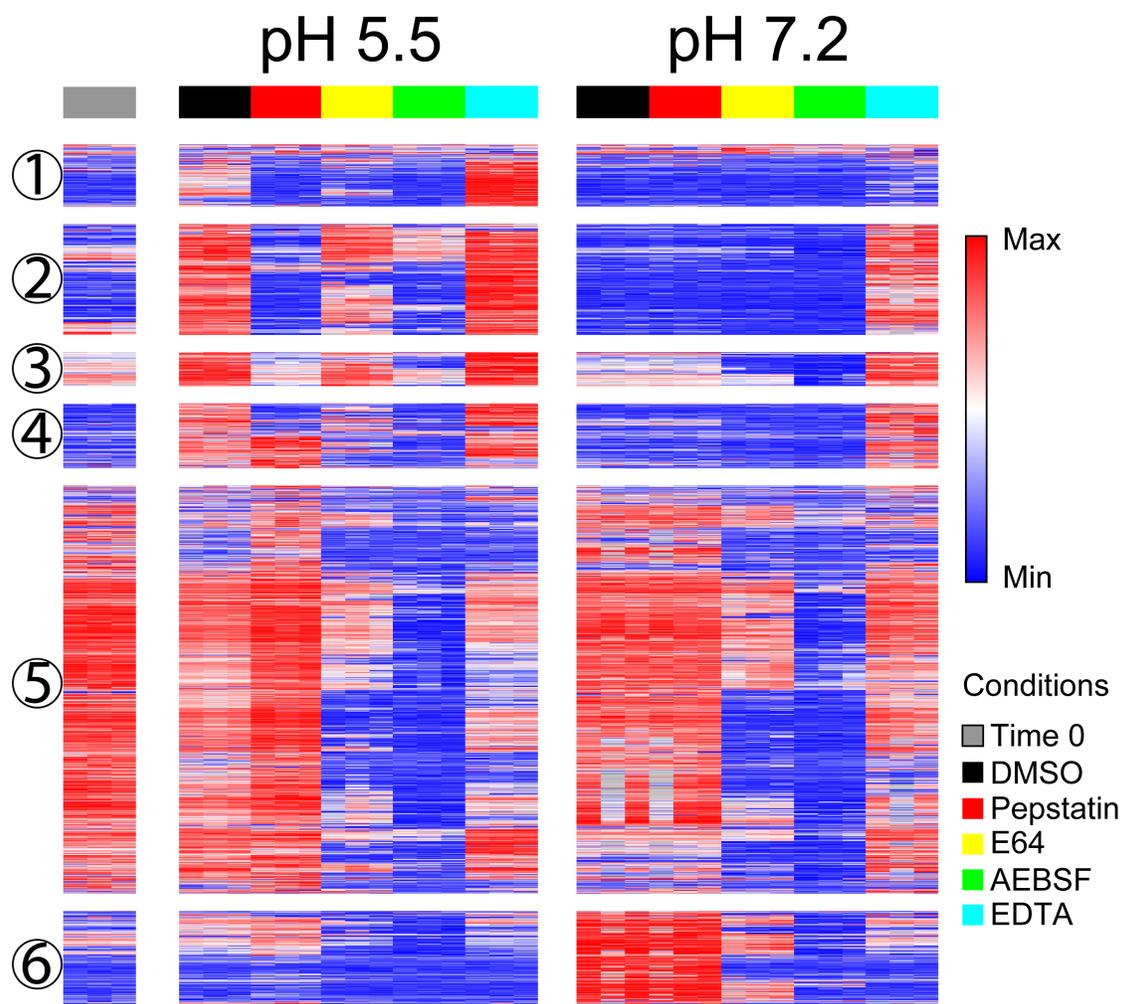
Figure 3.4. Effect of protease inhibitors on CG neuropeptide production. The heatmap shows the abundance of newly formed CG neuropeptides before and after 90-minute incubation with addition of class-specific inhibitors. Neuropeptides colored in red are at higher abundance and the ones colored in blue are at lower abundance. Hierarchical clustering was performed one row (each row represents a neuropeptide) and neuropeptides were grouped based on their inhibition profiles.
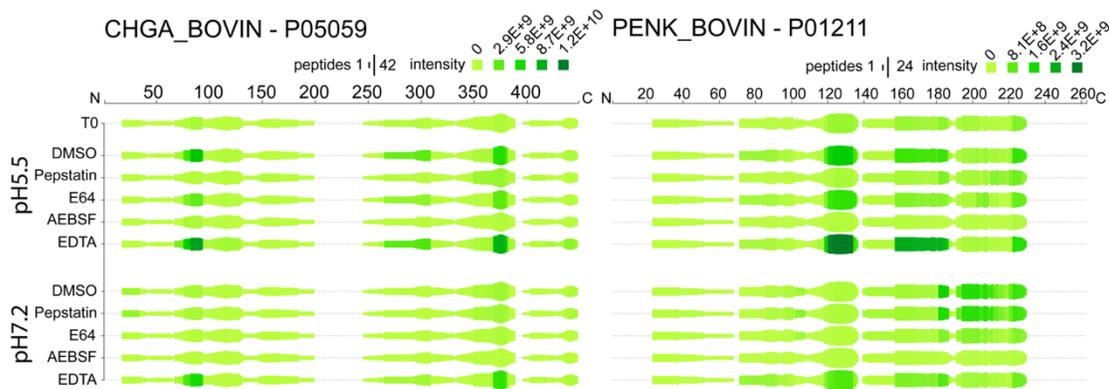
Figure 3.5. Heatmaps of 2 sample proneuropeptides peptidome profiles compiling peptidomics data after 90 min incubation with addition of different inhibitors. For each amino acid of proneuropeptides, the height of the green bars is proportional to the number of amino acids overlapping this region. The darkness of the color is proportional to the sum of the peptide intensities.

3. Characterize substrate specificities of proteases in bovine chromaffin granules via MSP-MS

Though proteomics studies identified and quantified proteases and proneuropeptides in CG, and peptidomics, together with *in vitro* degradation assays, provided us with peptidome data to reveal proteolytic processing of endogenous substrates, it's still challenging to associate individual proteases to cleavages due to several reasons. First, many substrates were undetectable due to their physical properties and the limitation of the bottom-up mass spectrometry. Some peptides were hard to ionize or carry charges. Some could be too large to fall within the scan range of a mass spectrometer. Second, it was demonstrated that endogenous substrates could be cleaved multiple times by multiple proteases before yielding the final cleavage products (Figure 3.4 and 3.5). Third, the identification and quantification of proteases in a complex biological sample were often siloed from their activities due to the presence of endogenous protease inhibitors and other regulators. They can also remain in their inactive forms and, thus, have no proteolytic function before activation. In order to characterize proteolytic activities and identify individual active proteases in CGs, we performed a functional analysis of proteolytic activities by evaluating the degradation profiles of a synthetic peptide library. This library consists of an equimolar mixture of 228 tetradecapeptides that were rationally designed for substrate specificity profiling of individual proteases[27, 38, 39] and complex biological samples[35, 40, 41]. Although a substrate library of this size has fewer cleavable peptide bonds compared to the endogenous proteins and peptide substrates in chromaffin granules, it offers several advantages for characterizing protease activity: 1) The substrates are sufficiently long to be able to quantify both the substrate and the cleaved product in the same sample, but are also unlike the endogenous substrates which are often proteins and therefore too large for peptidomics analysis; 2) The

limited length of substrates also prevents excessive proteolytic processing by multiple proteases, giving us a clear view of proteases' specificities; 3) The library substrates have a defined length and sequences that allow us to clearly distinguish aminopeptidase, endopeptidase and carboxypeptidase activities; 4) The library substrates were rationally designed to be orthogonal to most proteome. And every cleavage product that is greater than 4 amino acids can be directly linked to only one single substrate; 5) These substrates were mixed in equal molar to rule out the bias of preferred proteolysis on higher concentration substrates.

When the CG samples was incubated with the diverse library of synthetic peptides for 30 minutes in pH 5.5 assay buffer, 600 of the 2,960 peptide bonds were cleaved. When the same sample was incubated in pH 7.2 assay buffer, 204 cleavage sites were identified, with 148 sites being commonly cleaved under these different pH conditions (Figure 3.6A). The substrate specificity profile associated with the pH 5.5 cleavage sites showed a preference for hydrophobic amino acids such as Nle (n) and Phe at P2, P1, P1' and P2'. In addition, Leu, Tyr and Arg were also frequently found at the P1 site while Leu, Tyr and Ile were found at P1' (Figure 3.6B). The most frequently cleaved sites within the 14-mer peptides occurred between position 13 and 14 and between position 12 and 13 that are associated with mono- and di-carboxypeptidase activity, respectively (Figure 3.6D). When assayed at pH 7.4, the cleavage profile was distinct from the pH 5.5 profile with some notable differences being an increased frequency of Gln at P4, Arg and Lys at P1 and P2, Ser and Ala at P1' and Trp at P2' (Figure 3.6C). Under these assay conditions, the most frequently cleaved sites occurred between position 2 and 3 indicating the presence of di-aminopeptidases. In addition, the peptides were also cleaved by mono- and di-carboxypeptidases (Fig 3.6E).

As we did in peptide degradation assays, four class specific protease inhibitors, pepstatin, E64-c, AEBSF and EDTA, were incubated with CG samples prior to performing the MSP-MS to deconvolute the complexity of proteolytic activities. When CG samples were assayed at pH 5.5 in the presence of pepstatin the number of cleavage products was greatly reduced (Figure 3.7). In total, 124 sites were found to be sensitive to pepstatin as the cleavage products were detectable in the DMSO control assays but not in the pepstatin-treated assays. Further analysis showed that pepstatin-sensitive sites were generally located within the central regions of the 14-mer substrate between residues 4 and 12, indicating that the aspartic proteases have endopeptidase activities. In addition, these enzymes have a preference for cleaving hydrophobic residues such as Phe, Leu, Nle and Tyr at P1 and Nle, Phe, Tyr and Ile at P1′ (Figure 3.8A). Only two aspartic proteases were found in our proteomics analysis, with cathepsin D being present 33.9-fold higher in protein abundance than presenilin-1. These data indicate that under acid conditions, bovine cathepsin D in the CGs is very likely to be responsible for endopeptidase cleavage between hydrophobic amino acids. Additionally, pre-treatment of the CG samples with pepstatin at pH 7.2 assay conditions had no effect on peptide cleavage compared to the DMSO control (Figure 3.7). The result also supported our hypothesis as Cathepsin D is not active at pH 7.2. Therefore, pepstatin was unlikely to show inhibitory effect under that pH condition.
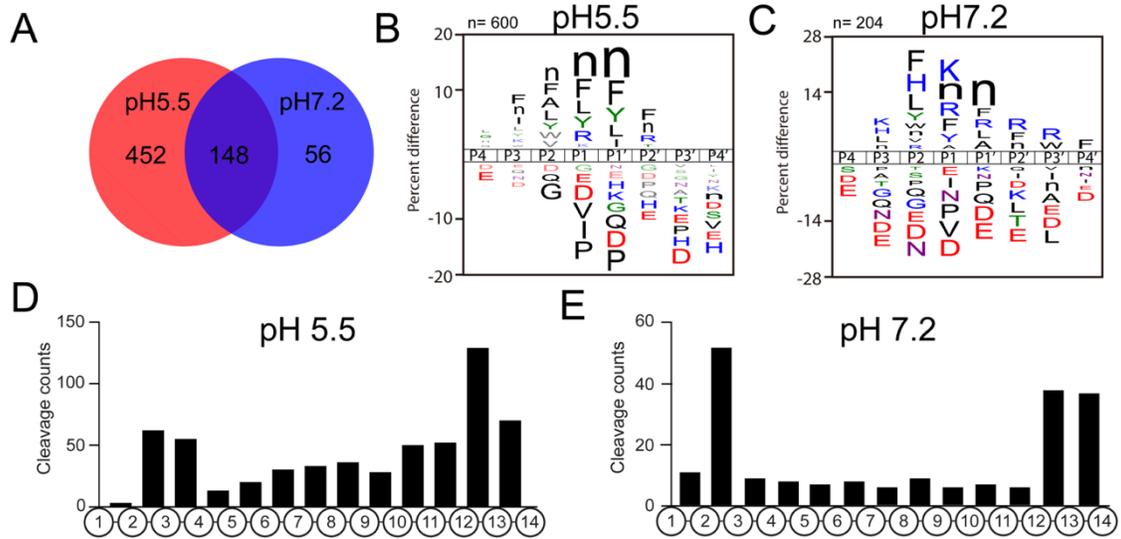
Figure 3.6. MSP-MS uncovered protease substrate specificity of CG. A. MSP-MS was performed at pH 5.5 and pH 7.2 to study protease activities inside CG and in extracellular space, respectively. B. Substrate specificity profiles of CG at pH 5.5 and C. at pH 7.2. D. Bars charts summarize the cleavage frequency at each peptide bond at pH 5.5 and E. pH 7.2.
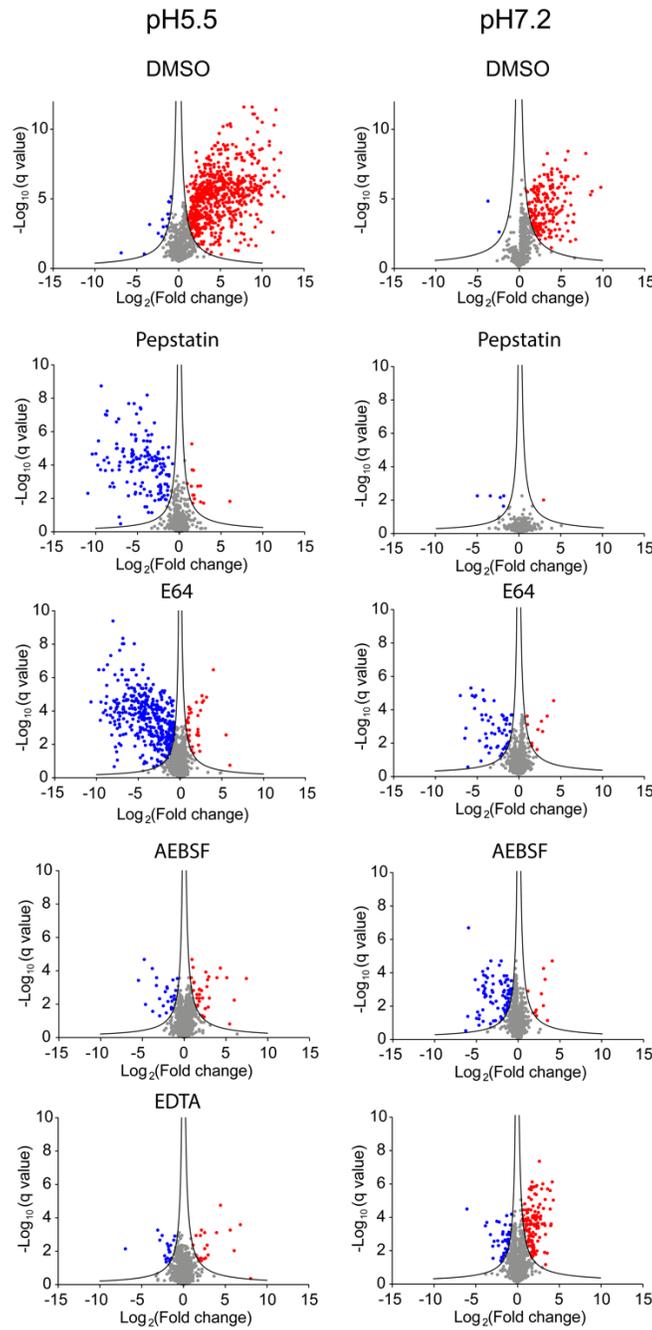
Figure 3.7. Volcano plots showing inhibition of cleavages by adding 4 class specific protease inhibitors. Peptide abundance were compared between 30min incubation and time 0. Peptides with pi value > 2.6138 or < -2.1618(significance level α < 0.05) were considered to be significantly changed. Abundance of peptides colored in grey are unchanged, peptides colored in red are significantly increased and peptides colored in blue are significantly decreased.

Treatment of CG samples with E64-c at pH 5.5 and comparison to the DMSO control revealed that at least three distinct cysteine proteases were present (Figure 3.8B, C). One enzyme is a di-aminopeptidase which primarily active at pH 5.5 and cleaves between residues 2 and 3; while another is a di-carboxypeptidase which is active at both pHs with a preference for cleaving between residues 12 and 13. A third cysteine endopeptidase is likely to be responsible for cleavage between residues 4 and 12 and mainly active at pH 5.5. Looking at the overall substrate profiles associated with the E64-c sensitive activities at pH 5.5, Arg and Lys were frequently found in the P1 position, with Phe, Val and Tyr at P2 and Nle, Leu and Phe at P1′. When the CG samples were incubated at pH 7.2 in the presence of E64-c, the total number of sensitive cleavage products was reduced however the dominant enzyme was a di-carboxypeptidase. In the proteomic data for chromaffin granules, we identified 17 cysteine proteases with cathepsin B being the most abundant in this group. Cathepsin B has been shown to have di-carboxypeptidase activity and is active at pH 5.5 and pH 7.2. In addition, cathepsin C and cathepsin L were also found in the granules and are possible candidates for the di-aminopeptidase and endopeptidase activity, respectively.

When chromaffin granules were assayed with AEBSF, the majority of AEBSF-sensitive serine protease activity occurred between residues 13 and 14 indicating a mono-carboxypeptidase, while some aminopeptidase and endopeptidase activities were also observed (Figure 3.8D, E). This mono-carboxypeptidase was likely to be cathepsin A, which is the most abundant serine mono-carboxypeptidase in our proteomics list.
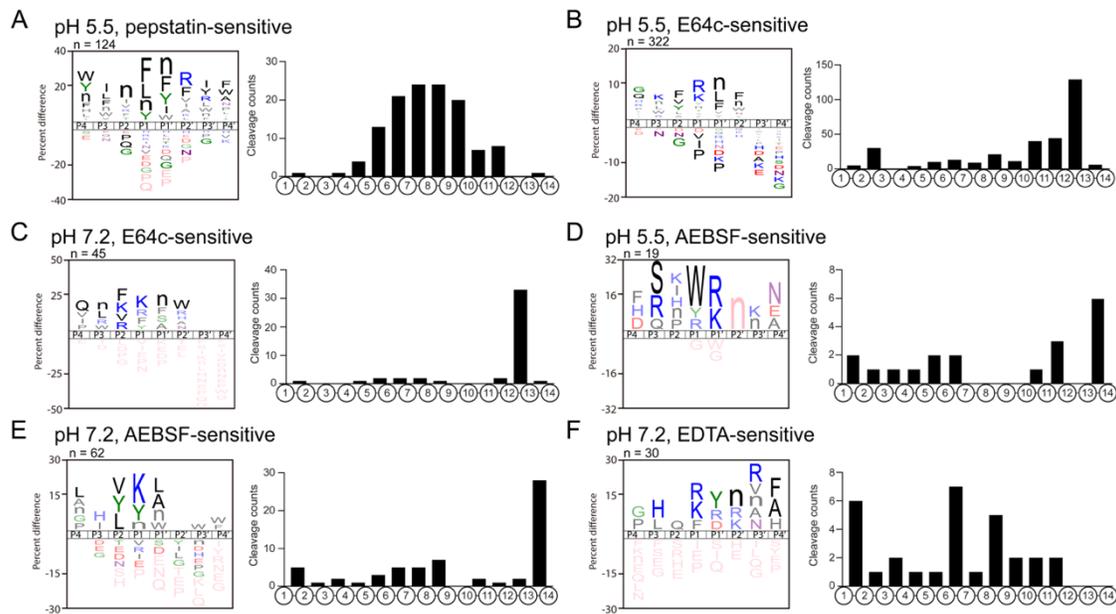
Figure 3.8. Inhibitor-sensitive protease specificity profiles. A. Substrate specificity profile of pepstatin-sensitive aspartic proteases at pH 5.5 B. Substrate specificity profile of E64-c-sensitive cysteine proteases at pH 5.5 and C. at pH 7.2. D. Substrate specificity profile of AEBSF-sensitive serine proteases at pH 5.5 and E. at pH7.2. F Substrate specificity profile of EDTA-sensitive serine proteases at pH 7.2.

