Title

Identification and exploration of apoptotic and caspase proteolytic substrates.

Author

Seaman, Julia

Publication Date

2016

Peer reviewed|Thesis/dissertation

Identification and exploration of apoptotic and caspase proteolytic substrates.

by

Julia Elizabeth Seaman

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedication

To my husband, Nate, and my family, Mom, Dad and Chris.

Acknowledgements

I am very grateful for the advice, assistance and support from the Wells Lab. My advisor, Jim Wells, has created a wonderful laboratory environment for collaborative research and friendships. The lab, and members in it, could not have survived without the innumerable skills of Marja Tarr. My graduate career has been greatly assisted by sharing the experiences with fellow students Ashley Smart and Charlie Morgan in the lab. It has been fun to explore science and solve the big and small problems with them and the other graduate students in lab- Justin Rettenmaier, Hai Tran, Sam Pollock, and Alex Martinko. I have relied heavily on everyone else who counts himself or herself as a Wellsome member. The postdocs- especially Kazu Shimbo, Nick Agard, Arun Wiita, Jason Porter, Juan Diaz, JT Koerber, Nathan Thomsen, Duy Ngyuen, Peter Lee, Min Zhuang, Zach Hill and Olivier Julien- have helped improved my scientific skills and graduate school experience throughout my time in the lab. Finally, my scientific work relies heavily on many members who came before me.

UCSF has been a great environment to learn and explore during my PhD. I am very thankful my program, PSPG and its staff- Deana Kroetz, Debbie Acoba and Rebecca Brown. The PSPG program also introduced me to my classmates who have helped relieve the stress of science throughout the years. Additionally, working at UCSF has exposed me to a great many number of core facilities and collaborators who have contributed to all facets of my work. First, I would like to thank the Mass Spectrometry Facility and its team under Al Burlingame, especially Dave Maltby and Robert Chalkley. I have also have worked with the Preclinical Therapeutic Core under Byron Hann and Don Hom and the MMTI with Blake Aftab.

Finally, I would like to thank, Seth Ginsberg and Louis Tharp of the Global Healthy Living Foundation, for their fellowship and guidance.

Abstract

Apoptosis, a type of programmed cell death, is a universal and essential cellular function. There are numerous homeostatic biological roles of apoptosis and many diseases have or cause mys- or dis- regulated apoptosis, most notably cancer's escape from apoptotic signals. Apoptosis has many distinctive characteristics including membrane blebbing, chromatin condensation, and most importantly for these studies, the activation of a class of protease proteins called caspases. Caspases are cysteine aspartic proteases with a unique preference to cleave hundred of substrates after acidic residues, especially aspartate. These cleavage events can activate, modify, or inhibit the substrate's function, which leads to the dismantling of the cell during apoptosis. This process is well conserved throughout metazoan evolution, with parallel pathways in coral, fish, flies, worms and mice.

To study apoptotic protease activity, the Wells Lab has developed a proteomic-based technique. This technique is an unbiased positive enrichment labeling mass spectrometry protocol. While the protocol was developed for studying caspase substrates during apoptosis in human cell culture, it is very versatile, allowing for use in almost any protein sample like perturbed cellular lysates, different species and primary samples. Chapter 1 covers the protocol specifics and uses, summarizing a decade's worth of optimization and application.

The DegraBase is the compilation of 44 different experiments using the N-terminal labeling method to examine apoptotic proteolytic activity described in Chapter 2. While much of the experimentation work was completed before I started the project, I completed compilation and standardization of raw data, and worked with Emily Crawford on the analysis. This global analysis reveals there is a large increase in proteolytic activity after apoptotic induction, and caspases account for 25% of the newly created fragments. Within caspase substrates, there is no

single biological process or sub-cellular location that is targeted, as caspases appear to cleave substrates throughout all the different pathways of the cell. Additionally, this database is also a good resource for endogenous proteolysis, including free methionines, and signal and transis peptide processing.

Analysis of the DegraBase also reveals evolutionary and biological discoveries. As the apoptotic pathway to activate caspases is highly conserved throughout metazoans, we wanted to investigate the conservation of caspase substrates. In Chapter 3, the comparison of caspase substrates in worm, fly, mouse and human reveals a hierarchal structure. My contribution includes the murine dataset and comparison analysis in collaboration with Emily Crawford. We found caspase cleavage is highly conserved at the pathway level, while individual targets and sites are not as well conserved the more distant the animals.

The unbiased and large size of the DegraBase also reveals broader caspase activity than previously described. Caspases had been assumed to have absolute specificity for aspartate and no activity for glutamate. However, with the large dataset, in Chapter 4, I reveal significant activity after glutamate and even potential activity after phosphoserine. This activity is verified through biochemical assays and x-ray crystallography, and expands the number of apoptotic caspase substrates by 15%.

As many chemotherapeutics induce apoptosis, apoptotic, especially caspase, proteolytic substrates may be good biomarkers of treatment efficacy. The development of mouse models and a positive control are discussed in Chapter 5. These models utilize the DegraBase and labeling technology to identify peptides enriched in the blood and tumor specific to treatment responding mice as potential biomarkers.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 1

Global analysis of cellular proteolysis by selective enzymatic labeling of protein N-termini

This work was published in the Methods in Enzymology: Regulated Cell Death:
Wiita AP, Seaman JE, Wells JA. Global analysis of cellular proteolysis by selective enzymatic labeling of protein N-termini. Methods Enzymol. 2014;544:327–358.

Abstract:

Proteolysis is a critical modification leading to alteration of protein function with important outcomes in many biological processes. However, for the majority of proteases we have an incomplete understanding of both the cleavage substrates and downstream effects. Here, we describe detailed protocols and applications for using the engineered peptide ligase, subtiligase, to specifically label and capture protein N-termini generated by proteases either induced or added to biological samples. This method allows identification of the protein targets as well as their precise cleavage locations. This approach has revealed >8000 proteolytic sites in healthy and apoptotic cells including >1700 caspase cleavages. One can further determine substrate preferences through rate analysis with quantitative mass spectrometry, physiological substrate specificities, and even infer the identity of proteases operating in the cell. In this chapter we also describe how this experimental method can be generalized to investigate proteolysis in any biological sample.

1. Introduction

1.1 Importance of proteolysis

Proteolysis, the hydrolysis of peptide bonds by proteases, is an essential activity in a wide range of cellular functions. Proteases exist in virtually all forms of life, and are classified into five mechanistic categories: serine, threonine, cysteine, acid, and metallo (1). In humans alone there are over 550 identified proteases, but their precise roles and substrates are generally poorly understood. Protease specificity differs widely, ranging from a single site on a single substrate to cleaving a broad swath of the proteome. In eukaryotes, proteolysis is involved in digesting and recycling proteins, irreversible post-translation modification via N-terminal methionine processing, signal or transit peptide removal, cleavage of single polypeptide chains into their multiple components, and removal of precursor domains. In addition to their role in protein maturation and function, proteolysis is critical for physiological processes from blood clotting to apoptosis, a form of programmed cell death. Endoproteolysis, cleavage within a protein, can lead to activation, inhibition, or a change of substrate function, allowing proteases to play an important role in signaling. Mis- and dysregulation of proteolysis activities are hallmarks of many pathological states such as arthritis, inflammation, and cancer. Additionally, proteases are used as tools in the laboratory, industrial manufacturing, and commercial products.

While some proteases have been intensely studied, we have a very limited understanding of the majority of these significant enzymes. A first step to understanding a protease's biological role is identification and validation of substrates and cleavage locations. After identification, the significance of protease activation and individual cleavage events can be investigated, leading to novel targets for drug therapy and greater understanding of these ubiquitous biological mechanisms. Such information leads naturally to quantitation of the rates of cleavage and examining the specific functional consequences for individual targets. These investigations facilitate understanding the importance of these events in biological processes. There has been a surge in the development of technologies for global and unbiased characterization of proteolysis in complex biological samples (2-5). We will briefly review the state-of-the-art in this field and then focus on the detailed implementation and applications of the N-terminomics technology developed in our lab using subtiligase.

1.2 Approaches to substrate cleavages and identification

Historically, the identification of proteases responsible for specific cleavage events has often been driven by knowledge of important substrate proteins that were found cleaved in a biological process. For example, insulin was known to be produced from the precursor pro-insulin and this motivated the discovery of furin, the responsible protease (6). The processing of pro-IL-1β to IL-1β further led to the discovery of the protease caspase-1 (7). Similarly, until recently, most substrates have been found in a labor intensive, candidate-based approach using a range of focused biochemical approaches.

Recently, proteomic methods have allowed for unbiased searches of proteolytic substrates in complex samples (Table 1). These global studies have the goal to identify all cleavage events in a cell during a particular process or, alternatively, all possible substrates of a protease. These aims have been greatly aided by the advancements of analytical instruments, specifically liquid chromatography coupled to mass spectrometry (LC-MS). Most methods enrich for proteolytically cleaved peptides by taking advantage of the newly created α-carboxy-

or α-amino-termini on either side of the cleavage site. This allows for capture and purification of substrates through specific chemical or enzymatic modification. A single global experiment can generate over a thousand peptide identifications that can be scored and mapped to a specific protein and/or cleavage site.

Method	**Description**	**Proteolytic substrates reported**	**Reference**
Subtiligase	Positive selection of free N-termini α-amines through subtiligase enzymatic labeling with an ester peptide tag.	8090 peptides (1706 caspase) from untreated and apototic human cells	(8): Crawford et al., 2013; (9): Mahrus et al., 2008

COFRADIC	COmbined FRActional DIagonal Chromatography uses negative selection with a chemical modification at free N-termini (or other modification of interest) to enable separation of modified from unlabeled peptides during chromatorgraphy.	68 caspase substrates from recombinant caspase -2, -3, -7; 9729 carboxypeptidase substrates from *in vitro* peptide library	(10): Staes et al., 2008;
			(11): Tanco et al., 2013;
			(12): Wejda et al., 2012

TAILS	Terminal Amine Isotopic Labeling uses chemica modifications of protein amines and thiols, sample trypsinization and negative selection to enrich for neo N- or C-termini.	288 MMP-2 cleavage sites; >100 GluC cleavage sites	(13): Kleifeld et al., 2010; (14): Schilling et al., 2011
N-CLAP	N-terminalomics by Chemical Labeling of the α-Amines of Proteins. Uses Edman degradation chemistry to block lysine amines to label N-terminal amines with a biotinylated tag for positive selection.	278 peptides (23 caspase) in apoptotic jurkat cells	(15): Xu and Jaffrey, 2010; (16): Xu et al., 2009
PROTOMAP	PROtein TOpography and Migration Analysis Platform creates visual peptographs from 1D SDS gel migration patterns and sequence coverage from MS of in-gel digestions to identify cleavages from mass shifts.	744 proteins with cleavages in apoptotic Jurkat cells	(17): Dix et al., 2008; (18): Dix et al., 2012
2D DiGE + MS	Two-dimensional differential gel electrophoresis (2D-DiGE) separates complex mixtures using orthogonal electrophoresis methods and comparison of induced proteolysis and control sample gels reveal shifted spots dut to proteolysis.	21 caspase substrates in Jurkat cells	(19): Tonge et al., 2001

1D gel + MS	Lysates harvested from in vivo induced proteolysis are run on a large gel, separated into 100 slices and prepared for mass spectrometry with in gel trypsinization. Substrates are identified as those with less mass than expected values indicating cleavage events.	37 peptides in apoptotic Jurkat cells	(20): Thiede et al., 2005
2D SDS PAGE	2D SDS PAGE gel electrophoresis with protease addition as an intermediate step to look for spots that differentially migrate compared to control indicating proteolysis.	41 Caspase-1 substrates in THP cells	(21): Shao et al., 2007
ProC-TEL	Positive selection through carboxy termini tagging using transpeptidation enzymatic reaction.	76 peptides from E. coli lysates	(22): Xu et al., 2011

Table 1. A summary of current methods for proteolytic cleavage site and substrate identification.

2. Applications

2.1 Introduction

Here, we describe a global "N-terminomics" positive enrichment method using the engineered enzyme subtiligase. This method allows one to specifically tag and identify with LC-MS new N-termini generated by endogenous or exogenous proteases (Figure 1A). With this approach we have identified over 8000 unique α-amines in healthy and apoptotic cell lysate (publicly available at wellslab.ucsf.edu/degrabase) as well as quantitatively monitor kinetics of individual cleavage events. The subtiligase method is easily applied to many different sample types and biological questions involving proteolysis.

2.2 Types of samples and peptide tags

The method has been successfully used on purified proteins, cell cultures, peripheral blood plasma, and tissue samples from humans, mice, insects and worms. In general, proteins can be solubilized into an appropriate buffer for subtiligase activity, as described below, and labeling can occur in essentially any biological sample. Additionally, one is free to design the chemical structure of the peptide ester to facilitate downstream purification, identification and quantitation for customization in specific applications (Figure 1B).

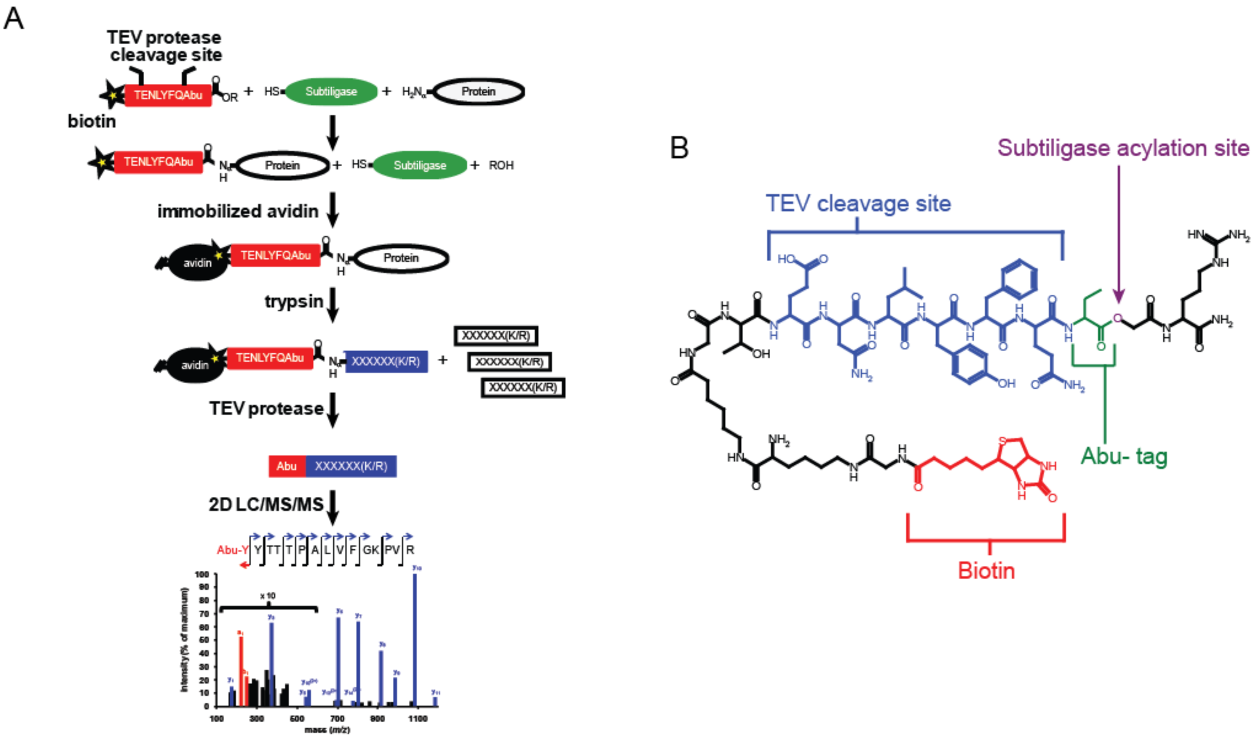


Figure 1. An overview of the subtiligase N-terminal labeling method. A. Proteins with free N-termini in a mixture are selectively tagged using the engineered enzyme, subtiligase. Whole protein samples are incubated with subtiligase and the peptide ester containing a biotin tag. After enzymatic labeling, free N-termini are captured on avidin beads. Proteins are digested by trypsin. The final N-terminal peptide is released from beads via TEV protease cleavage and identified by mass spectrometry. **B.** The current peptide ester contains an ester subtiligase acylation site, Abu-tag for positive mass spectrometry identification, a TEV protease site and a biotin label. The peptide ester can be further modified for specific experimental needs.

2.3 Sample setup introduction

2.3.1 Discovery vs Targeted protocols

There are two experimental protocols we currently use to provide the most complete information for global N-terminomics. The initial "Discovery" experiments are designed to identify which proteins and specific sites are cleaved. These discovery experiments are qualitative and focus on high confidence identification of tagged peptides. These experiments optimize the labeling procedures, determine background, and establish a list of high confidence peptide identification. These high confidence peptides from Discovery experiments are then used in "Targeted" mass spectrometry experiments. Targeted experiments allow for the specific monitoring of a subset of peptides in a more sensitive and/or quantitative manner across a wider range of samples. Examples of both types of experiments will be discussed in more detail below.

2.3.2 Forward vs Reverse experiments

There are two main experimental strategies for subtiligase labeling, which we term "Forward" and "Reverse" (Figure 2). Forward experiments use intact biological systems where endogenous proteolysis is either induced or monitored at baseline, followed by protein isolation and N-terminal labeling. In contrast, Reverse experiments are *in vitro* controlled experiments where an exogenous protease is added to the total protein from a sample where endogenous proteases have been inactivated. The Forward experiments allow for the identification of biologically relevant protease cleavage events but may not be able to identify the specific protease responsible. The Reverse setup specifically identifies the activity of the added protease or inducer, but it can include non-biologically relevant events that may not occur in an intact cell or organism.

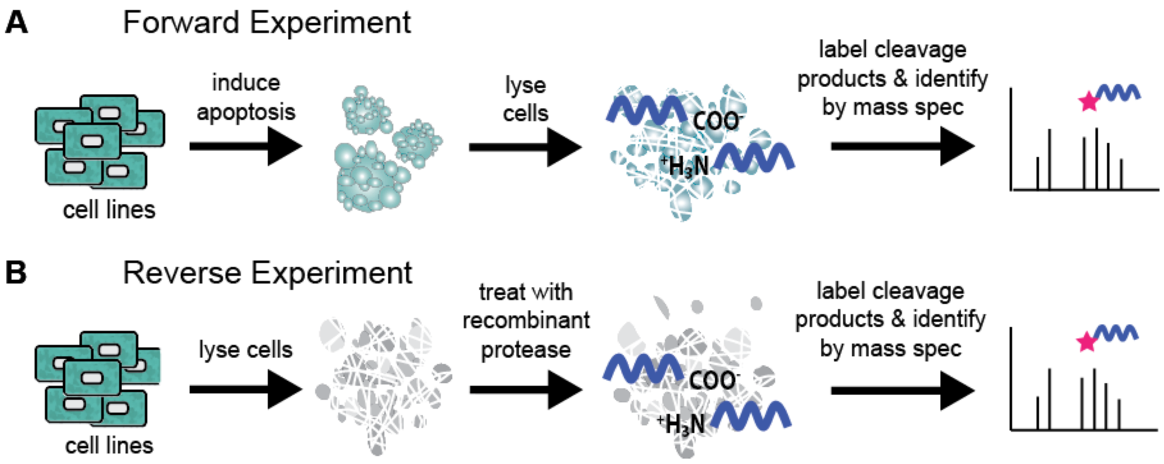

A
Forward Experiment
induce apoptosis
lyse cells
label cleavage products & identify by mass spec
cell lines
COO^-
^+H_3N
B
Reverse Experiment
lyse cells
treat with recombinant protease
label cleavage products & identify by mass spec
cell lines
COO^-
^+H_3N

Figure 2. A schematic difference between Forward and Reverse experiments. Forward experiments use samples from intact biological systems, either perturbed or unperturbed, that are then harvested, lysed, and labeled. Reverse experiments involve exogenous addition of protease to whole cell or tissue lysate of interest followed by labeling.

3. Subtiligase-Based Labeling Method

3.1 Overview of Method

The subtiligase protocol is designed to positively enrich N-termini from newly cleaved proteolytic substrates. Subtiligase itself is an engineered version of the bacterial serine protease BPN' subtilisin. Two point mutations simultaneously abolish protease activity and allow ligase activity (23). With these modifications subtiligase can covalently link free peptide α-amines with an ester-containing synthetic peptide. Importantly, the subtiligase enzyme is exquisitely selective for peptide α-amines over the ε-amines of lysine residues (24). Furthermore, acetylated N-termini present on 80-90% of native eukaryotic proteins (25) are ignored by the labeling process, greatly reducing background identifications.

The protocol workflow follows a catch-and-release strategy (Figure 1A). In combination with a designed synthetic peptide ester (Figure 1B), subtiligase is first used to selectively biotinylate free α-amines in a sample. Following avidin bead-mediated immobilization, proteins are then digested with trypsin and non-biotinylated protein fragments are washed away. The most N-terminal peptide from each substrate is then released from avidin beads through Tobacco Etch Virus (TEV) protease cleavage. TEV is an extremely specific plant virus protease that can be readily purified (26) or purchased commercially. TEV recognizes the amino acid sequence ENLYFQ|S, which importantly is not found in the mammalian proteome. After TEV cleavage, all labeled peptides have a non-natural amino acid mass tag (α-aminobutyric acid, or Abu-) remaining on the N-terminus. This tag, compatible with both subtiligase and TEV, greatly enhances confidence for identifying subtiligase-labeled peptides over non-specifically bound background. In our experience, >90% of peptides observed by LC-MS incorporate the Abu-mass tag, providing evidence for the specificity of the labeling procedure and recovery method.

3.2 Specialized Reagents for Subtiligase Labeling and Enrichment

3.2.1 Subtiligase enzyme

Plasmid vectors and detailed instructions for subtiligase expression are available on request from the Wells laboratory. Subtiligase is expressed in *B. subtilis* and the enzyme is secreted to the media. The enzyme is purified through ammonium sulfate precipitation, anion exchange, thiopropyl resin capture (for the catalytic cysteine residue in subtiligase), and gel filtration. The enzyme is stored at -80°C and retains activity for at least two years after purification. Activity can be tested and quantified using FRET ester reporters (27, 28).

3.2.2 Peptide ester label

The synthetic ester used for labeling is customizable for different experimental goals (28). The current version contains four distinct features: (i) an ester linkage for subtiligase acylation and transfer to the free peptide α-amine, (ii) the unique Abu- tag to facilitate MS identification, (iii) the TEV protease cleavage site for elution from avidin beads, and (iv) biotin for initial capture (Figure 1B). The peptide ester is synthesized in house using solid phase fMOC chemistry modified for the more reactive ester bond (24, 29). Since each amino acid is added individually, it is possible to change any part of the sequence after the ester bond so long as the first four amino acids can be recognized by subtiligase. A wide range of other N-terminal modifications are therefore also possible.

4. Experimental applications of subtiligase-based N-terminomics

4.1 Application to cell culture systems undergoing apoptosis (Forward Discovery experiments)

Below we describe a general protocol using subtiligase-based labeling to identify proteolytic substrates generated during apoptosis in a cell culture system.

4.1 Experimental Design

4.1.1 Sample size and expected yield.

The extent of subtiligase labeling varies but we estimate about 10-15% of the α-amines in a sample are routinely labeled. Hydrolysis of the ester by subtiligase is the biggest impediment to higher labeling efficiency. While the enzyme is very suitable for N-terminomics despite the hydrolysis side-reaction, it requires that the use of larger sample volumes. For initial Forward Discovery experiments we typically use a minimum of 5 x 10^8 cells (and up to 24 x 10^9 cells for large-scale Discovery experiments (27)) to maximize our number of peptides identified by mass spectrometry. The use of more sensitive mass spectrometers or an experimental system which does not require deep coverage allows for smaller amounts of starting sample.

4.1.2 Choice of proteolysis inducer for Forward experiments.

The specific proteolysis inducer chosen will depend on the system of interest and research question. Apoptosis can be induced in a cell culture system using a variety of agents. For example, we have used both small molecule cytotoxic agents (doxorubicin, bortezomib, staurosporine) to activate the intrinsic pathway of apoptosis and protein-based agents (TRAIL, Fas-ligand) that bind to extracellular surface death receptors and trigger the extrinsic pathway of apoptosis. (9, 27, 30).

4.1.3 Monitoring proteolysis and sample harvest

Once the desired cell culture system and inducer are chosen, it is recommended to perform validation experiments on a small scale to identify a concentration and time point where proteolysis is most extensive. For apoptosis, we find the maximal number of caspase cleavage events when the extent of apoptosis is over 90%. We have primarily used biochemical assays to monitor caspase activity (Caspase-Glo, Promega) and cell viability (Cell-Titer Glo, Promega), though there are a number of other experimental methods also available (31). Figure 3A demonstrates a typical time course of cell viability and caspase activiation with two different doses of apoptotic inducers in two different human malignancy-derived cell lines.

After cells grown to scale for Forward Discovery have undergone apoptosis to the desired extent, cell bodies and debris are pelleted by centrifugation and washed once with ice-cold PBS. The washed cell pellet can then be lysed directly or flash-frozen, stored at -80°C, and thawed prior to lysis. For comparison to background cellular proteolysis, it is desirable to include a control sample not exposed to proteolysis inducers.

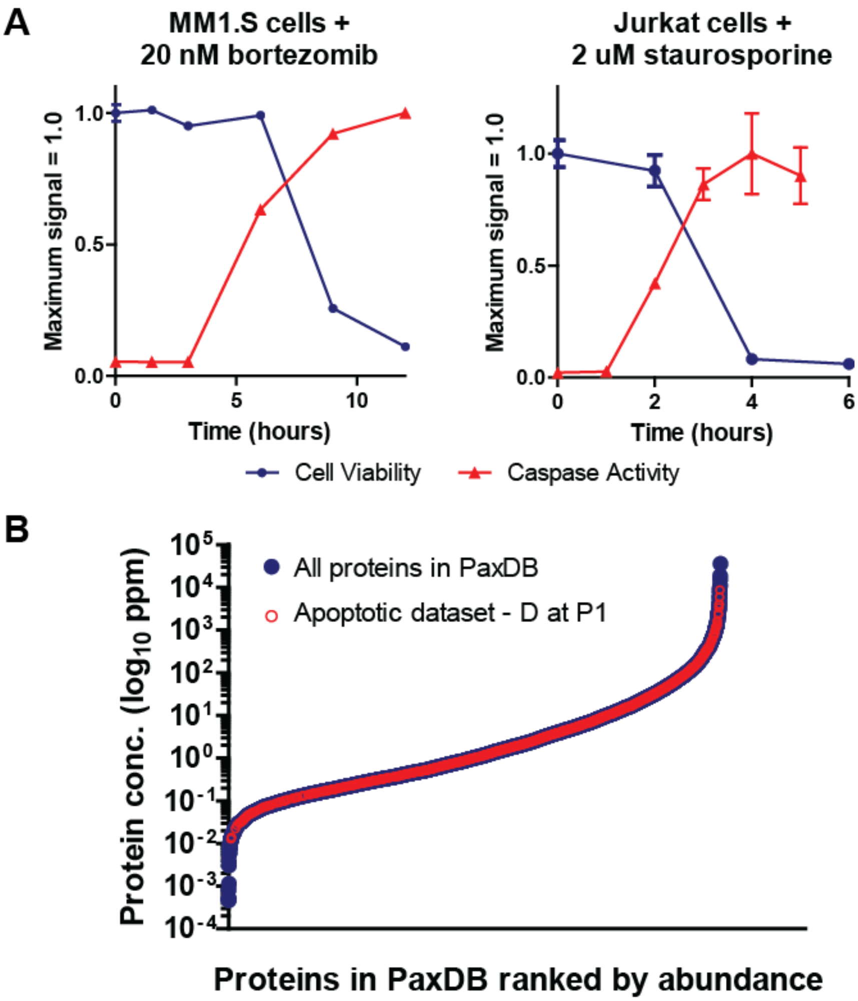


Figure 3. Monitoring Apoptosis and Proteomic Distribution of Cleavage Substrates. A. Biochemical monitoring of cell viability and caspase activation**.** It is important to monitor apoptosis vs time after exposure to drug, as the rate of apoptosis can vary substantially. Caspase activity appears before cell viability decreases. **B.** Comparison of caspase substrates identified versus broad range of baseline protein abundance. Protein abundance estimated derived from PaxDB. Extensive distribution overlap indicates that subtiligase-based N-terminomics leads to broad coverage across the proteome. Figure adapted from (8) with permission of the authors.