Pretreatment of CG samples with the metalloprotease inhibitor, EDTA, caused a reduction in cleavage of some peptides compared to control. However, many other peptides increased in concentration. This phenomenon occurred in both pH 5.5 and pH 7.2 assay conditions (Figure 3.7), making it difficult to interpret the cleavage data and identify the target enzymes. However, carboxypeptidase E (CPE), a metalloprotease, was the most abundant protease found in CGs and this enzyme was predicted to cleave between residues 13 and 14 in the peptide library. Though the substrate specificity profile (Figure 3.8F) didn't match CPE, we decided to include CPE in our further analysis.

To support our hypothesis, we performed MSP-MS assays on recombinant proteases to check if their specificity profiles correlate. Results showed that purified proteases have similar profiles as the inhibitor sensitive ones do, especially for cathepsin D and cathepsin B (Figure 3.9). Pearson correlation scores were calculated for each cleavages site between CG inhibition profiles and purified protease profiles (Table 3.1). We further compared the cleavages identified in MSP-MS assays with CG samples and recombinant proteases. We found that these 6 selected proteases altogether were capable of covering 67.3% and 88.5% of cleavages by CG proteases at pH 5.5 and pH 7.2 respectively (Figure 3.10).
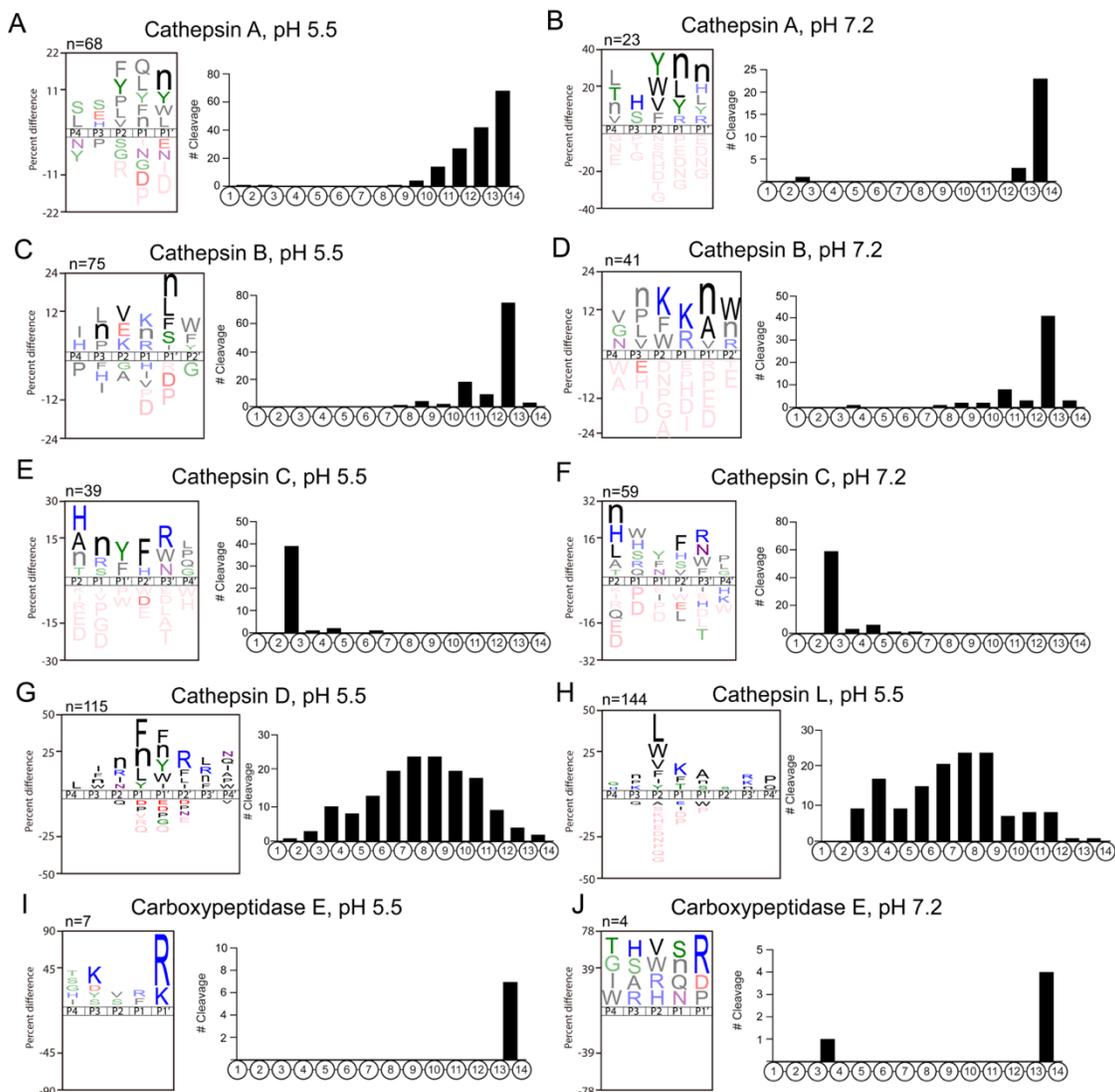
Figure 3.9. MSP-MS profiles of recombinant proteases. A. Cathepsin A at pH5.5 and B. at pH7.2, C. Cathepsin B at pH 5.5 and D. at pH 7.2, E. Cathepsin C at pH5.5 and F. at pH 7.2, G. Cathepsin D at pH 5.5, H. Cathepsin L at pH 7.2, I. Carboxypeptidase E at pH 5.5 and J. at pH 7.2.

Table 3.1. Pearson correlation of inhibitor sensitive specificity profiles and recombinant protease specificity profiles. Numbers are Pearson R.

E64 sensitive pH5.5	P4	P3	P2	P1	P1'	P2'	P3'	P4'
CatB	0.4164	0.704	0.7933	0.8697	0.9507	0.7439	0.3205	0.4234
CatC	-0.0687	-0.4356	0.0469	0.5783	0.6411	0.2647	0.2581	0.0667
CatL	0.1735	0.5676	0.5954	0.7607	0.1569	0.5025	0.1933	0.0038

E64 sensitive pH7.2	P4	P3	P2	P1	P1'	P2'	P3'	P4'
CatB	0.6417	0.637	0.8098	0.868	0.7119	0.7543	0.2608	-0.156
CatC	-0.0831	-0.0824	-0.1852	0.4481	0.5626	0.1979	-0.0548	-0.041

Pepstatin sensitive pH5.5	P4	P3	P2	P1	P1'	P2'	P3'	P4'
CatD	0.6417	0.637	0.8098	0.868	0.7119	0.7543	0.2608	-0.156

AEBSF sensitive pH7.2	P4	P3	P2	P1	P1'	P2'	P3'	P4'
CatA	0.5668	0.1387	0.623	0.3258	0.4924	0.3451	0.073	0.0558
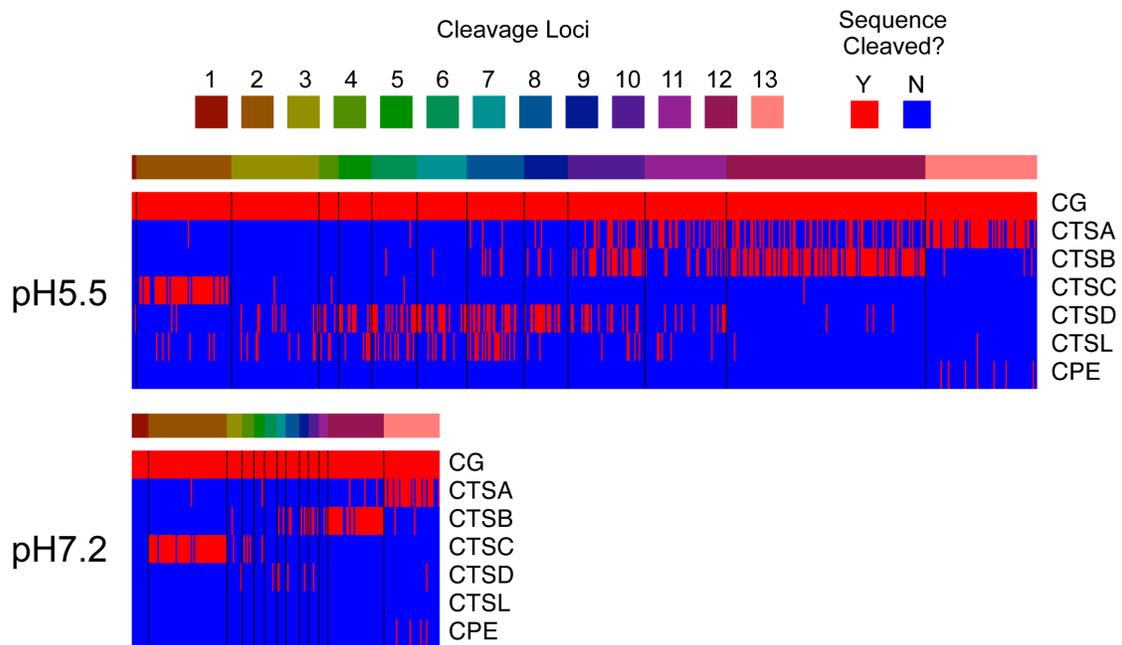
Figure 3.10. Comparison the cleavages identified in MSP-MS assays with CG and recombinant proteases.

4. Predict the roles of proteases in proenkephalin processing in bovine chromaffin granules

In peptide degradation assays, we were able to identify cleavage sites by mapping sequences back to precursor proneuropeptides. We also obtained the information of what classes of proteases were involved in generating these peptides with the help of adding exogenous inhibitors. Here, we demonstrated the example of proenkephalin (Figure 3.11). We observed several cleavage sites at two sides of dibasic residues that are sensitive to AEBSF. These were likely to be the classic PC1/2 cleavage sites that trim off the dibasic residues for the final step of peptide maturation. We also identified other endopeptidase cleavages, such as the one between 71L and 72T and the one between 157L and 158L. These peptides' formation was all sensitive to pepstatin, indicating cathepsin D may be responsible for the cleavages as it's the most abundant aspartic protease in CGs that preferentially cleaves leucine at P1 and P1'. The missing of these cleavage sites at pH 7.2 also supports our prediction.

It's evident that exopeptidase activities were strong in CGs as we observed a lot of "ladderings" of peptides, with processive cleavages of 1 or 2 amino acids from either or both termini. These "ladderings" are clearly results of exopeptidases, such as cathepsin A, B, C and carboxypeptidase E, which trim peptides step by step into a stable mature form. For example, at pH 5.5, exopeptidases remove N-terminal residues staring from K131 to M137 and C-terminal residues starting from G188 to Y182 and, revealing dibasic KK (138-139) and KR (180-181) residues at two termini. Then these two dibasic residues were further cleaved off, allowing the functional peptide, PENK (140-179) to be generated. On the other hand, the processive exopeptidase cleavages on PENK (111-130) at pH 5.5 will completely degrade the functional peptide to inactive fragments. Another interesting finding is that met-enkephalins and leu-

enkephalins were frequently found to be cleaved between F and M/L amino acids and the cleavages were often sensitive to AEBSF. We hypothesized that cathepsin A was the candidate protease that removes a single methionine/leucine from enkephalins to metabolize them into inactive fragments.

The distinct peptide alignment maps at pH 5.5 and pH 7.2 suggested different proteases being active inside CGs and at synaptic cleft. Clearly the pH change during exocytosis inactivates many CG proteases to restrict their activities only within CGs as we saw fewer protease cleavages at pH 7.2 comparing to pH 5.5 in both peptidomics and MSP-MS assays. Interestingly, here, we also observed some new cleavages on PENK at pH 7.2, mainly in the non-functional regions such as PENK (25-70) and PENK (190-220). These activated proteases were probably included in CGs to recycle amino acids after exocytosis.
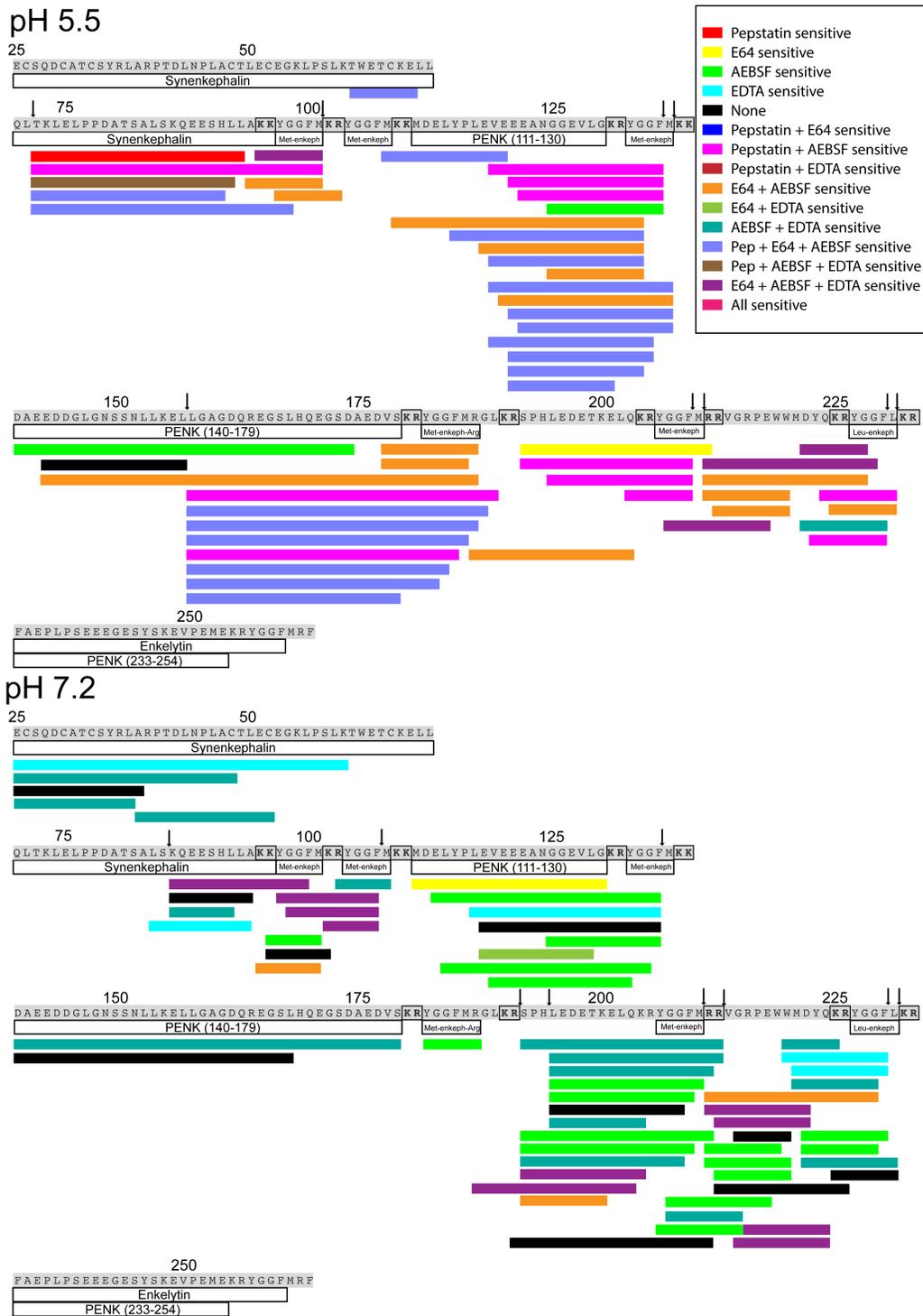
Figure 3.11. Peptide alignment maps of proenkephalin at pH 5.5 and pH 7.2. Newly formed peptides in peptide degradation assays that were derived from proenkephalin (PENK) were mapped to the sequence of PENK and illustrated by colored boxes. Colors indicate which inhibitors inhibit the formation of the peptides. Functional neuropeptides are shown in white boxes with annotations and dibasic cleavage sites are highlighted. Arrows show putative cleavage sites.

D. Conclusions

Neuropeptides are important macromolecules which are associated numerous biological processes and their production and degradation are highly regulated by proteases. Here, we demonstrated that the studies of using a combination of proteomics, peptidomics and substrate profiling methods to generate identify and quantify proteases, their endogenous substrates and cleavage products. To assist in deconvolution of sample complexity, class-specific protease inhibitors were introduced in protein/peptide degradation assays. Results of bioinformatic analysis of the data indicate 6 active proteases, cathepsin A, B, C, D, L and carboxypeptidase E, are present in CGs and may be responsible for processing proneuropeptides. To further demonstrate our method, we performed detailed analysis of proenkephalin processing. We identified multiple canonical dibasic cleavage sites as well as endopeptidase cleavages which were likely to be results of cathepsin D activity. We also observed processive trimming of several peptides from both N and C termini, which are results of exopeptidase activities. These exopeptidases are prevalent in secretory vesicles and play important roles in processing neuropeptide-intermediates into final forms of mature neuropeptides after cleavages of endopeptidases.

This multi-omics platform has been demonstrated as a powerful tool in studying protease-substrate interactions. It allowed us to uncover proteolytic repertoire on each proneuropeptide and associate protease activities with proneuropeptide processing. Of course, though we proved several proteases are capable of cleaving proneuropeptides at specific sites, these proteases are not necessarily the ones that are responsible for the cleavages. To fully connect proteolytic activities to proneuropeptide cleavages, further studies with more specific protease inhibitors, immuno-depletion or gene knockout will be needed. However, the integration of proteomics,

peptidomics and MSP-MS data enabled us to identify 6 candidate proteases from a list of 64 proteases, which significantly reduced the amount of work in downstream validation. The platform can be further employed to study not only proneuropeptide processing, but also many other protease-substrate networks. Our ultimate objective is to apply this strategy to decipher the molecular mechanisms of protease-associated diseases and develop protease inhibitors and protease-activated drugs for the treatment of diseases.

E. Acknowledgement

F. References

1.   Strand, F. L.; Rose, K. J.; Zuccarelli, L. A.; Kume, J.; Alves, S. E.; Antonawich, F. J.; Garrett, L. Y., Neuropeptide hormones as neurotrophic factors. *Physiol Rev* **1991,** *71* (4), 1017-46.

2.   Kastin, A. J.; Pan, W., Concepts for biologically active peptides. *Curr Pharm Des* **2010,** *16* (30), 3390-400.

3.   van den Pol, A. N., Neuropeptide transmission in brain circuits. *Neuron* **2012,** *76* (1), 98-115.

4.   Grimmelikhuijzen, C. J.; Hauser, F., Mini-review: the evolution of neuropeptide signaling. *Regul Pept* **2012,** *177 Suppl*, S6-9.

5.   Hallberg, M., Neuropeptides: metabolism to bioactive fragments and the pharmacology of their receptors. *Med Res Rev* **2015,** *35* (3), 464-519.