4.1.4 Protocol for Forward Discovery Labeling

1. Cell lysis. Prepare 1 mL per sample of 4x lysis buffer (ratio 4:4:2 of 10% SDS (w/v):1 M bicine pH 8.5:ddH_2O). Also prepare stocks of protease inhibitors (Sigma) to quench ongoing endogenous proteolysis: 10 mM z-VAD-fmk caspase inhibitor in DMSO; 10 mM E-64 cysteine protease inhibitor in DMSO; 100 mM AEBSF serine protease inhibitor in DMSO; 0.5 M EDTA pH 8.0 in ddH_2O; 100 mM PMSF (freshly prepared) in isopropanol. Add 5 µL of each protease inhibitor stock per 1 mL of 4x lysis buffer. Add 1 mL 4x lysis buffer to cell pellet and lyse completely by probe ultrasonication. Use lysate sample for protein concentration determination.

2. Cysteine reduction and alkylation. In all proteomic experiments, it is important to first reduce and then irreversibly block free thiol groups on cysteines to prevent formation of mixed oxidation products that hinder identification by MS. Prepare a fresh stock of 100 mM TCEP in ddH_2O. Add 80 µL of TCEP stock to each lysed sample and mix. Heat at 95°C for 15 min to reduce cysteines. Allow to cool to room temperature (RT). During cooling, prepare a fresh stock of 200 mM iodoacetamide (IAM). Add 80 µL of IAM stock to each sample and mix. Incubate 1 hr in the dark at RT. After incubation, add 40 µL of 1M dithiothreitol (DTT) stock to quench remaining IAM, as any free IAM will block catalytic cysteines of subtiligase at next step. Vortex briefly. Add 400 µL of 10% Triton-X (v/v) to form micelles with SDS (note: subtiligase labeling with not work without removing detergent in this manner).

3. Subtiligase labeling. Centrifuge each sample for 10 min, ~4000 x *g* to pellet out any insoluble debris and transfer supernatant to new tube. Add ddH_2O to final volume of 3.6 mL. Check and adjust pH to 8.5. Add 400 µL 10 mM peptide ester stock in DMSO to final concentration of 1 mM. Vortex briefly. Add 40 µL 100 µM subtiligase stock to final concentration of 1 µM. Vortex briefly. Incubate for 1 hr at RT (note: labeling for >1 hr generally does not improve yields as peptide ester is either ligated to N-termini or hydrolyzed by this time). Labeling can be confirmed through a western blot against biotin using NeutrAvidin-HRP (Pierce) with comparison to a pre-subtiligase sample.

4. Removal of excess peptide ester and exchange into denaturing conditions by protein precipitation. Biotin moieties on excess and hydrolyzed peptide ester will occupy binding sites on avidin beads and render them unavailable for capturing labeled proteins. Therefore, precipitate proteins by adding labeled sample dropwise to 35 mL acetonitrile at RT; short peptides, including excess peptide ester, will remain in solution. Vortex gently. Incubate on ice for at least 15 min up to overnight. Centrifuge 8000 x *g* for 30 min at 4° C. Carefully decant supernatant to waste. Let precipitated protein pellet air dry for ~15 min. Add 1 mL 8 M Guanidine HCl over pellet and let dissolve at RT for 30 min-1 hr. Swirl gently and pipet up and down to dissolve pellet. Add another 1 mL 8 M Guanidine HCl to fully dissolve; use ultrasonication if necessary. Precipitate protein a second time by adding dropwise to 30 mL ice-cold ethanol in new 50 mL conical vial. Incubate at -80°C overnight. The next day, centrifuge sample at 30 min, 8000 x *g*, 4°C to pellet precipitated protein. Decant supernatant and air dry pellet for 15-20 min (note: can also freeze pellet at -80° C for later use). Add 3 mL 8 M Guanidine HCl over pellet and

let dissolve at RT for 20 min. Add an additional 2 mL Guanidine HCl to complete dissolution. Transfer dissolved protein to new 15 mL conical vial. Rinse prior 50 mL conical vial with 2.5 mL ddH_2O, and transfer rinsed solution to same 15 mL conical. Take 8 μL aliquot of total dissolved protein sample for dot blot (below) and store at 4°C.

5. Capture on NeutrAvidin resin. We use NeutrAvidin High Capacity resin (Pierce) to maximize capture of biotinylated proteins. For protein from 5×10^8 cells we will typically add 1 mL of 50% bead slurry. After adding beads, place overnight at RT with gentle agitation or rotation. A dot blot against NeutrAvidin-HRP (Pierce) is recommended to confirm complete capture of labeled peptides compared to the pre-bead aliquot, indicated by disappearance of luminescence signal in the post-bead sample. If capture is not complete, add additional NeutrAvidin resin and incubate for 2h up to overnight, dot blot again, and repeat as necessary. Note that incompletely removed peptide ester will increase the amount of beads necessary for complete capture.

6. On-bead trypsinization. After complete peptide capture, transfer beads to empty polypropylene chromatography column with frit at outlet. Attach the column to vacuum set up and remove the supernatant. Wash beads (add buffer, vortex, flow through) three times with 2 mL biotin wash buffer (10 mM bicine pH 8.0, 1 mM biotin) to occupy unbound avidin sites. Wash beads 5-10x with 5M Guanidine HCl to remove non-specifically bound protein from beads. Wash beads 3x in trypsin wash buffer (100 mM bicine pH 8.0, 200 mM NaCl, 20 mM $CaCl_2$, 1M Guanidine HCl). Add 10-100 μg sequencing grade modified trypsin (trypsin should be added at 1:50 (w:w) to estimated

amount of captured protein) in trypsin wash buffer to each sample. Incubate overnight at 37°C with gentle agitation.

7. N-terminal peptide elution with TEV protease. Freshly prepare TEV protease buffer (50 mM ammonium bicarbonate pH 8.1, 2 mM DTT, 1 mM EDTA). Remove trypsinization supernatant to waste. Wash beads 5-10x with 5 M Guanidine HCl to remove non-specifically bound peptides. Wash beads 5x with TEV protease buffer to completely remove guanidine. For each sample mix 50 µg TEV protease with 1.5 mL TEV protease buffer and add to beads. Incubate overnight at RT with agitation or rotation. The next day, elute supernatant with TEV-cleaved peptides into 1.5 mL tubes. Evaporate to dryness.

8. Sample clean-up by ZipTip. Resuspend sample in a total of 100 µL 5% TFA to achieve pH ≤ 3. Let stand >10 min at RT. Spin 10 min at 14,000 x *g* at room temperature to pellet precipitated TEV protease. Transfer supernatant to new tube. For clean-up we use C18 ZipTips (Millipore) performed with manufacturer protocol with elution into low-protein retention 500 µL tubes. Evaporate to dryness. Peptides may now be stored at -80°C, resuspended in 0.1% FA to run directly on mass spectrometer, or used for fractionation.

9. Fractionation using reverse-phase high pH chromatography (optional). To obtain the greatest depth of peptide coverage and greatest number of substrate identifications in Discovery experiments it is advisable to perform a separation step prior to MS analysis.

We use reverse-phase high pH chromatography as it has been shown to offer similar separation capabilities to strong cation exchange and does not require additional ZipTip clean up of fractions (Yang, F., Shen, Y., Camp, D.G., 2nd, and Smith, R.D., 2012). If desired, after separation fractions can be pooled for analysis. Evaporate to dryness. Store at -80°C or resuspend each fraction in 0.1% FA for evaluation by MS.

4.1.5 Mass Spectrometry Analysis and Bioinformatics

Mass spectrometry analysis of samples is carried out essentially like any other proteomic-based method, incorporating low-pH reverse phase chromatography in-line with the mass spectrometer, as described in detail by others (for review see (32)). Samples are analyzed in data-dependent acquisition mode, with exact parameters dependent on instrument used.

1. <u>General protein database search to identify substrates.</u> To identify N-termini, MS data must be searched against a database of known proteins specific for the organism of interest. Such searches can be performed with a variety of resources (33); we typically use Protein Prospector (http://prospector.ucsf.edu). This search algorithm is able to search a semi-tryptic peptide space: while the C-terminus has trypsin cleavage (at Arg/Lys), the N-terminus is allowed to be any amino acid in order to capture all potential proteolytic cleavages. In addition to this feature, Protein Prospector also allows a search with a constant N-terminal Abu- modification, which has a mass orthogonal from any naturally-occurring amino acid. The completed database search across all analyzed fractions (typically at a false discovery rate of <1% based on a decoy database set) results in a list of N-terminal peptides identified in the sample. Further analysis of the sequence

prior to the identified N-terminus (derived from database protein sequence) can reveal proteolytic specificity. In our prior studies, we have found that after apoptosis there is a large increase in caspase-related Asp residues at the P1 position immediately N-terminal to the identified cleavage site. In non-apoptotic samples we find a preponderance of Arg or Lys at P1 (8).

We have found it best for downstream analysis to collect and store all experimental data in a central database. For each experiment we upload the relevant details (date, cell line, Forward or Reverse, inducer, etc.), links to the raw MS files, MS run parameters, protein database search parameters, and the protein database output file. We designed a FileMakerPro database (available at wellslab.ucsf.edu/degrabase) that also contains lookups to UniProtKb (www.uniprot.org) to add further details about the protein. This central database creates an easy and consistent workflow for importing and storing datasets and allows for easy export of data for analysis in another program or publication.

2. Comparison to other substrate databases. The list of potential proteolytic substrates can be compared to existing databases that focus on protease specificity or already identified substrates. Examples include MEROPS (34), TopFind (35), and the DegraBase (8).

3. Analysis of biological function. Additionally, biological function of substrates can be analyzed using tools such as GoMiner (36) or Ingenuity Pathway Analysis (www.ingenuity.com). Comparison to a general database of human cellular protein

abundance, PaxDB (37), demonstrates that subtiligase offers substantial coverage of proteolytic substrates across the concentration range of the cellular proteome (Figure 3B).

4.2 Identifying specific substrates of recombinant proteases in cell lysates (Reverse Discovery experiments)

4.2.1 Introduction to Method

In a Reverse Discovery experiment, the subtiligase-based method is modified to examine cleavage events catalyzed by the addition of a specific recombinant protease. The Reverse protocol is ideal for the study of individual proteases of biological interest and for systems that may not be amenable to cell culture-based experiments.

4.2.2 Choice of sample and preparation

4.2.2.1. Preparation of cell lysate. Like the Forward experiment, the choice of an appropriate cellular system of interest is extremely important. For reverse experiments we also typically start with 5 x 10^8 to 1 x 10^9 cells. Our protocol has been optimized for caspase experiments but can be customized for the requirements of other enzymes.

4.2.2.2. Inhibition of endogenous proteases. In order to reduce background protease activity, we include protease inhibitors in our lysis buffer. However, as the lysis buffer is not typically exchanged before protease addition, it is important to not include any inhibitors that would target the protease of interest. Therefore, for Reverse experiments with caspases, we include EDTA, AEBSF and PMSF but omit the cysteine protease inhibitors z-VAD-fmk and E-64.

.

4.2.3 Preparation of proteases and characterization in lysate system

The purity and activity of the protease of interest is the most important part of a Reverse experiment. For example, for caspase-3, we express and purify our own enzymes and then perform kinetic activity assays using z-DEVD-AFC fluorescent substrate (CalBioChem) in either a small volume (50 µL-1 mL) of cellular lysate or optimal enzyme buffer. We use a range of caspase-3 concentrations from 1 nM-1000 nM, near the expected physiological caspase concentration of ~50-200 nM. Ideally, activity in both lysate and buffer will be similar. Experiments at this small scale allows for lysis buffer modifications as necessary to enhance activity. These experiments can also determine appropriate enzyme concentration and time points where proteolysis has reached completion for full-scale studies. For other proteases, it is important to include any required cofactors, salts or accessory proteins.

4.2.4 Reverse Experimental Protocol

1. Cell lysis and caspase inactivation. For 1 x 10^9 cells in a caspase experiment, lyse using 10mL of cold 100mM HEPES pH 7.4, 0.1% Triton X-100, 10 mM IAM, and 5 mM EDTA, 1 mM AEBSF, and 1 mM PMSF. Use 50 µL of lysed sample for protein concentration determination. The lysate is kept in the dark for 15 minutes at 4°C for irreversible alkylation of endogenous caspase catalytic cysteines by IAM. Excess IAM is quenched by the addition of 20 mM DTT to prevent inhibition of recombinant caspase added in the next step. The lysate is cleared 2x by centrifugation at 4000 x *g* for 5 minutes.

2. Protease addition and quenching. Prepare a 1mL volume of active protease buffer (for caspases: 10 mM DTT, 0.1% CHAPS, 20 mM Pipes pH 7.2, 100 mM NaCl, 1 mM EDTA and 10% sucrose). Add active protease to the 1mL buffer, adjusted for the desired final concentration in ~15 mL after addition of enzyme in buffer to lysate. After a desired amount of time (determined in validation experiments) appropriate protease inhibitors are added to quench proteolysis. For caspases, we will use 10-1000 nM enzyme concentrations for 10 minutes to 3 hours and quench with 100 μM z-VAD-fmk.

3. Subtiligase labeling. To the lysate, add 10 mM ester peptide stock in DMSO to final concentration of 1 mM. Vortex briefly. Check and adjust pH to 8.5. Add 100 μM subtiligase stock to final concentration of 1 μM. Vortex briefly. Incubate for 1 hr at RT. Labeling can be confirmed through a western blot against biotin.

The remaining steps are similar to those in the Forward protocol (Section 4.1.4). Samples are desalted by acetonitrile precipitation. After resuspension of the pellet in 8 M Guanidine HCl, cysteines are reduced by TCEP and alkylated by IAM. Protein is then precipitated again in EtOH overnight. After resuspension of protein in 8 M Guanidine HCl, all steps from NeutrAvidin capture to MS analysis (Section 4.1.4, Steps 5 to 9) are identical.

4.3 Quantitation and Kinetics

Thus far we have described methods by which to identify proteolytic substrates, either in intact cells or as substrates of a specific recombinant enzyme. For additional insight into protease biology, we describe methods of quantitative proteomics combined with subtiligase-based labeling.

4.3.1 MS quantitation methods

There are a number of methods to quantify peptides by mass spectrometry, whose relative advantages and disadvantages have been reviewed in detail by others (38-41). These quantitative methods can be classified as either labeling or label-free, as well as either untargeted or targeted. Notably, essentially all can be implemented in combination with subtiligase-based N-terminomics. Frequently used untargeted labeling approaches include metabolic stable isotope labeling (SILAC) (42) and isobaric mass tags coupled to peptides after tryptic digestion (38). However, these methods suffer from labeling artifacts and a limit on the number of samples that can be compared simultaneously.

Various untargeted, label-free approaches have been implemented, including the largely imprecise method of "spectral counting" as well as others that rely on the peak area in the MS extracted ion chromatogram (43). Like all untargeted methods, a significant limitation is that these methods frequently do not identify all peptides of interest across all tested conditions. Targeted, label-free quantification, while limited to a few hundred proteins per run, offers highly sensitive and reproducible quantification. The most common method in this category is Selected Reaction Monitoring (SRM; also known as Multiple Reaction Monitoring (MRM)) (44). SRM methods can typically be run on unfractionated samples, leading to greatly reduced MS instrument time compared to other methods described. In addition, SRM assays can easily be applied to an unlimited number of samples (45, 46). Below we outline our approach to development of SRM assays to monitor proteolysis during apoptosis.

4.3.2 Design of a Kinetic Timecourse

We have primarily used SRM to quantify kinetics of proteolysis across a time course (27, 30, 47). Similar to the Forward and Reverse Discovery experiments, SRM kinetic analysis can also be performed in intact cells as well as with recombinant proteases. Furthermore, similar approaches can also be taken to examine the kinetics of proteolysis in other systems. In terms of experimental design, it is important to identify time points which will provide the most information on the proteolytic process of interest. A small scale experiment and Western blotting for cleavage of known substrates or a fluorescent reporter assay can fill this role.

4.3.3 Development and Use of Selected Reaction Monitoring (SRM) Assays

The SRM method relies on the use of a triple quadrupole mass spectrometer. In this system, a total peptide sample is injected onto an LC directly in-line with the MS instrument (Figure 4). As peptides are eluted, the first quadrupole is used as a mass filter to only isolate peptides with a targeted *m*/*z*. The second quadrupole serves as a collision cell to break the peptide into fragments. The third quadrupole functions as a second mass filter for specified *m/z* fragments from the initial parent peptide. Each of these parent-fragment ion pairs is termed a “transition,” and transition intensity is recorded by the detector. The co-elution of multiple fragments from a single parent peptide indicates the specific identification of the peptide of interest. The total peak area reflects the relative abundance of the peptide across conditions (Figure 4B).

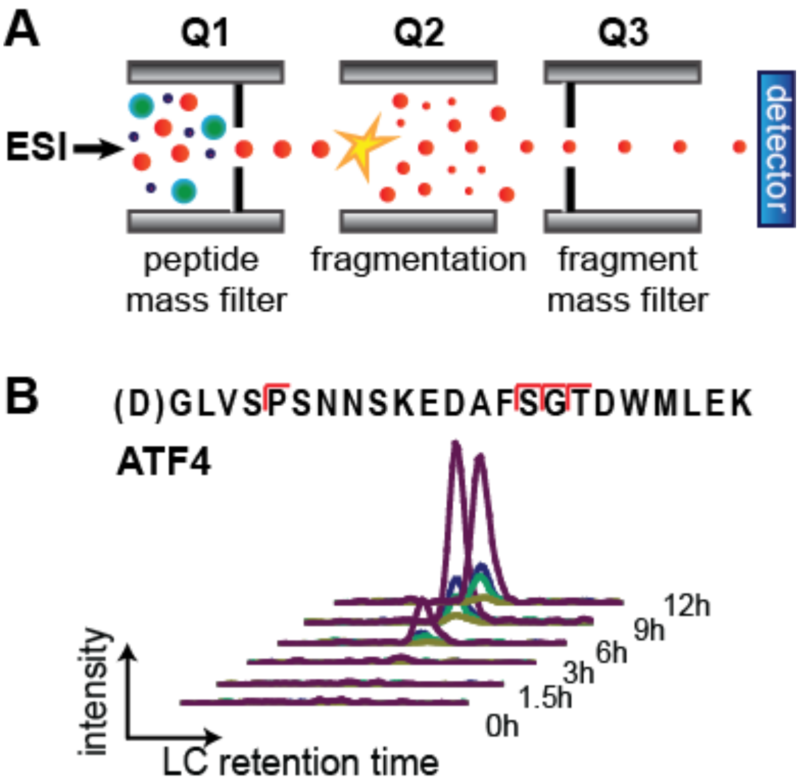


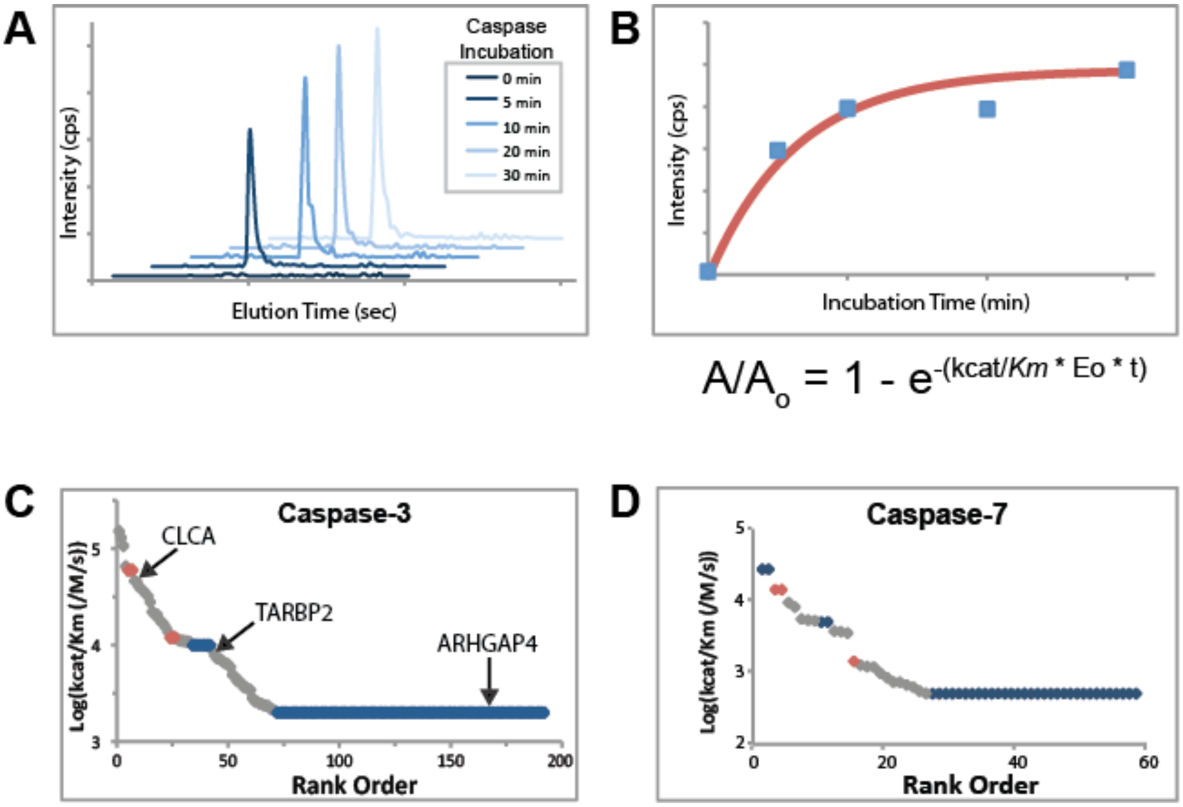
Figure 4. Selected Reaction Monitoring (SRM) for proteolytic substrate quantification. A. Schematic diagram of triple quadrupole (Q1-Q2-Q3) mass spectrometer used for SRM. ESI = electrospray ionization, representing ionized peptides eluted from liquid chromatograpy column into the mass spectrometer. **B.** Example data from SRM monitoring caspase-cleaved peptide from ATF4 protein during bortezomib-induced apoptosis. Each individual trace represents a parent-fragment ion pair (transition). Co-elution of multiple transitions from the same peptide confirms peptide identity. Peak area can be used for quantification, with kinetic parameters derived based on change across the time course.

1. Development of a spectral library. As SRM is a targeted method, Forward or Reverse Discovery experimental MS identification is required to develop an SRM assay. From this Discovery dataset, a "spectral library" of identified peptides is generated, including the parent peptide mass, fragment ions, and MS signal intensity information. From this spectral library peptides of further interest are specified for quantitative study. As an alternative source of a peptide data, our laboratory has made mass spectra available for peptides identified in Forward Discovery experiments in apoptotic and non-apoptotic cells (wellslab.ucsf.edu/degrabase).

2. SRM method development. Once a spectral library is obtained and peptides of interest are selected, the next step of method development is selection of the optimal transitions. One of the biggest recent advances in SRM is the implementation of the open-source, freely available Skyline software (48). Skyline quickly generates transition lists from the imported spectral library. For targeted peptides, we first run unscheduled SRM validation runs (no LC retention time information) monitoring up to 7 transitions per protein and ~200 transitions per run, ideally using the same Discovery samples. For peptides matching criteria (a minimum of 5 of 7 transitions and a retention time consistent with Discovery experiments; synthetic peptide standards can also aid in confirmation), we apply collision energy optimization and then create a scheduled method incorporating LC retention time information. Scheduling allows for monitoring significantly more peptides and transitions per run. We have found with the AB SCIEX QTRAP 5500 instrument up to ~250 peptides (~1000 transitions) can be included in a

single SRM run. The method development process typically can be completed in less than a week.

3. SRM sample preparation and analysis. Samples are prepared according to the same protocols as in Sections 4.1 or 4.2. The main difference is we have had success with smaller sample input, on the scale of 2 x 10^8 cells. Samples are not fractionated and the entire population of N-terminal labeled peptides is analyzed directly in a single SRM run. Skyline software is then used to determine peak area for each identified peptide in each sample. Overall sample intensity must be normalized to account for differences in labeling efficiency and MS conditions across runs. Prior to subtiligase labeling, we typically spike in purified proteins not endogenously present in the sample as internal normalization standards. Using SRM to monitor recombinant caspase cleavage of substrates in cell lysates, we were also able to derive plots of substrate appearance vs. time, analogous to more traditional enzymology experiments but capable of tracking hundreds of substrates simultaneously (30) (Figure 5).



$$A/A_o = 1 - e^{-(kcat/Km * Eo * t)}$$

Figure 5. Monitoring kinetics of recombinant caspase cleavage. A. SRM transitions show increase in intensity across timecourse after caspase addition. **B.** Peptide intensities are fit to pseudo-first order kinetic equations to determine kinetic efficiency (k_{cat}/K_M) for each substrate. **C-D.** Rank order of catalytic efficiencies for substrates of caspase-3 and -7 span at least two orders of magnitude. Caspase-3 plot indicates substrates with rapid, medium, and slow cleavage. Figure adapted from (30) with permission of the authors.

4.4 Applications to human plasma and serum

N-terminal labeling by subtiligase can be modified for human plasma and serum samples to reveal insights into the complex proteolysis in circulation (49). Blood collection tubes should be centrifuged as soon as possible after sampling to separate plasma/serum from cellular components. Plasma/serum can be stored at -80°C or used immediately for labeling. The labeling process is somewhat simplified compared to cell culture samples. To plasma/serum volumes of 0.5-2 mL, add 1 M bicine pH 8.5 to a final concentration of 100 mM. Then add AEBSF to a final concentration of 1 mM to inhibit plasma serine proteases. Incubate at RT for 10 min. Add DTT to a final concentration of 2 mM and ester peptide to a final concentration of 1 mM. Vortex briefly. Add subtiligase enzyme to a final concentration of 1 μM and incubate at RT for 1h. For plasma samples we have had success removing excess ester with NAP-25 chromatography columns (GE Healthcare) per manufacturer protocol with equilibration buffer of 50 mM bicine pH 8.0. After elution in 2.5 mL 50 mM bicine pH 8.0, add 7.5 mL 8M Guanidine HCl to desalted sample. Reduce and alkylate cysteines as in the protocols above. Then add NeutrAvidin beads, typically at a slurry volume similar to the initial plasma volume. Sample preparation now proceeds identically as from Step 5 in Section 4.1.4. Enriched N-terminal peptides can be used for either Discovery or Targeted MS. Of note, general plasma proteomic studies are often confounded by the extremely high abundance of serum albumin, as signal intensity from albumin precludes detection of biologically interesting changes in low-abundance proteins (50). Fortunately, subtiligase-labeling leads to extremely limited pull-down of albumin, enabling detection of even very low-abundance plasma proteins (Figure 6).

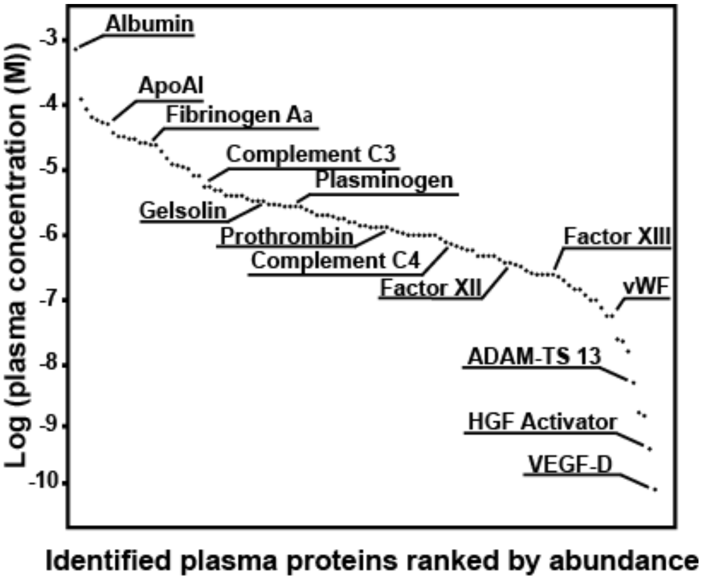


Figure 6. Plasma N-terminomics. Proteins identified with free N-termini in plasma demonstrate over 6-order of magnitude range of abundance, demonstrating ability of subtiligase labeling to track low-abundance plasma proteins. Figure adapted from (Wildes and Wells, 2010) with permission of the authors.

4.5 Application of N-terminal labeling by subtiligase to any biological sample

While we have focused on intact cells, cellular lysates, and human plasma, subtiligase-based enrichment can easily be applied to study proteolysis in any biological sample. The general protocol would be highly similar to those shown above. The key step is obtaining the total protein sample in a buffer compatible with subtiligase labeling: non-denaturing conditions (i.e. no free detergent or denaturant), pH ~8.5, and eliminating any additional free amines that compete for labeling (such as Tris). Similar to plasma labeling, labeling of biological fluids (CSF, urine, etc.) can likely be achieved without any significant sample manipulation. Tissue samples could be processed similarly to cell culture samples. Alternatively, we have had success using trichloroacetic acid precipitation of total protein following by resuspension in guanidine-containing buffer. Buffer is then exchanged with desalting columns into subtiligase-compatible conditions. From as little as 40 µg of starting protein we can identify ~50-200 proteolytically cleaved peptides released into the culture media after cellular apoptosis (A.P.W., unpublished data).

5. Limitations to method

Subtiligase labeling clearly has many advantages in proteolysis research: (i) it is an unbiased, enzymatically-driven method without chemical protein modification, (ii) it can identify thousands of peptides over 6 orders of magnitude in abundance from complex biological samples, (iii) it can be combined with highly reproducible, label-free quantification, and (iv) single labeling with positive enrichment allows not only identification of the target, but reveals the precise site of proteolysis. However, there are limitations to this approach.

As described above, subtiligase labeling efficiency is the biggest limitation to application to all systems of interest. To counter this inefficiency, we use large amounts of starting material. This is relatively easy for cell culture based experiments, but can be more complicated for animal or human samples, where obtaining sufficient sample input may require pooling strategies. Additionally, there are a few N-terminal amino acids, notably proline, valine and isoleucine, that subtiligase is relatively slow to ligate *in vitro* (51). However, we have found that subtiligase can indeed label these N-terminal amino acids in our cell-based experiments. Thus, we believe this is a small bias and does not significantly affect the utility of the method.

For Discovery experiments, these methods suffer the same limitation as all MS experiments: there is only limited overlap between peptides identified in one run to another, even under the same conditions. This is due to the stochastic nature of sequencing of low-abundance peptides by the MS instrument. This limitation may be circumvented through the use of biological and technical replicates. Importantly, a particular limitation for our strategy is that short peptides labeled by subtiligase will not be precipitated with other proteins (in Step 4 of protocol in 4.1.4) and will be discarded. Furthermore, the tryptic digestion of biotinylated proteins may lead to N-terminal peptides either too long or too short to be identified by MS, leading to missed identifications.

6. Summary of findings from subtiligase-based N-terminomics

As discussed in the introduction, there are multiple methods for enrichment and identification of proteolytically-cleaved peptides in biological samples (Rogers and Overall, 2013). Our laboratory's unique method of enzymatically-driven labeling has primarily been applied to the study of caspases. These studies showed that caspases cleave hundreds of cellular substrates during apoptosis, greatly expanding the scope of caspase biology. Our initial study revealed that caspases have a general preference for disordered structural elements in substrates (9). Further experiments across multiple cell lines allowed us to compile over 1,700 caspase cleavage sites in nearly 1,300 substrates as well as over 6,000 non-caspase proteolytic sites (8). With this large database, publicly available at http://wellslab.ucsf.edu/degrabase, we identified conserved motifs of caspase cleavage (Figure **7**) as well as sequence features of non-caspase endoproteases in non-apoptotic cells (8). Furthermore, by combining recombinant enzyme purification with subtiligase labeling, we characterized specific substrates of the inflammatory caspases-1, -4, and -5 as well as the apoptotic caspases -3, -7, -8, and -9 (30, 52). A combination of Forward and Reverse experiments allowed us to find evolutionarily conserved relationships between caspase cut site, protein substrates, and pathway-level relationships across organisms (53). With a similar experimental combination we probed proteolytic cleavage in human plasma which may relate to disease signatures (49). With the discovery of a broad range of caspase catalytic efficiencies across hundreds of substrates in parallel (30), we have pioneered the combination of label-free quantification with N-terminomics to determine catalytic efficiencies of natural substrates in complex mixture. Furthermore, we showed that quantitative signatures of caspase cleavage can be used to monitor chemotherapeutic effects in cancer cells (27, 47). In summary, this method has revealed extensive information about proteolytic cleavage during apoptosis.

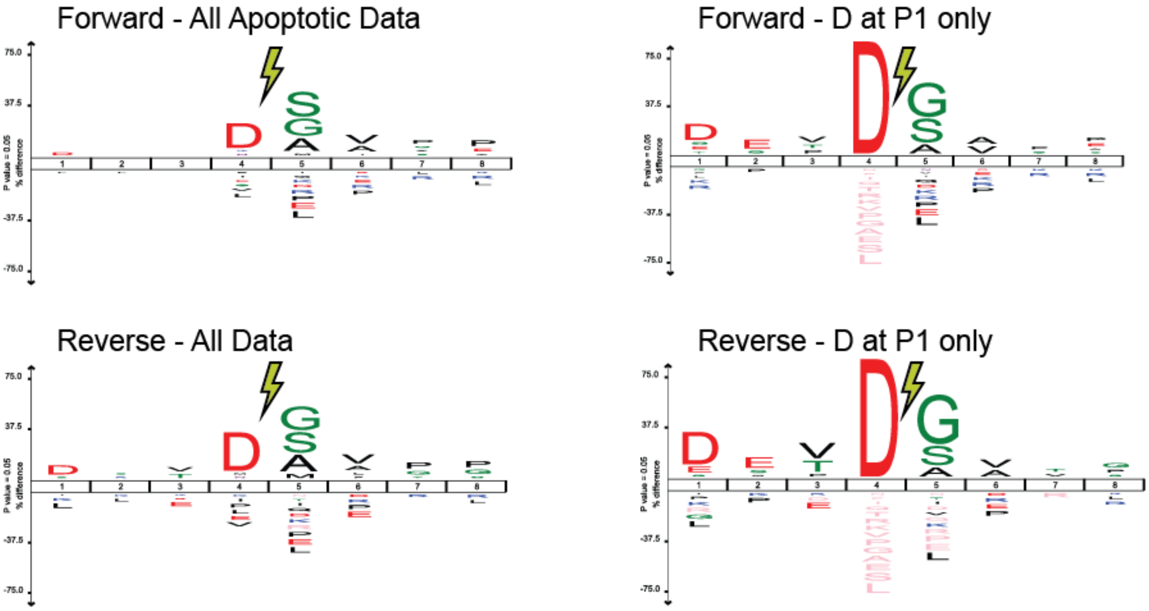


Figure 7. Sequence features of identified N-termini. Aggregate plots across all N-termini shown using IceLogo, where amino acids favored at a given site are above the baseline while those disfavored are below (54). Cleavage occurs between position P1 (left of lightning bolt) and P1'. **A.** Across all peptides identified in forward discovery experiments in apoptotic cells, Asp at P1 is highly enriched. **B.** Focusing on only peptides shown in **A** with Asp at P1, a signature of caspase cleavage, we identify the canonical D-E-V-D cleavage motif for caspases from the P4 to P1 site. **C-D.** This motif is also conserved in N-termini identified in cell lysate incubated with recombinant caspase-3 during reverse discovery experiments.

7. Future directions

To broaden the range of applications of our methods, we are using protein engineering to improve subtiligase labeling efficiency and decrease the need for large amounts of starting material. In tandem, new highly sensitive mass spectrometers will allow for both more comprehensive substrate identification and potentially label-free quantification with less protein input (55-57). Both of these advances will allow for in-depth analysis of many more recombinant proteases and biological systems. In intact cells, caspases have only been fully profiled for substrate generation during apoptosis; their substrate profiles in processes such as differentiation or non-apoptotic cell stress (58) have yet to be elucidated. Others have recently shown that there is significant cross-talk between protein phosphorylation and caspase cleavage (18). N-terminomics can further be combined with new enrichment methods to investigate the relation between caspase cleavage and other post-translational modifications, such as ubiquitination and lysine acetylation (59). Alternatively, isolating intracellular organelles or secreted domains from cell membrane proteins will allow one to monitor proteolysis specific to different cellular perturbations in different cellular compartments. This knowledge may lead to information relevant to therapeutic and diagnostic development. As described here, subtiligase-based N-terminomics has already revealed significant new insight into caspase biology. This technique is now poised for wide use in proteolysis research.

Acknowledgements

This project was supported by grants from National Institutes of Health R01 GM081051, R01 GM097316, and RO1 CA154802 (J.A.W.), The Rogers Family Foundation (J.A.W.), an award from the Sandler Program in Biological Sciences (J.A.W.), and the UCSF Stephen and Nancy Grand Multiple Myeloma Translational Initiative (J.A.W.). A.P.W. is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG 111-12). J.E.S is supported by National Institutes of Health Training Grant T32 GM007175 and the Global Healthy Living Foundation. Mass spectrometry was performed at the Bio-Organic Biomedical Mass Spectrometry Resource at UCSF, which is supported by grants from the National Center for Research Resources (5P41RR001614) and the National Institute of General Medical Sciences (8P41GM103481 and 1S10RR015804) from the National Institutes of Health. We thank our long time collaborator Professor Alma Burlingame for providing expert advice and open access to the UCSF Mass Spectrometry Center.

References

1. Lopez-Otin C, Matrisian LM. Emerging roles of proteases in tumour suppression. Nat Rev Cancer. 2007;7(10):800-8.

2. Agard NJ, Wells JA. Methods for the proteomic identification of protease substrates. Curr Opin Chem Biol. 2009;13(5-6):503-9.

3. Impens F, Colaert N, Helsens K, Plasman K, Van Damme P, Vandekerckhove J, et al. MS-driven protease substrate degradomics. Proteomics. 2010;10(6):1284-96.

4. Klingler D, Hardt M. Profiling protease activities by dynamic proteomics workflows. Proteomics. 2012;12(4-5):587-96.

5. Rogers LD, Overall CM. Proteolytic post-translational modification of proteins: proteomic tools and methodology. Mol Cell Proteomics. 2013;12(12):3532-42.

6. Smeekens SP, Montag AG, Thomas G, Albiges-Rizo C, Carroll R, Benig M, et al. Proinsulin processing by the subtilisin-related proprotein convertases furin, PC2, and PC3. Proc Natl Acad Sci U S A. 1992;89(18):8822-6.

7. Black RA, Kronheim SR, Sleath PR. Activation of interleukin-1 beta by a co-induced protease. FEBS Lett. 1989;247(2):386-90.

8. Crawford ED, Seaman JE, Agard N, Hsu GW, Julien O, Mahrus S, et al. The DegraBase: a database of proteolysis in healthy and apoptotic human cells. Mol Cell Proteomics. 2013;12(3):813-24.

9. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell. 2008;134(5):866-76.

10. Staes A, Van Damme P, Helsens K, Demol H, Vandekerckhove J, Gevaert K. Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). Proteomics. 2008;8(7):1362-70.

11. Tanco S, Lorenzo J, Garcia-Pardo J, Degroeve S, Martens L, Aviles FX, et al. Proteome-derived peptide libraries to study the substrate specificity profiles of carboxypeptidases. Mol Cell Proteomics. 2013;12(8):2096-110.

12. Wejda M, Impens F, Takahashi N, Van Damme P, Gevaert K, Vandenabeele P. Degradomics reveals that cleavage specificity profiles of caspase-2 and effector caspases are alike. J Biol Chem. 2012;287(41):33983-95.

13. Kleifeld O, Doucet A, auf dem Keller U, Prudova A, Schilling O, Kainthan RK, et al. Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. Nat Biotechnol. 2010;28(3):281-8.

14. Schilling O, Huesgen PF, Barre O, Overall CM. Identification and relative quantification of native and proteolytically generated protein C-termini from complex proteomes: C-terminome analysis. Methods Mol Biol. 2011;781:59-69.

15. Xu G, Jaffrey SR. N-CLAP: global profiling of N-termini by chemoselective labeling of the alpha-amine of proteins. Cold Spring Harb Protoc. 2010;2010(11):pdb prot5528.

16. Xu G, Shin SB, Jaffrey SR. Global profiling of protease cleavage sites by chemoselective labeling of protein N-termini. Proc Natl Acad Sci U S A. 2009;106(46):19310-5.

17. Dix MM, Simon GM, Cravatt BF. Global mapping of the topography and magnitude of proteolytic events in apoptosis. Cell. 2008;134(4):679-91.

18. Dix MM, Simon GM, Wang C, Okerberg E, Patricelli MP, Cravatt BF. Functional interplay between caspase cleavage and phosphorylation sculpts the apoptotic proteome. Cell. 2012;150(2):426-40.

19. Tonge R, Shaw J, Middleton B, Rowlinson R, Rayner S, Young J, et al. Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. Proteomics. 2001;1(3):377-96.

20. Thiede B, Treumann A, Kretschmer A, Sohlke J, Rudel T. Shotgun proteome analysis of protein cleavage in apoptotic cells. Proteomics. 2005;5(8):2123-30.

21. Shao W, Yeretssian G, Doiron K, Hussain SN, Saleh M. The caspase-1 digestome identifies the glycolysis pathway as a target during infection and septic shock. J Biol Chem. 2007;282(50):36321-9.

22. Xu G, Shin SB, Jaffrey SR. Chemoenzymatic labeling of protein C-termini for positive selection of C-terminal peptides. ACS Chem Biol. 2011;6(10):1015-20.

23. Abrahmsen L, Tom J, Burnier J, Butcher KA, Kossiakoff A, Wells JA. Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. Biochemistry. 1991;30(17):4151-9.

24. Braisted AC, Judice JK, Wells JA. Synthesis of proteins by subtiligase. Methods Enzymol. 1997;289:298-313.

25. Polevoda B, Sherman F. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. J Mol Biol. 2003;325(4):595-622.

26. Lucast LJ, Batey RT, Doudna JA. Large-scale purification of a stable form of recombinant tobacco etch virus protease. Biotechniques. 2001;30(3):544-6, 8, 50 passim.

27. Shimbo K, Hsu GW, Nguyen H, Mahrus S, Trinidad JC, Burlingame AL, et al. Quantitative profiling of caspase-cleaved substrates reveals different drug-induced and cell-type patterns in apoptosis. Proc Natl Acad Sci U S A. 2012;109(31):12432-7.

28. Yoshihara HA, Mahrus S, Wells JA. Tags for labeling protein N-termini with subtiligase for proteomics. Bioorg Med Chem Lett. 2008;18(22):6000-3.

29. Jackson DY, Burnier J, Quan C, Stanley M, Tom J, Wells JA. A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. Science. 1994;266(5183):243-7.

30. Agard NJ, Mahrus S, Trinidad JC, Lynn A, Burlingame AL, Wells JA. Global kinetic analysis of proteolysis via quantitative targeted proteomics. Proc Natl Acad Sci U S A. 2012;109(6):1913-8.

31. Galluzzi L, Aaronson SA, Abrams J, Alnemri ES, Andrews DW, Baehrecke EH, et al. Guidelines for the use and interpretation of assays for monitoring cell death in higher eukaryotes. Cell Death Differ. 2009;16(8):1093-107.

32. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198-207.

33. Kapp E, Schutz F. Overview of tandem mass spectrometry (MS/MS) database search algorithms. Curr Protoc Protein Sci. 2007;Chapter 25:Unit25 2.

34. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2012;40(Database issue):D343-50.

35. Lange PF, Huesgen PF, Overall CM. TopFIND 2.0--linking protein termini with proteolytic processing and modifications altering protein function. Nucleic Acids Res. 2012;40(Database issue):D351-61.

36. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, et al. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). BMC Bioinformatics. 2005;6:168.

37. Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, et al. PaxDb, a database of protein abundance averages across all three domains of life. Mol Cell Proteomics. 2012;11(8):492-500.

38. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem. 2012;404(4):939-65.

39. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007;389(4):1017-31.

40. Liebler DC, Zimmerman LJ. Targeted quantitation of proteins by mass spectrometry. Biochemistry. 2013;52(22):3797-806.

41. Nikolov M, Schmidt C, Urlaub H. Quantitative mass spectrometry-based proteomics: an overview. Methods Mol Biol. 2012;893:85-100.

42. Bushell M, Stoneley M, Kong YW, Hamilton TL, Spriggs KA, Dobbyn HC, et al. Polypyrimidine tract binding protein regulates IRES-mediated gene expression during apoptosis. Mol Cell. 2006;23(3):401-12.

43. Nahnsen S, Bielow C, Reinert K, Kohlbacher O. Tools for label-free peptide quantification. Mol Cell Proteomics. 2013;12(3):549-56.

44. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 2012;9(6):555-66.

45. Huttenhain R, Soste M, Selevsek N, Rost H, Sethi A, Carapito C, et al. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. Sci Transl Med. 2012;4(142):142ra94.

46. Li XJ, Hayward C, Fong PY, Dominguez M, Hunsucker SW, Lee LW, et al. A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. Sci Transl Med. 2013;5(207):207ra142.

47. Wiita AP, Ziv E, Wiita PJ, Urisman A, Julien O, Burlingame AL, et al. Global cellular response to chemotherapy-induced apoptosis. Elife. 2013;2:e01236.

48. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26(7):966-8.

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

50. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1(11):845-67.

51. Chang TK, Jackson DY, Burnier JP, Wells JA. Subtiligase: a tool for semisynthesis of proteins. Proc Natl Acad Sci U S A. 1994;91(26):12544-8.

52. Agard NJ, Maltby D, Wells JA. Inflammatory stimuli regulate caspase substrate profiles. Mol Cell Proteomics. 2010;9(5):880-93.

53. Crawford ED, Seaman JE, Barber AE, 2nd, David DC, Babbitt PC, Burlingame AL, et al. Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis. Cell Death Differ. 2012;19(12):2040-8.

54. 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.

55. Gallien S, Duriez E, Crone C, Kellmann M, Moehring T, Domon B. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. Mol Cell Proteomics. 2012;11(12):1709-23.

56. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, et al. The one hour yeast proteome. Mol Cell Proteomics. 2014;13(1):339-47.

57. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol Cell Proteomics. 2012;11(11):1475-88.

58. Kuranaga E. Beyond apoptosis: caspase regulatory mechanisms and functions in vivo. Genes Cells. 2012;17(2):83-97.

59. Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, Mani DR, et al. Integrated proteomic analysis of post-translational modifications by serial enrichment. Nat Methods. 2013;10(7):634-7.

Chapter 2

The DegraBase: a database of proteolysis in healthy and apoptotic human cells

This work was published in Molecular and Cellular Proteomics:
Crawford ED, Seaman JE, Agard N, Hsu GW, Julien O, Mahrus S, et al. The DegraBase: a database of proteolysis in healthy and apoptotic human cells. Mol Cell Proteomics. 2013;12(3):813-24.

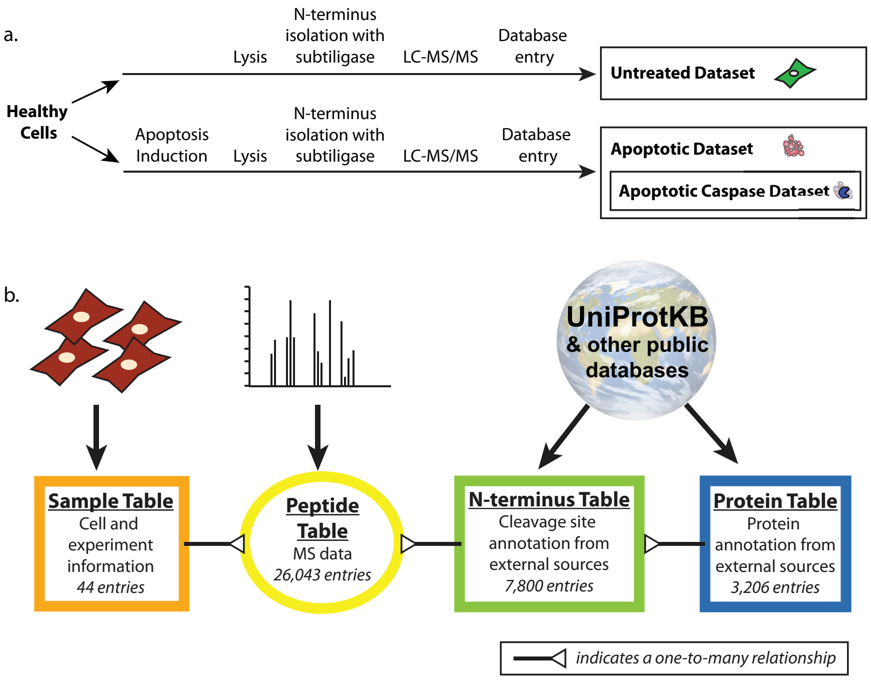
Abstract

Proteolysis is a critical post-translational modification for regulation of cellular processes. Our lab has previously developed a technique for specifically labeling unmodified protein N-termini, the α-aminome, using the engineered enzyme, subtiligase. Here we present a database, called the DegraBase (http://wellslab.ucsf.edu/degrabase/), that compiles 8090 unique N-termini from 3206 proteins directly identified in subtiligase-based positive enrichment mass spectrometry experiments in healthy and apoptotic human cell lines. We include both previously published and unpublished data in our analysis, resulting in a total of 2144 unique α-amines identified in healthy cells, and 6990 in cells undergoing apoptosis. The N- termini derive from three general categories of proteolysis with respect to cleavage location and functional role: translational N-terminal methionine processing (~10% of total proteolysis), sites close to the translational N-terminus that likely represent removal of transit or signal peptides (~25% of total), and lastly, other endoproteolytic cuts (~65% of total). Induction of apoptosis causes relatively little change in the first two proteolytic categories, but dramatic changes are seen in endoproteolysis. For example, we observed 1706 putative apoptotic caspase cuts, more than double the total annotated sites in the CASBAH and MEROPS databases. In the endoproteolysis category, there are a total of nearly 3000 non-caspase non-tryptic cleavages that are not currently reported in the MEROPS database. These studies significantly increase the annotation for all categories of proteolysis in human cells and allow public access for investigators to explore interesting proteolytic events in healthy and apoptotic human cells.

Introduction:

Annotation of the human α-aminome, the full set of unmodified protein N-termini, can provide a wealth of information regarding protein turnover, protein trafficking, and protease activity (*1*). The vast majority of protein N-termini in eukayotic cells are co-translationally blocked by acetylation through the action of N-acetyl transferases (*2*). Free α-amines occur on some proteins that are never N-terminally acetylated, and can also be regenerated by signal or transit peptide removal during protein trafficking, and endo- or exoproteolysis during protein maturation and signaling. Thus, there has been considerable effort to develop unbiased proteomic methods to characterize the α-aminome in healthy and diseased states (*3-8*).

We have developed a positive enrichment method in which the α-amines of intracellular (*8*) or extracellular proteins (*9*) can be specifically and directly tagged and captured, without pretreatment or protection, using subtiligase, an engineered peptide ligase **(Figure 1A)** (*10, 11*). Following purification, tryptic digestion, and LC-MS/MS, the protein sequence and exact site of proteolysis are readily identified. We have applied this approach to study proteolysis by caspases, cysteine-class aspartyl specific proteases, during cellular apoptosis (*8, 12-14*), and inflammatory response (*15*). These studies in a variety of cell types and apoptotic inducers, have revealed much about the targets, substrate recognition, timing, logic, and evolution of caspase cleavage events. These efforts have generated a huge amount of data that requires systematic compilation, organization, and normalization so that it can be shared and queried easily by all investigators and compared with other databases describing proteolytic events (*16-18*).


a.
Healthy Cells
Lysis
N-terminus isolation with subtiligase
LC-MS/MS
Database entry
Untreated Dataset
Apoptosis Induction
Lysis
N-terminus isolation with subtiligase
LC-MS/MS
Database entry
Apoptotic Dataset
Apoptotic Caspase Dataset
b.
UniProtKB & other public databases
Sample Table
Cell and experiment information
44 entries
Peptide Table
MS data
26,043 entries
N-terminus Table
Cleavage site annotation from external sources
7,800 entries
Protein Table
Protein annotation from external sources
3,206 entries
indicates a one-to-many relationship

c.

	Total	Apoptotic	Untreated
Experiments	44	33	11
Cell Lines	7	5	5
Inducers	7	7	N/A
Peptides	26,043	22,311	3,732
Unique N-termini	8,090	6,990	2,144
Unique Proteins	3,206	3,020	1,239
Unique Caspase-Cleaved N-termini	1,717	1,706	140
Unique Caspase-Cleaved Proteins	1,275	1,268	127

Figure 1: Experimental schema, database design and database summary. **A**. For all experiments, human cells were grown under standard conditions, either with or without treatment with apoptosis inducing agents. Cells are lysed and proteins biotinylated on their free α-amines using subtiligase, followed by purification and identification by LC-MS/MS. N-termini identifications from every experiment were entered into the database to create the untreated and apoptotic datasets, and a subset apoptotic caspase-cleaved dataset for apoptotic N-termini following aspartic acid cleavage. **B.** The DegraBase database is structured around 4 main tables linking the experimental data to the MS identifications and external database information at both the N-terminus and protein level (for more details see **Supplemental File 1**). **C.** Summary statistics of the DegraBase for all experiments in the DegraBase and for both the untreated and apoptotic datasets (more details in **Supplemental Table 1A**). The blue box highlights the apoptotic caspase-cleaved dataset within the apoptotic dataset.

Here we present the results of both previously published and new experiments that detect α-amines in both untreated and apoptotic human cells. These studies reveal new translational N-terminal processing, signal and transit peptide removal, and other proteolytic events associated with normal protein maturation and function in healthy cells. Comparing these data to the apoptotic dataset reveals that the greatest changes in apoptosis are due to endoproteolysis, owing to the induction of caspases as well as other proteases. We find a total of 1706 putative caspase sites in nearly 1300 different human proteins. We further find an additional 2900 non-caspase, non-tryptic, non-transit and non-signal peptide cleavage sites in 1415 proteins.

In addition to the analyses described here, we provide a publically available database, the DegraBase, that is dynamic, expandable, searchable, and readily accessible (http://wellslab.ucsf.edu/degrabase/). With this database, investigators can query all 8,090 unique α-amines detected with high confidence from 26,043 peptide observations in both previously published (*8, 12, 13*) and new subtiligase α-aminome labeling experiments. The DegraBase substantially expands annotated intracellular proteolytic events in healthy and apoptotic cells.

Experimental Procedures:

Cell Cultures

Jurkat, THP-1, DB, RPMI 8226, MM1-S and U266 human cell lines were acquired from the American Type Culture Collection (ATCC, Manassas, Virginia, USA) and were cultured under the recommended conditions. When cells reached a density of 1x10^6 cells/mL, an apoptotic inducer (doxorubicin, etoposide, bortezomib, FasL, CD95, staurosporine, or TRAIL) was added from 1000x stock (for individual experimental details, see **Supplemental Table 1A**). Cell viability and caspase activity were monitored by CellTiter-Glo, Caspase-Glo (Promega, Madison, Wisconsin, USA) and Ac-DEVD-AFC activity assays. Cells were harvested by centrifugation after 0-40 hours, washed with phosphate buffered saline solution, pelleted, and stored at -80°C. For untreated experiments, healthy cells cultured under the same conditions were harvested without any inducer added.