6.   Hook, V.; Lietz, C. B.; Podvin, S.; Cajka, T.; Fiehn, O., Diversity of Neuropeptide Cell-Cell Signaling Molecules Generated by Proteolytic Processing Revealed by Neuropeptidomics Mass Spectrometry. *J Am Soc Mass Spectrom* **2018,** *29* (5), 807-816.

7.   Hook, V.; Funkelstein, L.; Lu, D.; Bark, S.; Wegrzyn, J.; Hwang, S. R., Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. *Annu Rev Pharmacol Toxicol* **2008,** *48*, 393-423.

8.   Lazure, C.; Seidah, N. G.; Pelaprat, D.; Chretien, M., Proteases and posttranslational processing of prohormones: a review. *Can J Biochem Cell Biol* **1983,** *61* (7), 501-15.

9.   Li, J. Y.; Dahlstrom, A., Axonal transport of neuropeptides: Retrograde tracing study in live cell cultures of rat sympathetic cervical ganglia. *J Neurosci Res* **2007,** *85* (12), 2538-45.

10.   Sasaki, K.; Satomi, Y.; Takao, T.; Minamino, N., Snapshot peptidomics of the regulated secretory pathway. *Mol Cell Proteomics* **2009,** *8* (7), 1638-47.

11.   Wegrzyn, J. L.; Bark, S. J.; Funkelstein, L.; Mosier, C.; Yap, A.; Kazemi-Esfarjani, P.; La Spada, A. R.; Sigurdson, C.; O'Connor, D. T.; Hook, V., Proteomics of dense core secretory vesicles reveal distinct protein categories for secretion of neuroeffectors for cell-cell communication. *J Proteome Res* **2010,** *9* (10), 5002-24.

12.   Gupta, N.; Bark, S. J.; Lu, W. D.; Taupenot, L.; O'Connor, D. T.; Pevzner, P.; Hook, V., Mass spectrometry-based neuropeptidomics of secretory vesicles from human adrenal medullary pheochromocytoma reveals novel peptide products of prohormone processing. *J Proteome Res* **2010,** *9* (10), 5065-75.

13.   Fried, G.; Wikstrom, L. M.; Franck, J.; Rokaeus, A., Galanin and neuropeptide Y in chromaffin granules from the guinea-pig. *Acta Physiol Scand* **1991,** *142* (4), 487-93.

14.   Gozes, I.; O'Connor, D. T.; Bloom, F. E., A possible high molecular weight precursor to vasoactive intestinal polypeptide sequestered into pheochromocytoma chromaffin granules. *Regul Pept* **1983,** *6* (2), 111-9.

15.   Taupenot, L.; Harper, K. L.; O'Connor, D. T., The chromogranin-secretogranin family. *N Engl J Med* **2003,** *348* (12), 1134-49.

16.   Kirchmair, R.; Hogue-Angeletti, R.; Gutierrez, J.; Fischer-Colbrie, R.; Winkler, H., Secretoneurin--a neuropeptide generated in brain, adrenal medulla and other endocrine tissues by proteolytic processing of secretogranin II (chromogranin C). *Neuroscience* **1993,** *53* (2), 359-65.

17.   Soszynski, D.; Metz-Boutigue, M. H.; Aunis, D.; Bader, M. F., Secretogranin II: regulation of synthesis and post-translational proteolysis in bovine adrenal chromaffin cells. *J Neuroendocrinol* **1993,** *5* (6), 655-62.

18.   Fischer-Colbrie, R.; Frischenschlager, I., Immunological characterization of secretory proteins of chromaffin granules: chromogranins A, chromogranins B, and enkephalin-containing peptides. *J Neurochem* **1985,** *44* (6), 1854-61.

19.   Kuramoto, H.;   Kondo, H.; Fujita, T., Neuropeptide tyrosine (NPY)-like immunoreactivity in adrenal chromaffin cells and intraadrenal nerve fibers of rats. *Anat Rec* **1986,** *214* (3), 321-8.

20.   Hwang, S. R.; O'Neill, A.; Bark, S.; Foulon, T.; Hook, V., Secretory vesicle aminopeptidase B related to neuropeptide processing: molecular identification and subcellular localization to enkephalin- and NPY-containing chromaffin granules. *J Neurochem* **2007,** *100* (5), 1340-50.

21.   Hook, V. Y.; Schiller, M. R.; Azaryan, A. V.; Tezapsidis, N., Proenkephalin-processing enzymes in chromaffin granules: model for neuropeptide biosynthesis. *Ann N Y Acad Sci* **1996,** *780*, 121-33.

22.   Parmer, R. J.; Miles, L. A.; Xi, X. P.; Gill, B. M.; Wu, H. J.; O'Connor, D. T., Processing of chromaffin granule proteins: a profusion of proteases? *Neurochem Int* **1993,** *22* (4), 361-7.

23.   Azaryan, A. V.; Krieger, T. J.; Hook, V. Y., Purification and characteristics of the candidate prohormone processing proteases PC2 and PC1/3 from bovine adrenal medulla chromaffin granules. *J Biol Chem* **1995,** *270* (14), 8201-8.

24.   Wolkersdorfer, M.; Laslop, A.; Lazure, C.; Fischer-Colbrie, R.; Winkler, H., Processing of chromogranins in chromaffin cell culture: effects of reserpine and alpha-methyl-p-tyrosine. *Biochem J* **1996,** *316 (Pt 3),* 953-8.

25.   Doucet, A.; Butler, G. S.; Rodriguez, D.; Prudova, A.; Overall, C. M., Metadegradomics: toward in vivo quantitative degradomics of proteolytic post-translational modifications of the cancer proteome. *Mol Cell Proteomics* **2008,** *7* (10), 1925-51.

26.   Menschaert, G.; Vandekerckhove, T. T.; Baggerman, G.; Schoofs, L.; Luyten, W.; Van Criekinge, W., Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J Proteome Res* **2010,** *9* (5), 2051-61.

27.   O'Donoghue, A. J.; Eroy-Reveles, A. A.; Knudsen, G. M.; Ingram, J.; Zhou, M.; Statnekov, J. B.; Greninger, A. L.; Hostetter, D. R.; Qu, G.; Maltby, D. A.; Anderson, M. O.; Derisi, J. L.; McKerrow, J. H.; Burlingame, A. L.; Craik, C. S., Global identification of peptidase specificity by multiplex substrate profiling. *Nat Methods* **2012,** *9* (11), 1095-100.

28.   Smith, A. D.; Winkler, H., A simple method for the isolation of adrenal chromaffin granules on a large scale. *Biochem J* **1967,** *103* (2), 480-2.

29.   Hook, V. Y.; Eiden, L. E., Two peptidases that convert 125I-Lys-Arg-(Met)enkephalin and 125I-(Met)enkephalin-Arg6, respectively, to 125I-(Met)enkephalin in bovine adrenal medullary chromaffin granules. *FEBS Lett* **1984,** *172* (2), 212-8.

30.   Prusiner, S. B., Shattuck lecture--neurodegenerative diseases and prions. *N Engl J Med* **2001,** *344* (20), 1516-26.

31.   Gratzl, M.; Krieger-Brauer, H.; Ekerdt, R., Latent acetylcholinesterase in secretory vesicles isolated from adrenal medulla. *Biochim Biophys Acta* **1981,** *649* (2), 355-66.

32.   Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2007,** *2* (8), 1896-906.

33.   Rawlings, N. D.; Barrett, A. J.; Bateman, A., MEROPS: the peptidase database. *Nucleic Acids Res* **2010,** *38* (Database issue), D227-33.

34.   Kim, Y.; Bark, S.; Hook, V.; Bandeira, N., NeuroPedia: neuropeptide database and spectral library. *Bioinformatics* **2011,** *27* (19), 2772-3.

35.   Lapek, J. D., Jr.; Jiang, Z.; Wozniak, J. M.; Arutyunova, E.; Wang, S. C.; Lemieux, M. J.; Gonzalez, D. J.; O'Donoghue, A. J., Quantitative Multiplex Substrate Profiling of Peptidases by Mass Spectrometry. *Mol Cell Proteomics* **2019**.

36.   Xiao, Y.; Hsiao, T. H.; Suresh, U.; Chen, H. I.; Wu, X.; Wolf, S. E.; Chen, Y., A novel significance score for gene selection and ranking. *Bioinformatics* **2014,** *30* (6), 801-7.

37.   Manguy, J.; Jehl, P.; Dillon, E. T.; Davey, N. E.; Shields, D. C.; Holton, T. A., Peptigram: A Web-Based Application for Peptidomics Data Visualization. *J Proteome Res* **2017,** *16* (2), 712-719.

38.   Roncase, E. J.; Moon, C.; Chatterjee, S.; Gonzalez-Paez, G. E.; Craik, C. S.; O'Donoghue, A. J.; Wolan, D. W., Substrate Profiling and High Resolution Co-complex Crystal Structure of a Secreted C11 Protease Conserved across Commensal Bacteria. *ACS Chem Biol* **2017,** *12* (6), 1556-1565.

39.   Xu, J. H.; Jiang, Z.; Solania, A.; Chatterjee, S.; Suzuki, B.; Lietz, C. B.; Hook, V. Y. H.; O'Donoghue, A. J.; Wolan, D. W., A Commensal Dipeptidyl Aminopeptidase with Specificity for N-Terminal Glycine Degrades Human-Produced Antimicrobial Peptides in Vitro. *ACS Chem Biol* **2018,** *13* (9), 2513-2521.

40.   O'Donoghue, A. J.; Jin, Y.; Knudsen, G. M.; Perera, N. C.; Jenne, D. E.; Murphy, J. E.; Craik, C. S.; Hermiston, T. W., Global substrate profiling of proteases in human neutrophil extracellular traps reveals consensus motif predominantly contributed by elastase. *PLoS One* **2013,** *8* (9), e75141.

41.   Ivry, S. L.; Sharib, J. M.; Dominguez, D. A.; Roy, N.; Hatcher, S. E.; Yip-Schneider, M. T.; Schmidt, C. M.; Brand, R. E.; Park, W. G.; Hebrok, M.; Kim, G. E.; O'Donoghue, A. J.; Kirkwood, K. S.; Craik, C. S., Global Protease Activity Profiling Provides Differential Diagnosis of Pancreatic Cysts. *Clin Cancer Res* **2017,** *23* (16), 4865-4874.

Chapter IV

Quantitative Multiplex Substrate Profiling of

Proteases by Mass Spectrometry

A. Introduction

Proteases play an important role in maintaining cellular health. These enzymes generally catalyze irreversible reactions and thus it is imperative that their activity is tightly controlled. Unregulated proteolysis results in protein and peptide substrates being cleaved at faster rates or at unusual sites. In human disease, an increase in proteolytic activity may be due to upregulation or mis-localization of proteases or down-regulation of endogenous inhibitors [1-3]. During infection, invading microbes or host immune response can also cause an increase in proteolytic activity. Irrespective of the source of the activity, it is important to understand the substrate preference of disease associated proteases as it will facilitate the development of tools that can selectively monitor, enhance, or inhibit their activities. Many assays have been developed to reveal the substrate specificity of proteases, including the use of synthetic fluorogenic peptides [4-7], bacterial and phage displayed peptides [8, 9], and proteome derived proteins and peptides [10-14]. Peptide display technologies are capable of generating vast numbers of diverse substrate sequences, however mapping the exact cleavage locus is not trivial and requires downstream mutagenesis or mass spectrometry studies. Technologies that use cellular proteins or peptides as substrates generally require enrichment steps to separate cleaved from uncleaved proteins [15, 16]. In addition, the concentration of each substrate in the protein extract can differ greatly, making it difficult to compare the cleavage rate between substrates. These technologies are semi-quantitative and limited in their ability to distinguish the cleavage efficiency of each substrate. Fluorescent methods, such as positional scanning synthetic combinatorial libraries [4, 17, 18], are generally more quantitative and can take advantage of natural and non-natural amino acids. However, these substrates lack extended amino acid sequence on the carboxy-terminal side of

the scissile bond, and the location of the fluorophore adjacent to the scissile bond can impede enzyme-substrate binding interactions.

Our group previously developed a peptide degradation assay that uses tandem mass spectrometry to identify cleavage products within an equimolar mixture of tetradecapeptide substrates [19]. These peptides were rationally designed such that all possible neighbor and near-neighbor amino acid pairs are present in the library and hydrolysis of any of the 2,964 peptide bonds can be detected by tandem mass spectrometry. In addition, every substrate has a unique dipeptide sequence on the amino and carboxyl termini to facilitate the characterization of amino- and carboxy-protease substrate specificity. The amino acid sequences of the synthetic substrates are distinct from endogenous proteins and peptides; therefore, when complex biological samples are assayed with this library, cleaved products are easily discriminated from endogenous peptides. This assay has been used to uncover the substrate specificity profile of enzymes from every protease family, and the data generated from these assays has guided the development of optimized fluorescent reporters [20, 21], peptide inhibitors [22, 23], activity-based probes [24], therapeutic peptides [25] and activity based biomarkers [26]. All peptide substrates are present at equimolar concentration and this assay identifies cleavage products that appear at defined time-intervals following addition of enzyme.

In this study, we utilized isobaric tandem mass tags (TMT) [27, 28] to quantify all cleaved and uncleaved peptides within the tetradecapeptide substrate library following addition of a protease or a complex sample containing multiple proteases. The reaction was incubated for up to 20 hours and different TMTs were used to label each of the assay time-points. Changes in peptide abundance were quantified and proteolytic kinetic constants were calculated. In addition to providing quantification, TMT labeling can also minimize experimental and instrument

variation while allowing us to gain crucial temporal information. This combination greatly improves the reproducibility and accuracy, while also allowing for a reduction in instrument time by 10-fold. We have named this approach Quantitative Multiplex Substrate Profiling by Mass Spectrometry (qMSP-MS).

First, we used a well characterized protease, papain, to validate the qMSP-MS assay. In addition, the substrate specificity of two integral membrane proteases were characterized, and a fluorescent reporter was developed based on the preferred peptide substrate, which can be used for high-throughput biochemical assays. Finally, we analyzed lung cancer cell secretions using our workflow and found that they generally produced exopeptidases that can degrade the peptides at the amino and carboxyl termini. These results highlight the potential of applying quantitative mass spectrometry to proteolytic assays which can lead to the development of novel therapeutic delivery systems for important human diseases.

B. Materials and Methods

1. Experimental design and statistical rationale

A total of 144 samples were analyzed in this study using the qMSP-MS assay, corresponding to 12 different protease samples; papain, PsAarA, HiGlpG, *Aspergillus phoenicis* extract and secretions from 8 lung cancer cell lines (Table 4.1). For each protease reaction, time points were taken after 0.25, 1, 4 and 20 hours of incubation. Papain, *Aspergillus phoenicis* extract, PsAarA and HiGlpG samples were assayed in triplicate enzyme reactions while triplicate lung cancer secretion samples were generated from three separate culture flasks. A total of 27 samples were analyzed by proteomics corresponding to secretions from 8 lung cancer cell lines and *Aspergillus phoenicis* extract in triplicate. Replicate lung cancer cell lines were labeled with

different TMTs and assayed via mass spectrometry. The *Aspergillus phoenicis* extract was not labelled with TMT. Significance was assessed by ANOVA and Student's T-test; variance was assessed by an F-test to ensure the correct statistical assumptions were used. q-values of q≤0.05 were considered significant.

Table 4.1. Two cell lines from each subtype of lung cancer were used in the study.

Cell line	Disease	Gender
BEN	Squamous carcinoma	Male
H520	Squamous carcinoma	Male
H460	Large cell carcinoma	Male
H661	Large cell carcinoma	Male
H1944	Adenocarcinoma	Female
H2228	Adenocarcinoma	Female
DMS273	Small cell carcinoma	Female
SHP77	Small cell carcinoma	Male

2. Expression and purification of rhomboid proteins

The purification of all rhomboid proteins was similar to the previous report [29]. Briefly, rhomboid genes were cloned into pBAD-Myc/HisA plasmid (Invitrogen), having a C-terminal tobacco etch virus (TEV) protease cleavage site, Myc-epitope and His6-tag, were expressed in TOP10 *E. coli* cells. The protein was induced with 0.002% w/v arabinose and expressed at 24 °C for 8 hours in LB media. The cells were harvested, resuspended in 50 mM Tris-HCl pH 8.0, 150 mM NaCl and lysed under high pressure using an EmulsiFlex-C3 (Avestin). The membranes were isolated by ultracentrifugation at 95,800 *g* for 2 hours, solubilized in 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 10 mM imidazole, 20% glycerol, 1% (w/v) DDM and applied onto a Ni-NTA column (Qiagen). The proteins were eluted with 250-500 mM of imidazole, 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20% glycerol, 0.1% DDM. From 1L of cell culture, purified protein

yield was 1-2 mg for PsAarA and 2-3 mg for HiGlpG. The His-tag was removed by TEV protease (1 mg per 100 mg of protein, overnight, 16 °C) and a subsequent Ni-NTA column was performed to remove uncleaved protein and TEV protease. The flow-through was collected and concentrated using 30,000 MWCO concentrators (Millipore). The protein samples (Figure 4.1) were flash-frozen and stored at -80 °C prior to analysis.
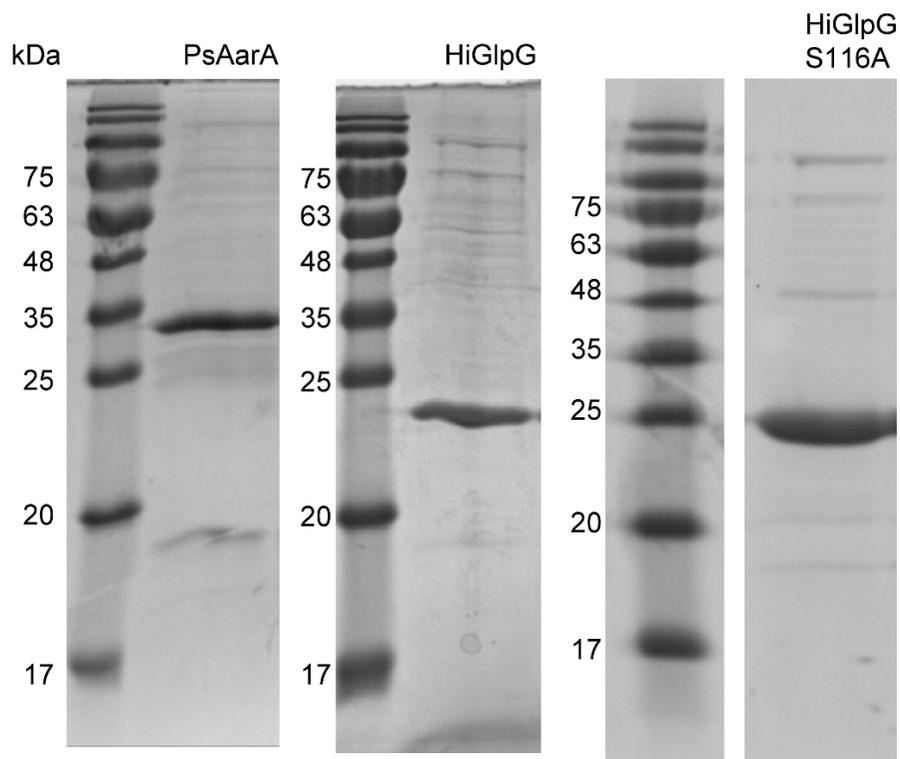
Figure 4.1. SDS-PAGE gel images of rhomboid proteases PsAarA, HiGlpG and HiGlpG-S116A (inactive mutant).

3. Collection of lung cancer secretions

Lung cancer cell lines BEN, H520, H2228, H460, H661, DMS273, and SHP77 were cultured in RPMI-1640 (Corning) supplemented with 10% fetal bovine serum (Corning) and 100 U/mL penicillin-streptomycin (HyClone). The cell line H1944 was cultured in RPMI-1640 supplemented with 10% fetal bovine serum, 1.5 g/L $NaHCO_3$ (HyClone), 4.5 g/L glucose, 10 mM HEPES (GE Healthcare Life Sciences), 1 mM sodium pyruvate (HyClone) and 100 U/mL penicillin-streptomycin (HyClone). All cells were maintained at 37 °C in an atmosphere of 5% CO_2 and grown to ~80% confluence in triplicate T-175 flasks. Culture media was removed, and cells were washed twice with Dulbecco's PBS (Thermo) and twice with RPMI-1640. Cells were then incubated with serum-free RPMI-1640 for 24 hours. The conditioned media was removed, filtered (0.22 μm) and subsequently buffer-exchanged and concentrated into Dulbecco's PBS using an Amicon Ultra centrifugal filter (EMD Millipore) with 10-kDa cutoff. Protein concentration was determined by BCA assay (Thermo).

4. Protein identification of *Aspergillus phoenicis* extract

200 μg of *Aspph* extract (Sigma, P2143) was incubated with 1% sodium deoxycholate (Thermo), 10 mM tris(2-carboxyethl)phosphine (Sigma), 40 mM chloroacetamide (Sigma), and 100 mM Tris pH 8.0, (Research Products International) at 90 °C for 10 min. Samples were cooled to room temperature and diluted 2× in 100 mM Tris (pH 8.0). Trypsin (sequencing grade, Promega V5113) was added at 1:50 trypsin:total protein for digestion overnight at 37 °C. Reactions were quenched by adding 10% TFA to bring the pH down to ~2 and desalted using C18 LTS tips (Rainin). 2 μg of the extracted peptides were analyzed on a Q Exactive Mass Spectrometer (Thermo) equipped with an Ultimate 3000 HPLC (Thermo). Peptides were

separated by reverse phase chromatography on a C_{18} column (1.7 μm bead size, 75 μm x 20 cm, heated to 65 °C) at a flow rate of 300 nL/min using a 56-minute linear gradient from 4% B to 17% B followed by a 20-minute gradient from 17% B to 25% B, with solvent A: 0.1% formic acid (Thermo) in water and solvent B: 0.1% formic acid in acetonitrile (Thermo). Survey scans were recorded over a 350–1200 m/z range at 35,000 resolution at 200 m/z. MS/MS was performed in data-dependent acquisition mode with HCD fragmentation (28 normalized collision energy) on the 20 most intense precursor ions at 17,500 resolution at 200 m/z. Data was processed using PEAKS 8.5 (Bioinformatics Solutions Inc.). MS^2 data were searched against *Aspergillus phoenicis* proteome and annotated by InterPro (JGI Project ID: 1020378). Fixed modifications of carbamidomethylation of cysteines (+57.02146 Da), variable modification of acetylation of protein N-termini (+42.0106) and oxidation of methionine (+15.99492 Da) were specified. A precursor tolerance of 20 ppm and 0.01 Da for MS^2 fragments was defined. Data were filtered to 1% peptide and protein level false discovery rates with the target-decoy strategy.

5. Protein identification and quantification in lung cancer secretions

Secreted proteins were denatured in a buffer containing 75 mM NaCl (Sigma), 3% sodium dodecyl sulfate (SDS, Thermo), 1 mM NaF (Sigma), 1 mM β-glycerophosphate (Sigma), 1 mM sodium orthovanadate (Sigma), 10 mM sodium pyrophosphate (Sigma), 1 mM phenylmethylsulfonyl fluoride (Sigma), and 1X Complete Mini EDTA free protease inhibitors (Roche) in 50 mM HEPES (Sigma), pH 8.5 [30]. Insoluble debris was pelleted by centrifugation for 5 minutes at 18,000 x g at 22 °C. Supernatants were transferred into fresh tubes, and an equal volume of 8 M urea (Thermo) in 50 mM HEPES, pH 8.5 was added to each sample. Samples

were then vortexed briefly and sonicated for 5 minutes in a sonicating water bath to maximize protein denaturation.

Proteins were reduced and alkylated as previously described [31]. Proteins were then precipitated via a methanol-chloroform procedure [32]. Precipitated proteins were re-solubilized in 300 μL of 1 M urea (Thermo) in 50 mM HEPES (Thermo), pH 8.5. Solubility was aided through vortexing, sonicating, and manual grinding. Proteins were then digested in a two-stage process. First, 3 μg of LysC (Wako 129-02541) was added to each sample and allowed to incubate overnight at room temperature. Next, 3 μg of trypsin (Promega V5113) was added, and samples were allowed to digest for six hours at 37 °C. Digestion was quenched by addition of trifluoroacetic acid (TFA, Pierce). Peptides were desalted with C18 Sep-Paks (Waters) as previously described [33]. Peptide concentration was determined with a PepQuant Assay (Thermo), and peptides were aliquoted into 50 μg portions which were dried under vacuum and stored at -80 °C until they were labeled with TMT reagents.

Peptides were labeled with 10-plex TMT reagents (Thermo) [27, 28] as previously described [34]. TMTs were reconstituted at a concentration of 20 μg/μL in dry acetonitrile (Sigma). Dried peptides were reconstituted in 30% dry acetonitrile in 200 mM HEPES, pH 8.5, and 8 μL of the appropriate TMT reagent was added to peptides. Reagents 126 and 131 were used to bridge between mass spectrometry runs [35] while the remaining TMT reagents were used to label samples as listed in Table 4.2. Labeling was carried out for one hour at room temperature and was quenched by adding 9 μL of 5% hydroxylamine (Sigma). Samples were acidified with 50 μL of 1% TFA then pooled into appropriate 10-plex TMT samples, with pooled standard samples labeled with 126 and 131. Pooled 10-plex samples were desalted with C18 Sep-Paks.

Samples were separated into eight fractions by basic pH reverse-phase chromatography using spin columns (Pierce). Fractions were dried then reconstituted in 5% formic acid/5% acetonitrile and analyzed on an Orbitrap Fusion Tribrid Mass Spectrometer (Thermo) equipped with an EASY-nLC 1000 (Thermo) for identification and quantitation. MS^2/MS^3 analysis for identification was carried out with chromatographic and mass spectrometry acquisition settings as previously defined [36, 37]. Briefly, peptides were separated on a 100 μm ID x 30 cm home-pulled home-packed column (0.5 cm C_4 5μm, 100Å and 29.5 cm C_{18} 1.8 μm, 120Å). A linear gradient from 11 to 30% Acetonitrile in 0.135% formic acid over 165 min at a flow rate of 300 nL/min with the column heated to 60°C was used for separation. The mass spectrometer was operated in data dependent mode, with a survey scan across the mass to charge range of 500-1,200 at 120,000 resolution in the Orbitrap. Automatic Gain Control (AGC) target was set to 5 x 10^5 with a max ion time of 100 ms. The s-lens was set at an RF of 60. Top speed mode was used to select the most abundant ions for MS^2/MS^3 in a 5 s experimental cycle. For MS^2 analysis, precursors were isolated with the quadrupole at a width of 0.5 m/z. CID normalized energy was 30% and fragments were detected in the ion trap at rapid scan rate. AGC was set to 1 x 10^4 with a max ion time of 35 ms. For MS^3 analysis, SPS was used with a maximum of 10 ions isolated for MS^3 analysis. These ions were fragmented with HCD at a normalized energy of 50% and detected in the Orbitrap at 60,000 resolution with a low mass of 110 m/z. AGC was set to 5 x 10^4 with a maximum ion time of 150 ms. MS^2 precursors were excluded in a range of 40 m/z below and 15 m/z above the MS^1 precursor. All data (MS^1, MS^2, MS^3) were centroided. Data were processed using Proteome Discoverer 2.1 (Thermo). MS^2 data were searched against Uniprot Human (05/11/2017) using the Sequest algorithm [38]. A decoy search was also conducted with sequences in reverse order [39-41]. A precursor tolerance of 50 ppm [42, 43] and 0.6 Da for MS^2

fragments was defined. Static modifications of TMT 10-plex tags on lysine and peptide N-termini (+229.162932 Da) and carbamidomethylation of cysteines (+57.02146 Da) were specified. Variable oxidation of methionine (+15.99492 Da) was included in the search parameters as well. A maximum of two missed cleavages of trypsin was allowed. Data were filtered to 1% peptide and protein level false discovery rates with the target-decoy strategy through Percolator [44, 45]. TMT reporter ion intensities were extracted from MS^3 spectra for quantitative analysis, and signal-to-noise ratios were used for quantitation. Spectra were filtered, summed, and normalized as previously described [35, 37].

Table 4.2. Schematic view of TMT labeling of lung cancer proteomic samples. The TMT used to label each sample was indicated in the header. All samples in the same row were combined and subjected to one injection.

126	127N	127C	128N	128C	129N	129C	130N	130C	131
Bridge	H661	SHP77	DMS273	H520	H1944	BEN	H460	H2228	Bridge
Bridge	H2228	H661	SHP77	DMS273	H520	H1944	BEN	H460	Bridge
Bridge	H460	H2228	H661	SHP77	DMS273	H520	H1944	BEN	Bridge

6. Quantitative multiplex substrate profiling

All qMSP-MS protease assays were performed in triplicate using an equimolar mixture of 228-synthetic tetradecapeptides. The final concentration of each peptide was 0.5 µM. Papain (3.3 µg/mL, Sigma, P4762) and *Aspph* extract (5 µg/mL, Sigma, P2143) were incubated with the peptide mixture in 20 mM Citrate-Phosphate buffer, pH 5.0, 100 mM NaCl, 1 mM CaCl$_2$ and 2 mM DTT. Membrane proteases PsAarA (175 µg/mL) and HiGlpG (5 µg/mL) were assayed in 20 mM Citrate-Phosphate buffer, pH 6.0, 150 mM NaCl, 2 mM DTT, 20% Glycerol, 0.1% n-

Dodecyl β-D-maltoside (DDM). All lung cancer cell line secretions were assayed at 160 μg/mL of total protein in 20 mM Citrate-Phosphate buffer, pH 6.5, 100 mM NaCl, 1 mM CaCl$_2$, 2 mM DTT at 37 °C. For each assay, 10% of the reaction mixture was removed after 0.25, 1, 4, and 20 hours of incubation. Enzyme activity was quenched by addition of GuHCl to a final concentration of 6.4 M and samples were immediately stored at -80 °C. Control samples consisted of enzyme incubated with 6.4 M GuHCl prior to addition of the peptide library. All samples were desalted with STAGE tips as previously described [46] and dried under vacuum.

Samples were reconstituted in 55 μL of 30% dry acetonitrile in 200 mM HEPES, pH 8.5 and assembled into groups of nine. 5 μL from each sample was removed and combined with samples from within the group (45 μL total in each tube + 5 μL 30% Dry acetonitrile in 200 mM HEPES, pH 8.5). A unique TMT reagent was used to label peptides in each tube as outlined in Table 4.3. The tube containing a mixture of samples was labeled with TMT-131 and used to generate bridge channels for inter-run normalization as described previously [35]. Labeled samples were combined into appropriate 10-plex experiments and desalted with Sep-Paks. 10-plex experiments were dried and stored at -80 °C until mass spectrometry analysis. Samples were reconstituted in 8 μL of 5% formic acid in 5% acetonitrile and analyzed on an Orbitrap Lumos Tribrid Mass Spectrometer (Thermo) equipped with an EASY-nLC 1000 (Thermo) with 3 μL of sample analyzed for each 10-plex. Acquisition parameters were the same as those outlined above for lung cancer secretion proteomics. Data were processed using Proteome Discoverer 2.1 (Thermo) as outlined above with the exception that MS^2 data were searched against the 228 synthetic peptides library with no enzyme digestion specified and a decoy search was conducted with peptide sequences in reverse order [39-41]. Data were filtered to 1% peptide level false discovery rates with the target-decoy strategy.

Table 4.3. Schematic view of TMT labeling of qMSP-MS samples. The TMT used to label each sample was indicated in the header. All samples in the same row were combined and subjected to one injection. Control samples consisted of enzyme incubated with 6.4 M GuHCl prior to addition of the peptide library. Other samples (in grey) were additional qMSP-MS samples that were not included in this manuscript.

126	127N	127C	128N	128C	129N	129C	130N	130C	131
Control 1	Papain 15	Papain 60	Papain 240	Papain 1200	Aspph 15	Aspph 60	Aspph 240	Aspph 1200	Bridge
Control 1	Aspph 1200	Papain 15	Papain 60	Papain 240	Papain 1200	Aspph 15	Aspph 60	Aspph 240	Bridge
Control 1	Aspph 240	Aspph 1200	Papain 15	Papain 60	Papain 240	Papain 1200	Aspph 15	Aspph 60	Bridge
Control 2	HiGlpG 15	HiGlpG 60	HiGlpG 240	HiGlpG 1200	Other_Sample	Other_Sample	Other_Sample	Other Sample	Bridge
Control 2	HiGlpG 15	HiGlpG 60	HiGlpG 240	HiGlpG 1200	PsAarA 15	PsAarA 60	PsAarA 240	PsAarA 1200	Bridge
Control 2	HiGlpG 15	HiGlpG 60	HiGlpG 240	HiGlpG 1200	Other_Sample	Other_Sample	Other_Sample	Other Sample	Bridge
Control 2	Other_Sample	Other_Sample	Other_Sample	Other_Sample	PsAarA 15	PsAarA 60	PsAarA 240	PsAarA 1200	Bridge
Control 2	PsAarA 15	PsAarA 60	PsAarA 240	PsAarA 1200	Other_Sample	Other_Sample	Other_Sample	Other Sample	Bridge
Control 3	BEN 15	BEN 60	BEN 240	BEN 1200	H520 15	H520 60	H520 240	H520 1200	Bridge
H520 1200	Control 3	BEN 15	BEN 60	BEN 240	BEN 1200	H520 15	H520 60	H520 240	Bridge
H520 240	H520 1200	Control 3	BEN 15	BEN 60	BEN 240	BEN 1200	H520 15	H520 60	Bridge
Control 4	H460 15	H460 60	H460 240	H460 1200	H661 15	H661 60	H661 240	H661 1200	Bridge
H661 1200	Control 4	H460 15	H460 60	H460 240	H460 1200	H661 15	H661 60	H661 240	Bridge
H661 240	H661 1200	Control 4	H460 15	H460 60	H460 240	H460 1200	H661 15	H661 60	Bridge
Control 5	H1944 15	H1944 60	H1944 240	H1944 1200	H2228 15	H2228 60	H2228 240	H2228 1200	Bridge
H2228 1200	Control 5	H1944 15	H1944 60	H1944 240	H1944 1200	H2228 15	H2228 60	H2228 240	Bridge
H2228 240	H2228 1200	Control 5	H1944 15	H1944 60	H1944 240	H1944 1200	H2228 15	H2228 60	Bridge
Control 6	DMS273 15	DMS273 60	DMS273 240	DMS273 1200	SHP77 15	SHP77 60	SHP77 240	SHP77 1200	Bridge
SHP77 1200	Control 6	DMS273 15	DMS273 60	DMS273 240	DMS273 1200	SHP77 15	SHP77 60	SHP77 240	Bridge
SHP77 240	SHP77 1200	Control 6	DMS273 15	DMS273 60	DMS273 240	DMS273 1200	SHP77 15	SHP77 60	Bridge

7. qMSP-MS Data Analysis

Proteolytic progress of each peptide was modeled by performing non-linear least-squares regression to pseudo-first-order enzymatic kinetics: $Y = (plateau-Y_0) \times (1-exp(-t \times k_{cat}/K_M \times [E_0])) + Y_0$, where E_0 is the total enzyme concentration. Proteolytic efficiency was reported as k_{cat}/K_M for purified enzymes or k ($k = k_{cat}/K_M \times [E_0]$) for samples where the concentration of protease is unknown. Non-linear fitting was performed on cleavage products only if peptides were detected in at least 2 of the 3 replicates and passed ANOVA test ($p<0.05$). Cleaved peptides were clustered using Jenks natural breaks optimization and peptides in the group with highest proteolytic efficiencies were used to generate specificity profiles. Peptides in Group 2 are slowly formed cleavage products and proteolysis did not go to completion within the 20-hour assay time. The catalytic efficiency values for these peptides are less accurate as their progression curves are non-parabolic (linear) and cannot be well fitted with pseudo-first-order enzymatic kinetics. Protease activity correlates with protease stability and therefore it is likely that enzyme activity decreases during the 20-hour assay. For exopeptidases activities, a Student's t-test was performed on cleavage products only if peptides were detected in at least 2 of the 3 replicates and were significant by ANOVA ($q < 0.05$). Peptides with a significant increase (t-test $q < 0.05$) of at least 2-fold compared to controls were considered significantly increased and used to generate specificity profiles. Specificity profiles of proteases were generated by iceLogo software [47] to visualize amino-acid frequency surrounding the cleavage sites. Amino acids that were most frequently observed (above axis) and least frequently observed (below axis) from P4 to P4′ positions were illustrated ($p < 0.3$). Norleucine (Nle) was represented as 'n' in the reported profiles. Amino acids in opaque black text were statistically significant ($p < 0.05$).

8. Fluorescent substrates screening

Real-time fluorescence measurements were performed using a Synergy HTX Multi-Mode Microplate Reader (BioTek). PsAarA (175 µg/mL), HiGlpG (100 µg/mL) and HiGlpG S116A (100 µg/mL) were assayed with 5 µM of IQ1 (Table 4.4) in 20 mM Citrate-Phosphate, pH 6.0, 150 mM NaCl, 2 mM DTT, 20% Glycerol, 0.1% DDM. *Aspph* extract was assayed with 5 µM of IQ2 (Table 4.4) in 20 mM Citrate-Phosphate buffer, pH 5.0, 100 mM NaCl, 1 mM $CaCl_2$, 2 mM DTT, 1 mM EDTA (MP Biomedicals), 10 µM E64 (Sigma), 1 mM AEBSF (Tocris) and +/− 1µM Pepstatin (MP Biomedicals). Fluorescence was read at Ex/Em= 320/400 nm. The cleavage sites of the fluorescent substrates were mapped by mass spectrometry as described previously [48].

Lung cancer secretions (160 µg/mL) were mixed with 5 µM AMC substrates (Table 4.4) in 20 mM Citrate-Phosphate, pH 6.5, 100 mM NaCl, 1 mM $CaCl_2$, 2 mM DTT at 37 °C. Fluorescence was read at Ex/Em= 360/460 nm.

Table 4.4. Peptide sequences of internally quenched and AMC fluorescent substrates. Mca: (7-methoxycoumarin-4-yl)acetyl, DNP: 2,4-dinitrophenol, AMC: 7-amino-4-methylcoumarin, n: Norleucine

Substrate	Protease	Sequence
IQ1	Rhomboids	Mca-VKLFRFNWMK(DNP)-NH$_2$
IQ2	*Aspph* extract	Mca-HQKLVFFAK(DNP)-NH$_2$
AMC1	Lung cancer secretion	Gn-AMC
AMC2	Lung cancer secretion	GR-AMC
AMC3	Lung cancer secretion	n-AMC
AMC4	Lung cancer secretion	R-AMC
AMC5	Lung cancer secretion	A-AMC

9. Data deposition

All mass spectrometry data have been deposited in MassIVE and proteomeXchange with accession numbers: MassIVE: MSV000082187; ProteomeXchange: PXD009269.