N-terminal Labeling

The lysis and N-terminal labeling were performed as described previously (*8, 12, 13, 15*). For experiments not previously published, the following protocol was used. Cells were lysed in a bicine buffer with triton or SDS containing the protease inhibitors EDTA, PMSF, E-64, z-VAD-fmk and AEBSF. Proteins were reduced with 2mM TCEP at 90°C for 15min and alkylated once cooled with 4mM IAM in dark for 1 hour, then quenched with 10mM DTT. Labeling was performed with 1mM of a biotinylated synthesized peptide ester called TEVest and 1 μM subtiligase for at least an hour at room temperature (*11, 19*). There were 4 different TEVest peptide esters used to facilitate the identification of the labeled products; they were identical except for the small tag left after processing to aid in mass spectrometry recognition: serine-tyrosine (SY), glycine-tyrosine (GY), phenylalanine (Phe), or 2-aminobutyric acid (Abu). The biotinylated proteins were separated by gel filtration or precipitation and captured on NeutrAvidin agarose beads (Pierce, Rockford, Illinois, USA). The samples were digested with sequence grade modified Trypsin (Promega, Madison, Wisconsin, USA) before or after capture. After capture, the labeled peptides were released with recombinant TEV protease and collected. Samples were desalted by chromatography with C18 ZipTip Pipette Tips (Millipore, Billerica, Massachusetts, USA) or C18 HPLC (Waters, Milford, Massachusetts, USA). Further offline strong cation exchange fractionation was performed on some samples. For further individual experimental details, see **Supplemental Table 1A.**

LC-MS/MS and Peptide Identification

For all experiments, samples were separated by reverse phase HPLC coupled to a mass spectrometer: QSTAR Pulsar, QSTAR XL, QSTAR Elite (Applied Biosystems, Foster City, California, USA), LTQ-Orbitrap XL or QExactive (Thermo Fisher Scientific, San Jose, California, USA). Spectra were converted into peak lists for database searching using the mascot dll in Analyst for QSTAR instruments or using an in-house script based on the Raw_Extract script in Xcalibur v2.4 (Thermo Fisher Scientific). Peptide identification was performed using Protein Prospector version 5.10.0 (*20*). Search parameter mass allowances were tailored for each instrument: 100 ppm precursor and 0.15 Da fragment for QSTAR instruments, 20 ppm precursor and 0.6 Da fragment for LQT-Orbitrap XL, and 20 ppm precursor and 0.8 Da for QExactive. All searches were performed with constant modification of the peptide N-terminus with the appropriate TEVest tag, variable modifications of carbamidomethylation of cysteines and oxidation of methionine, and allowing for up to 3 missed tryptic cleavages. All datasets were searched assuming tryptic specificity at the peptide C-terminus, but no cleavage specificity at the N-terminus. All fractions (including re-analysis of previously published data) were searched against the human SwissProt library release 2012_03 (20,255 entries) in order to provide consistent accession number annotations for all data. Maximum expectation value scores for protein and peptide of 0.02 were employed as acceptance criteria. Searches against a decoy library of random and reversed protein sequences revealed an average false discovery rate (FDR) across all datasets of 0.55%.

Data Analysis

The DegraBase framework was created using FileMakerPro version 9.0, and houses three types of data: the sample, peptide and N-terminus/protein tables **(Figure 1B).** Experimental parameters are entered by investigators, mass spectrometry data are imported from files created by Protein Prospector, and protein- and cleavage site-specific annotation data are imported from a number of external databases including UniProtKB (*21*), the CASBAH (*16*), and MEROPS (*17*). Full documentation, including FileMakerPro scripts for data analysis and Perl scripts for processing of UniProtKB data prior to input, is available as **Supplemental File 1**. The DegraBase also exists as an HTML-based website (http://wellslab.ucsf.edu/degrabase/) to allow for more accessible searching.

Abundance data were taken from PaxDB version 2.1 using the integrated dataset called “Weighted average of 'H. sapiens PeptideAtlas Build May 2010'(weighting 50%), 'H. sapiens PeptideAtlas Build March 2009',(weighting 50%)” available from the downloads tab at www.pax-db.org (*22*). Sequence logos were made using iceLogo with the whole human SwissProt library as background (*23*). All logo images were made with the percent difference scoring system, except when stated as “Filled Logos” representing amino acid frequency, not information content. Significance was determined by chi-square analyses. Data for methionine processing, mitochondrial transit peptide removal and signal peptide removal were compared with SwissProt library release 2012_03. Mitochondrial localization was determined based on the MitoCarta database (*24*).

GO term enrichment was determined using the GO::TermFinder software (*25*). A list of unique proteins for each dataset was created and uploaded to the database and tested for enrichment against the human SwissProt background using all evidence codes except ND (No biological Data available) and IEA (Inferred from Electronic Annotation). Enriched terms were defined using a corrected p-value cutoff of less than 0.01. To compare terms between datasets, a pairwise chi-square test was performed using the Benjamini-Hochberg multiple testing correction procedure.

Results:

The DegraBase:

Given the massive amount of data generated from multiple experiments under different conditions, it was necessary to create a simple and normalized database. The DegraBase is a relational database built to house our α-aminomics data **(Figure 1B)**. It is available in three formats (see Supplemental Information): a FileMakerPro file (**Supplemental File 2**), an excel document containing worksheets for each of the major tables (**Supplemental File 3**), and a web interface (http://wellslab.ucsf.edu/degrabase/) where users may search by substrate name or accession number. Full documentation of the database is available in **Supplemental File 1**.

Data and Datasets

The current DegraBase contains a total of 26043 independent peptide identifications from 44 different proteomic labeling experiments (11 untreated and 33 apoptotic) **(Figure 1C** and **Supplemental Table 1A)**. There are a total of 8090 unique N- terminus identifications from 3206 proteins. We subdivided our data into three sets: (1) untreated, (2) apoptotic and (3) apoptotic caspase-cleaved. In a separate study using our labeling method, we have seen that there is cell line- and drug-specific variability in the data, but most differences show up in detected abundance of cleavage product over time rather than the presence or absence (reported here) of the specific identified N-termini (*13*). Therefore, we were comfortable pooling our multiple apoptotic experiments together in order to compare all proteins detected in all untreated cells tested versus those undergoing apoptosis.

The untreated dataset contains all observations from the 11 experiments performed in 5 different cell lines **(Supplemental Table 1B)**. This dataset has 3732 identified N- termini corresponding to 2144 unique N-terminus start sites from 1239 proteins. The apoptotic dataset consists of all observations from the 33 experiments using 7 different chemotherapeutic inducers in 5 cell lines **(Supplemental Table 1C)**. This generated a total of 22311 independent peptide identifications, corresponding to 6990 unique N-terminus sites from 3020 different proteins. This reflects the dramatic activation of caspases following the induction of apoptosis our samples, also observed with caspase activity and cell death assays (data not shown). We defined the third dataset, the apoptotic caspase-cleaved dataset, as a subset of the apoptotic dataset that includes all apoptotic aspartic acid-cleaved N-termini **(Supplemental Table 1D)**. This dataset includes 1706 unique N-termini from 1268 proteins, and in combination with our previous studies, MEROPS and CASBAH, increases the number of published human caspase-cleavage events to over 2200. The apoptotic dataset contains 1706 aspartate cleaved peptides compared to the 140 seen in untreated dataset, reflecting a dramatic induction of caspase activity.

To estimate to what degree the α-aminome MS data are biased by protein abundance in cells, we compared the datasets to PaxDB (*22*), a database that provides an independent estimate of relative protein abundance based on MS spectral counting data. All three α-aminome datasets cover more than 6 orders of magnitude of ppm **(Supplemental Figure 1 A-C)**. Only for the small set of low abundance proteins did our α-aminome identification tail off, which presumably reflects the limits of detection of the methodology. There is a slight enrichment for higher abundance proteins overall **(Supplemental Figure 1 D)**.

At the protein level, there is a large overlap between the untreated and apoptotic datasets; 1053 of the 1239 proteins (85%) from untreated cells were also found in the apoptotic dataset **(Figure 2A)**. In contrast, we observed a smaller overlap between datasets when considering the particular N-termini within each protein **(Figure 2B)**; only1328 of the 2144 untreated N-termini (62%) were labeled under apoptotic conditions. There is a small set of 361 proteins, but only 129 N-termini, that overlap between the untreated and apoptotic caspase-cleaved datasets. The presence of caspase-cleaved products in healthy cells likely reflects low levels of apoptosis that occurs in any healthy cell population, and make up a very small portion of the total untreated set. The protein overlap may represent apoptotic caspase substrates that also undergo endoproteolysis in healthy cells by non-caspases. Interestingly, many of the proteolytic substrates in healthy cells are cleaved at different positions upon induction of apoptosis.

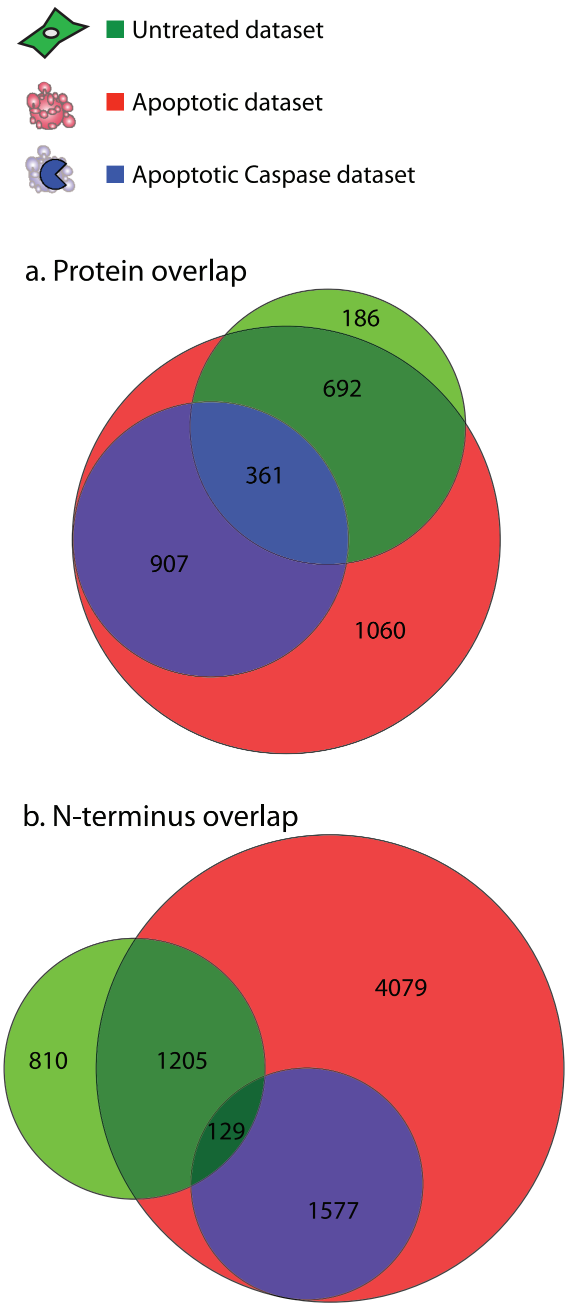


Figure 2: Venn diagrams show a larger overlap of unique protein (**A**) than peptide (**B**) identifications between untreated, apoptotic and apoptotic caspase-cleaved datasets. The apoptotic caspase-cleaved dataset is defined as a subset of the apoptotic dataset, and is therefore contained wholly within the apoptotic set.

To compare the functional properties of the different datasets, we performed Gene Ontology (GO) term enrichment using GO::TermFinder **(Supplemental Table 2)** (*25*). We looked at the terms unique to each dataset to identify specific process, function or component annotations related to healthy or dying cellular states. The untreated dataset was enriched in terms related to homeostatic functions like metabolic and biosynthetic processes (specifically related to ribosomal, coenzyme, amino acids and fatty acids, NADH dehydrogenase and isomerase functions), the mitochondrial proton-transporting ATP synthase complex, and organelle envelope lumen (prominently related to the endoplasmic reticulum). In the apoptotic set, we compared the significant terms from caspase substrates to the non-caspase apoptotic terms, and to the terms unique to the apoptotic set only. The caspase substrates are enriched in the regulation of transcription, and there were many terms related to cell morphogenesis, specifically chromosome and microtubule structure, which are known to change and break down during apoptosis. The non-caspase apoptotic enriched terms in process, function and component ontologies relate to chromatin assembly (especially DNA binding, vesicle coating and targeting), signal transduction involved in DNA damage and cell cycle checkpoints, and nucleotide catabolic processes. We also saw enrichment in the non-caspase apoptotic set for proteins associated with terms for proteolysis and cell death.

We next analyzed the precise sequences surrounding the N-termini identified in each dataset. We used iceLogo (*23*) to visualize the sequence specificity for cleavage events for each dataset using the human SwissProt database to establish background amino acid frequencies **(Figure 3)**. The cleavage sites are presented in the standard Schechter-Berger form, with the scissile bond between the P1 residue and the P1' residue (*26*). All three logos show a strong preference for small amino acids (glycine, serine, or alanine) at the P1' position, but significant differences at the P1 position. In healthy cells, there is enrichment for cleavage sites following lysine, arginine, and methionine. The methionine cleavages mainly represent N-terminal methionine processing. The large number of cuts following basic residues is consistent with a high activity of trypsin-like activities in both healthy and apoptotic cells. In apoptotic cells this tryptic-like activity is overshadowed by the large number of caspase cleavages following aspartic acid residues. The apoptotic caspase-cleaved dataset shows a degenerate specificity with moderate enrichment for aspartic acid-glutamic acid-valine in the P4-P2 positions, matching the classic "DEVD" substrate preference for executioner caspases-3 and -7, the signature proteases of apoptosis (*17*).

Three categories of proteolysis: translational N-terminus processing, signal/transit peptide removal, and endoproteolysis

In the global analysis presented above, we have discussed all the proteolytic events in healthy and apoptotic cells without distinguishing between the different kinds of proteolytic processing known to occur in cells. We now look more closely at three important areas of proteolytic processing: (i) processing around the methionine at the translational N-terminus (N-termini labeled at residues 1 and 2), (ii) cleavage of possible secretory or transit peptides during organelle trafficking (labeled residues 3-65), and (iii) other endoproteolytic events (labeled residues 66+) **(Figure 4)**. We chose to define possible signal or transit peptides within residues 3-65 based on patterns from previously published datasets (*27, 28*) and from our own data (see below). Subdividing each dataset into these different groups, we see that the majority of cuts results from endoproteolysis (55-80%), then putative signal or transit peptide removal (20-35%), and finally processing around the initiator methionine (<10%)

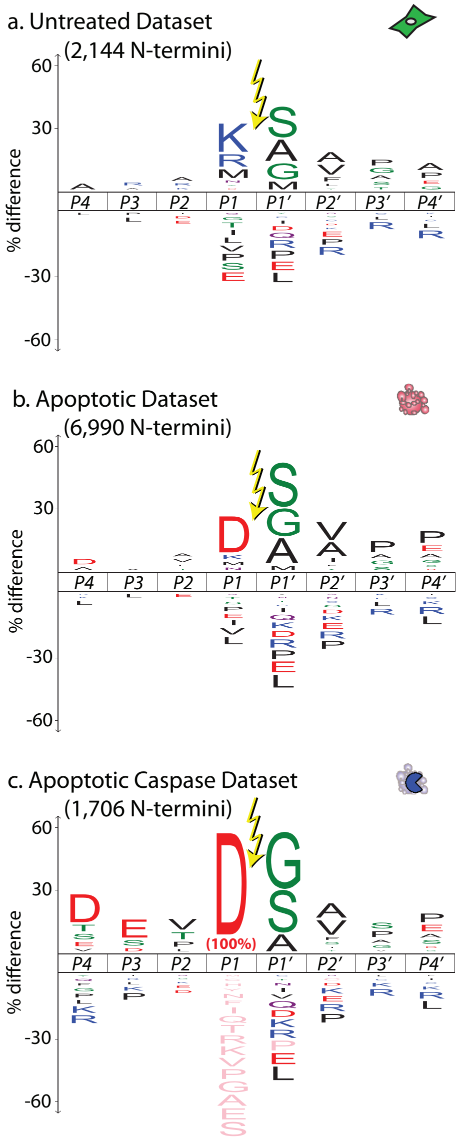


Figure 3: The untreated (**A**) and apoptotic (**B**) datasets show distinct patterns of amino acid frequency for the 8 positions surrounding the labeled α-amine (P4-P4') from all unique N-termini, as shown in the iceLogo diagrams. Enrichment (above the line) or reduction (below the line) of amino acid frequency is determined using the human SwissProt as background. The apoptotic caspase-cleaved iceLogo (**C**) represents all cleavages following aspartic acid (P1=D) in the apoptotic dataset.

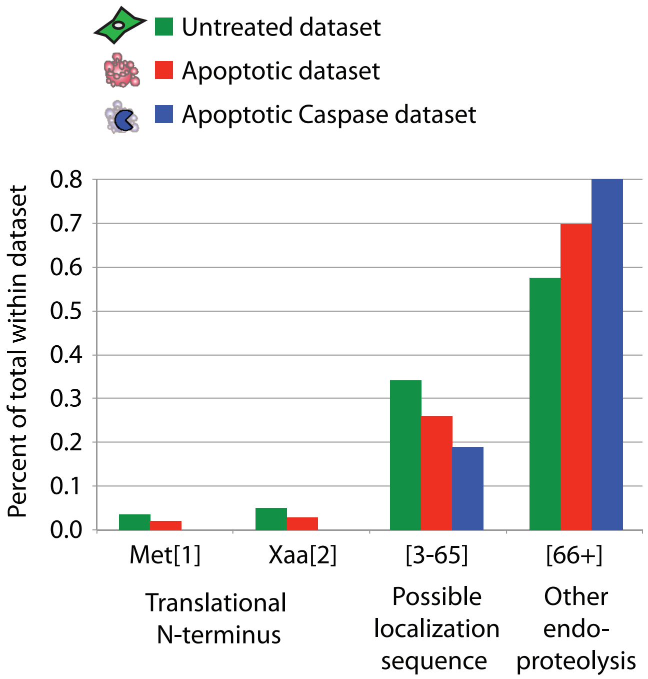


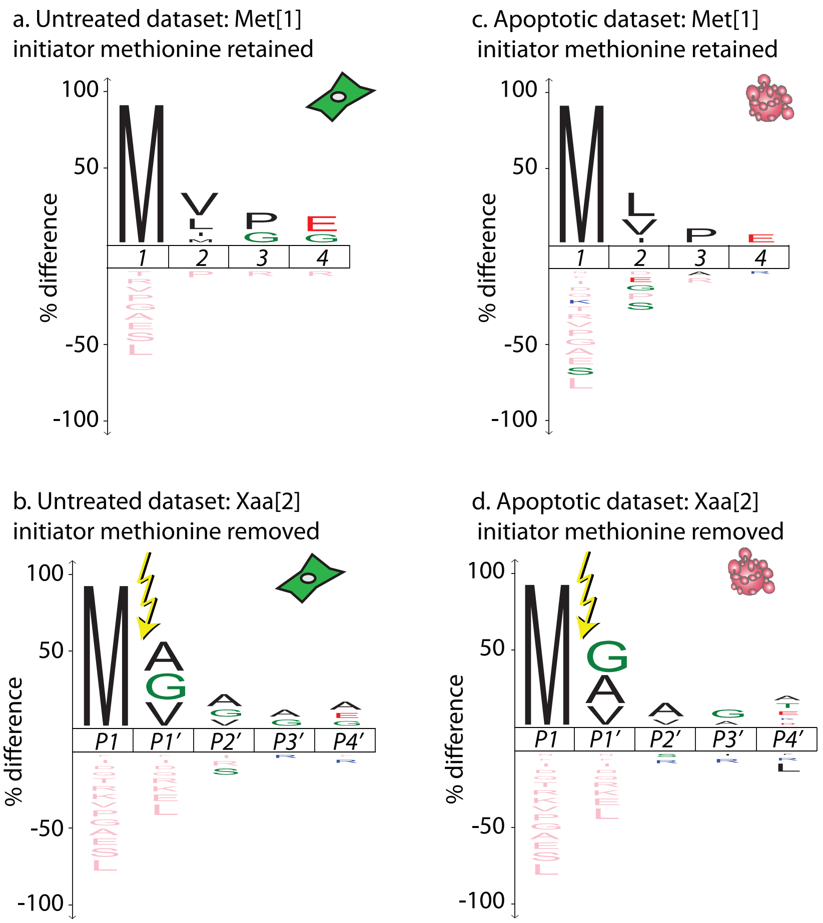
Figure 4: There are three distinct groups of proteolytic processing within the data: (i) processing around the methionine at the translational N-terminus (N-termini observed at residues 1 and 2), (ii) cleavage of possible secretory or transit peptides during organelle trafficking (observed residues 3-65), and (iii) other endoproteolytic events (considered as cuts after residue 65). The untreated and apoptotic datasets had similar levels of translational N-terminus labeling (~10%), but differed for the latter categories, with the apoptotic datasets having more cleavage events past residue 65. The apoptotic caspase- cleaved set is shifted even more towards endoproteolytic cleavages than the apoptotic set.

(i) Initiator Methionine Processing

Eukaryotic proteins are typically co-translationally acetylated, rendering the translational N-terminus inaccessible to the subtiligase labeling technique (*29, 30*). Recent work suggests that this acetylation is largely irreversible (*31*). However, there are some proteins, both with and without initiator methionine removal, that do not undergo co-translational acetylation; these have translational N-termini that are accessible to labeling. In the full database (both untreated and apoptotic samples), we observed labeling of the initiator methionine in 154 proteins (12% of identified proteins) **(Supplemental Table 3A)**, and labeling of the second residue, indicating methionine removal, in 198 proteins (15% of identified proteins) **(Supplemental Table 3B)**. It is noteworthy that the majority of these translational N-termini are not yet annotated in SwissProt (**Supplemental Figure 2**).

The sequence logo for proteins in the untreated dataset retaining the initiator methionine suggests enrichment for a hydrophobic amino acid at the second residue (**Figure 5A),** whereas proteins with the initiator methionine removed had enrichment for small amino acids at the second residue (**Figure 5B**). This is largely consistent with previous studies showing that methionine removal is most efficient for proteins with small amino acids following the initiator methionine (*32*). Almost identical patterns were seen for the proteins in the apoptotic set (**Figures 5C-D**). Out of 182 methionine processing events seen in healthy cells, only 35 were not also found in the apoptotic set, suggesting there is little change in the translational N-terminal processing events during apoptosis.

Figure 5


a. Untreated dataset: Met[1] initiator methionine retained

c. Apoptotic dataset: Met[1] initiator methionine retained

b. Untreated dataset: Xaa[2] initiator methionine removed

d. Apoptotic dataset: Xaa[2] initiator methionine removed

Figure 5: The iceLogos for the untreated (**A**, **B**) and apoptotic (**C**, **D**) sets are very similar for processing around the methionine at the translational N-terminus. Each iceLogo contains unique N-termini against the SwissProt human background within each dataset. (**A**) and (**C**) represent retention (but not acetylation) of the initiator methionine (Met[1]); (**B**) and (**D**) represent removal of the initiator methionine without acetylation of the second residue (Xaa[2]). Two proteins are labeled at residue 1 but annotated in UniProt as not containing an initiator methionine (Ig lambda chain V-IV region HII (P01717, serine) and 40S ribosomal protein S30 (P62861, lysine)) were removed from the datasets for the iceLogo creation.

(iia) Mitochondrial Transit Peptide Removal

Proteins expressed from nuclear genes but destined for the mitochondria generally contain positively charged N-terminal regions that direct them to the mitochondrial import machinery (*33*). Once inside the mitochondria, these mitochondrial transit peptides (mTPs) are removed by the mitochondrial processing peptidase (MPP), and in some cases the truncated proteins are further processed by other proteases that remove one or a few additional residues (*27*). While the mitochondrial proteome has been well characterized, data on the precise location of mTP cleavage sites remains minimal. Moreover, sequence specificities of the proteases involved are only partially understood (*24, 28, 34*). Considering only the untreated dataset to avoid possible apoptosis-induced cleavages, we identified roughly 250 labeled N-termini from the approximately 1,000 human SwissProt proteins that are in MitoCarta, a highly curated database of mitochondrial proteins (*24*).

The distribution of N-terminus placement found in mitochondrial proteins is quite different from that of non-mitochondrial proteins. We see a significant spike in the range of position 10 to position 65, roughly the location of most known mTP cleavage sites **(Figure 6A)** (*27*). We therefore focused our examination on the 171 N-termini seen in this range in MitoCarta proteins. For the purpose of these analyses, in cases where one protein had more than one N-terminus in this region, we chose the site closest to the translational N- terminus. We found that 58 had no mTP cleavage site annotated in SwissProt, and 67 had an annotated site different from the one we observed **(Supplemental Figure 3A and Supplemental Table 3C)**. Additionally, 24 peptide removal start sites are within one residue of their SwissProt annotations. This may reflect the secondary proteolysis known to occur in the mTP removal pathway (*27*), and may be a part of a mitochondrial version of the N-end rule (*35*). It is notable that 46 out of the 67 cases (69%)

where our data disagrees with SwissProt annotation involve cleavage sites that SwissProt describes as “By Similarity,” “Potential,” or “Probable,” indicating a lack of strong evidence for these cuts. In contrast, in only 17 out of the 46 cases (37%) where our data agree with SwissProt does the annotation contain this qualifying language (Supplemental Table 3C). We generated an iceLogo for the 16 residues before and 4 residues after the 171 mTP cleavage sites **(Figure 6B)**. As expected based on previous studies, the iceLogo shows enrichment for arginine and de-enrichment for acidic residues throughout the transit peptide. The strongest arginine signal is in the P2 and P3 positions, as expected from previous studies (*28*) and is the most enriched signal in our untreated position 10-65 logo **(Figure 6C)**.

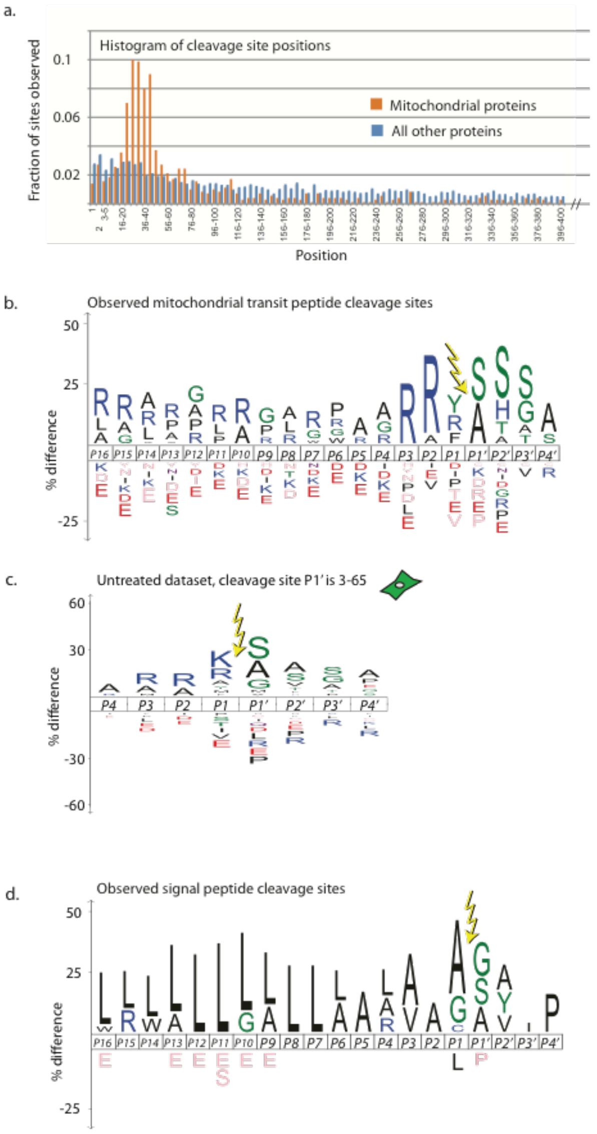


Figure 6: **A**. Annotated mitochondrial proteins were greatly enriched for labeling between residues 10-65 compared to all other proteins in the untreated dataset, reflecting labeling at mitochondrial transit peptide removal sites. **B-D**. iceLogos for subsets of the set of 28 cleavages at positions 3 to 65 in the untreated dataset: (**B**) all unique N-termini from proteins thought to be mitochondrial; (**C**) all N-termini in the untreated dataset labeled between 10-65; and (**D**) all unique N-termini thought to contain signal peptides.