C. Results

1. Workflow of qMSP-MS

In our previous use of the substrate profiling assay, the peptide library was incubated with a protease and an aliquot of the reaction volume was removed at four time-intervals (T1 to T4), quenched with acid or denaturant, and analyzed by mass spectrometry. Although this assay has been successfully used to characterize many enzymes, the method is largely qualitative, where the time interval at which a cleavage product is first detected is correlated with cleavage efficiency. For example, a cleavage product discovered at T3 but absent in T1 and T2 was considered to be the result of a slower protease cleavage reaction than a product that is first detected at T1. In this study, peptide substrates are incubated with a protease, however, a unique TMT reagent is used to label cleaved and uncleaved peptides at each time-interval. Each labeled sample can then be combined into a single mass spectrometry experiment. This multiplex approach allows for accurate comparison of peptide abundance at each time-point so that kinetic parameters such as k_{cat}/K_M can be directly calculated for the time dependent accumulation of each cleavage product (Figure 4.2).
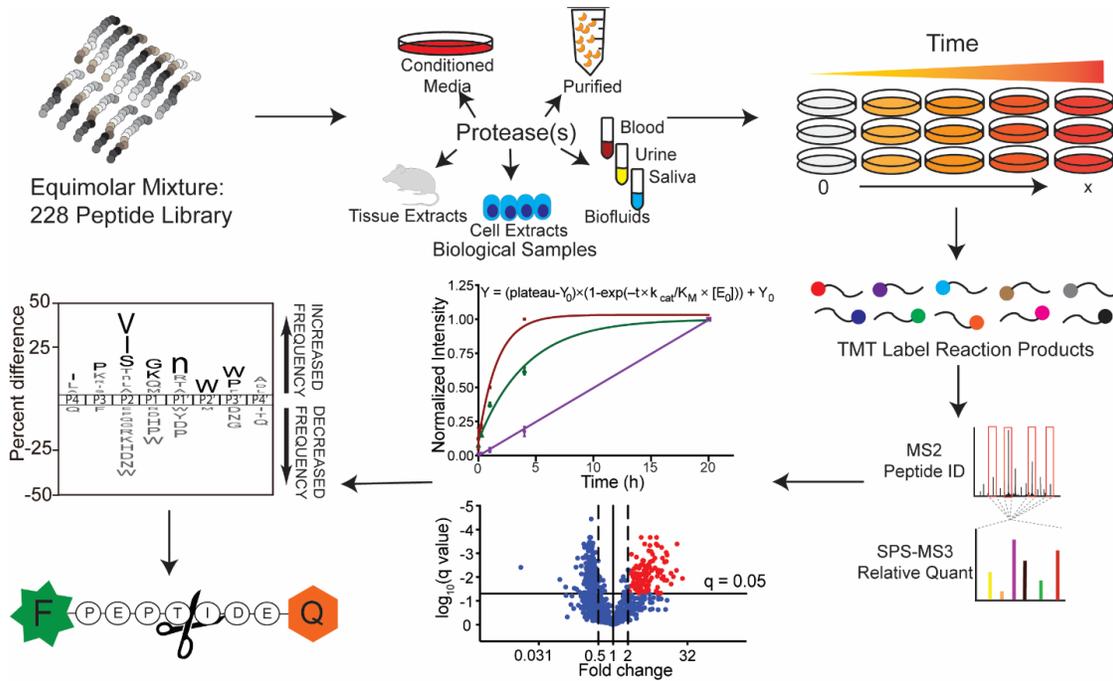
Figure 4.2. Workflow of qMSP-MS. An equimolar mixture of 228 tetradecapeptides is incubated with a purified protease or a biological sample containing multiple proteases. Aliquots of this assay are removed at multiple time intervals and the enzymes are inactivated with a denaturant. Quenched samples are labeled with specific TMTs and pooled prior to mass spectrometry analysis. Cleavage products are identified by MS/MS and quantified by TMT labels. Proteolytic progression curves or relative abundance changes of each cleavage product are calculated, and peptides are ranked by their proteolytic efficiency. Cleavage products with highest proteolytic efficiencies are selected to generate specificity profiles and design fluorescent substrates.

2. Validation of the method using papain

To validate this method, we incubated the 228-peptide library with papain, a prototypical cysteine protease isolated from papaya fruit. This enzyme is homologous to all C1 family enzymes, the predominant cysteine protease found in microbes, plants, and animals [49]. We detected cleavage at 275 of the 2,964 available peptide bonds within the tetradecapeptide library over the course of a 20-hour incubation. The use of TMT in this assay allowed us to determine cleavage efficiency for the formation of every peptide product. The proteolytic constant k_{cat}/K_M was calculated from progress curves using the first-order kinetics formula: $Y = (plateau-Y_0) \times (1-exp(-t \times k_{cat}/K_M \times [E_0])) + Y_0$. The k_{cat}/K_M values ranged from 0.041 M^{-1} s^{-1} to 1.34×10^6 M^{-1} s^{-1} and were separated into 2 groups using the Jenks natural breaks algorithm (Figure 4.3A). The progress curves for peptides in Group 1 consisting of the fastest cleaved substrates were parabolic and therefore the reaction was generally complete within 20 hours (Figure 4.3B). However, the progress curves for peptides in Group 2 (e.g. THATPGIHFVL*LRP) were linear and therefore these slow reactions were still progressing even after 20 hours incubation (Figure 4.3B). All cleaved products were matched to the parent 14-mer substrate in order to determine the cleavage location and identify the amino acids at both sides of the scissile bond. Papain hydrolyzed peptide bonds at each of the 13 sites within the substrates. However, cleavage of bonds between residues 1 to 3, and 12 to 14 occurred at lower frequency than at all other sites, indicating that papain is an endopeptidase and has low preference for hydrolysis of bonds near the termini of peptides (Figure 4.3C).

For each cleavage site found in this study, the amino acids in the P4 to P4′ positions were identified, and two distinct substrate specificity profiles were generated for papain representing the fastest and slowest substrates (Figure 4.3D). The profile for the fastest substrates revealed

that short aliphatic (Val, Ile, Leu, Ala) and hydroxylic (Ser, Thr) amino acids are preferred in the P2 position whileGly, Asn and charged amino acids such as Asp, Glu, Arg, Lys and His are not well tolerated. In the P3 position, papain has preference for Pro, Leu, Lys and Ile. Our results are consistent with previous studies [50, 51], and the P3 and P2 described here strongly correlate with the P3 (Spearman = 0.857) and P2 (Spearman = 0.755) specificity of papain when assayed with a combinatorial library of 160,000 fluorescent tetrapeptide substrates [52]. In the P1 position, Lys, Gly and Nle (norleucine; lowercase 'n' in iceLogo) are favored while Pro, His, Asp, Ile and Val are disfavored. This specificity weakly, but positively, correlates with the P1 preference found using the combinatorial library (Spearman = 0.44), most likely due to the position of the bulky fluorescent reporter molecule in the P1′ position, directly adjacent to the P1 amino acid. However, when internally quenched fluorescent substrates that do not have the fluorophore in the P1′ position were assayed with papain, Gly was found to be the preferred amino acid in the P1 position [53]. On the prime side, Nle and Trp are preferred at P1′ and P2′ positions, respectively.

We next generated a specificity profile of papain using the amino acids surrounding the cleavage sites found in Group 2 (Figure 4.3E). Peptides in Group 2 are cleaved slowly most likely due to one or more unfavorable amino acids present in the substrate sequence. In the P2 position, bulky hydrophobic amino acids, such as Tyr, Nle, Pro and Phe are found instead of the small aliphatic and hydroxyl amino acids present in the Group 1 profile. Asp in the P1′ was the most favored amino acid in Group 2 but this residue is strongly disfavored in Group 1. Likewise, Gln in P4′ was favored in Group 2 and disfavored in Group 1. At other positions many of the same residues are common between Group 2 and Group 1. These data show that papain has broad specificity and the rate of substrate cleavage is primarily driven by the amino acid in the

P2 position. In the absence of small aliphatic and hydroxyl amino acids, peptide hydrolysis does occur, but the rate is considerable slower than when these preferred amino acids are present.

Taken together, the qMSP-MS assay data allowed us to rank the papain cleavage events based on efficiency and generate a substrate specificity profile corresponding to the bonds that are hydrolyzed rapidly compared to bonds that are slowly hydrolyzed. In addition, we show that papain is an endopeptidase and therefore preferentially hydrolyzes peptide bonds distal to the amino and carboxyl termini. These results demonstrate that qMSP-MS provides important kinetic information for protease mediated hydrolysis of peptide bonds.
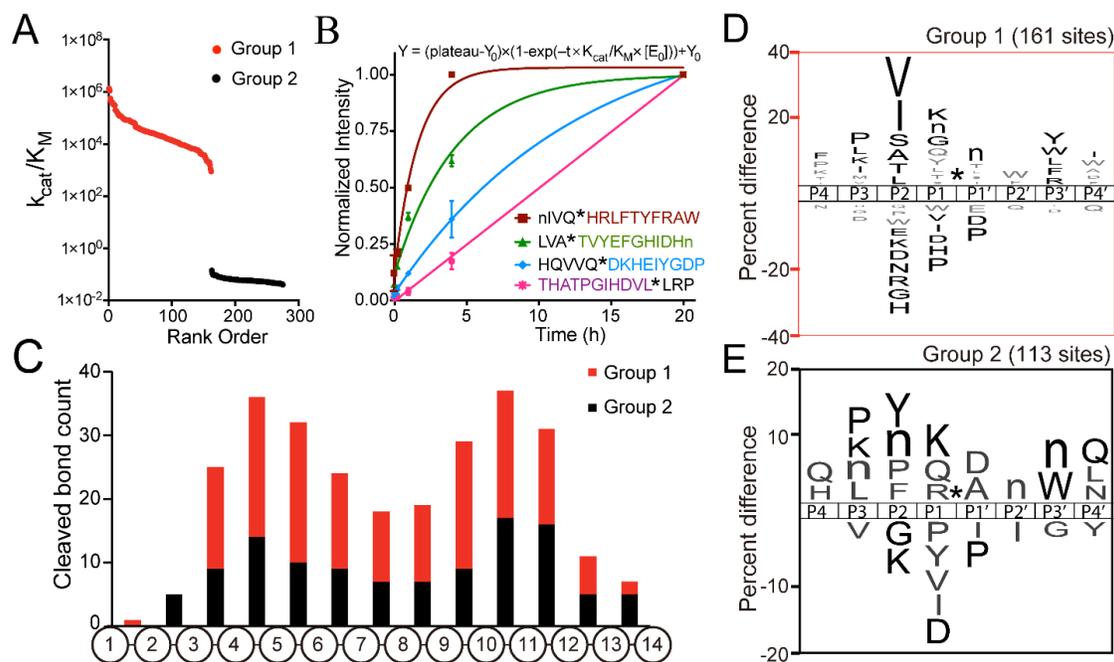
Figure 4.3. Quantitative multiplex substrate profiling of papain. A. Proteolytic constant k_{cat}/K_M of each cleaved peptide was calculated, ranked, and clustered into Group 1 (red) and Group 2 (black). B. Progress curves for sample cleavage products from Group 1 (Brown, Green and Blue) and Group 2 (Pink). The sequence of the full-length substrate is shown, and the cleaved product quantified in this assay is colored. C. Distribution of Group 1 and 2 cleavage sites within 14-mer peptides. D. A frequency plot showing the P4 to P4′ specificity profile of substrates that a rapidly cleaved by papain (Group1). E. A frequency plot showing the P4 to P4′ specificity profile of substrates that a slowly cleaved by papain (Group 2). Lowercase 'n' corresponds to norleucine and non-grayed residues have $p \leq 0.05$. Cleavage sites are indicated with *.

3. Uncovering the substrate specificity of membrane embedded intramembrane proteases

After validation of qMSP-MS assay with papain, we studied an intramembrane serine protease from *Providencia stuartii*, PsAarA. This enzyme has been defined as a bacterial rhomboid protease, and related enzymes are also found in mammalian cells. Rhomboid proteases are known to control many cellular functions and play important roles in human diseases [54], however, few studies have been conducted to elucidate their substrate preferences. PsAarA is known to cleave the transmembrane protein TatA between a pair of Ala residues in the following sequence, IATAAFGS. Strisovsky and colleagues expressed more than 130 mutant variants of the TatA protein substrate that have different amino acid sequence surrounding the scissile bond and discovered that the P4, P1 and P2′ positions were most important for substrate specificity [55]. We expressed and isolated PsAarA from *E. coli* (Supplemental Figure S1) as described previously [29] and incubated it with the 228-member peptide library for 0.25, 1, 4 and 20 hours. In total, cleavage products from 29 hydrolyzed peptide bonds were identified and progress curves were generated. The rate constants were calculated to range from 1.04×10^3 M^{-1}s^{-1} to 1.47×10^{-1} M^{-1}s^{-1} (Figure 4.4A). PsAarA frequently cleaved at sites that have small aliphatic amino acids in the P1 position and bulky hydrophobic amino acids such as Phe in the P2′ and P4 positions. In addition, Leu was frequently found in the P2 position while Arg and Ser were present in P1′ (Figure 4.4B). Although we were only able to quantify a low number of cleavage products, the overall substrate preference is similar to the substrate preference previously reported by Strisovsky and colleagues [55].

We next assayed a rhomboid protease from *Haemophilus influenzae* (HiGlpG) that had not previously been characterized for substrate specificity. This enzyme was also expressed and

purified from *E. coli* however only two substrates were hydrolyzed following 20 hours of incubation with the peptide mixture. The k_{cat}/K_M values for PQnIGHVKLFRFN*W and KWLIHPTF*SYnRWP were calculated to be 8.97×10^3 $M^{-1}s^{-1}$ and 5.61×10^3 $M^{-1}s^{-1}$, respectively, where * indicates the cleavage site. Interestingly, although the substrate preference of HiGlpG is more stringent than PsAarA, PQnIGHVKLFRFN*W was the most efficiently hydrolyzed substrate for both enzymes. This substrate likely contains sequence features that promote favorable interaction with the enzyme active site that is buried within the detergent micelle. A peptide corresponding to the P7 to P1′ amino acids (VKLFRFN*W) was synthesized flanked on the N-terminus with a fluorescent reporter molecule, 7-methoxycoumarin-4-acetic acid and on the C-terminus a quenching group, 2,4-dinitrophenol (Table 4.4). This soluble fluorescent substrate was then assayed with HiGlpG and PsAarA and the rate of cleavage was determinedin a microplate assay. In the qMSP-MS assay, HiGlpG cleaved the 14-mer peptide substrate six times faster than PsAarA (Figure 4.4C). This rate difference was also seen for the fluorescent substrate (Figure 4.4D). In addition, we assayed a catalytically inactive HiGlpG, consisting of a Ser to Ala mutation at position 116, with the fluorescent substrate and failed to detect hydrolysis. This mutant protein was expressed and purified under identical conditions to the active enzyme, thereby providing a control sample to ensure that proteolytic activity associated with this wild-type enzyme was not due to a contaminating *E. coli* protease. Finally, we sequenced the cleavage products of the fluorescent substrate and determined that hydrolysis occurred between Asn and Trp confirming that addition of the fluorophore and quencher molecules does not affect the cleavage site specificity of this substrate (Figure 4.4E, Figure 4.5). Taken together, these studies revealed that HiGlpG has narrow specificity but higher catalytic activity than PsAarA.
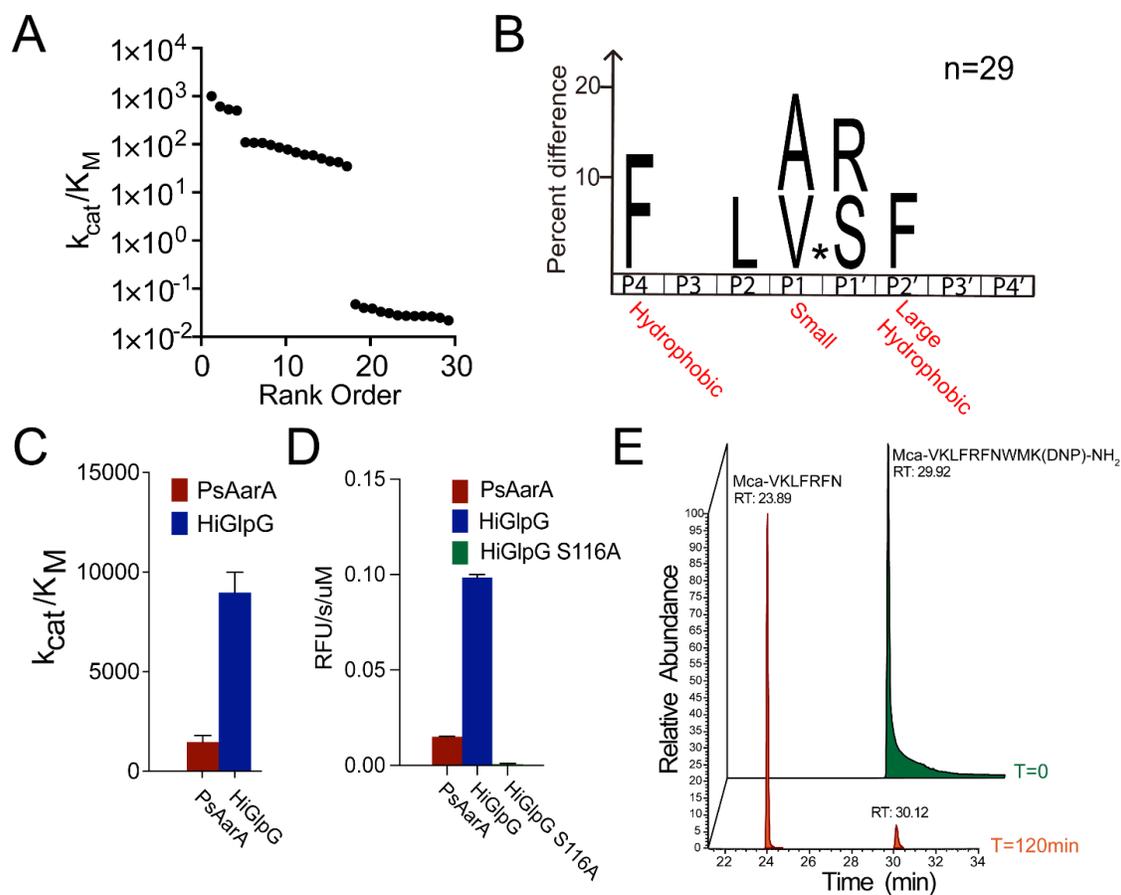
Figure 4.4. Substrate specificities and fluorescent assays of rhomboid proteases. A. Proteolytic constant k_{cat}/K_M of each cleaved peptide was calculated and ranked. B. PsAarA specificity profile was generated using 29 cleavages (p<0.05). Due to the low number of cleavage sites, no amino acid was found to be significantly de-enriched. Text in red was previously described to be the specificity of PsAarA [55]. C. PQnIGHVKLFRFNW was cleaved by PsAarA and HiGlpG between Asn and Trp and cleavage efficiencies were calculated. D. Fluorescent substrate screening of PsAarA, HiGlpG and HiGlpG S116A with the rationally designed internally quenched substrate IQ1. E. Cleavage of the fluorescent substrate between Asn and Trp was confirmed by mass spectrometry.
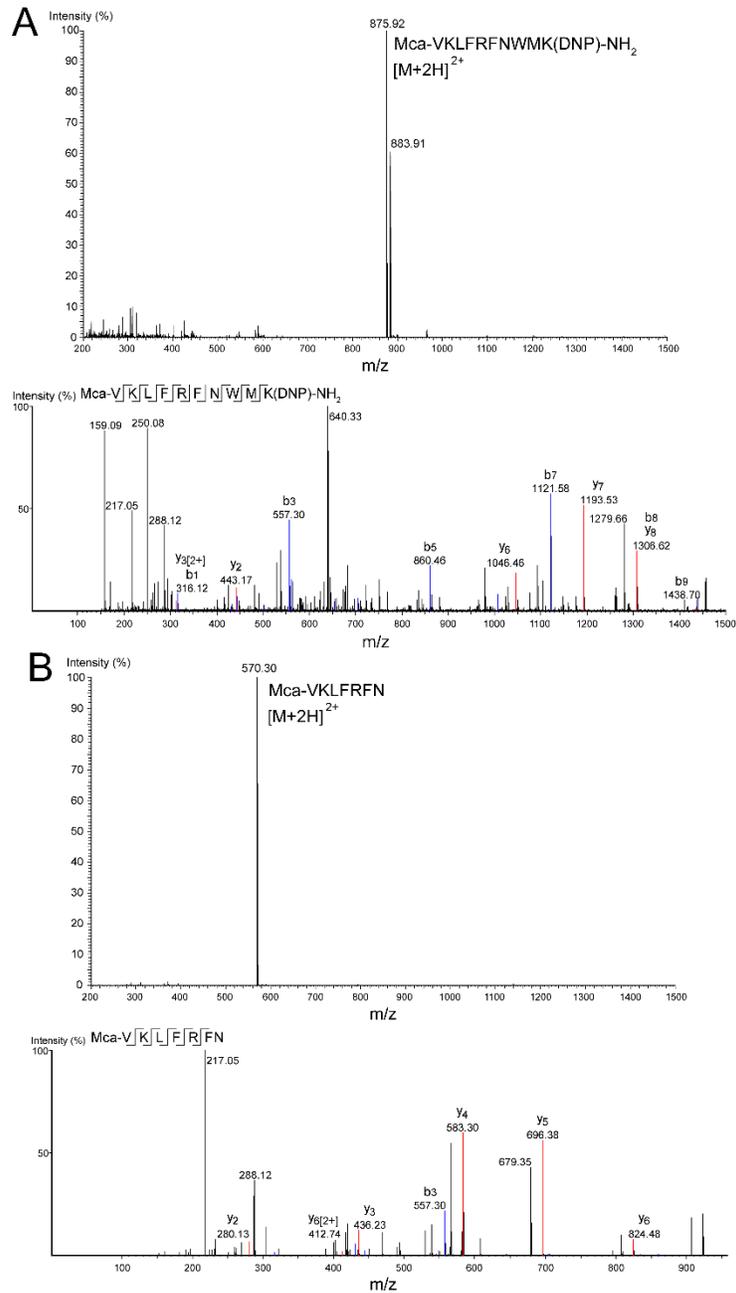
Figure 4.5. IQ1 was confirmed to be cleaved by Rhomboid proteases between Asn and Trp. A. MS and MS/MS spectra of full length IQ1 Mca-VKLFRFNWMK(DNP)-NH$_2$. B. MS and MS/MS spectra of IQ2 cleavage product Mca-VKLFRFN.

127

4. Characterization of proteolytic activities of a complex protease sample

We next investigated the utility of the qMSP-MS assay for characterizing complex biological samples containing proteases from diverse families. Using a commercial extract from *Aspergillus phoenicis* that is known to contain α-galactosidase and 1,3-β-D-glucanse activities [56, 57] we confirmed by SDS-PAGE and proteomics that this sample contained many fungal proteins (Figure 4.6). In total we identified 528 proteins of which, 49 were proteases that could be classified into 4 families, namely, serine-, cysteine-, aspartyl and metalloproteases (Figure 4.7A). Serine proteases were the most frequently found proteolytic enzymes in the extract while cysteine proteases were the least frequent. However, by spectral counting, aspartyl proteases were the most abundant proteolytic enzymes found (Figure 4.7A). We assayed this sample at pH 5.0 with the synthetic peptide library and quantified 243 cleaved products. As the concentration of each protease within this sample is unknown, proteolytic efficiencies of each cleavage were calculated as k values where $k = k_{cat}/K_M \times [E_0]$ and clustered into two groups (Figure 4.7B). Proteases in this extract rarely cleaved bonds between amino acids 1 and 3 indicting that no aminopeptidases were active under these assay conditions. In addition, cleavage of the peptide bonds between position 12 and 13 was slow and only found in Group 2. These data indicate that acid-acting endopeptidases rapidly hydrolyze peptide bonds between position 3 and 12 while one or more acid-acting carboxypeptidase cleaves single amino acids from the carboxyl terminus of these peptides (Figure 4.7C).
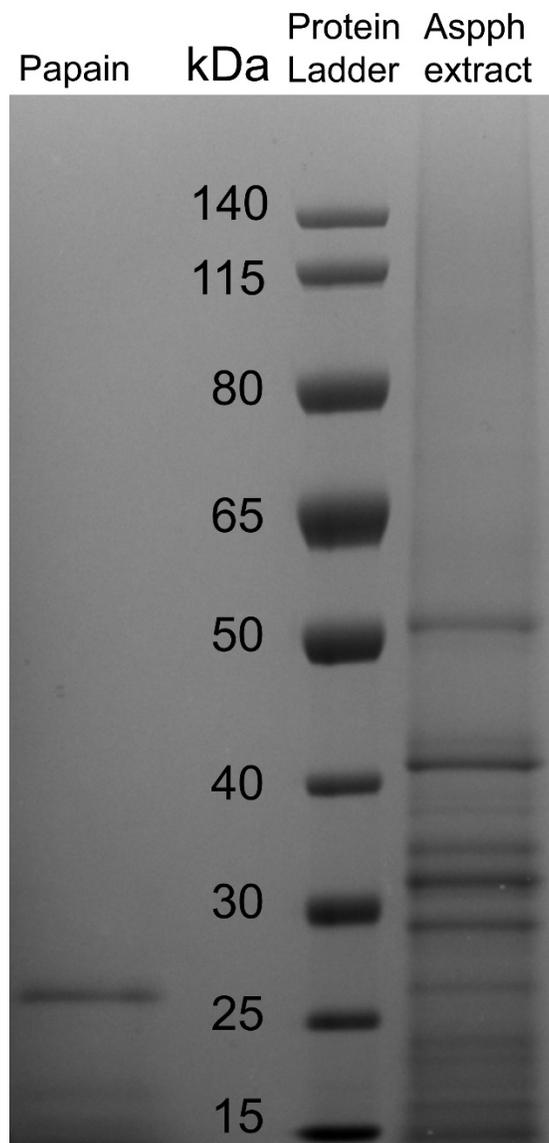
Figure 4.6. SDS-PAGE gel images of *Aspph* extract (Sigma, P2143) and papain (Sigma, P4762) samples used in our studies.
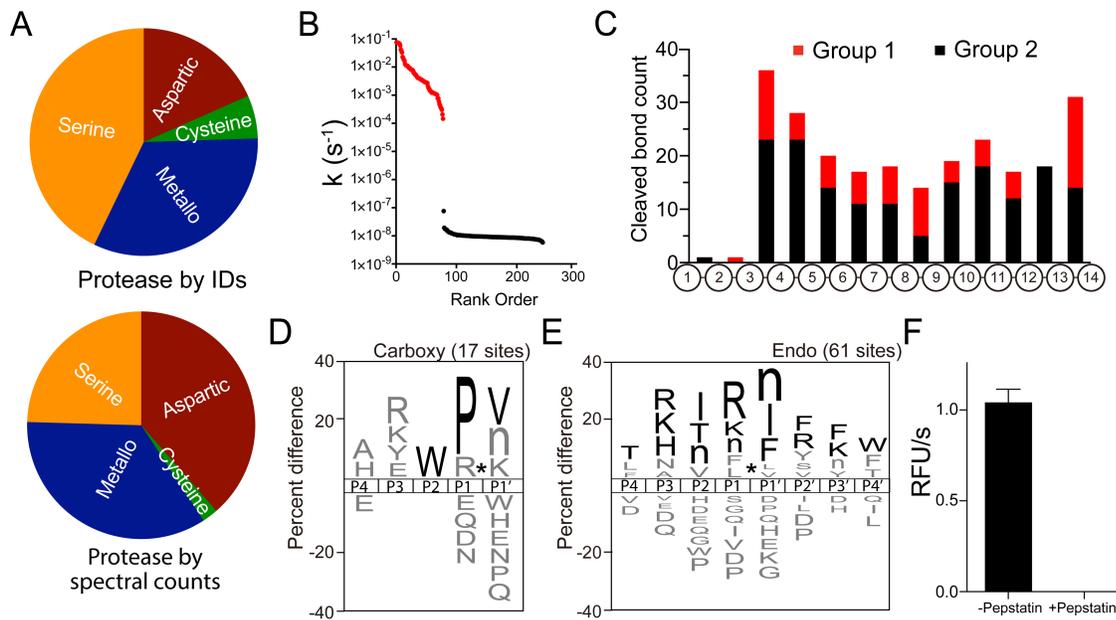
Figure 4.7. Proteomic analysis and Quantitative multiplex substrate profiling of *Aspergillus phoenicis (Aspph)* extract. A. Proteomic analysis of *Aspph* extract revealed the presence of 51 proteases. Spectral counts ranking determined that aspartyl proteases were abundant in *Aspph* extract. B. Proteolytic efficiency, $k = k_{cat}/K_M \times [E_0]$, for 243 cleavage products were calculated, ranked, and clustered into Group 1 (red) and Group 2 (black). C. Distribution of Group 1 and 2 cleavage sites within 14-mer peptides. D. Substrate specificity profile of mono-carboxypeptidase activities. E. Substrate specificity profile of endopeptidases activities. F. *Aspph* extract cleaved fluorescent substrate IQ2 and the activity is completely inhibited with addition of pepstatin.

We generated a specificity profile of the carboxypeptidase activity and found that these enzymes in the fungal extract remove C-terminal Val, Nle and Lys when Pro or Arg are in the penultimate position and Trp is in the P2 position (Figure 4.7D). The cleavage profile of the endopeptidases showed a preference for hydrophobic and positively charged amino acids in the P1 positions and while hydrophobic amino acids were exclusively found in the P1′ position. In addition, positively charged amino acids were frequently present in the P3 position (Figure 4.7E). This substrate profile was similar to the cleavage specificity generated by secreted aspartyl proteases from *Malassezia glabosa* [48] and *Candida albicans* [58]. Therefore, we assayed the *Aspph* extract with a fluorescent substrate IQ2 (Supplemental Table S4) used for aspartic acid proteases [59] and found that it was efficiently hydrolyzed. Cleavage occurred between the Phe-Phe bond (Figure 4.8) and this activity was completely inhibited by pepstatin, an aspartyl protease inhibitor (Figure 4.7F).

These studies show that the qMSP-MS assay can be used to distinguish between endo- and exopeptidases within complex biological samples. In addition, knowledge of protease class and understanding of the cleavage pattern facilitated the identification of fluorescent reporter substrates that can be used for subsequent biochemical assays.
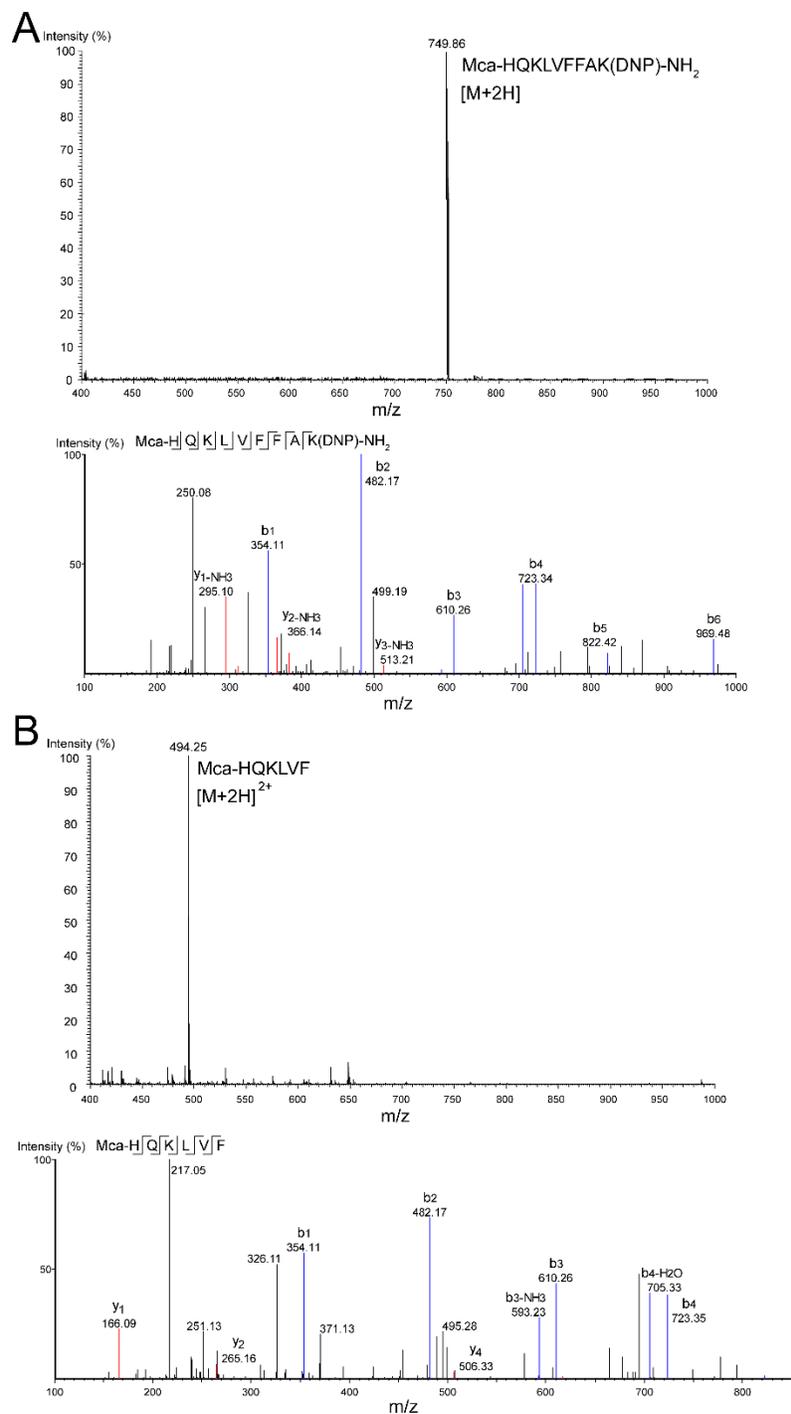
Figure 4.8. MS and MS/MS spectra of IQ2 and its cleavage product showed *Aspph* extract sample cleaved the fluorescent substrate IQ2 between Phe and Phe. A. MS and MS/MS spectra of full length IQ2 Mca-HQKLVFFAK(DNP)-NH$_2$. B. MS and MS/MS spectra of IQ2 cleavage product Mca-HQKLVF.

5. Application of qMSP-MS in studying lung cancer pericellular protease activities

In-depth transcriptional analysis of cell lines and tumors has identified candidate proteases that may play a role in disease progression. However, such expression-based strategies to study proteases cannot account for the effects of translational regulation, post-translational modifications, stability and the presence of endogenous inhibitors in the pericellular environment. Therefore, increased expression of a protease may not result in a concomitant increase in proteolytic activity. New, function-based technologies with high information content are needed to detect dysregulated proteolysis in the tumor microenvironment. Given that we have demonstrated qMSP-MS can be used to uncover the proteolytic activity in complex protein samples mixture, we sought to apply this method to study the secreted protease activity from a panel of cell lines representing four sub-types of lung cancer (Table 4.1). The long-term goal is to identify key protease activities that can be targeted in the tumor microenvironment for diagnostics, imaging or prodrug activation.

Using qMSP-MS, we characterized the secreted proteolytic activity in conditioned media from these lung cancer cell lines. However, unlike previous assays with papain and *Aspergillus phoenicis* extract, we found that most peptides were sequentially hydrolyzed at the amino and carboxyl terminus and the initial cleavage products were subsequently degraded at later time points. For example, the substrate FRIHGFDEAHNAWL was hydrolyzed to GFDEAHNAWL by sequential removal of F, R, I and H from the amino terminus by proteases secreted from DMS273 cells (Figure 4.9A). Therefore, the progression curve of each cleavage product was not monotonic and therefore did not conform to pseudo-first order kinetics. To address this issue, we calculated the change in abundance of each cleavage product between two-time intervals (Figure

4.10). Our results showed that most cleavage events occurred at the N-terminus of the tetradecapeptides, indicating the presence of strong aminopeptidases activities in lung cancer secretions (Figure 4.9B). In addition, BEN, H661, H1944 and SHP77 secrete proteases with a preference for removing single amino acids from the C-terminus of the tetrapeptides. In parallel, quantitative shotgun proteomics of secretions from each of the cell lines was performed to identify proteases that were present in the conditioned media. In total, 6,343 unique proteins were identified, 55 of which were proteases, including 14 aminopeptidases, 12 carboxypeptidases and 29 endopeptidases (Figure 4.9C). The qMSP-MS assay data clearly indicated that aminopeptidases were more active than the endopeptidases and carboxypeptidases in these lung cancer secretions and therefore we focused on developing fluorescent substrates to further characterize the aminopeptidase activities.
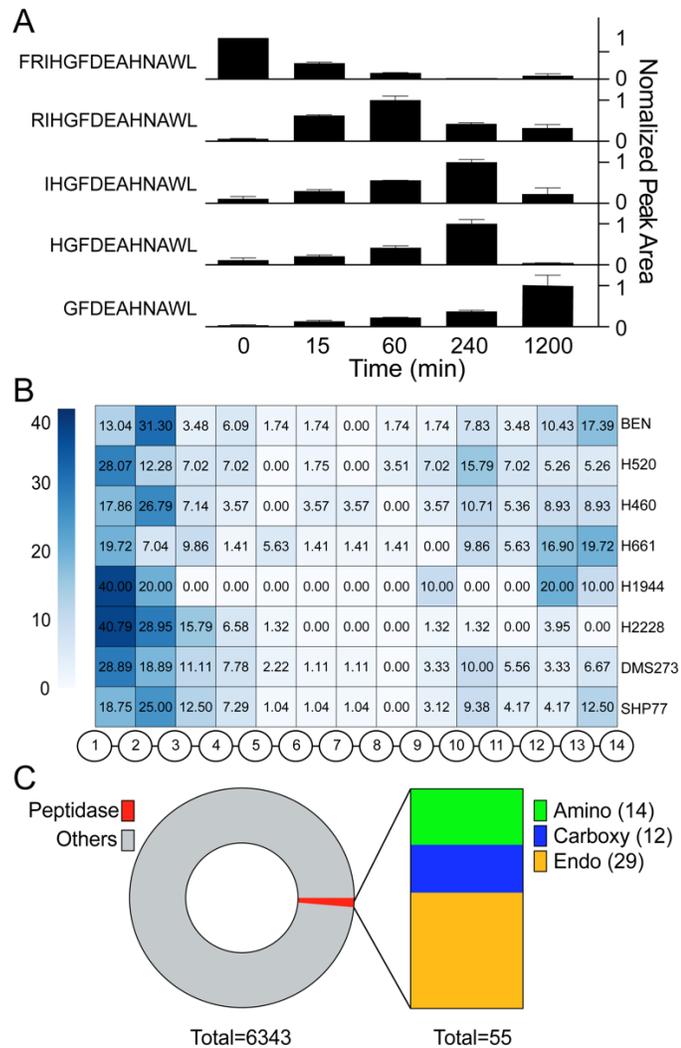
Figure 4.9. Application of qMSP-MS to lung cancer secretions. A. An example showing the N-terminus of a substrate being sequentially cleaved by aminopeptidases in DMS273 secretion. B. Analysis of cleavage loci within the tetradecapeptide library. Numbers represent the percentage of cleavage sites at each position relative to the total cleavage sites generated by secreted proteases in each cell line. C. 55 proteases identified in lung cancer secretions of which 14 are predicted to be aminopeptidases and 12 are predicted to be carboxypeptidases.
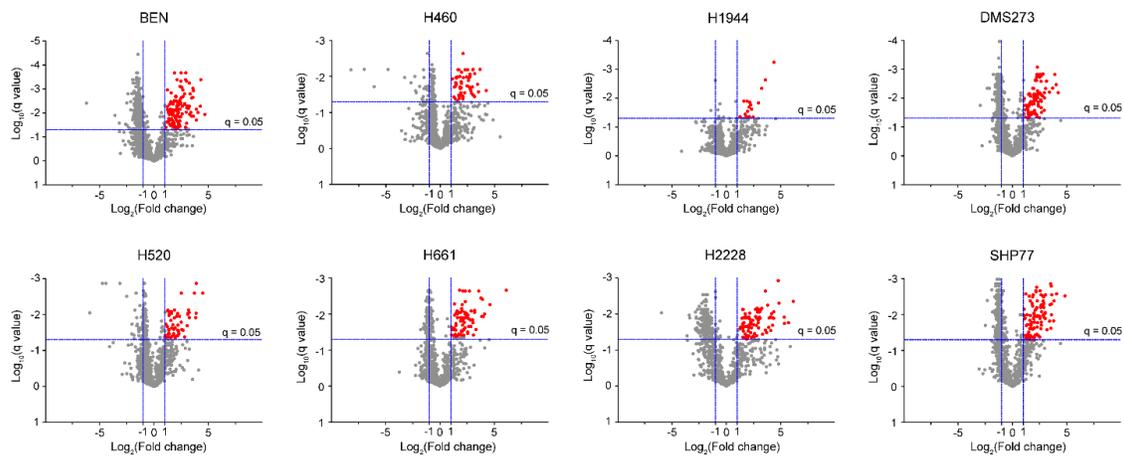
Figure 4.10. Volcano plot of lung cancer qMSP-MS results showing fold change of peptide between 15 min time interval and control samples. Red symbols indicate cleavage products that were significantly increased (fold change > 2, Student's T-test q value < 0.05).