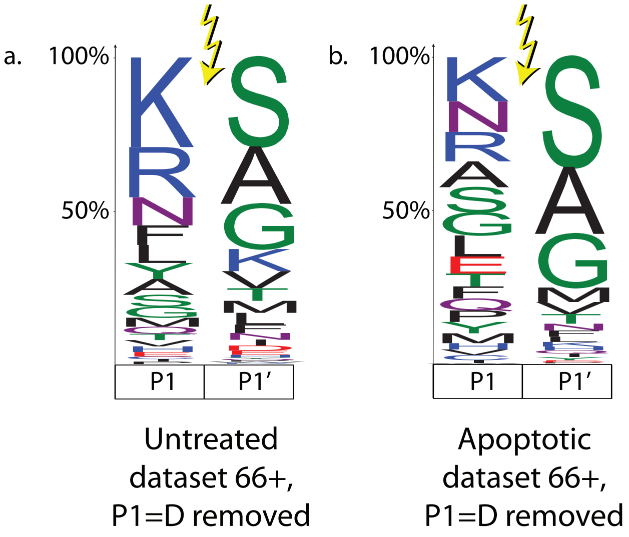
(iib) Signal Peptide Removal

Proteins destined for the secretory pathway are subjected to proteolytic removal of signal peptides (SPs) near the N-terminus (*36*). As with mTPs, we looked for these cleavages only in the untreated dataset to avoid any apoptosis-specific events. In this case, we focused on the 63 proteins that were annotated in SwissProt as having a signal peptide removal site between positions 10 and 65 **(Supplemental Table 3D)**. Most of these were ER or Golgi resident proteins; our technique does not efficiently capture secreted proteins after they have dissociated from the cell. In 30 of these proteins, the cleavage site we observed matched the one annotated in SwissProt **(Supplemental Figure 3B)**. In this case, the qualifiers “By Similarity” and “Potential” were used by SwissProt to describe at least 60% of their signal peptide removal site annotations both in the set that agrees with our data, and the set that does not. We cannot be sure what accounts for the differences between our data and SwissProt, but we hope that these data will prove useful to researchers who work in this area. As with the mTP sites, we generated an iceLogo for positions P16 to P4’ relative to the cleavage site **(Figure 6D)**. In this case, we saw a substantial enrichment for leucine residues in the P16– P6 positions, which is consistent with the known requirement for hydrophobicity in the signal peptide to facilitate interaction with the ER membrane (*36, 37*).

(iii) Endoproteolysis before and after apoptosis

We next consider the endoproteolytic events that occur after residue 65, which represent the majority of our identified N-termini **(Figure 4)**. Although many of the cleavages that occur before or at residue 65 are endoproteolytic, we chose to focus on the 66+ set to reduce the contamination of this set with possible signal or transit peptide removal sites. Considering the apoptotic set cleaved after residue 65, 28% of them occur with an aspartic acid residue at the P1 position, suggesting a caspase cleavage event **(Supplemental Figure 4)**. In order to visualize non-caspase protease activity, we removed all cleavage sites with aspartic acid at the P1 position and then used iceLogo to generate filled logos (where letter height represents amino acid frequency) for the P1 and P1' positions in the untreated and apoptotic datasets **(Figure 7)**. Both logos show the predominance of basic amino acids (arginine and lysine) at P1 and small amino acids (glycine, serine or alanine) at P1'. However, there is an overall decrease in the fraction of cleavages following arginine or lysine in the apoptotic dataset. This likely reflects the induction of other non-caspase and non-tryptic proteases during apoptosis that are different from proteases in healthy cells.

Many human proteins undergo degradation by the proteasome, however we do not expect to see very many proteasome products using our subtiligase method because the majority are quickly degraded into their amino acid components (*38*). Those that remain intact (for example, in order to be displayed on an MHC class I complex) are mostly less than 10 amino acids long; the fractionation steps required to separate the small biotin label from the larger labeled protein fragments causes significant loss of short peptides. Furthermore, our mass spectrometry data searches only considered peptides with a tryptic site on the C-terminus, and only a subset of proteasome peptide products would fit this criteria.



Untreated dataset 66+, P1=D removed

Apoptotic dataset 66+, P1=D removed

Figure 7: Filled logos for endoproteolysis occurring at residue 66 or above for the untreated (**A**) and apoptotic (**B**) datasets with all aspartic acid cleavages removed (9% of the untreated and 28% of the apoptotic N-termini). The size of each letter represents its relative frequency within the dataset. Distributions are very similar for the P1' position, but show differences at the P1 position.

The N-end rule before and after apoptosis

The P1' position of a cleavage site is important for the half-life of the resulting protein fragment, as determined by the Arg/N-end rule pathway (*35*). Many proteases, including caspases (*39*), prefer the small amino acids glycine, serine or alanine in the P1' pocket; in fact, these three residues make up 32% of the P1' residues of the approximately 56,000 protease cleavages (of both native and synthetic substrates) described in substrate section of the MEROPS database. However, small amino acids are not a requirement, as we saw all twenty possible amino acids in the P1' position in the untreated, apoptotic and apoptotic caspase-cleaved datasets **(Table 1)**. The untreated dataset had a greater proportion of cleavages with charged amino acids (lysine, arginine, aspartic acid, and glutamic acid) at the P1' position than apoptotic or apoptotic caspase-cleaved datasets, while the apoptotic and apoptotic caspase-cleaved datasets had more cleavages with serine and glycine in the P1' position.

The Arg/N-end rule pathway degrades proteins and proteolysis products through ubiquination and targeting to the proteasome. Different N-terminal amino acids may be stabilizing or destabilizing, and thus affect the half-life of the protein (*35*). To investigate the potential for a biological effect of the proteolysis products, we analyzed the data with respect to the theoretical half-lives for each P1' amino acid (*40*). We grouped the amino acids into stabilizing P1' amino acids (half-life greater than 20 hours), destabilizing (half-life less than 1.5 hours), and intermediate (half-live between 1.5 and 20 hours) **(Table 1)**. For all three datasets, the majority (53-60%) of N-termini were found in the intermediate group. However, there is a clear difference in the pattern of stabilizing and destabilizing cuts depending on cell condition. Almost 14% of all untreated cleavage events occurred before destabilizing amino acids. In contrast, only 7% of apoptotic cleavages and 4% of apoptotic caspase cleavages occurred before destabilizing amino acids, with most shifting into the intermediate range of predicted half-lives. In general, untreated proteolysis events were more destabilizing and had shorter theoretical half-lives than apoptotic events, and caspase cleavages leave particularly stable and longer lasting predicted N-termini.

			🟩			🟥			🟦		
	P1'	**N-End Half-Life (hr)**	**# seen in untreated**	**% of total**	**Group %**	**# seen in apoptotic (including caspase)**	**% of total**	**Group %**	**# seen in apoptotic caspase-cleaved**	**% of total**	**Group %**
Destabilizig	**Q**	0.8	17	0.82	14.31	63	0.92	7.16	3	0.18	4.04
	R	1	7	0.34		8	0.12		1	0.06	
	E	1	26	1.26		31	0.45		3	0.18	
	F	1.1	55	2.66		151	2.20		34	1.99	
	D	1.1	25	1.21		22	0.32		2	0.12	
	C	1.2	2	0.10		7	0.10		2	0.12	
	K	1.3	123	5.94		93	1.36		10	0.59	
	N	1.4	41	1.98		116	1.69		14	0.82	
↓	**S**	1.9	531	25.66	55.49	2185	31.87	61.28	521	30.54	55.51
	Y	2.8	19	0.92		101	1.47		36	2.11	
	W	2.8	6	0.29		11	0.16		5	0.29	
	H	3.5	21	1.01		76	1.11		17	1.00	
	A	4.4	409	19.77		1461	21.31		286	16.76	
	L	5.5	80	3.87		141	2.06		29	1.70	
	T	7.2	82	3.96		227	3.31		53	3.11	
Stabilizing	**P**	20	14	0.68	30.21	25	0.36	31.56	0	0.00	40.45
	I	20	32	1.55		75	1.09		18	1.06	
	M[1]	30	84	4.06		266	3.88		23	1.35	
	G	30	340	16.43		1504	21.93		609	35.70	
	V	100	155	7.49		294	4.29		40	2.34	
	Total		***2069***	***100***	***100***	***6857***	***100***	***100***	***1706***	***100***	***100***

[1] Initiator methionines removed

Table 1: P1’ amino acid frequency and Arg/N-End Rule Pathway half-lives.

Discussion:

The subtiligase N-terminal labeling method yields high selectivity for α-amines with accurate peptide and protein identifications and very low false discovery rate (FDR < 1%). Overall, we observed 3206 proteins, corresponding to almost 16% of the entire SwissProt human proteome and covering over 6 logs in protein abundance. We see internally consistent labeling between the data sets. For example, the same mitochondrial transit sites and initiator methionine sites were labeled in 35-40 out of 44 untreated and apoptotic experiments. These cleavage events are independent of apoptosis, and therefore show the consistency in our labeling and detection method. Additionally, we believe that little bias originates due to subtiligase specificity. For example, all 20 amino acids were labeled at P1', including proline and valine, which are known to be slow substrates for subtiligase *in vitro* (10).

We are confident that the distinctions between healthy and apoptotic datasets reflect the biology of the human cell lines, as the apoptotic samples showed decreased cell viability and increased caspase activity (as measured both in cell culture and in observed caspase-cleaved N-termini relative to the untreated samples). In fact, 129 of the 140 aspartic-cleaved N-termini identified in healthy cells were also in the 1706 apoptotic N- terminus set, and likely reflects a small population of apoptotic cells within the healthy cell population. Comparing our apoptotic protein dataset to other published datasets shows that we capture most known apoptosis-related proteins. We have labeled more than 60% of the proteins listed in the ApoptoProteomics database (*41*), an apoptotic proteomics database. Additionally, 75 of our caspase substrates overlap with the literature-curated CASBAH database (*16*), and 79 overlap with the caspase-3, -6, or -7 substrates listed in MEROPS (*17*) (in both of these comparisons, we excluded >200 entries present in these databases but derived from the original subtiligase study (*8*)). These 1706 apoptotic caspase-cleaved sites, in combination with MEROPS, the CASBAH, and other work from our lab (*8, 12, 13*), bring the total known human caspase cleavage sites to more than 2200. Importantly, the DegraBase contains a larger number of new non-caspase proteolytic events that have yet to be assigned to a specific protease. Surprisingly, there are only 45 sites in the 2900 non-caspase, non-tryptic (not cleaved after lysine or arginine) endoproteolytic events (residue 66+) that are present in the MEROPS database (release 9.6), and only 10 of these sites are annotated as “physiological” cleavages in MEROPS **(Supplemental Table 4).** Interestingly, there is little exoproteolysis of these intracellular proteins, whereas laddering produced by sequential exoproteolysis was observed in 24% of all proteins identified in a subtiligase-based study of human serum (*9*).

It is probable that only a subset of the total identified apoptotic proteolytic substrates needs to be cleaved to complete apoptosis. For example, there are 1706 putative caspase cleavages, but 784 was the largest number of sites labeled in any single cellular experiment. Additionally, about 50% of all apoptotic N-termini identified were only seen in one experiment. This may reflect the diversity in drug induction and cell types chosen as well as the expected stochasticity in mass spectrometry and our labeling technology. Some of the apoptotic labeling patterns may also be due to induced polyspecific proteases, like caspases, with large and diverse sets of possible substrates. Only 110 sites (87 caspase-cleaved) from 109 proteins were seen consistently in at least 10 apoptotic experiments and only one or zero untreated experiments (**Supplemental Table 5**). Interestingly, these common cleavages have a wide kinetic range. Some are shown to be cleaved by up to 3 different apoptotic caspases (*12*), and many have homologs in mouse and fly that are also known caspase substrates (*14*). These apoptosis-enriched sites may represent important apoptotic nodes, while other cleavage sites may be unique to the experimental conditions, or possibly to be bystander cleavages.

Remarkably, the protease actors in healthy and apoptotic cells appear to target an overlapping set of substrates, but not always at the same cleavage sites (**Figure 2**). This could allow for different regulation of these targets depending on the cellular conditions. An important difference between the untreated and apoptotic datasets is the theoretical half-lives of the newly created N-termini. We found the neo-N-termini created by caspase cleavages have a higher proportion of stabilizing N-terminal amino acids than those in the untreated and apoptotic datasets (**Table 1**). A similar conclusion that apoptotic fragments tended to persist during apoptosis was reached using the PROTOMAP method (*42*). Stable apoptotic cleavage products, in particular protein fragments of caspase substrates, may function in a different manner from the parent protein. We realize that these half-life values are largely dependent on the assay method and may not be representative of specific *in vivo* half-lives of a given protein fragment. However, Piatkov et al. have recently showed the greater extent that caspases and the Arg/N-End Rule pathway interact: proapoptotic protein fragments contain evolutionarily conserved destabilizing N-terminal amino acids, targeting them for quick degradation by the proteasome in healthy cells; caspases cleave and inactivate members of the proteasome pathway, preventing peptide degradation (*43*). Indeed, we do see caspase cleavages in UBR4 and UBR5 which function as the Arg/N-End Rule pathway E3 Ubiquitin ligases.

During apoptosis, caspase cleavages may result in loss-of-function, gain-of-function, or no functional effect on the substrate. As many caspase cleavage events occur between protein domains (*8*), many substrates have the potential for gain of function events in which a catalytic domain is relocated or an inhibitor removed. Several such cleavages have been thoroughly studied in kinases, as reviewed by Kurokawa and Kornbluth (*44*). In a preliminary search through our database, we see enrichment for caspase cleavages between annotated domains. For example, in the kinase family, 52 of the 57 caspase- derived N-termini occurred between domains, compared to 51 of 76 non-caspase apoptotic N-termini and 15 of 26 untreated N-termini (*45, 46*) **(Supplemental Table 6).** This is consistent with a recent study by Dix, et al., that demonstrated crosstalk between caspases and kinases, where phosphorylation can direct caspase cleavages on kinases that may lead to a change in kinase activity (*47*).

In sum, we provide and unbiased and global annotation of the human cellular α-aminome. The data are consolidated into a searchable database, the DegraBase, revealing a large amount homeostatic and apoptotic proteolysis in cells. To our knowledge, these untreated and apoptotic datasets are the most extensive published to date using a single methodology. We confidently identified many new sites related to protein processing, including initiator methionine retention or removal, and the specific cleavage locations for signal or transit peptide removal during protein trafficking. Additionally, our dataset shows the abundance of healthy homeostatic and non-caspase apoptotic endoproteolytic events that occur in cells. We hope that our colleagues across many areas of biology will find the DegraBase to be a useful resource for further understanding and characterization of proteolytic events in cells.

Acknowledgements

We thank David Wildes, Alan Barber, Jason Porter, Gemma Holliday and Aenoch Lyn for their helpful discussions. We thank David Maltby, Jonathan Trinidad, Al Burlingame and the rest of the UCSF Mass Spectrometry Facility, plus Iman Mohtashemi and Yan Chen of Thermo Scientific, for mass spectrometry assistance. This project was supported by a grant from the UCSF Stephen and Nancy Grand Multiple Myeloma Translational Initiative. It was supported by the National Science Foundation GRFP (E.D.C.), the University of California Systemwide Biotechnology Research & Education Program GREAT Training Grant #2007- 20 (E.D.C.), NIH Training Grant T32 GM007175 (J.E.S.), and Banting Postdoctoral Fellowships funded by the Government of Canada (O.J.). Additional support from National Institutes of Health R01 GM081051, R01 GM097316 and RO1 CA154802 (J.A.W.), The Rogers Family Foundation (J.A.W.), and an award from the Sandler Program in Biological Sciences (J.A.W.). Mass spectrometry was performed at the Bio-Organic Biomedical Mass Spectrometry Resource at UCSF, which is supported by grants from the National Center for Research Resources (5P41RR001614) and the National Institute of General Medical Sciences (8P41GM103481 and 1S10RR015804) from the National Institutes of Health.

References

1. Arnesen, T. (2011) Towards a functional understanding of protein N-terminal acetylation. *PLoS Biol* 9, e1001074
2. Starheim, K. K., Gevaert, K., and Arnesen, T. (2012) Protein N-terminal acetyltransferases: when the start matters. *Trends Biochem Sci* 37, 152-161
3. van den Berg, B. H., and Tholey, A. (2012) Mass spectrometry-based proteomics strategies for protease cleavage site identification. *Proteomics* 12, 516-529
4. Staes, A., Impens, F., Van Damme, P., Ruttens, B., Goethals, M., Demol, H., Timmerman, E., Vandekerckhove, J., and Gevaert, K. (2011) Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc* 6, 1130-1141
5. Impens, F., Colaert, N., Helsens, K., Plasman, K., Van Damme, P., Vandekerckhove, J., and Gevaert, K. (2010) MS-driven protease substrate degradomics. *Proteomics* 10, 1284-1296
6. auf dem Keller, U., and Schilling, O. (2010) Proteomic techniques and activity-based probes for the system-wide study of proteolysis. *Biochimie* 92, 1705-1714
7. Drag, M., Bogyo, M., Ellman, J. A., and Salvesen, G. S. (2010) Aminopeptidase fingerprints, an integrated approach for identification of good substrates and optimal inhibitors. *J Biol Chem* 285, 3310-3318
8. Mahrus, S., Trinidad, J. C., Barkan, D. T., Sali, A., Burlingame, A. L., and Wells, J. A. (2008) Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* 134, 866-876
9. Wildes, D., and Wells, J. A. (2010) Sampling the N-terminal proteome of human blood. *Proc Natl Acad Sci U S A* 107, 4561-4566
10. Chang, T. K., Jackson, D. Y., Burnier, J. P., and Wells, J. A. (1994) Subtiligase: a tool for semisynthesis of proteins. *Proc Natl Acad Sci U S A* 91, 12544-12548
11. Jackson, D. Y., Burnier, J., Quan, C., Stanley, M., Tom, J., and Wells, J. A. (1994) A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. *Science* 266, 243-247
12. Agard, N. J., Mahrus, S., Trinidad, J. C., Lynn, A., Burlingame, A. L., and Wells, J. A. (2012) Global kinetic analysis of proteolysis via quantitative targeted proteomics. *Proc Natl Acad Sci U S A* 109, 1913-1918
13. Shimbo, K., Hsu, G. W., Nguyen, H., Mahrus, S., Trinidad, J. C., Burlingame, A. L., and Wells, J. A. (2012) Quantitative profiling of caspase-cleaved substrates reveals different drug-induced and cell-type patterns in apoptosis. *Proc Natl Acad Sci U S A* 109, 12432-12437
14. Crawford, E. D., Seaman, J. E., Barber, A. E., 2nd, David, D. C., Babbitt, P. C., Burlingame, A. L., and Wells, J. A. (2012) Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis. *Cell Death Differ* 19, 2040-2048
15. Agard, N. J., Maltby, D., and Wells, J. A. (2010) Inflammatory stimuli regulate caspase substrate profiles. *Mol Cell Proteomics* 9, 880-893
16. Luthi, A. U., and Martin, S. J. (2007) The CASBAH: a searchable database of caspase substrates. *Cell Death Differ* 14, 641-650
17. Rawlings, N. D., Barrett, A. J., and Bateman, A. (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 40, D343-350

18. Lange, P. F., Huesgen, P. F., and Overall, C. M. (2012) TopFIND 2.0--linking protein termini with proteolytic processing and modifications altering protein function. *Nucleic Acids Res* 40, D351-361

19. Yoshihara, H. A., Mahrus, S., and Wells, J. A. (2008) Tags for labeling protein N-termini with subtiligase for proteomics. *Bioorg Med Chem Lett* 18, 6000-6003

20. Chalkley, R. J., Baker, P. R., Medzihradszky, K. F., Lynn, A. J., and Burlingame, A. L. (2008) In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Mol Cell Proteomics* 7, 2386-2398

21. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40, D71-75

22. Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., and von Mering, C. (2012) PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. *Mol Cell Proteomics* 11, 492-500

23. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* 6, 786-787

24. Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., and Mootha, V. K. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 134, 112-123

25. Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20, 3710-3715

26. Schechter, I., and Berger, A. (1967) On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 27, 157-162

27. Vogtle, F. N., Wortelkamp, S., Zahedi, R. P., Becker, D., Leidhold, C., Gevaert, K., Kellermann, J., Voos, W., Sickmann, A., Pfanner, N., and Meisinger, C. (2009) Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* 139, 428-439

28. Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2, 953-971

29. Brown, J. L., and Roberts, W. K. (1976) Evidence that approximately eighty per cent of the soluble proteins from Ehrlich ascites cells are Nalpha-acetylated. *J Biol Chem* 251, 1009-1014

30. Van Damme, P., Arnesen, T., and Gevaert, K. (2011) Protein alpha-N-acetylation studied by N-terminomics. *Febs J* 278, 3822-3834

31. Hwang, C. S., Shemorry, A., and Varshavsky, A. (2010) N-terminal acetylation of cellular proteins creates specific degradation signals. *Science* 327, 973-977

32. Bradshaw, R. A., Brickey, W. W., and Walker, K. W. (1998) N-terminal processing: the methionine aminopeptidase and N alpha-acetyl transferase families. *Trends Biochem Sci* 23, 263-267

33. Neupert, W., and Herrmann, J. M. (2007) Translocation of proteins into mitochondria. *Annu Rev Biochem* 76, 723-749

34. Taylor, A. B., Smith, B. S., Kitada, S., Kojima, K., Miyaura, H., Otwinowski, Z., Ito, A., and Deisenhofer, J. (2001) Crystal structures of mitochondrial processing peptidase reveal the mode for specific cleavage of import signal sequences. *Structure* 9, 615-625

35. Varshavsky, A. (2011) The N-end rule pathway and regulation by proteolysis. *Protein Sci* 20, 1298-1345

36. Rapoport, T. A. (1992) Transport of proteins across the endoplasmic reticulum membrane. *Science* 258, 931-936

37. Imai, K., and Nakai, K. (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10, 3970-3983

38. Kisselev, A. F., Akopian, T. N., Woo, K. M., and Goldberg, A. L. (1999) The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. *J Biol Chem* 274, 3363-3371

39. Stennicke, H. R., Renatus, M., Meldal, M., and Salvesen, G. S. (2000) Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8. *Biochem J* 350 Pt 2, 563-568

40. Gonda, D. K., Bachmair, A., Wunning, I., Tobias, J. W., Lane, W. S., and Varshavsky, A. (1989) Universality and structure of the N-end rule. *J Biol Chem* 264, 16700-16712

41. Arntzen, M. O., and Thiede, B. (2012) ApoptoProteomics, an integrated database for analysis of proteomics data obtained from apoptotic cells. *Mol Cell Proteomics* 11, M111 010447

42. Dix, M. M., Simon, G. M., and Cravatt, B. F. (2008) Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell* 134, 679-691

43. Piatkov, K. I., Brower, C. S., and Varshavsky, A. (2012) The N-end rule pathway counteracts cell death by destroying proapoptotic protein fragments. *Proc Natl Acad Sci U S A* 109, E1839-1847

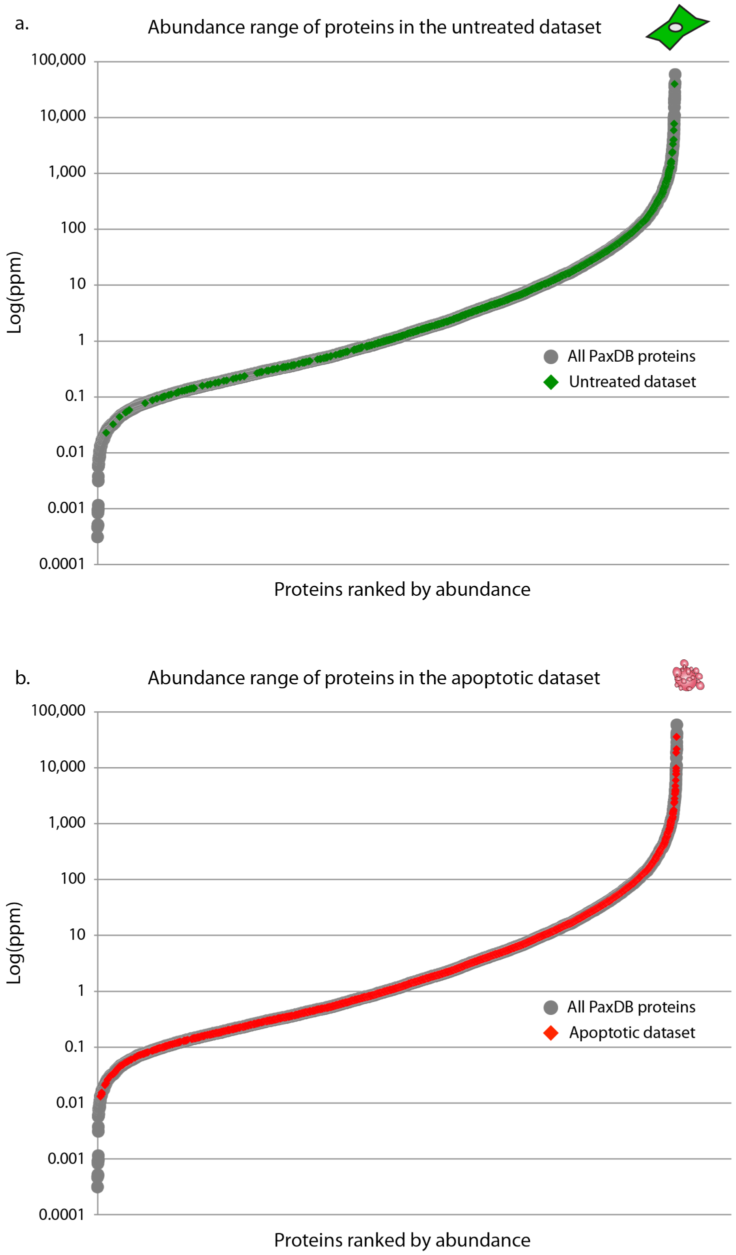
44. Kurokawa, M., and Kornbluth, S. (2009) Caspases and kinases in a death grip. *Cell* 138, 838-854

45. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* 298, 1912-1934

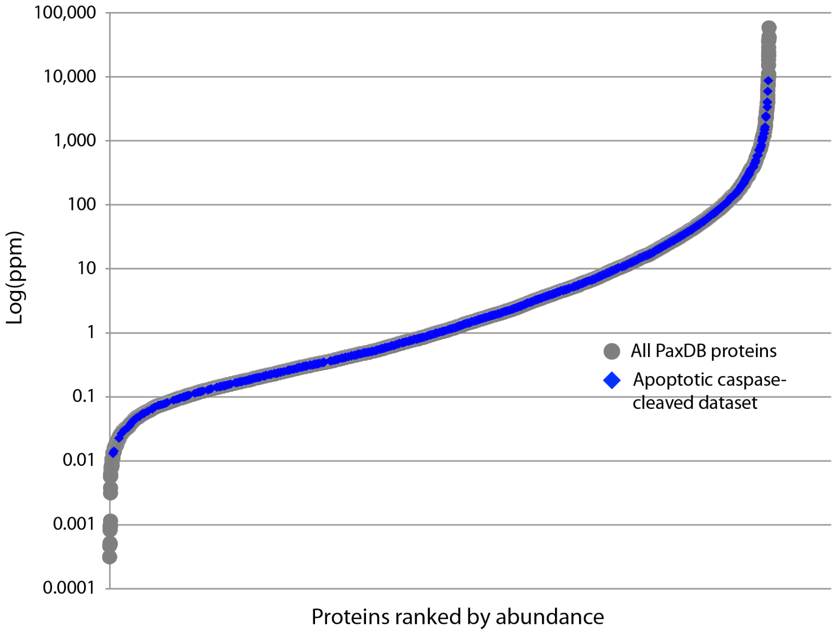
46. Miranda-Saavedra, D., and Barton, G. J. (2007) Classification and functional annotation of eukaryotic protein kinases. *Proteins* 68, 893-914

47. Dix, M. M., Simon, G. M., Wang, C., Okerberg, E., Patricelli, M. P., and Cravatt, B. F. (2012) Functional Interplay between Caspase Cleavage and Phosphorylation Sculpts the Apoptotic Proteome. *Cell* 150, 426-440