We generated substrate specificity profiles of proteolysis between residue 1 and 2 corresponding to mono-aminopeptidases activity for all cell lines (Figure 4.11A) and discovered that these human aminopeptidases frequently removed Ala, Trp, Phe, Tyr and Arg from the amino terminus while rarely or never cleaving Asp, Pro and Gln. We subsequently evaluated a selection of mono-aminopeptidase substrates consisting of Ala, Tyr and Arg linked to the fluorescent reporter molecule, 7-amino-4-methylcoumarin (AMC). These substrates were hydrolyzed by aminopeptidases in conditioned media from all cell lines (Figure 4.11B). The rate of cleavage of Ala-AMC and Tyr-AMC was similar in all cell lines while hydrolysis of Arg-AMC was generally 1.5 to 3-fold faster. In the substrate profiling assay, tetradecapeptides with amino terminal Asp are rarely hydrolyzed. Therefore, we designed Asp-AMC as a negative control to assay these secretions and found there was no detectable cleavage of this substrate.

In some tetradecapeptide substrates, removal of two amino acids from the amino terminus was commonly observed. In most cases, this appears to be due to di-aminopeptidase activity rather than two sequential mono-aminopeptidase cleavage events because some 12-mer cleavage products were found but not the related 13-mer intermediate product. In addition, we cannot rule out that these substrates are also cleaved by endopeptidases that have little or no specificity beyond the P2 position. We generated a substrate specificity profile of peptide cleavage sites between residue 2 and 3 (Figure 4.11C). From this specificity profile, we designed two dipeptide substrates that are unlikely to be hydrolyzed by mono-aminopeptidases. To achieve this, we included a terminal Gly residue because this amino acid is significantly enriched in the P2 position but not preferentially cleaved by mono-aminopeptidases. In the P1 position, we included Nle or Arg, two amino acids that are frequently enriched in the substrate profile. The results showed that both fluorescent substrates were commonly hydrolyzed by lung cancer

cell secretions but at different rates (Figure 4.11D). In addition, H460, H661 and H1944 cell secretions had reduced specific activity for both mono- and di-amino fluorescent substrates, indicating the aminopeptidases composition in these cell secretions might be lower compared to the others. Taken together, the fluorescent substrate screening assays supported the qMSP-MS specificity profile and demonstrates that this assay is not limited to characterizing purified proteases but can be applied to studying complex biological samples. The specificity information generated from these lung cancer cell lines will be utilized for developing activity-based diagnostics or prodrugs that will be activated in the tumor microenvironment by one or more proteases.
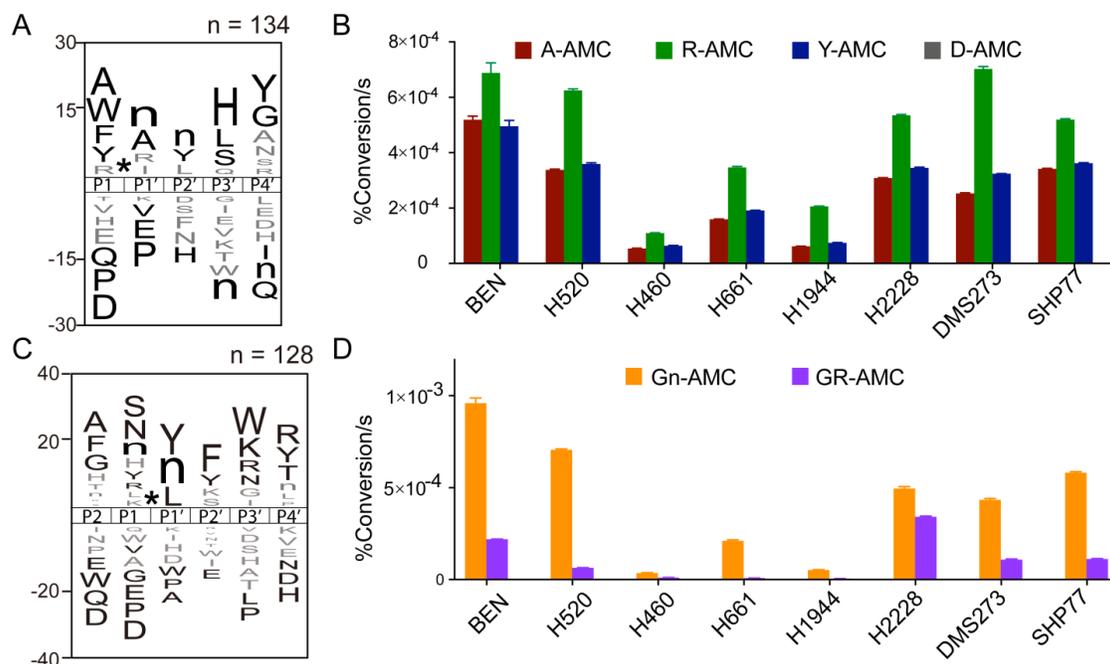
Figure 4.11. Global specificity profiles of aminopeptidases activities in lung cancer secretions and fluorescent substrate screenings. A. qMSP-MS showed the specificity profile of mono-aminopeptidases activities using the pool of cleavages from all lung cancer secretions. B. Mono-amino fluorescent substrates were designed based on the profiles for orthogonal validation. C. qMSP-MS showed the specificity profile of di-aminopeptidases activities using the pool of cleavages from all lung cancer secretions. D. Di-amino fluorescent substrates were designed based on the profiles for orthogonal validation. Data presented as mean ± sd of triplicate assays.

D. Discussion

Several studies have used isobaric labels to quantify peptides that are modified by proteolysis, phosphorylation, or methylation [60-66]. These labels are generally used to directly compare the peptide abundance between multiple biological samples or within a single biological sample that has been subjected to different treatment regimes. However, in this study, we utilize isobaric labels to quantify and rank hundreds of peptide cleavage reactions that take place simultaneously. Previously, we developed a library of 124 tetradecapeptides composed of all neighbor and near-neighbor amino acid pairs and monitored the time dependent appearance of cleavage products upon addition of a protease [19]. This method used tandem mass spectrometry to identify new products but similar to many other substrate profiling methods [10, 14], it was largely qualitative. We therefore incorporated TMT labels into our MSP-MS workflow to minimize experimental and instrumentation derived variance while improving throughput of the assay. More importantly, by labeling samples at multiple time intervals, we were able to perform accurate quantification of peptides, allowing us to calculate the turnover rate of each proteolytic event.

To validate our workflow, we demonstrated that qMSP-MS is able to accurately characterize papain substrate specificity. We highlighted the necessity of calculating the cleavage efficiency of different peptides as it can vary by eight orders of magnitude. This information is crucial when studying proteases as it allows us to identify individual amino acids within the substrate sequence that promote rapid hydrolysis of the peptide. Previously, papain was found to preferentially cleave many hydrophobic amino acids at the P2 position [51, 52]. However, we demonstrated that Val, Ile, Ala and Leu are strongly favored over other hydrophobic amino acids such as Tyr, Nle, Phe and Pro.

To further demonstrate the versatility of this method, we assayed two related intramembrane proteases whose active sites are buried in the lipid bilayer of cell membranes. It is unknown how many of the 228 peptides are able to access the active site and therefore substrate specificity may be influenced by sequence features of the peptide that promote or prevent accessibility to the active site. Using qMSP-MS, we found that HiGlpG has stricter specificity and higher catalytic activity than PsAarA and a single fluorescent substrate was designed that could monitor activity of both enzymes. Compared to the standard gel-based methods for detecting rhomboid protease activity [55, 67], this fluorescent substrate can be used in microplate assays and therefore facilitates rapid and quantitative activity assays for integral membrane proteases. Five members of this rhomboid protease family are found in mammalian cells and these enzymes are associated with tumor metastasis, tumor angiogenesis, type 2 diabetes and neurodegenerative disorders [68]. In general, these integral membrane serine proteases are much less characterized than their soluble serine protease counterparts and no substrates have been discovered for RHBDL1 and RHBDL3. Therefore, quantitative substrate specificity technologies like qMSP-MS can be used to uncover the substrate specificity of these enzymes, facilitating the discovery of their biological substrates.

From these findings, we predicted that qMSP-MS would have an impact on designing substrates for monitoring proteolytic activities and developing protease-activated drugs where potency and selectivity are essential. The qMSP-MS assay can pinpoint which sequences are preferred by proteases, allowing us to design optimal substrates for protease screening. Our studies on the *Aspph* extract mixture showed the power of qMSP-MS in studying complex samples and designing individual fluorescent substrates that can be used to monitor their activities. As a final evaluation of qMSP-MS performance in the context of a complex protease

system, we applied it to a panel of lung cancer cell line secretions, representing the first global characterization and comparison of proteolytic activity in the cancer pericellular environment.

Lung cancer represents 25% of all cancer diagnoses, corresponding to 200,000 new cases in the United States each year. Furthermore, it is the second most deadly cancer with only an 18% 5-year survival rate (58). Increased pericellular proteolysis is a hallmark of aggressive cancers (59). Secreted proteases can modulate tumor progression and metastasis through highly regulated processes that involve shedding of cell adhesion molecules and processing of growth factors and cytokines (60-65). Various strategies have been developed to utilize proteases for drug delivery for cancer chemotherapy [69, 70] and surgery [71]. Traditionally, the discovery of protease targets for prodrugs or antibody-drug conjugate activation is largely on the expression level and further downstream characterizations are usually required. However, our qMSP-MS method allows us to functionally characterize proteases specificities, which narrows the gap between biological discovery and drug design, offering valuable information for translational studies. In this study, we showed the application of qMSP-MS to this panel of lung cancer cell lines, demonstrating the feasibility of this approach for studying dynamic diseases.

Aminopeptidase activity has previously been detected in blood [63] and inflamed skin [72] using N-terminal enrichment strategies, while carboxypeptidase activity has been detected in mammalian [73] and bacterial [74] cell lysates using C-terminal enrichment strategies. In the lung cancer secretions, we detected both N-terminal and C-terminal trimming of many peptide substrates that we predict are the result of aminopeptidase and carboxypeptidase activity. In addition, we developed a panel of fluorescent substrates to quantify mono- and di-aminopeptidase activity in microplate assays. Our findings, together with previous studies, indicated that aminopeptidases and carboxypeptidases are associated with tumor invasion and

metastasis, but the development of exopeptidase inhibitors for cancer treatment is in a relatively early stage [75-78]. The logical next step from the outcomes of our study would be to design protease-activated prodrugs targeted to global pericellular proteolysis in tumor microenvironment. As MS technology and multiplexing abilities continue to advance, driving down costs and instrumentation time requirements, it is easy to foresee qMSP-MS as a potential first step in the development of personalized medicine geared towards not only tumors, but other protease-related diseases as well.

E. References

Chapter IV is a full reprint of the material as it appears in the article published in Molecular & Cellular Proteomics, 18(5), pp.968-981 by John D. Lapek, Zhenze Jiang, Jacob M. Wozniak, Elena Arutyunova, Steven C. Wang, M. Joanne Lemieux, David J. Gonzalez, and Anthony J. O'Donoghue. The thesis author is the primary researcher and author of the article.

F. References

1.   Rawlings, N. D.; Salvesen, G., Handbook of proteolytic enzymes. 3rd ed.; Academic Press,: London ; Boston, 2013; pp. 1 online resource (3 volumes (liv, 3,932 pages)).

2.   Turk, B., Targeting proteases: successes, failures and future prospects. *Nat Rev Drug Discov* **2006,** *5* (9), 785-99.

3.   Choi, K. Y.; Swierczewska, M.; Lee, S.; Chen, X., Protease-activated drug development. *Theranostics* **2012,** *2* (2), 156-78.

4.   Harris, J. L.; Backes, B. J.; Leonetti, F.; Mahrus, S.; Ellman, J. A.; Craik, C. S., Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proc Natl Acad Sci U S A* **2000,** *97* (14), 7754-9.

5.   Poreba, M.; Kasperkiewicz, P.; Snipas, S. J.; Fasci, D.; Salvesen, G. S.; Drag, M., Unnatural amino acids increase sensitivity and provide for the design of highly selective caspase substrates. *Cell Death Differ* **2014,** *21* (9), 1482-92.

6.   Wysocka, M.; Lesner, A.; Guzow, K.; Mackiewicz, L.; Legowska, A.; Wiczk, W.; Rolka, K., Design of selective substrates of proteinase 3 using combinatorial chemistry methods. *Anal Biochem* **2008,** *378* (2), 208-15.

7.   Oliveira, L. C.; Silva, V. O.; Okamoto, D. N.; Kondo, M. Y.; Santos, S. M.; Hirata, I. Y.; Vallim, M. A.; Pascon, R. C.; Gouvea, I. E.; Juliano, M. A.; Juliano, L., Internally quenched fluorescent peptide libraries with randomized sequences designed to detect endopeptidases. *Anal Biochem* **2012,** *421* (1), 299-307.

8.   Boulware, K. T.; Daugherty, P. S., Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proc Natl Acad Sci U S A* **2006,** *103* (20), 7583-8.

9.   Matthews, D. J.; Wells, J. A., Substrate phage: selection of protease substrates by monovalent phage display. *Science* **1993,** *260* (5111), 1113-7.

10.   Schilling, O.; Overall, C. M., Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol* **2008,** *26* (6), 685-94.

11.   Kleifeld, O.; Doucet, A.; auf dem Keller, U.; Prudova, A.; Schilling, O.; Kainthan, R. K.; Starr, A. E.; Foster, L. J.; Kizhakkedathu, J. N.; Overall, C. M., Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat Biotechnol* **2010,** *28* (3), 281-8.

12.   Gevaert, K.; Goethals, M.; Martens, L.; Van Damme, J.; Staes, A.; Thomas, G. R.; Vandekerckhove, J., Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol* **2003,** *21* (5), 566-9.

13.   Mahrus, S.; Trinidad, J. C.; Barkan, D. T.; Sali, A.; Burlingame, A. L.; Wells, J. A., Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* **2008,** *134* (5), 866-76.

14.   Vizovisek, M.; Vidmar, R.; Fonovic, M., FPPS: Fast Profiling of Protease Specificity. *Methods Mol Biol* **2017,** *1574*, 183-195.

15.   Staes, A.; Impens, F.; Van Damme, P.; Ruttens, B.; Goethals, M.; Demol, H.; Timmerman, E.; Vandekerckhove, J.; Gevaert, K., Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc* **2011,** *6* (8), 1130-41.

16.   Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J. N.; Overall, C. M., Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat Protoc* **2011,** *6* (10), 1578-611.

17.   Thornberry, N. A.; Rano, T. A.; Peterson, E. P.; Rasper, D. M.; Timkey, T.; Garcia-Calvo, M.; Houtzager, V. M.; Nordstrom, P. A.; Roy, S.; Vaillancourt, J. P.; Chapman, K. T.; Nicholson, D. W., A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem* **1997,** *272* (29), 17907-11.

18.   Backes, B. J.; Harris, J. L.; Leonetti, F.; Craik, C. S.; Ellman, J. A., Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. *Nat Biotechnol* **2000,** *18* (2), 187-93.

19.   O'Donoghue, A. J.; Eroy-Reveles, A. A.; Knudsen, G. M.; Ingram, J.; Zhou, M.; Statnekov, J. B.; Greninger, A. L.; Hostetter, D. R.; Qu, G.; Maltby, D. A.; Anderson, M. O.; Derisi, J. L.; McKerrow, J. H.; Burlingame, A. L.; Craik, C. S., Global identification of peptidase specificity by multiplex substrate profiling. *Nat Methods* **2012,** *9* (11), 1095-100.

20.   O'Donoghue, A. J.; Knudsen, G. M.; Beekman, C.; Perry, J. A.; Johnson, A. D.; DeRisi, J. L.; Craik, C. S.; Bennett, R. J., Destructin-1 is a collagen-degrading endopeptidase secreted by Pseudogymnoascus destructans, the causative agent of white-nose syndrome. *Proc Natl Acad Sci U S A* **2015,** *112* (24), 7478-83.

21.   Winter, M. B.; La Greca, F.; Arastu-Kapur, S.; Caiazza, F.; Cimermancic, P.; Buchholz, T. J.; Anderl, J. L.; Ravalin, M.; Bohn, M. F.; Sali, A.; O'Donoghue, A. J.; Craik, C. S., Immunoproteasome functions explained by divergence in cleavage specificity and regulation. *Elife* **2017,** *6*.

22.   Roncase, E. J.; Moon, C.; Chatterjee, S.; Gonzalez-Paez, G. E.; Craik, C. S.; O'Donoghue, A. J.; Wolan, D. W., Substrate Profiling and High Resolution Co-complex Crystal Structure of a Secreted C11 Protease Conserved across Commensal Bacteria. *ACS Chem Biol* **2017,** *12* (6), 1556-1565.

23.   Li, H.; O'Donoghue, A. J.; van der Linden, W. A.; Xie, S. C.; Yoo, E.; Foe, I. T.; Tilley, L.; Craik, C. S.; da Fonseca, P. C.; Bogyo, M., Structure- and function-based design of Plasmodium-selective proteasome inhibitors. *Nature* **2016,** *530* (7589), 233-6.