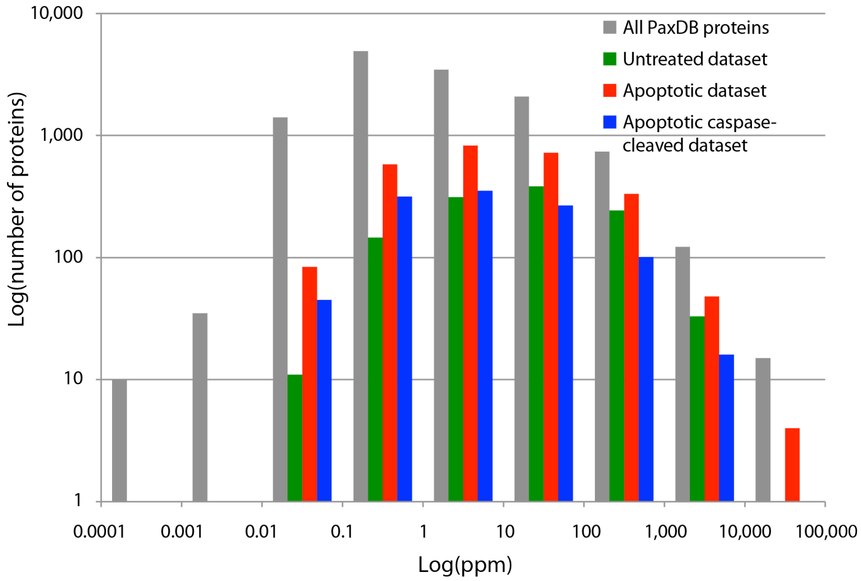
Supplemental Figure 1: Protein abundance



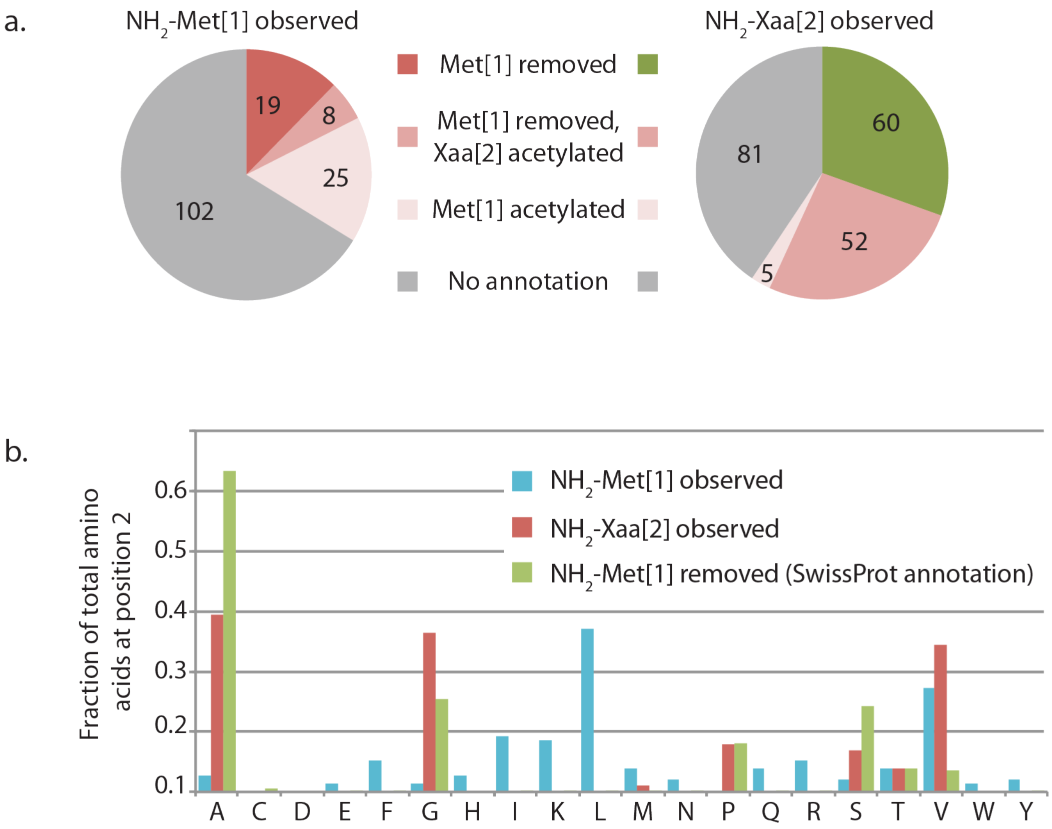
c. Abundance range of proteins in the apoptotic caspase-cleaved dataset



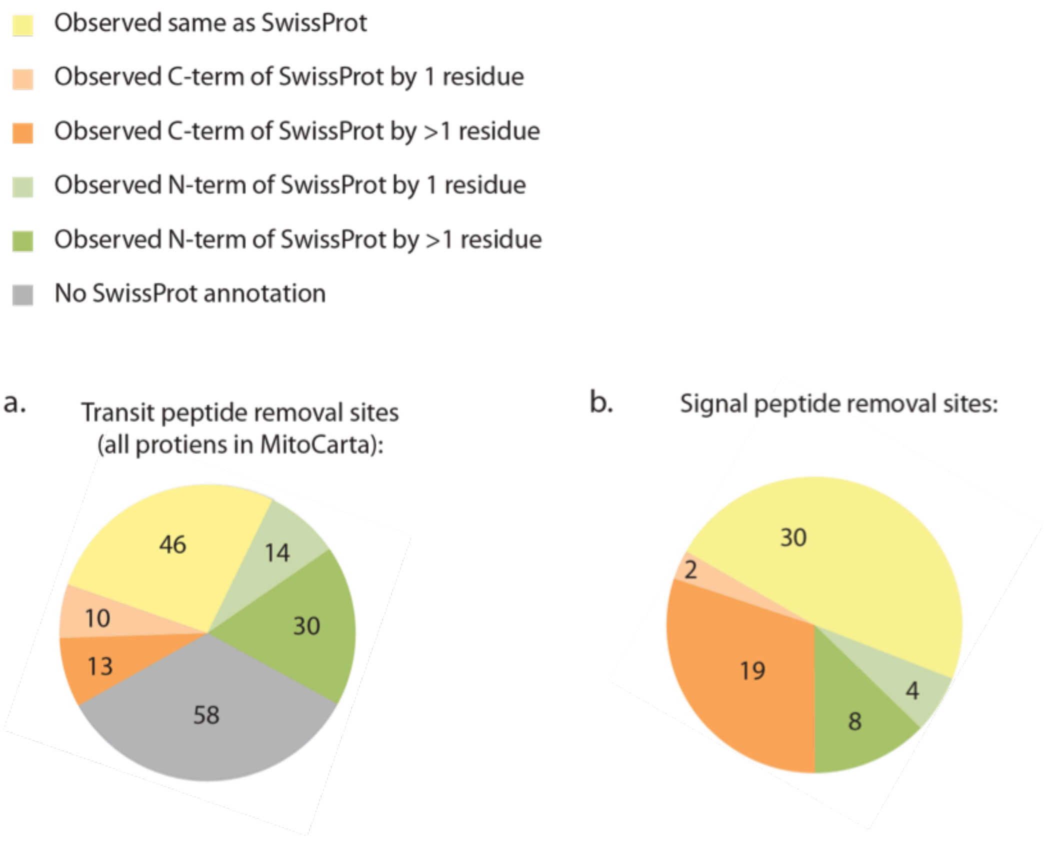
d. Abundance historgram for all datasets



Supplemental Figure 2



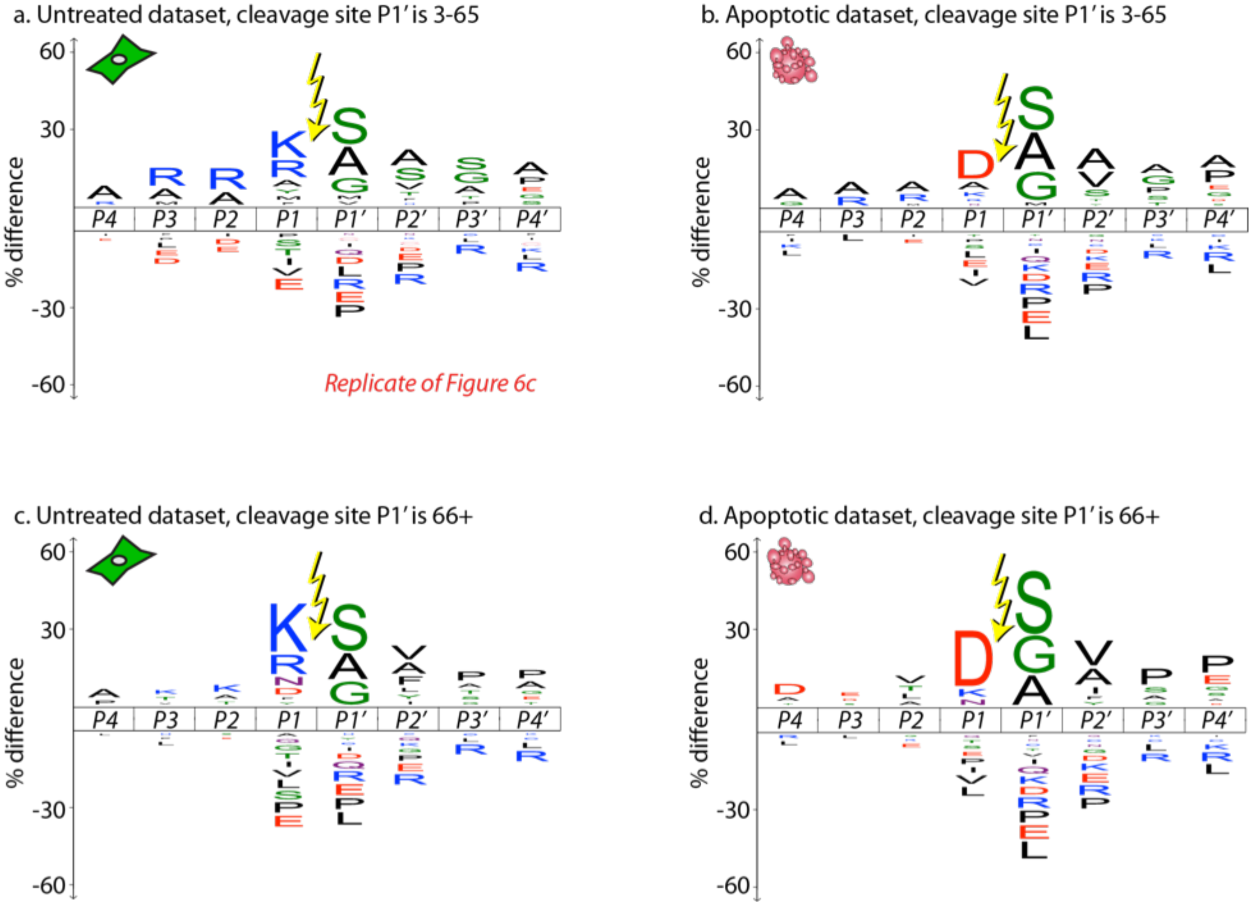
Supplemental Figure 3



a. Transit peptide removal sites (all protiens in MitoCarta):

b. Signal peptide removal sites:

Supplemental Figure 4: IceLogos for untreated and apoptotic datasets, P1′ site 3-65 and 66+



a. Untreated dataset, cleavage site P1′ is 3-65

b. Apoptotic dataset, cleavage site P1′ is 3-65

c. Untreated dataset, cleavage site P1′ is 66+

d. Apoptotic dataset, cleavage site P1′ is 66+

Supplemental Tables and Files

For Supplemental Tables and Files, see http://www.mcponline.org/content/12/3/813.

Supplemental File 1 - Filemaker Pro info and scripts

Supplemental File 2 - Filemaker Pro file database

Supplemental File 3 - Excel version of database

Supplemental Table 1 - Untreated and apoptotic datasets.

Supplemental Table 2 - Summary of Go Term analysis

Supplemental Table 3 - N-terminal, mitochondrial transit and signal peptides from the database.

Supplemental Table 4 - Overlap of database with MEROPS.

Supplemental Table 5 - List of most frequently seen peptides throughout analyses.

Supplemental Table 6 - Kinase cleavage patterns in the database.

Chapter 3

Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis.

This work was published in Cell Death and Differentiation:
Crawford ED, Seaman JE, Barber AE, 2nd, David DC, Babbitt PC, Burlingame AL, et al. Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis. Cell Death Differ. 2012;19(12):2040-8.

Abstract

Caspases, cysteine proteases with aspartate specificity, are key players in programmed cell death across the metazoan lineage. Hundreds of apoptotic caspase substrates have been identified in human cells. Some have been extensively characterized, revealing key functional nodes for apoptosis signaling and important drug targets in cancer. But the functional significance of most cuts remains mysterious. We set out to better understand the importance of caspase cleavage specificity in apoptosis by asking which cleavage events are conserved across metazoan model species. Using N-terminal labeling followed by mass spectrometry, we identified 257 caspase cleavage sites in mouse, 130 in *Drosophila*, and 50 in *C. elegans*. The large majority of the caspase cut sites identified in mouse proteins were found conserved in human orthologs. However, while many of the same proteins targeted in the more distantly related species were cleaved in human orthologs, the exact sites were different. Furthermore, similar functional pathways are targeted by caspases in all four species. Our data suggest a model for the evolution of apoptotic caspase specificity that highlights the hierarchical importance of functional pathways over specific proteins, and proteins over their specific cleavage site motifs.

Introduction

Apoptosis, a form of programmed cell death, is conserved across the entire metazoan lineage, and is crucial for removing harmful or unneeded cells. Apoptosis has been extensively studied in several model organisms, including mouse, *Drosophila*, and *C. elegans*. While some differences in the early stages of the pathway exist (1), in all cases the later stages of apoptosis involve strong activation of one or more caspases, cysteine class proteases with exquisite specificity for cleavage after aspartic acid.

To better understand how caspase activation leads to cell death, there has been considerable interest in identifying the protein targets of apoptotic caspases in human cells. Recently, the number of observed human apoptotic caspase cleavage targets has grown to several hundred. (2-5) (For a recent review, see (6)). Usually these proteins are cleaved at a single site in a loop or disordered region, and the cut may cause a gain or loss of function in the substrate protein, or it may have no functional effect. Detailed molecular studies have established important roles of some substrates; many are key players, and sometimes good drug targets or biomarkers, in cancer and other diseases of apoptosis misregulation (6). Nonetheless, for the vast majority of cleavages, functional significance has not yet been established.

In the present study, we have used a subtiligase N-terminal labeling strategy to identify caspase cleavages across four metazoan model organisms representing roughly 600 million years of evolution (7) (See Figure 1A and B for a basic schematic). We base our analysis of evolutionary conservation on a dataset of human caspase substrates generated in our lab. Inevitably, there are some substrates that we have not detected. While this limits the completeness of our results, we believe the dataset is large enough to allow us to see important general patterns. Our model organism datasets are much smaller, so we are unable to evaluate the conservation of every known human caspase substrate. However, we did find that the majority of caspase cleavage sites discovered in mouse, roughly half in *Drosophila*, and one third in *C. elegans*, were in orthologs of human proteins known to be caspase substrates, suggesting that a large number of caspase cleavages are indeed strongly conserved and thus likely to be functional. Notably, in many cases the cleavages occurred at different places in the two orthologs, often with the Asp itself having been lost at the aligned site in human. This is consistent with the tendency of short linear motifs to be lost over long evolutionary distances (8).

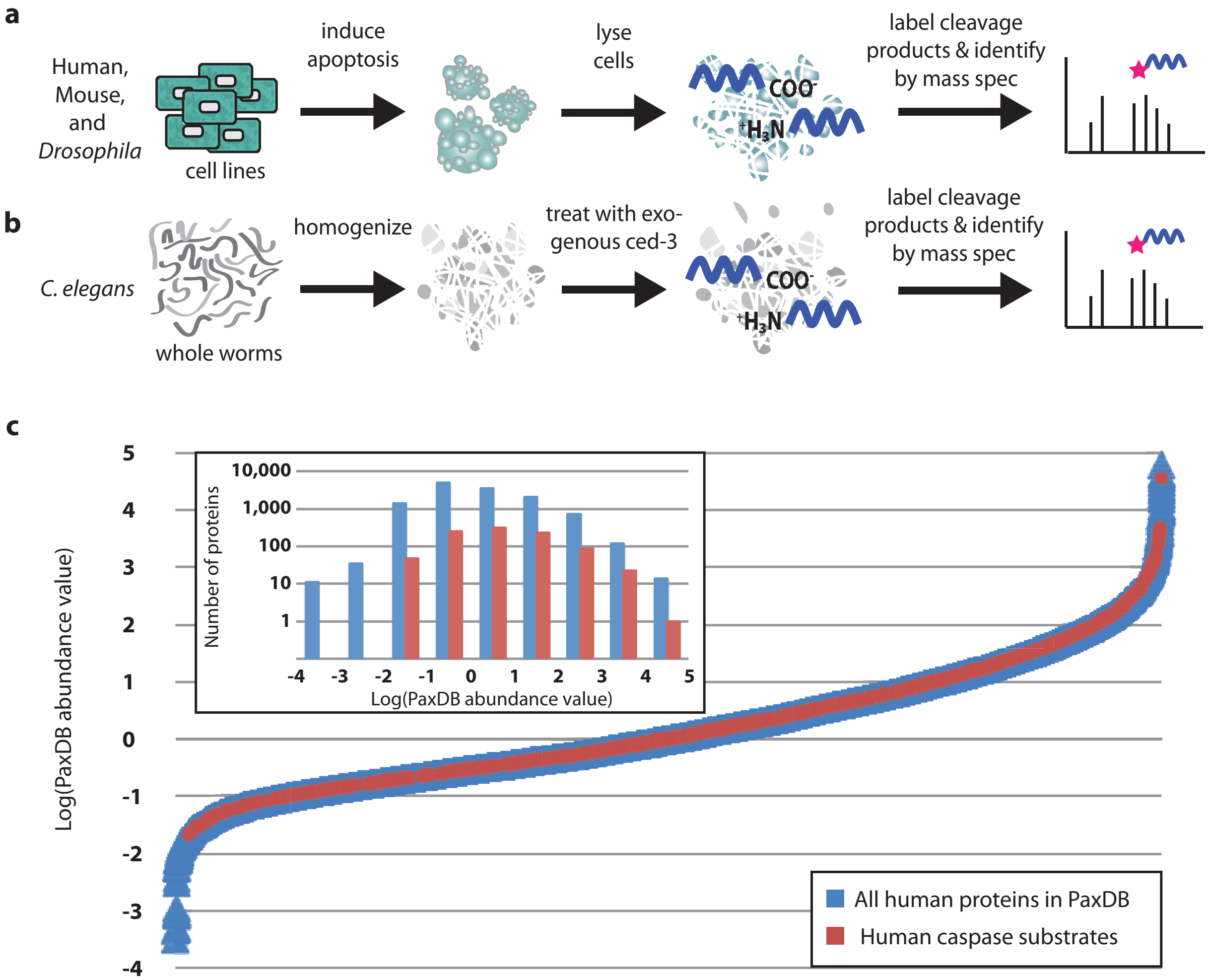


Figure 1: Experimental scheme. a) For human, mouse, and *Drosophila* experiments, cell lines were grown under standard culture conditions, and induced to apoptose using various toxic agents. The apoptotic cells were lysed, and the subtiligase labeling method was used to enrich for N-termini generated during apoptosis. b) For *C. elegans*, whole worms were homogenized, cuticles and other debris were spun out, and the resulting lysate was treated with exogenously expressed CED-3 for 2 hours at room temperature. The subtiligase labeling method was again used to enrich for unblocked N-termini, in this case many of them derived from the added protease.

Furthermore, bioinformatic pathway analysis showed that when a mouse, *Drosophila*, or *C. elegans* caspase substrate was not conserved at the protein level, it often had a human ortholog that was part of a pathway known to be heavily targeted by caspases in human. This finding led us to a hierarchical model of functional evolution: Caspase cleavages at specific motifs tend to be conserved only over short evolutionary distances; the set of proteins targeted by caspases is more conserved over longer distances; and the set of pathways targeted by caspases, even more conserved over even longer distances. Throughout this report, we refer to these three evolutionary modes as motif level, protein level, and pathway level conservation. A similar hierarchical conservation model has recently been proposed for the case of phosphorylation, another important post-translational modification (9).

Results

DATA COLLECTION

Human Reference Dataset. We previously developed a technique for proteome-scale positive enrichment of free N-termini that are liberated by proteolysis, based on labeling with subtiligase (4). This method was employed using various combinations of eight different human cell lines treated with any of five different apoptosis-inducing compounds, leading to the identification of 2021 Asp cleavage sites in 1444 proteins (Published in (4), Shimbo et al. PNAS 2012 (in press), and in preparation). We are confident in implicating caspases in these cleavages, as caspases have a strong and virtually unique preference for cleaving sites with Asp at the P1 position. (The Schechter & Berger protease notation (10) is used throughout; proteolysis occurs between the P1 residue on the N-terminal side and the P1' residue on the C-terminal side).

To determine how robust our human reference dataset is to variations in protein abundance, we consulted the PaxDB database (11, 12). This database provides relative protein abundance levels in several species based on spectral counting data drawn from many mass spectrometry datasets. As Figure 1C shows, our coverage extends over more than six orders of magnitude of ppm, although representation is somewhat thinner at the lower abundance levels.

At least 300 additional caspase substrates were not detected in our data but have been described elsewhere. Many of these are listed in the MEROPS database (5), a thorough and frequently updated repository of data on proteases and their substrates. Since cleavage site location validation varies greatly among MEROPS's sources, we have chosen not to include these data directly in our analyses. However, in Supplementary Table 1 we have made note of the 20 human orthologs of our non-human caspase substrates that are identified in MEROPS as substrates of one of the apoptotic effector caspases (caspases-3, -6, and -7). In only one case did a MEROPS cleavage site change an orthologous pair from not conserved, to conserved at the motif level (IL16_HUMAN and IL16_MOUSE). Only two additional orthologous pairs were changed from not conserved (or pathway-level conserved) to protein-level conserved (BRCA1_HUMAN and BRCA1_MOUSE; TOP1_HUMAN and TOP1_DROME). In the remaining 17 cases, the MEROPS site either matched one observed in our human data, or matched an additional unaligned site.

Mouse and *Drosophila* Datasets. Three mouse cell lines were used: A20, Tk-1, and wild-type MEFs. Apoptosis was induced in multiple experiments using bortezomib, doxorubicin, etoposide, or staurosporine. For the *Drosophila* experiments, *Drosophila* S2 cells were treated with doxorubicin, cyclohexamide, or actinomycin D. For both species, Cell Titre Glo and

Caspase Glo assays from Promega (Madison, Wisconsin, USA) were performed on separate experimental samples and confirmed that the drugs, concentrations, and time points used did induce both cell death and caspase activation (data not shown). Cells were harvested at a series of time points surrounding the time of maximal caspase activity, from 4 to 24 hours after treatment, and processed with the N-terminal labeling and mass spectrometry identification protocol as described (13). The data were searched with Protein Prospector (http://prospector.ucsf.edu/prospector/ms_home.htm). For mouse, the SwissProt database was searched; for *Drosophila*, the UniProtKB database was searched. The basic workflow is shown in Figure 1A, and results are listed in Supplemental Table 1.

***C. elegans* Dataset**. *C. elegans* is a classic cell death model organism, and has only one functional caspase, called CED-3. Unfortunately, no *C. elegans* immortalized cell lines have been established. In a gel-based proteomics study, Taylor *et al* (14) identified 22 CED-3 substrates by treating an extract made from homogenized whole worms with recombinant CED-3. Following their protocol, we grew Bristol N2 worms in liquid culture, homogenized them, isolated the soluble protein fraction, and incubated it at room temperature for two hours with a CED-3 preparation made from *E. coli*. We then subjected the CED-3-digested extract to our in vitro N-terminal labeling protocol, as described (13) (Figure 1B). The data were searched with Protein Prospector, using the UniProtKB database. Results are listed in Supplemental Table 1.

DATA ANALYSIS

Primary Structure Specificity. Our N terminomics technology allows direct determination of the primary sequence flanking the P1 Asp residue for each caspase-derived cleavage site. We used IceLogo to generate visual representations of sequence specificity for all four species (Figure 2A). In each case, the appropriate SwissProt or UniProtKB database was used to establish background amino acid frequencies. Synthetic peptide-based experiments suggest a strong preference for Asp at P4 and Glu at P3 for human executioner caspases-3 and -7 (ref (15)). In contrast, physiological sets of whole human proteins cut by caspases consistently show diminished preference for these non-P1 acidic residues (2, 4, 16), suggesting that primary sequence alone is not sufficient to explain caspase-substrate recognition in cells. The additional datasets here show that this holds true for mouse, *Drosophila*, and *C. elegans* caspase cleavage sites as well. The issue of specificity is complicated somewhat by the mixture of different caspases that are active during apoptosis. However, studies focusing on single caspases (including the CED-3 experiment described here) have yielded similar sequence logos as the physiological datasets (13).

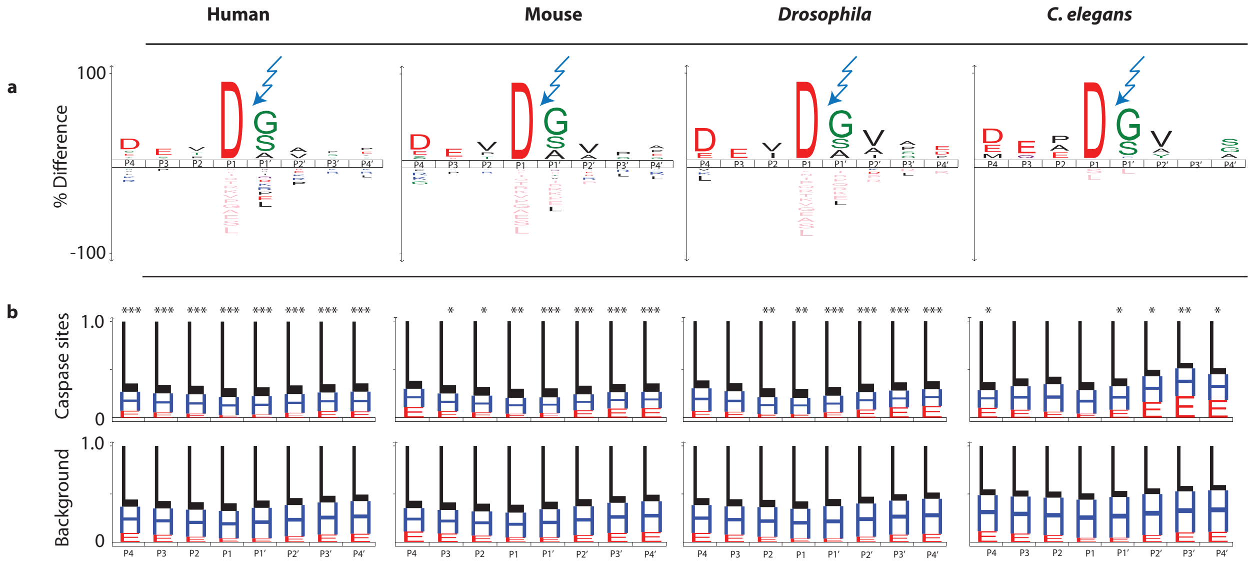


Figure 2: a) IceLogo diagrams depicting primary structure preferences for the P4-P4' residues for cleavages following aspartic acid in each of the four species. Letters above the axis indicate residues enriched over background, and letters below the axis indicate residues depleted with respect to background. b) Secondary structure predictions for the P4-P4' residues for cleavages following aspartic acid in each of the four species. L=loop, H=alpha helix, and E=beta sheet. The height of the letter indicates the fraction of sites with the corresponding secondary structure prediction, based on predictions using the NetSurfP server (17). Top row represents caspase cleavage sites; bottom row represents all 8mers with D at the P1 position in the same proteins. Asterisks indicate statistically significant enrichment of loops in caspase sites compared with background. *, p<0.05. **, p<0.01. ***,p<0.001.

As in other proteome-scale studies of caspase substrates (4, 18), the P1' position shows a strong preference for a small amino acid. In all four species, >75% of P1' residues are either Ala, Gly, or Ser. This preference for small amino acids at the P1' position is not explained by any bias from the subtiligase N-terminal labeling procedure because subtiligase has a slight bias towards large residues on the labeled N terminus (19).

Secondary Structure Specificity. Examining predicted secondary structure for the caspase cleavage sites in our data confirms a statistically significant preference for sites that occur in loop regions. Caspases, like most proteases, require substrate cleavage sites to be in an extended conformation while binding to the active site (20). These sites must therefore either be in loop regions lacking secondary structure, or in helix or sheet regions that are dynamic or flexible enough to give caspases reasonable access to subpopulations in unstructured conformations. To extend this analysis, we used NetSurfP (17) to predict secondary structure for all proteins targeted by caspases in our datasets. For each species, we compared predictions for the P4-P4' sequences surrounding all cleavage sites to a background made up of all 8mers in the same set of proteins with Asp at the P1 position. Secondary structure prediction surrounding the cut sites shows enrichment in loop structures in all four species, although the p-values were strongest for the larger human dataset (Figure 2B).

Establishing Orthologous Relationships. We next determined the level of conservation of caspase substrates among the four metazoan species using the strategy shown in Figure 3. The EggNOG database (21) was used to determine orthologous relationships between caspase substrates in human and those in mouse, *Drosophila* or *C. elegans*. EggNOG was built using non-supervised clustering methods to assign proteins from 1133 species to over 700,000 orthologous groups. Closely related paralogs in the same species are all represented in the same orthologous group, meaning that in some cases a single protein in one species has more than one ortholog in another species. We looked mainly at orthologous groups on the metazoan level (called meNOGs), since caspases are restricted to this lineage (22).

The caspase substrate sets we identified in mouse, *Drosophila*, and *C. elegans* are remarkably highly enriched in proteins whose human orthologs are also known caspase substrates. Figure 3 depicts the data analysis pipeline used to classify the non-human substrates, and Table 1 summarizes the results described here; the full data, including the meNOG and human orthologs associated with each cleavage, are shown in Supplementary Table 1. A total of 217 mouse (84%), 64 *Drosophila* (49%), and 17 *C. elegans* (34%) caspase cleavage sites have human orthologs that are caspase substrates. The sizes of the meNOG group overlaps between all four species are shown in Figure 4A. For 38 mouse, 36 *Drosophila*, and 22 *C. elegans* proteins, human orthologs were present in corresponding meNOGs, but these human proteins are not known to be caspase substrates. The remaining 2 mouse, 30 *Drosophila*, and 11 *C. elegans* proteins were either not present in any meNOG, or were present in a meNOG that had no human members (for example, the vitellogenin protein meNOG, which contains egg yolk nutrient proteins that are not present in placental mammals). Considering that the 1444 human caspase substrates used as the reference set for this study make up only about 7% of the SwissProt human proteome, it is clear that the nonhuman datasets are extremely highly enriched for proteins with human caspase substrates as orthologs.

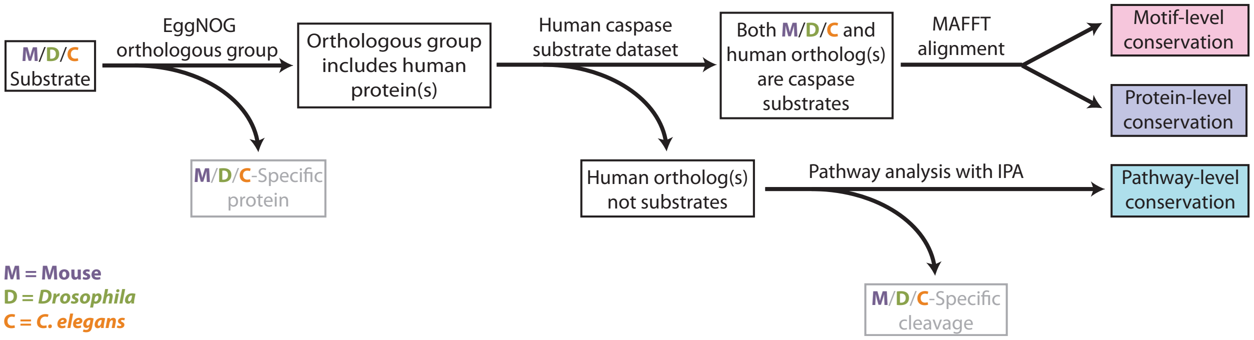


Figure 3: Data analysis pipeline. For each caspase cleavage observed in mouse, *Drosophila*, or *C. elegans*, we searched the EggNOG database for human orthologs. If the human ortholog found was also present in our human caspase substrate database, an alignment was created to determine whether the orthologs were cleaved at the same site or different sites. If the human ortholog found was not known to be a caspase substrate, we searched IPA's list of "Canonical Pathways" to determine whether it functioned in any pathway(s) known to be enriched for caspase substrates. Each mouse, *Drosophila*, or *C. elegans* protein is thus assigned to one of five categories: 1. No human ortholog (species-specific protein), 2. Human ortholog is not a substrate (species-specific caspase cleavage), 3. Pathway-level conservation, 4. Protein-level conservation, or 5. Motif-level conservation. Supplementary Table 1 shows all mouse, *Drosophila* and *C. elegans* data, organized into these categories.

Table 1: Summary of Results

							Level of Conservation		
Species	**# of cell lines**	**# of experi-ments**	**Proteins cut by caspases**	**Total Caspase Cuts**	**No Human Ortholog**	**Human Ortholog Not Substrate**	**Pathway**	**Protein**	**Motif**
Human	8	>50	1444	2021	-	-	-	-	-
Mouse	2	7	221	257	2	30	8	41	176
Drosophila	1	7	118	130	30	20	16	49	15
C. elegans	N/A	1	46	50	11	13	9	11	6

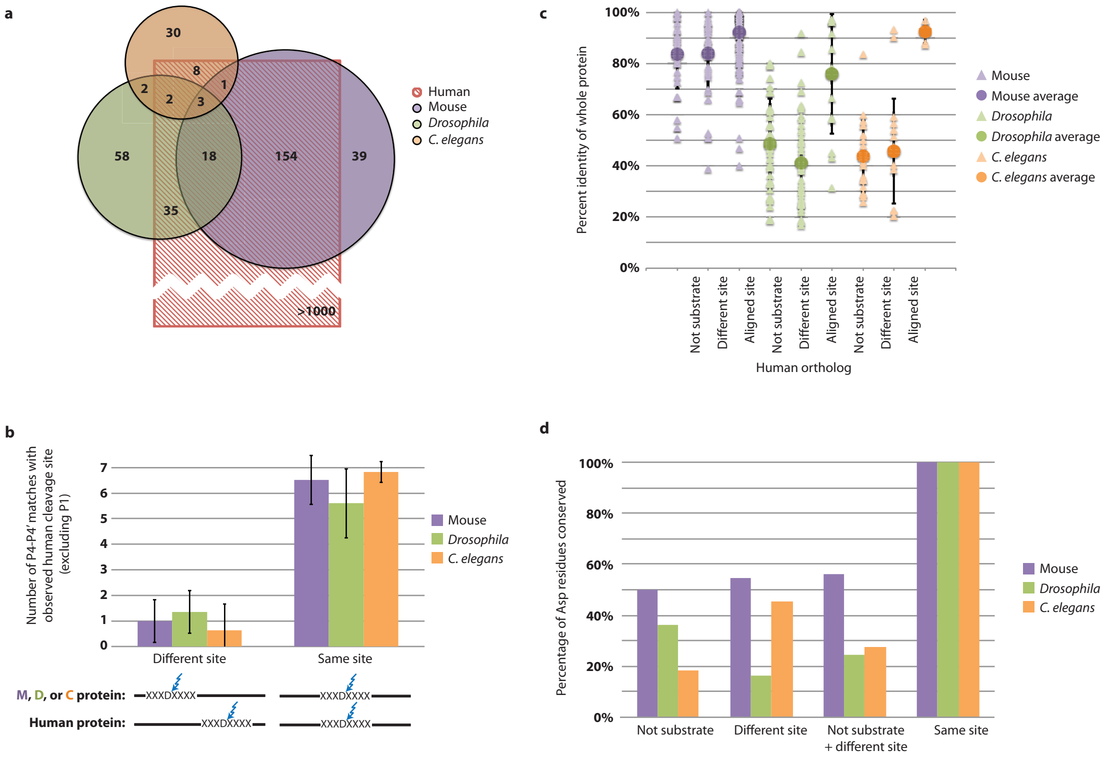


Figure 4: a) Venn diagram indicating the overlap in the meNOGs associated with caspase substrates in human, mouse, *Drosophila* and *C. elegans*. b) When Mouse, *Drosophila*, or *C. elegans* caspase cleavage sites aligned with known cleavage sites in human orthologs, the aligned sites shared an average of between 5 and 7 identical residues (considering the P4, P3, P2, P2', P3', and P4' positions, since the P1 position is fixed to Asp). In contrast, when a mouse, *Drosophila*, or *C. elegans* protein has a human ortholog with a cleavage site at a different location, the two observed cleavage sites share only an average of between 0 and 2 identical residues (considering the same 7 positions). Error bars represent standard deviation. c) The protein %ID was calculated for pairwise alignments of each mouse, *Drosophila*, or *C. elegans* protein with its closest human ortholog. Pairwise alignments were calculated in Jalview. Triangles represent individual pairwise alignments, while large circles represent the average %ID for each category, with error bars representing the standard deviation.

Alignments. For meNOGs that contained a *Drosophila* or *C. elegans* caspase substrate and also contained at least one human caspase substrate, we created multiple sequence alignments using the EINS-i algorithm from MAFFT (23). These alignments were then examined by eye using Jalview 2.7 (24) to confirm relative placement of all caspase cleavage sites. The same procedure was used for mouse substrates, except in cases where the mouse substrate's meNOG contained only one human ortholog substrate. In these cases, most of which had extremely high sequence identity, we simply used pairwise ClustalW alignments performed with default parameters on the UniProt website (www.uniprot.org) (25). In the majority of these cases, the cleavage sites aligned. When they did not, we repeated the alignment using MAFFT's EINS-i algorithm.

We found that 176 out of 217 mouse sites, 15 out of 64 *Drosophila* sites, and six out of 17 *C. elegans* sites were aligned perfectly with cleavage sites previously observed in their human orthologs, indicating motif-level conservation. In the remaining cases, where the sites were not aligned, we wondered whether errors in the alignment method could have separated two sites that had evolved from a common ancestor site. For example, the sites could have become spatially displaced by insertion or translocation of a domain in one of the orthologs. If the site were simply shifted, we would expect a high level of sequence conservation. We compared the eight residues immediately surrounding the observed mouse, *Drosophila*, or *C. elegans* cleavage sites with the eight residues immediately surrounding the observed human cleavage site. We found that for substrates with motif-level conservation, the pairs of corresponding cleavage sites shared an average of more than 5.5 out of seven identical residues (considering only P4-P2 and P1'-P4'; P1 is fixed to Asp) (Figure 4B). In contrast, for substrates with different sites, the cleavage site pairs shared an average of less than 1.5 out of seven identical residues, suggesting that the majority of sites that are not aligned evolved separately and are not simply the result of one site shifting in the sequence.

We also looked at the overall protein sequence similarity between our mouse, *Drosophila*, and *C. elegans* substrates and their human orthologs. This was based on pairwise alignments calculated with Jalview. Not surprisingly, proteins with higher percent identity were more likely to share an aligned cleavage site (Figure 4C). In particular, for *Drosophila* and *C. elegans* proteins with >80% identity to their human orthologs, 10/12 (*Drosophila*) and 3/7 (*C. elegans*) shared aligned caspase cleavage sites (compared with 15/102 and 4/37 when considering all caspase substrates with human orthologs).

In cases where the human ortholog either was not a substrate or was cleaved at a different site from the mouse, *Drosophila* or *C. elegans* site, we asked whether the non-cleaved aligned site contained a cleavable Asp at the P1 position (Supplementary Table 1 and Figure 4D). The Asp was conserved in less than 30% of cases for both *Drosophila* and *C. elegans*, and less than 60% in mouse. For these cases, we cannot rule out the possibility that the human orthologs are true caspase substrates, but have never been observed. This could be due to low expression levels in the cell lines used, poor detectability of the peptide, or other factors. But in cases where the Asp is not conserved in human (and therefore the site cannot be cut by caspases), we are confident in categorizing these cleavage events as having been lost or moved during the course of evolution.

Pathway and Function Analysis. Our finding that caspase substrate sets are more broadly conserved at the protein level than at the motif level led us to wonder whether we could detect an even broader conservation at the level of functional pathways (Figure 3). It is known that caspase substrates are more likely to occur in protein-protein complexes with each other than are a randomly selected set of proteins.(4) We reasoned that in many cases caspases could achieve the same functional effect by cleaving one (or more) of several different proteins that operate in the same complex or pathway. We used the highly curated IPA (Ingenuity Pathway Analysis, Ingenuity Systems) analysis software to generate a list of predicted "Canonical Pathways" which were significantly enriched in our human caspase substrate dataset (Supplementary Table 2). We then asked whether any of the non-substrate human orthologs of mouse, *Drosophila*, or *C. elegans* substrates were members of these pathways. Eight out of 40 mouse, 16 out of 36 *Drosophila*, and nine out of 22 *C. elegans* substrates did indeed have non-substrate human orthologs that were members of one or more of these pathways (Supplementary Table 1). Although these proteins had no direct human orthologs that are known to be substrates, they are conserved at the pathway level. In some cases, the functional pathways themselves may not be conserved in the other organisms, but many of them are highly conserved, essential pathways involved in processes such as translation or proteasomal degradation.

One example that illustrates pathway conservation is the highly conserved EIF2 Signaling pathway, as defined by IPA (Supplementary Figure 1). In many cases within this pathway, caspases in more than one species cut the same protein; in other cases, different proteins in the same complex are cut in different species. For example, one component of the eIF4γ complex, EIF4G2, is a substrate in *Drosophila*; the same protein, plus two other members of this complex (EIF4G1 and EIF4G3) are also known human substrates. In contrast, in the 60S ribosomal subunit, RPL27 and RPL4 are substrates in *Drosophila*, but are not known to be substrates in human. Instead, three different 60S ribosomal subunit proteins, RPL5, RPL17, and RPLP2, are known substrates in human, and RPL5 is a known substrate in mouse.

To estimate the important biological functions for caspase-cleaved proteins, we examined GO term assignments for the caspase substrates of each organism. In all four species, there was enrichment in GO terms related to the nucleus, the cytoskeleton, transcription, and nucleic acid binding, consistent with previous GO term analysis on human caspase substrates (4) (see Supplementary Table 3 for full results).

Discussion

All multicellular organisms need mechanisms for eliminating cells that are either unneeded or are posing a danger to the whole body. The apoptosis pathway appears to be evolutionarily related, both phenotypically and molecularly, across all metazoans studied to date, indicating that its basic features were likely present in the last common metazoan ancestor (6, 22). While the triggers for apoptosis vary depending on species, cell type, and environmental conditions, all apoptotic pathways in all species studied lead eventually to caspase activation, followed by non-inflammatory dismantling of the apoptotic cell by phagocytes. Here, we begin to globally determine apoptotic caspase cleavage sites in three non-human model metazoans in order to investigate the evolutionary conservation of the common caspase stage of apoptosis.

The human dataset used here, collected over six years, contains 1444 protein targets. In contrast, these initial mouse, *Drosophila*, and *C. elegans* datasets are roughly an order of magnitude smaller. The human dataset was compiled from about ten times more samples than the other metazoans. Furthermore, eight different human cell lines were examined, representing a much larger swath of expressed proteomes compared with the two mouse cell lines, one *Drosophila* cell line, and whole *C. elegans*. Despite the incomplete sampling size of the non-human metazoans, we can make meaningful positive comparisons between these smaller sets against the much more complete human dataset because the low false positive rate (<3%, data not shown) provides high confidence identification. However, given the small and incomplete size of these nonhuman datasets, we urge readers to avoid making conclusions based upon the absence of observed cleavages except in cases where the aspartic acid site is simply missing in the comparisons.

All comparative proteomics studies come with a risk of bias towards high-abundance proteins. Proteins with high abundance tend to be less conserved than those with low abundance (26), so sampling bias towards high abundance proteins might falsely mimic a result of high conservation. However, our experimental technique reduces the effect of this problem. By enriching samples at the tryptic peptide level (leaving only one or a few peptides per protein), our N-terminal labeling method decreases the complexity of the peptide mixture further than traditional methods, which diminishes the chance of low abundance proteins being masked by extremely high abundance proteins like actin. In fact, studies using our method in human serum were able to detect proteins with known abundance values spanning six orders of magnitude (27). Using the PaxDB database (11, 12), we have also established high coverage of the abundance range of the caspase substrates identified in our cell line-based human reference dataset (Figure 1C). Given this result, we are confident that abundance bias does not play a large role in our high conservation results.

Our results suggest a hierarchical arrangement of conservation, depicted in Figure 5. Caspases target a specific set of conserved pathways across metazoans in order to successfully complete apoptosis. In some cases, this means that caspases in different species may target different proteins within the same pathways. In other cases, the same protein remains a caspase substrate across a broad lineage, and in still other cases, caspases even target the same cleavage site motif in different species. Our data are consistent with previous observations that short linear motifs tend to be poorly conserved over evolution,(8) but also shows that loss of the short linear motif recognized by caspases does not necessarily mean that the same target is not cleaved.

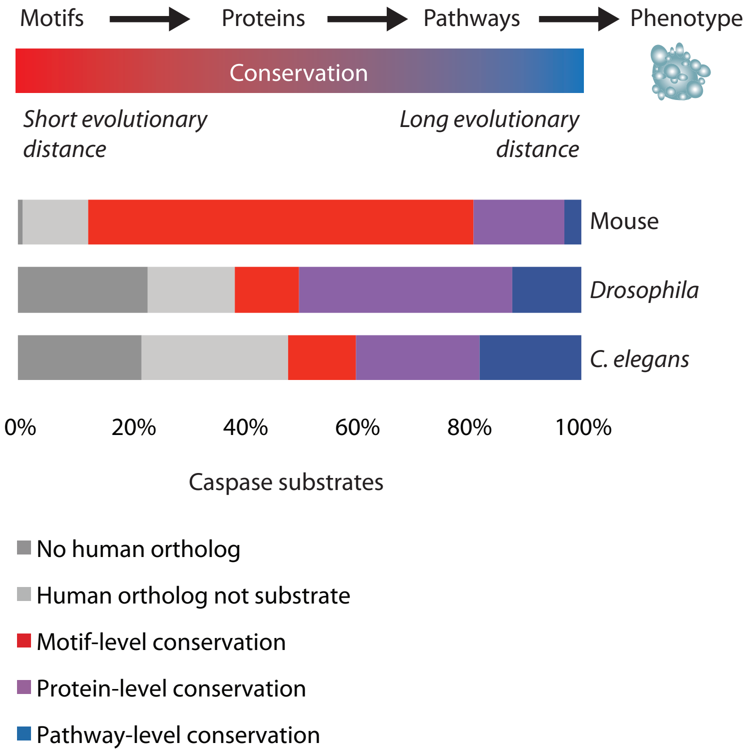


Figure 5: Results of this study show that caspases tend to recognize and cleave particular motifs only across short evolutionary distances (represented by the human-mouse comparison, spanning less than 100 million years of evolution(28)), but that the same proteins will remain targets across longer distances, and the same pathways over even longer distances (represented by the human-*Drosophila* and human-*C. elegans* comparisons, both represented by roughly 600 million years of evolution(7)). Ultimately, the phenotype of apoptosis is conserved across the whole metazoan lineage.

Our study has yielded similar conservation patterns to those observed for phosphorylation and transcription factors. Tan *et al*. (9) used mass spectrometry to identify phosphorylation sites in *Drosophila*, *C. elegans* and yeast, and compared them to known human sites. They found that the set of modifications conserved at the motif level was small, but that the networks defining relationships between kinases and substrates were conserved more broadly. Additional phosphorylation studies focused on computational analysis have also revealed that motif-level conservation is weaker than protein-level conservation (29, 30). Studies of transcription factor binding in yeast show a somewhat similar conservation pattern: while the binding sites on DNA are only modestly conserved, overall regulatory networks and the functions of the sets of genes effected by them are retained across hundreds of millions of years of evolution (31-33). Notably, three studies of single *Drosophila* caspase substrates have also demonstrated protein level (but not motif level) conservation (34-36).

Caspases, like all enzymes that catalyze specific post-translational modifications, must have mechanisms for recognizing their particular targets in the complex cellular milieu. How caspases achieve this is not well understood. Executioner caspases, like human caspase-3, are functionally polyspecific, meaning that they recognize several otherwise unrelated substrates, and yet are restrained from targeting all Asp sites in all proteins in a digestive manner (6). The extreme degeneracy of the physiological cleavage site consensus sequence seen in all four organisms (Figure 2A), plus the higher conservation at the protein level than at the motif level, suggest that other factors, such as exosite binding (37, 38) or subcellular compartmentalization, may influence what is cleaved in cells.

We examined UniProt, Flybase (39), and Wormbase (www.wormbase.org, release WS229) annotations for the 62 *Drosophila* and 19 *C. elegans* caspase substrate proteins that are specific to those species, either because they have no human orthologs, or their human ortholog(s) are not known substrates. Some of these are listed in Table 2. The majority have little or no functional annotation, but several in both species are either known or predicted to be involved in development. In *Drosophila*, nine proteins fall into this category: three function in spermatogenesis, three in neurogenesis, and three in other types of development. Five *C. elegans*-specific proteins are either known or predicted to be involved in embryo development. Programmed cell death is an important aspect of development in all species, but further studies are needed to determine whether these cleavages are relevant to developmental apoptosis. Three of the five *C. elegans* development proteins are known to be extracellular. There is no evidence of extracellular caspase activity, so it is likely that these proteins are not physiologically relevant – rather, they may result from the fact that the *C. elegans* extract was made from whole bodies, rather than a cell line. Another interesting finding is that three proteins cleaved in *Drosophila* are all members of the gypsy chromatin insulator complex. This complex is thought to regulate accessibility of certain chromosomal regions (40), so these three cleavages are in accordance with the GO analysis showing the tendency for caspases to cleave DNA- and chromatin-associated complexes (Supplementary Table 3). Our lab's previous work established the caspase cleavage of several components of a human chromatin-associated complex, the NCoR-SMART complex (4).

In aggregate, these studies represent a unique systematic comparison of apoptotic caspase substrates across the metazoan lineage, shedding light on what is functionally most important. In all three species, at least half of the substrates identified are conserved with human substrates on either the motif, protein, or pathway level, and given that our human caspase substrate dataset is likely not comprehensive, this could be an underestimate. This finding supports the view that a substantial fraction of targets do serve an important function that has been conserved by selective pressure across 600 million years of metazoan evolution.

Table 2: Subset of *Drosophila* and *C. elegans* Caspase Substrates with no Human Ortholog

Species	**Acc #**	**SwissProt ID**	**Description**	**Function**
Drosophila	Q95RU0	CUE_DROME	Protein cueball	spermatogenesis
	Q8MRY4	-	SD13619p	
	Q8T044	-	LD29665p	
	Q9VFE6	RRP15_DROME	RRP15-like protein	neurogenesis
	Q9VAF4	-	Dim gamma-tubulin 1	
	Q7K126	-	LD13864p	
	Q09024	IMPL2_DROME	Neural/ectodermal development factor IMP-L2	embryogenesis
	Q9VYV4	-	CG2446, isoform A	development, various
	Q9VI56	-	CG1943, isoform A	wing disc development
	Q24478	CP190_DROME	Centrosome-associated zinc finger protein CP190	gypsy chromatin insulator complex
	P08970	SUHW_DROME	Protein suppressor of hairy wing	
	Q86B87	MMD4_DROME	Modifier of mdg4	
C. elegans	P05690	VIT2_CAEEL	Vitellogenin 2	embryogenesis (extracellular)
	P06125	VIT5_CAEEL	Vitellogenin 5	
	Q18823	LAM2_CAEEL	Laminin-like protein lam-2	
	Q17796	-	hgrs-1	embryogenesis
	Q22469	-	Protein T13H5.4	

Materials and Methods

Cell lines: Mouse Embryonic Fibroblasts (MEFs) were a kind gift from Dr. Richard Flavell. TK-1 and A20 cells were purchased from ATCC (Manassas, Virginia, USA) and cultured in the recommended media with 10% FBS at 37°C. When cells were between 1x10^6 and 2x10^6 cells/mL, one of four apoptosis-inducing agents was added (bortezomib, doxorubicin, etoposide or staurosporine). Cells were harvested by centrifugation after 4-24 hours, washed once in 1X PBS, and then frozen at -80°C as pellets.

Drosophila S2 cells were purchased from ATCC and cultured in Schneider's *Drosophila* Medium from Gibco (Carlsbad, California, USA) with 10% FBS at 26°C. At a cell density of 2x10^6 cells/mL, one of three apoptosis-inducing agents was added (doxorubicin, cycloheximide, or actinomycin D). Cells were harvested by centrifugation after 5-24 hours, washed once in 1X *Drosophila* PBS (pH 6.7), and then frozen at -80°C as pellets.

Both mouse and *Drosophila* cell lysates were formed by resuspending pellets in an SDS solution containing protease inhibitors (including the pan-caspase inhibitor z-VAD-fmk, to prevent any post-harvest caspase activity) and then sonicating. The N-terminal labeling reaction and mass spectrometry prep was performed as described for apoptotic cells (13).

C. elegans: Bristol N2 worms were obtained from Dr. Cynthia Kenyon at UCSF. They were grown in 250mL volumes of S Complete Media liquid culture containing streptomycin and the fungicide carbendazim, with daily additions of streptomycin-resistant OP50 E. coli. When worm density reached 100 worms/µL (without regard to worm size or age), and the E. coli was visibly depleted from the culture flasks, the cultures were harvested by centrifugation, washed with M9 buffer, and left-over *E. coli* removed by centrifuging in 30% sucrose. They were immediately removed from the sucrose, washed three times with M9 buffer, and incubated on a nutator at room temperature for 30 minutes to allow for digestion of residual bacteria present in the intestine. They were then drip-frozen in liquid nitrogen, and stored at -80. The frozen worms were homogenized with a large mortar and pestle kept cold with liquid nitrogen. When substantial breakage of worm bodies was observed under a light microscope, the homogenate was collected and mixed with just enough M9 buffer (containing protease inhibitors PMSF, AEBSF, IAM and EDTA) to form a slurry. The slurry was sonicated at 4C, and then passed through low and then high gauge needles. Finally, the cuticles and other insoluble debris were removed by centrifugation, and the extract frozen at -80 until needed. Later, the extract was labeled and processed for mass spectrometry as described for in vitro cleavage assays (13).

The *ced-3* gene, with the prodomain (residues 1-220) removed, was cloned from a *C. elegans* cDNA library (kindly provided to us by Dr. Aimee Kao). The construct was inserted into various vectors with different configurations of His6, GST, and MBP tags at either end. We found in all cases that CED-3 solubility was extremely limited, with the majority of the enzyme lost in the *E. coli* pellet after lysis by microfulidization or sonication. The best yields came from expression in pPAL7 vector with N-terminal HIS and GST tags, and a single HIS column purification. Since further purification methods, and efforts at concentration, led to increased losses, we chose to use a low concentration, semi-pure CED-3 extract for our proteomics experiments. Since we did not produce enough enzyme to complete an active site titration, we estimated the concentration of active CED-3 by relating its activity on the peptide substrate DEVD-afc to the activity of a pure human caspase-3 sample of known concentration. The CED-3 extract had a V_{max} of 0.075μmol/s and a K_M of 8.4μM, which is equivalent to a 200nM sample of caspase-3 purified in our lab (with k_{cat}/K_M of $0.3M^{-1}s^{-1}$). The CED-3 extract was added to the worm preparation at a 1:10 volume ratio, so the final concentration of CED-3 was equivalent to 20 nM human caspase-3 by activity.

Mass Spectrometry: All mass spectrometry data were collected in the UCSF Mass Spectrometry Facility on a QStar Elite from Applied Biosystems (Carlsbad, California, USA), with the exception of one mouse dataset, which was collected on a QExactive instrument kindly made accessible to us by Thermo Scientific (Waltham, Massachusetts, USA). Data were searched with Protein Prospector (UCSF Mass Spectrometry Facility). For the human and mouse data, the search library contained only proteins from the SwissProt database. For *Drosophila* and *C. elegans*, the highly curated SwissProt database is limited in size, so the broader UniProtKB library was searched instead. For these two organisms, if one peptide could be mapped to either a SwissProt entry or a non-SwissProt UniProtKB entry, then only the SwissProt entry was considered.