24.   Lentz, C. S.; Ordonez, A. A.; Kasperkiewicz, P.; La Greca, F.; O'Donoghue, A. J.; Schulze, C. J.; Powers, J. C.; Craik, C. S.; Drag, M.; Jain, S. K.; Bogyo, M., Design of Selective Substrates and Activity-Based Probes for Hydrolase Important for Pathogenesis 1 (HIP1) from Mycobacterium tuberculosis. *ACS Infect Dis* **2016,** *2* (11), 807-815.

25.   Joshi, S.; Chen, L.; Winter, M. B.; Lin, Y. L.; Yang, Y.; Shapovalova, M.; Smith, P. M.; Liu, C.; Li, F.; LeBeau, A. M., The Rational Design of Therapeutic Peptides for Aminopeptidase N using a Substrate-Based Approach. *Sci Rep* **2017,** *7* (1), 1424.

26.   Ivry, S. L.; Sharib, J. M.; Dominguez, D. A.; Roy, N.; Hatcher, S. E.; Yip-Schneider, M. T.; Schmidt, C. M.; Brand, R. E.; Park, W. G.; Hebrok, M.; Kim, G. E.; O'Donoghue, A. J.; Kirkwood, K. S.; Craik, C. S., Global Protease Activity Profiling Provides Differential Diagnosis of Pancreatic Cysts. *Clin Cancer Res* **2017,** *23* (16), 4865-4874.

27.   McAlister, G. C.; Nusinow, D. P.; Jedrychowski, M. P.; Wuhr, M.; Huttlin, E. L.; Erickson, B. K.; Rad, R.; Haas, W.; Gygi, S. P., MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem* **2014,** *86* (14), 7150-8.

28.   Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003,** *75* (8), 1895-904.

29.   Arutyunova, E.; Panwar, P.; Skiba, P. M.; Gale, N.; Mak, M. W.; Lemieux, M. J., Allosteric regulation of rhomboid intramembrane proteolysis. *EMBO J* **2014,** *33* (17), 1869-81.

30.   Villen, J.; Gygi, S. P., The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* **2008,** *3* (10), 1630-8.

31.   Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P., Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **2006,** *5* (7), 1326-37.

32.   Wessel, D.; Flugge, U. I., A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **1984,** *138* (1), 141-3.

33.   Tolonen, A. C.; Haas, W., Quantitative proteomics using reductive dimethylation for stable isotope labeling. *J Vis Exp* **2014,** (89).

34.   Wang, Y.; Yang, F.; Gritsenko, M. A.; Wang, Y.; Clauss, T.; Liu, T.; Shen, Y.; Monroe, M. E.; Lopez-Ferrer, D.; Reno, T.; Moore, R. J.; Klemke, R. L.; Camp, D. G., 2nd;

Smith, R. D., Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **2011,** *11* (10), 2019-26.

35.   Lapek, J. D., Jr.; Greninger, P.; Morris, R.; Amzallag, A.; Pruteanu-Malinici, I.; Benes, C. H.; Haas, W., Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat Biotechnol* **2017,** *35* (10), 983-989.

36.   Lapek, J. D., Jr.; Fang, R. H.; Wei, X.; Li, P.; Wang, B.; Zhang, L.; Gonzalez, D. J., Biomimetic Virulomics for Capture and Identification of Cell-Type Specific Effector Proteins. *ACS Nano* **2017,** *11* (12), 11831-11838.

37.   Lapek, J. D., Jr.; Lewinski, M. K.; Wozniak, J. M.; Guatelli, J.; Gonzalez, D. J., Quantitative Temporal Viromics of an Inducible HIV-1 Model Yields Insight to Global Host Targets and Phospho-Dynamics Associated with Protein Vpr. *Mol Cell Proteomics* **2017,** *16* (8), 1447-1461.

38.   Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **1994,** *5* (11), 976-89.

39.   Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** *4* (3), 207-14.

40.   Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2005,** *2* (9), 667-75.

41.   Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2003,** *2* (1), 43-50.

42.   Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **2006,** *24* (10), 1285-92.

43.   Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villen, J.; Haas, W.; Sowa, M. E.; Gygi, S. P., A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **2010,** *143* (7), 1174-89.

44.   Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007,** *4* (11), 923-5.

45.   Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S., Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. *J Proteome Res* **2009,** *8* (7), 3737-45.

46.   Rappsilber, J.; Ishihama, Y.; Mann, M., Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **2003,** *75* (3), 663-70.

47.   Colaert, N.; Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K., Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* **2009,** *6* (11), 786-7.

48.   Li, H.; Goh, B. N.; Teh, W. K.; Jiang, Z.; Goh, J. P. Z.; Goh, A.; Wu, G.; Hoon, S. S.; Raida, M.; Camattari, A.; Yang, L.; O'Donoghue, A. J.; Dawson, T. L., Jr., Skin Commensal Malassezia globosa Secreted Protease Attenuates Staphylococcus aureus Biofilm Formation. *J Invest Dermatol* **2018,** *138* (5), 1137-1145.

49.   Bromme, D., Papain-like cysteine proteases. *Curr Protoc Protein Sci* **2001,** *Chapter 21*, Unit 21 2.

50.   Fu, Y.; Liu, J.; Hansen, E. T.; Bredie, W. L. P.; Lametsch, R., Structural characteristics of low bitter and high umami protein hydrolysates prepared from bovine muscle and porcine plasma. *Food Chem* **2018,** *257*, 163-171.

51.   Koga, H.; Yamada, H.; Nishimura, Y.; Kato, K.; Imoto, T., Comparative study on specificities of rat cathepsin L and papain: amino acid differences at substrate-binding sites are involved in their specificities. *J Biochem* **1990,** *108* (6), 976-82.

52.   Choe, Y.; Leonetti, F.; Greenbaum, D. C.; Lecaille, F.; Bogyo, M.; Bromme, D.; Ellman, J. A.; Craik, C. S., Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J Biol Chem* **2006,** *281* (18), 12824-32.

53.   Del Nery, E.; Alves, L. C.; Melo, R. L.; Cesari, M. H.; Juliano, L.; Juliano, M. A., Specificity of cathepsin B to fluorescent substrates containing benzyl side-chain-substituted amino acids at P1 subsite. *J Protein Chem* **2000,** *19* (1), 33-8.

54.   Urban, S., A guide to the rhomboid protein superfamily in development and disease. *Semin Cell Dev Biol* **2016,** *60*, 1-4.

55.   Strisovsky, K.;   Sharpe, H. J.; Freeman, M., Sequence-specific intramembrane proteolysis: identification of a recognition motif in rhomboid substrates. *Mol Cell* **2009,** *36* (6), 1048-59.

56.   Sugimoto, H.; Buren, J. P. v., Removal of oligosaccharides from soy milk by an enzyme from Aspergillus saitoi. *Journal of Food Science* **1970,** *35* (5), 655-660.

57.   Kasahara, S.; Nakajima, T.; Miyamoto, C.; Wada, K.; Furuichi, Y.; Ichishima, E., Characterization and mode of action of exo-1,3-β-d-glucanase from Aspergillus saitoi. *Journal of Fermentation and Bioengineering* **1992,** *74* (4), 238-240.

58.   Winter, M. B.; Salcedo, E. C.; Lohse, M. B.; Hartooni, N.; Gulati, M.; Sanchez, H.; Takagi, J.; Hube, B.; Andes, D. R.; Johnson, A. D.; Craik, C. S.; Nobile, C. J., Global Identification of Biofilm-Specific Proteolysis in Candida albicans. *MBio* **2016,** *7* (5).

59.   Farzan, M.; Schnitzler, C. E.; Vasilieva, N.; Leung, D.; Choe, H., BACE2, a beta - secretase homolog, cleaves at the beta site and within the amyloid-beta region of the amyloid-beta precursor protein. *Proc Natl Acad Sci U S A* **2000,** *97* (17), 9712-7.

60.   Jakoby, T.; van den Berg, B. H.; Tholey, A., Quantitative protease cleavage site profiling using tandem-mass-tag labeling and LC-MALDI-TOF/TOF MS/MS analysis. *J Proteome Res* **2012,** *11* (3), 1812-20.

61.   Shishkova, E.; Zeng, H.; Liu, F.; Kwiecien, N. W.; Hebert, A. S.; Coon, J. J.; Xu, W., Global mapping of CARM1 substrates defines enzyme specificity and substrate recognition. *Nat Commun* **2017,** *8*, 15571.

62.   Trinidad, J. C.; Thalhammer, A.; Specht, C. G.; Lynn, A. J.; Baker, P. R.; Schoepfer, R.; Burlingame, A. L., Quantitative analysis of synaptic phosphorylation and protein expression. *Mol Cell Proteomics* **2008,** *7* (4), 684-96.

63.   Wildes, D.; Wells, J. A., Sampling the N-terminal proteome of human blood. *Proc Natl Acad Sci U S A* **2010,** *107* (10), 4561-6.

64.   Schlage, P.; Egli, F. E.; Nanni, P.; Wang, L. W.; Kizhakkedathu, J. N.; Apte, S. S.; auf dem Keller, U., Time-resolved analysis of the matrix metalloproteinase 10 substrate degradome. *Mol Cell Proteomics* **2014,** *13* (2), 580-93.

65.   Schlage, P.; Kockmann, T.; Sabino, F.; Kizhakkedathu, J. N.; Auf dem Keller, U., Matrix Metalloproteinase 10 Degradomics in Keratinocytes and Epidermal Tissue Identifies Bioactive Substrates With Pleiotropic Functions. *Mol Cell Proteomics* **2015,** *14* (12), 3234-46.

66.   Plasman, K.; Van Damme, P.; Kaiserman, D.; Impens, F.; Demeyer, K.; Helsens, K.; Goethals, M.; Bird, P. I.; Vandekerckhove, J.; Gevaert, K., Probing the efficiency of proteolytic events by positional proteomics. *Mol Cell Proteomics* **2011,** *10* (2), M110 003301.

67.   Urban, S.; Freeman, M., Substrate specificity of rhomboid intramembrane proteases is governed by helix-breaking residues in the substrate transmembrane domain. *Mol Cell* **2003,** *11* (6), 1425-34.

68.   Dusterhoft, S.; Kunzel, U.; Freeman, M., Rhomboid proteases in human disease: Mechanisms and future prospects. *Biochim Biophys Acta* **2017,** *1864* (11 Pt B), 2200-2209.

69.   Zhong, Y. J.; Shao, L. H.; Li, Y., Cathepsin B-cleavable doxorubicin prodrugs for targeted cancer therapy (Review). *Int J Oncol* **2013,** *42* (2), 373-83.

70.   Vartak, D. G.; Gemeinhart, R. A., Matrix metalloproteases: underutilized targets for drug delivery. *J Drug Target* **2007,** *15* (1), 1-20.

71.   Nguyen, Q. T.; Olson, E. S.; Aguilera, T. A.; Jiang, T.; Scadeng, M.; Ellies, L. G.; Tsien, R. Y., Surgery with molecular fluorescence imaging using activatable cell-penetrating peptides decreases residual cancer and improves survival. *P Natl Acad Sci USA* **2010,** *107* (9), 4317-4322.

72.   auf dem Keller, U.; Prudova, A.; Eckhard, U.; Fingleton, B.; Overall, C. M., Systems-level analysis of proteolytic events in increased vascular permeability and complement activation in skin inflammation. *Sci Signal* **2013,** *6* (258), rs2.

73.   Tanco, S.; Lorenzo, J.; Garcia-Pardo, J.; Degroeve, S.; Martens, L.; Aviles, F. X.; Gevaert, K.; Van Damme, P., Proteome-derived peptide libraries to study the substrate specificity profiles of carboxypeptidases. *Mol Cell Proteomics* **2013,** *12* (8), 2096-110.

74.   Schilling, O.; Barre, O.; Huesgen, P. F.; Overall, C. M., Proteome-wide analysis of protein carboxy termini: C terminomics. *Nat Methods* **2010,** *7* (7), 508-11.

75.   Wickstrom, M.; Larsson, R.; Nygren, P.; Gullbo, J., Aminopeptidase N (CD13) as a target for cancer chemotherapy. *Cancer Sci* **2011,** *102* (3), 501-8.

76.   Hitzerd, S. M.; Verbrugge, S. E.; Ossenkoppele, G.; Jansen, G.; Peters, G. J., Positioning of aminopeptidase inhibitors in next generation cancer therapy. *Amino Acids* **2014,** *46* (4), 793-808.

77.   Tokuhara, T.; Hattori, N.; Ishida, H.; Hirai, T.; Higashiyama, M.; Kodama, K.; Miyake, M., Clinical significance of aminopeptidase N in non-small cell lung cancer. *Clin Cancer Res* **2006,** *12* (13), 3971-8.

78.   Novakova, Z.; Cerny, J.; Choy, C. J.; Nedrow, J. R.; Choi, J. K.; Lubkowski, J.; Berkman, C. E.; Barinka, C., Design of composite inhibitors targeting glutamate carboxypeptidase II: the importance of effector functionalities. *FEBS J* **2016,** *283* (1), 130-43.

Chapter V

Conclusions and Future Directions

A. Summary

My PhD studies primarily focused on developing and applying a mass spectrometry-based protease profiling technology named Multiplex Substrate Profiling by Mass Spectrometry, along with proteomics, peptidomics and enzyme kinetics assays to characterize proteases' activities and substrate specificities[1-6]. These knowledges enable us to identify their endogenous substrates as well as develop chemical tools, such as fluorescent substrates and inhibitors, to better understand their roles in biological processes and disease development. Additionally, we further extended our methods to investigate proteolytic activities in complex biological samples such as lung cancer secretions and chromaffin granules[6]. We are expecting to establish a platform to functionally profile proteolytic activities in disease and healthy sites, allowing us to 1) identify proteases as potential biomarkers of diseases and 2) develop protease inhibitors and protease-activated drugs for the treatment.

B. Future directions

Although MSP-MS has been demonstrated as a powerful technology to characterize protease specificity, it also suffers from some limitations. Firstly, even though substrate library was rationally designed and comprehensive, it still lacks certain amino acid combinations, especially amino acid pairs at two termini such as KR, RR, KK. The lack of these amino acid combinations limits its application in studying exopeptidases. A library with a more comprehensive coverage of amino acid combinations will be needed. Secondly, the library consists of peptides with unmodified termini, allowing it to be used to characterize exopeptidase specificity. However, when studying proteolytic activities in complex biological samples, endopeptidase cleavage products are often further cleaved by exopeptidases. Therefore, it is

difficult to isolate activities of endopeptidases and exopeptidases. A second MSP-MS library with termini-modified peptides (for example, N-terminus acetylated and C-terminus amidated) will be useful to only assaying endopeptidase activity in complex samples.

Many protease substrate profiling methods have been focusing only on purified or recombinant proteases. In Chapter III and IV, we have demonstrated our methods in studying complex biological samples containing multiple active proteases. To establish a platform to systematically study protease-substrate interactions, we need to not only develop better substrate profiling technologies but also develop better tools and methods to deconvolute the complexity. Ultimately, methods such as gene knockout, RNAi and protease inhibitors with high specificity are needed to be introduced to completely dissect activities into individual proteases. And bioinformatics methods also need to be developed to integrate results of different experiments.

MSP-MS has shown its success in characterizing protease specificity, and based on the specificity profile, we can design fluorescent substrates and peptide mimetic inhibitors. We have been collaborating with parasitologists and medicinal chemists to develop selective proteasome inhibitors for Schistosoma[7] and Trichomonas[8]. On the one hand, we are performing MSP-MS assays to determine the substrate specificities of parasite proteasomes as well as human proteasomes to identify the differences. We will also solve the structure of parasite proteasomes to gain structural insights of their binding pockets. On the other hand, we are conducting high throughput screenings of small molecule compounds for structure–activity relationship analysis. We will eventually combine the knowledge we obtained from both two paths together to guide our study. Similar to developing selective proteasome inhibitors, the same rationale can be applied in developing protease-activated drugs for diseases such as cancer. Functionally characterization of the proteolytic activity in both disease tissues and healthy tissues, along with

proteomics and genomics, will be conducted to identify disease-associated protease. MSP-MS assays can be performed on this disease-associated protease and its substrate specificity profile will be used to guide the design of the linker for protease-activated drugs.

C. References

1.   Leontovyc, A.; Ulrychova, L.; O'Donoghue, A. J.; Vondrasek, J.; Maresova, L.; Hubalek, M.; Fajtova, P.; Chanova, M.; Jiang, Z.; Craik, C. S.; Caffrey, C. R.; Mares, M.; Dvorak, J.; Horn, M., SmSP2: A serine protease secreted by the blood fluke pathogen Schistosoma mansoni with anti-hemostatic properties. *PLoS Negl Trop Dis* **2018,** *12* (4), e0006446.

2.   Li, H.; Goh, B. N.; Teh, W. K.; Jiang, Z.; Goh, J. P. Z.; Goh, A.; Wu, G.; Hoon, S. S.; Raida, M.; Camattari, A.; Yang, L.; O'Donoghue, A. J.; Dawson, T. L., Jr., Skin Commensal Malassezia globosa Secreted Protease Attenuates Staphylococcus aureus Biofilm Formation. *J Invest Dermatol* **2018,** *138* (5), 1137-1145.

3.   Xu, J. H.; Jiang, Z.; Solania, A.; Chatterjee, S.; Suzuki, B.; Lietz, C. B.; Hook, V. Y. H.; O'Donoghue, A. J.; Wolan, D. W., A Commensal Dipeptidyl Aminopeptidase with Specificity for N-Terminal Glycine Degrades Human-Produced Antimicrobial Peptides in Vitro. *ACS Chem Biol* **2018,** *13* (9), 2513-2521.

4.   Arutyunova, E.; Jiang, Z.; Yang, J.; Kulepa, A. N.; Young, H. S.; Verhelst, S.; O'Donoghue, A. J.; Lemieux, M. J., An internally quenched peptide as a new model substrate for rhomboid intramembrane proteases. *Biol Chem* **2018,** *399* (12), 1389-1397.

5.   Beekman, C.; Jiang, Z.; Suzuki, B. M.; Palmer, J. M.; Lindner, D. L.; O'Donoghue, A. J.; Knudsen, G. M.; Bennett, R. J., Characterization of PdCP1, a serine carboxypeptidase from Pseudogymnoascus destructans, the causal agent of White-nose Syndrome. *Biol Chem* **2018,** *399* (12), 1375-1388.

6.   Lapek, J. D., Jr.; Jiang, Z.; Wozniak, J. M.; Arutyunova, E.; Wang, S. C.; Lemieux, M. J.; Gonzalez, D. J.; O'Donoghue, A. J., Quantitative Multiplex Substrate Profiling of Peptidases by Mass Spectrometry. *Mol Cell Proteomics* **2019**.

7.   Bibo-Verdugo, B.; Wang, S. C.; Almaliti, J.; Ta, A. P.; Jiang, Z.; Wong, D. A.; Lietz, C. B.; Suzuki, B. M.; El-Sakkary, N.; Hook, V.; Salvesen, G. S.; Gerwick, W. H.; Caffrey, C. R.; O'Donoghue, A. J., The Proteasome as a Drug Target in the Metazoan Pathogen, Schistosoma mansoni. *ACS Infect Dis* **2019,** *5* (10), 1802-1812.

8.   O'Donoghue, A. J.; Bibo-Verdugo, B.; Miyamoto, Y.; Wang, S. C.; Yang, J. Z.; Zuill, D. E.; Matsuka, S.; Jiang, Z.; Almaliti, J.; Caffrey, C. R.; Gerwick, W. H.; Eckmann, L., 20S Proteasome as a Drug Target in Trichomonas vaginalis. *Antimicrob Agents Chemother* **2019,** *63* (11).