Abundance Analysis: Data on protein abundance were taken from the integrated human dataset from PaxDB version 2.0 (available in the archive section at www.pax-db.org (12).

Primary and Secondary Structure Analyses: Primary structure logos were generated using IceLogo (41) with the appropriate SwissProt or UniProtKB database as background. Secondary structure predictions were determined using NetSurfP (17). Secondary structure background was based on all 8mers with Asp in the fourth position (equivalent to P1 position) found in the same set of proteins. The logos for these data were created using the “Filled Logo” option on the IceLogo server, and significance was determined by chi-squared tests.

Orthology and alignments: Metazoan orthologous groups (meNOGs) containing the *Drosophila* and *C. elegans* proteins found in this study were retrieved from the downloads area of the EggNOG 3.0 website (21). Each meNOG, if it contained a human protein known to be a caspase substrate, was aligned using the EINS-i algorithm from MAFFT (23). Trees derived from these alignments were visualized by eye to confirm that sequences were well distributed across the metazoan phylogeny. Alignments were then displayed and analyzed using Jalview. For each mouse, *Drosophila* or *C. elegans* cleavage site, we assessed four points: (1) whether it aligned with a caspase cleavage site in a human ortholog, (2) the number of residues in common between it and the most similar human caspase cleavage site in the P4-P2 and P1'-P4' positions, (3) if the sites did not align, whether the P1 Asp of the mouse, *Drosophila,* or *C. elegans* cleavage site was conserved in any human ortholog, and finally (4) the pairwise % ID between the mouse, *Drosophila* or *C. elegans* protein and its closest human ortholog. This was calculated with the pairwise alignment tool in Jalview 2.7, which uses the BLOSUM62 matrix and gap opening and extending penalties of 12 and 2 respectively.

In some cases, the mouse, Drosophila or C. elegans peptide discovered in our mass spectrometry experiments was matched to more than one protein by the Protein Prospector program. We checked each of these peptides individually, and determined that in all cases, the results presented here were the same no matter which protein was chosen.

Pathway analysis: The set of 1444 human caspase substrates was uploaded to Ingenuity Pathway Analysis (IPA, Ingenuity Systems). A "Core Analysis" was performed to generate a list of all "Canonical Pathways" whose genes were significantly overrepresented in the human caspase substrate dataset, with p values <0.05. (Supplementary Table 2). We then took the list of non-substrate human orthologs for each of the three other species and checked the IPA database to see which, if any, canonical pathways they were associated with.

GO term enrichment was determined using the GO::TermFinder software (42). Lists of unique proteins were created for each species based on the discovered apoptotic peptides. The datasets were uploaded to the database and tested for enrichment against a background of the organism's SwissProt (mouse) or UniProtKB (*Drosophila* and *C. elegans*) database. The mouse, *Drosophila* and *C. elegans* tests used all GO evidence codes, and the human test used all evidence codes except ND ("No biological data available") and IEA ("Inferred from Electronic Annotation"). Enriched terms in human sets were defined as those with corrected p-values less than 0.01. The human set was then filtered to remove terms that were not statistically significantly different from untreated background (data not shown, manuscript in preparation).

The significant GO terms in mouse, *Drosophila* and *C. elegans* were compared to the top human terms in each ontology. The fold enrichment was calculated for each significant term; the percentage of proteins annotated with each term in the experimental set was divided by the percentage of proteins annotated with each term in the proteome background.

Conflict of Interest Statement:

The authors declare no conflict of interest.

Acknowledgements: We thank Sami Mahrus, David Wildes, Nicholas Agard, Hikari Yoshihara, Huy Nguyen, Gerald Hsu, and Kazutaka Shimbo for contributing to the human caspase substrate dataset. We also thank Cynthia Kenyon and Aimee Kao for assistance with the *C. elegans* work, and David Maltby, Jonathan Trinidad, and Shenheng Guan and the rest of the UCSF Mass Spectrometry Facility, plus Iman Mohtashemi and Yan Chen of Thermo Scientific, for mass spectrometry assistance. We thank Bent Petersen for assistance with the NetSurfP server. This project was supported by a grant from the UCSF Stephen and Nancy Grand Multiple Myeloma Translational Initiative. It was also supported by the National Science Foundation GRFP (E.D.C.), NIH Training Grant T32 GM007175 (J.E.S. and P.C.B.), the Pharmaceutical Research and Manufacturers of America (A.E.B.), National Institutes of Health R01 GM60595 (P.C.B.), and National Institutes of Health R01 GM081051 (J.A.W.). Mass spectrometry was performed at the Bio-Organic Biomedical Mass Spectrometry Resource at UCSF (A.L.B., Director), which is supported by grants from the National Center for Research Resources (5P41RR001614) and the National Institute of General Medical Sciences (8 P41 GM103481) from the National Institutes of Health.

References

1. Zmasek CM, Zhang Q, Ye Y, Godzik A. Surprising complexity of the ancestral apoptosis network. Genome biology. 2007;8(10):R226.

2. Van Damme P, Martens L, Van Damme J, Hugelier K, Staes A, Vandekerckhove J, et al. Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. Nature Methods. 2005;2(10):771-7.

3. Dix MM, Simon GM, Cravatt BF. Global mapping of the topography and magnitude of proteolytic events in apoptosis. Cell. 2008;134(4):679-91.

4. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell. 2008;134(5):866-76.

5. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. Nucleic Acids Research. 2010;38(Database issue):D227-33.

6. Crawford ED, Wells JA. Caspase substrates and cellular remodeling. Annual review of biochemistry. 2011;80:1055-87.

7. Erwin DH, Davidson EH. The last common bilaterian ancestor. Development (Cambridge, England). 2002;129(13):3021-32.

8. Neduva V, Russell R. Linear motifs: evolutionary interaction switches. FEBS Letters. 2005;579(15):3342-5.

9. Tan C, Bodenmiller B, Pasculescu A. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. Science (New York, NY). 2009.

10. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. Biochemical and biophysical research communications. 1967;27(2):157-62.

11. Schrimpf S, Weiss M, Reiter L, Ahrens C. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS Biology. 2009.

12. Weiss M, Schrimpf S, Hengartner M. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. 2010.

13. Agard NJ, Mahrus S, Trinidad JC, Lynn A, Burlingame AL, Wells JA. Global kinetic analysis of proteolysis via quantitative targeted proteomics. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(6):1913-8.

14. Taylor RC, Brumatti G, Ito S, Hengartner MO, Derry WB, Martin SJ. Establishing a blueprint for CED-3-dependent killing through identification of multiple substrates for this protease. The Journal of biological chemistry. 2007;282(20):15011-21.

15. Thornberry N, Rano T, Peterson E, Rasper D, Timkey T, Garcia-Calvo M, et al. A combinatorial approach defines specificities of members of the caspase family and granzyme B. Journal of Biological Chemistry. 1997;272(29):17907.

16. Lüthi AU, Martin SJ. The CASBAH: a searchable database of caspase substrates. Cell Death and Differentiation. 2007;14(4):641-50.

17. Petersen B, Petersen T, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Structural Biology. 2009;9(1):51.

18. Schilling O, Overall C. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nature Biotechnology. 2008;26(6):685-94.

19. Chang TK, Jackson DY, Burnier JP, Wells JA. Subtiligase: a tool for semisynthesis of proteins. Proc Natl Acad Sci USA. 1994;91(26):12544-8.

20. Fontana A, de Laureto PP, Spolaore B, Frare E, Picotti P, Zambonin M. Probing protein structure by limited proteolysis. Acta biochimica Polonica. 2004;51(2):299-321.

21. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research. 2012;40(Database issue):D284-9.

22. Lamkanfi M, Declercq W, Kalai M, Saelens X, Vandenabeele P. Alice in caspase land. A phylogenetic analysis of caspases from worm to man. Cell Death and Differentiation. 2002 May 01.

23. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics. 2008;9(4):286-98.

24. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics (Oxford, England). 2009;25(9):1189-91.

25. Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2012;40(Database issue):D71-5.

26. Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. Nature Reviews Genetics. 2006;7(5):337-48.

27. Wildes D, Wells JA. Sampling the N-terminal proteome of human blood. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(10):4561-6.

28. Huchon D, Madsen O, Sibbald MJJB, Ament K, Stanhope MJ, Catzeflis F, et al. Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes. Molecular biology and evolution. 2002;19(7):1053-65.

29. Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM, et al. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biology. 2009;7(6):e1000134.

30. Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. Science (New York, NY). 2009;325(5948):1682-6.

31. Tsong AE, Tuch BB, Li H, Johnson AD. Evolution of alternative transcriptional circuits with identical logic. Nature. 2006;443(7110):415-20.

32. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. The evolution of combinatorial gene regulation in fungi. PLoS Biology. 2008;6(2):e38.

33. Booth LN, Tuch BB, Johnson AD. Intercalation of a new tier of transcription regulation into an ancient circuit. Nature. 2010;468(7326):959-63.

34. Creagh E, Brumatti G, Sheridan C, Duriez P, Taylor R, Cullen S, et al. Bicaudal Is a Conserved Substrate for Drosophila and Mammalian Caspases and Is Essential for Cell Survival. PLoS ONE. 2009;4(3).

35. Yokoyama H, Mukae N, Sakahira H, Okawa K, Iwamatsu A, Nagata S. A novel activation mechanism of caspase-activated DNase from Drosophila melanogaster. The Journal of biological chemistry. 2000;275(17):12978-86.

36. Amarneh B, Matthews KA, Rawson RB. Activation of sterol regulatory element-binding protein by the caspase Drice in Drosophila larvae. The Journal of biological chemistry. 2009;284(15):9674-82.

37. Fuentes-Prior P, Salvesen GS. The protein structures that shape caspase activity, specificity, activation and inhibition. The Biochemical journal. 2004;384(Pt 2):201-32.

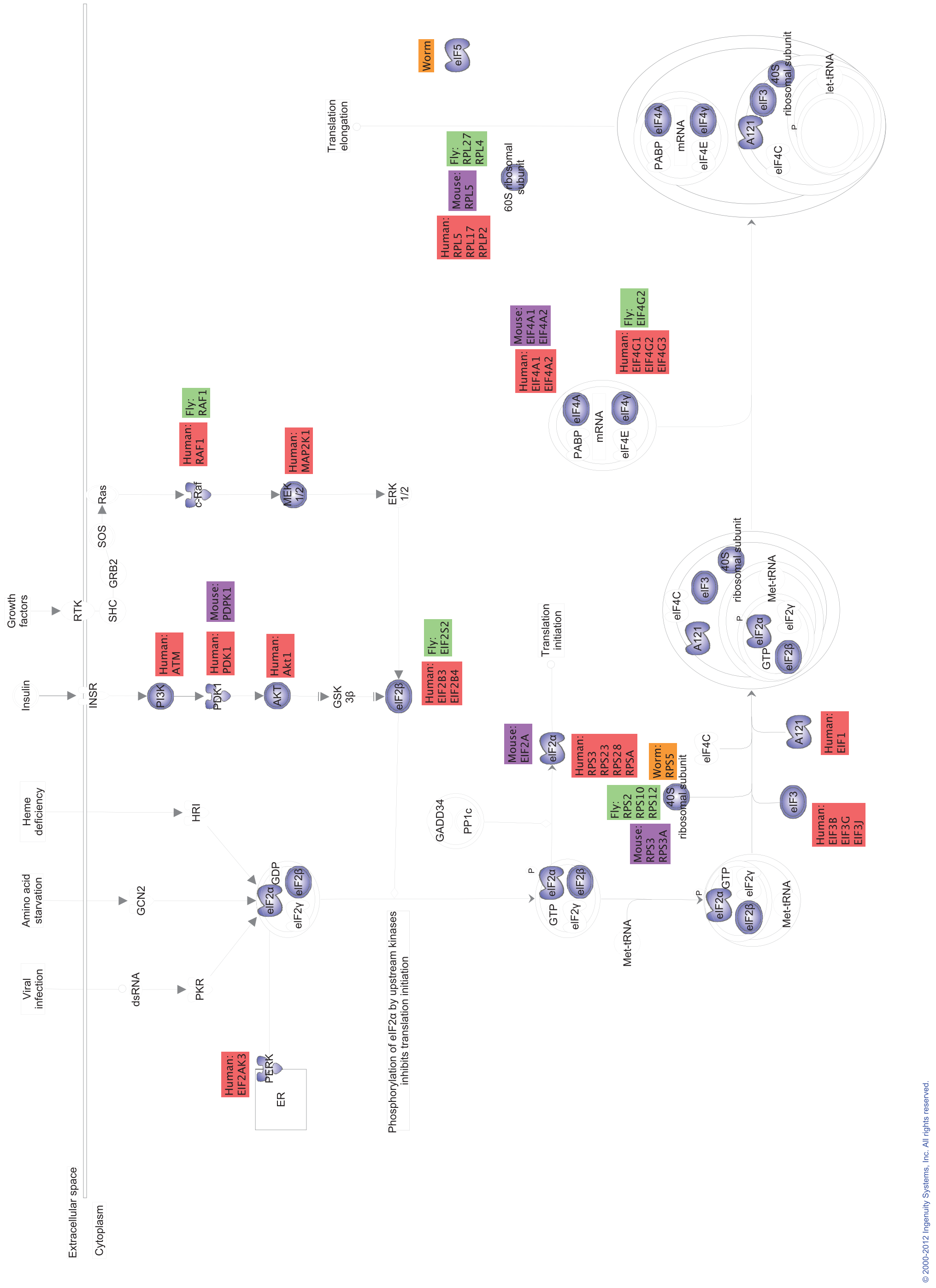
38. Boucher D, Blais V, Denault J-B. Caspase-7 uses an exosite to promote poly(ADP ribose) polymerase 1 proteolysis. Proceedings of the National Academy of Sciences of the United States of America. 2012.

39. McQuilton P, St Pierre SE, Thurmond J, Consortium F. FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Research. 2012;40(Database issue):D706-14.

40. Wallace JA, Felsenfeld G. We gather together: insulators and genome organization. Current opinion in genetics &amp; development. 2007;17(5):400-7.

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

42. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics (Oxford, England). 2004;20(18):3710-5.



Supplementary Figure 1: In cases where the human orthologs were either not substrates, or were cut at different sites, the Asp residue was conserved between 15% and 65% of the time.

Supplementary Tables and Document

For supplementary Tables and Information, see http://www.nature.com/cdd/journal/v19/n12/full/cdd201299a.html.

Supplementary Table 1 – List of all cleavage sites observed in mouse, Drosophila, and C. elegans, and their human orthologs.
Supplementary Table 2 – List of all Canonical Pathways from IPA significantly enriched in the human caspase substrate dataset.
Supplementary Table 3 - GO terms significantly overrepresented in caspase substrate datasets for human, mouse, *Drosophila* and *C. elegans*.

Chapter 4

Cacidases: caspases can cleave after aspartate, glutamate and phosphoserine residues

This work will be published in Cell Death and Differentiation:
Seaman J. E., Julien O., Lee P., Rettenmaier J., Thomsen, N., Wells, J. A., Cacidases: Apoptotic caspases can cleave after both acidic residues and phosphoserine. Cell Death & Dif. (Accepted May 2016).

Abstract:

Caspases are a family of proteases found in all metazoans, including a dozen in humans, that drive the terminal stages of apoptosis as well as other cellular remodeling and inflammatory events. Caspases are named because they are cysteine class enzymes shown to cleave after aspartate residues. In the past decade, we and others have developed unbiased proteomic methods that collectively identified ~2000 native proteins cleaved during apoptosis after the signature aspartate residues. Here, we explore non-aspartate cleavage events and identify 100s of substrates cleaved after glutamate in both human and murine apoptotic samples. The extended consensus sequence patterns are virtually identical for the aspartate and glutamate cleavage sites suggesting they are cleaved by the same caspases. Detailed kinetic analyses of the dominant apoptotic executioner caspases-3 and -7 show that synthetic substrates containing DEVD↓ are cleaved only two-fold faster than DEVE↓, which is well within the 500-fold range of rates that natural proteins are cut. X-ray crystallography studies confirm that the two acidic substrates bind in virtually the same way to either caspases-3 or -7 with minimal adjustments to accommodate the larger glutamate. Lastly, during apoptosis we found 121 proteins cleaved after serine residues that have been previously annotated to be phosphorylation sites. We found that caspase-3, but not caspase-7, can cleave peptides containing DEVpS↓ at only three-fold slower rate than DEVD↓, but does not cleave the unphosphorylated serine peptide. There are only a handful of previously reported examples of proteins cleaved after glutamate and none after phosphorserine. Our studies reveal a much greater promiscuity for cleaving after acidic residues and the name "cacidase" could aptly reflect this broader specificity.

Introduction:

Human caspases are a family of twelve homologous intracellular proteases known for driving cellular state changes such as apoptosis and differentiation, as well as inflammatory responses. Caspases are cysteine-class proteases named for their signature ability to cleave after aspartate residues, or P1 is aspartate using the Schetchter and Berger notation (1, 2). Synthetic peptide profiling for purified caspases show distinctive sub-site preferences extending from P1 to P4 or P5 (3-5). Sequence conservation and signal pathway analyses have further grouped the proteases into apoptotic initiators (-2, -8, -9, -10), apoptotic executioners (-3, -6, -7), regulators of inflammation (-1, -4, -5, -12), and keratinocyte differentiation (-14).

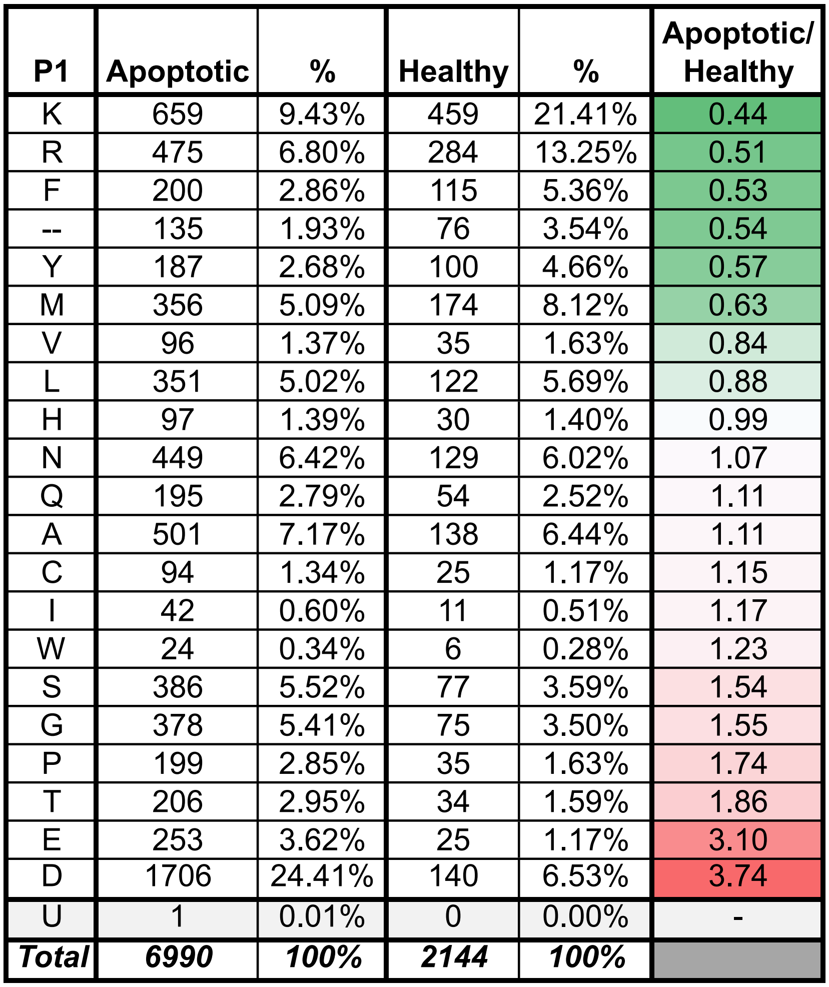
The past decade has seen a significant advancement in the use of LC-MS to define the spectrum of natural proteins cleaved by caspases in cells (6-21). In addition to providing unbiased information about which proteins are cleaved, in some cases, these experiments locate the precise sites and quantify the rates of cleavage (13-17). Using the subtiligase-based N-terminomics approach, we identified more than 1700 aspartate cleavage events distributed among about 1200 different protein substrates upon induction of apoptosis across seven human cell lines (http://wellslab.ucsf.edu/degrabase/)(18). These findings and others revealed structural preferences for cleavage in loops>helices>sheets (14, 19, 20). This large database of aspartate cleaved proteins and their conservation in metazoans has helped to reveal the pathways and nodes that drive apoptosis.

Here, we extend this analysis to proteins cleaved at non-aspartate sites during apoptosis. Surprisingly, we find enrichment of proteins that are cut after glutamate in apoptotic cells. We find these glutamate sites have similar subsite specificities, degrees of conservation, and Gene Ontology (GO) term enrichments as seen for aspartate cleavages. Remarkably, the catalytic efficiency for cleaving a glutamate substrate is only two-fold less than for cleaving the matched aspartate substrate by caspase-3 or -7. Structural studies show that both acidic residues are accommodated in the binding sites of the two enzymes. Finally, we identified 121 P1 serine sites that are literature annotated phosphorylation sites and cut in apoptosis, and show caspase-3 can cleave after phosphoserine. Previously, there have been a handful of studies reporting proteins cut after glutamate by caspases (17, 22-26). Our studies reveal a surprising promiscuity for caspases to cleave P1 acidic residues suggesting a more expanded range of substrates than previously appreciated.

Results:

Human and mouse apoptotic cells are similarly enriched for proteolysis after aspartate and glutamate. The DegraBase (http://wellslab.ucsf.edu/degrabase/) is a database comprising about 8000 unique proteolytic cuts identified by the subtiligase-based N-terminomics labeling technology from 33 apoptotic and 11 healthy cellular experiments (18). From this resource we ranked the percentage of instances that the 20 amino acids appear in the P1 position in apoptotic compared to healthy cells (Table 1). Consistent with major activation of caspases, the strongest enrichment ratio for substrates cut in apoptosis versus healthy cells is for aspartate residues (3.7-fold enriched). However, we were intrigued to find the second most enriched P1 residue during apoptosis is glutamate (3.1-fold enriched). From the Degrabase, we identified a total of 1706 P1 aspartate cleavage sites in human cells and 253 P1 glutamate sites (Table 2a). The 1706 P1 aspartate cleavages are distributed among 1268 proteins, which represent 1.3 cleavages per protein. Similarly, 253 P1 glutamate cleavages are found in 226 proteins, or 1.1 cuts per protein. We had previously generated a smaller apoptotic data set for mouse cells (27). From this dataset we find 38 unique P1 glutamate cleavages compared to 259 P1 aspartate cleavages (Table 2a).

We find a similar number of cut sites per protein from the mouse data: 38 cuts in 36 unique proteins for P1 glutamate compared to the 259 peptides in 221 proteins for P1 aspartate, reflecting 1.1 and 1.2 cuts per protein, respectively. These data suggest that cleavage after glutamate is quite frequent in both human and mouse cells and represents the same fraction of total cuts.



P1	Apoptotic	%	Healthy	%	Apoptotic/ Healthy
K	659	9.43%	459	21.41%	0.44
R	475	6.80%	284	13.25%	0.51
F	200	2.86%	115	5.36%	0.53
--	135	1.93%	76	3.54%	0.54
Y	187	2.68%	100	4.66%	0.57
M	356	5.09%	174	8.12%	0.63
V	96	1.37%	35	1.63%	0.84
L	351	5.02%	122	5.69%	0.88
H	97	1.39%	30	1.40%	0.99
N	449	6.42%	129	6.02%	1.07
Q	195	2.79%	54	2.52%	1.11
A	501	7.17%	138	6.44%	1.11
C	94	1.34%	25	1.17%	1.15
I	42	0.60%	11	0.51%	1.17
W	24	0.34%	6	0.28%	1.23
S	386	5.52%	77	3.59%	1.54
G	378	5.41%	75	3.50%	1.55
P	199	2.85%	35	1.63%	1.74
T	206	2.95%	34	1.59%	1.86
E	253	3.62%	25	1.17%	3.10
D	1706	24.41%	140	6.53%	3.74
U	1	0.01%	0	0.00%	-
Total	***6990***	***100%***	***2144***	***100%***	

Table 1. Enrichment for each of the 20 possible amino acids observed at the P1 position in apoptotic samples compared to healthy cells. The fold enrichment of each residue is calculated from the ratio of the percent of peptides found with each P1 residue in apoptosis to the percent found in healthy cells. The fold enrichment values (final column) is ranked from low (green) to high (red) reflecting higher prevalence in apoptotic cells. Note the high enrichment in apoptotic cells for substrates cleaved with P1 aspartate, glutamate and serine. There is one cleavage in apoptotic cells after selenocysteine (U).

a

	Human		Mouse	
	Count	Acidic %Glu	Count	Acidic %Glu
Asp N-termini	1706	12.9%	259	12.8%
Glu N-termini	253		38	
Asp Proteins	1268	15.1%	221	14.0%
Glu Proteins	226		36	

b

		Human		Mouse	
		Asp	Glu	Asp	Glu
Human	Asp	1268			
	Glu	100	226		
Mouse	Asp	179	31	221	
	Glu	14	18	10	36

Table 2: High conservation in mouse and human of P1 aspartate and glutamate cleavage sites and protein targets. **(a)** The number of unique N-termini and protein targets are shown for aspartate and glutamate sites for human and mouse. The percent of P1 glutamate sites is essentially the same between human and mouse datasets. **(b)** There is a significant overlap of aspartate and glutamate proteins within human and between human and mouse. Almost half the proteins cleaved after glutamate are also cleaved at an aspartate site.

We next compared the conservation of P1 glutamate and aspartate apoptotic sites between mouse and human (Table 2b). Of the 221 P1 aspartate cleaved mouse proteins, 179 human orthologs were also cleaved (27). Here we find that of the 36 glutamate cleaved mouse proteins identified, 18 proteins have human orthologs that are also cut. For the 179 targets that are cut at aspartate in both mouse and human, 149 are cleaved at precisely the same site (83% site conservation). Similarly for the 18 glutamate cut targets in common between human and mouse, 15 of the sites are conserved (83% site conservation). Thus, aspartate and glutamate sites appear equally conserved suggesting they are under roughly equivalent evolutionary pressure.

Proteins cleaved at P1 glutamate frequently have P1 aspartate cleavages. During apoptosis, some proteins are cut more than once by caspases. Thus, we compared the frequency that proteins are cut both after glutamate and aspartate (Table 2b). Of 1268 aspartate cleaved proteins, we find 100 are also cut at a P1 glutamate site, which is 8% of the aspartate cleaved proteins. In contrast, of 226 proteins that cut after glutamate, 100 are also cut after aspartate representing 45% of the total glutamate cleaved proteins. Similarly in mouse, for the 221 total aspartate cleaved targets only 10 had an additional P1 glutamate cut site (~4%), but for 36 glutamate targets 10 had an additional P1 aspartate cut site (~30%). Thus, proteins targeted for glutamate cleavage have a higher propensity to contain an additional aspartate cut site. Moreover, there are two populations of dual acid cut proteins. About 40% of human proteins cut at both aspartate and glutamate sites (112 proteins total) are cleaved within 20 amino acids of each other, suggesting the presence of hot spots, and another 40% cut over 100 residues apart, suggesting unique cleavage effects (Supplementary Figure 1a). In contrast, of the 726 cleavages from 287 proteins with multiple aspartate cleavages, only 11% are within the 20 amino acid hot spot and 70% are cut over 100 residues apart, suggesting aspartate sites are optimized for unique cleavage effects (Supplementary Figure 1b). We found no difference in the location of sites within the substrate for cutting after either aspartate or glutamate (Supplementary Figure 1c) as most cleavages appear after 60 residues of the protein. The glutamate cuts follow the same pattern of structural preference as aspartate cuts, loops>helices>sheets (Supplementary Figure 1d), and almost all of the glutamate substrates GO terms overlap with the aspartate substrates (Supplementary Figure 1e). Thus, from a secondary structural and functional bioinformatics perspective we cannot distinguish P1 glutamate from P1 aspartate cleaved targets.

Caspases can cleave after aspartate or glutamate residues in cells and *in vitro*. To determine the protease(s) responsible for glutamate cleavage, we compared the consensus cut site motifs with those surrounding aspartate cuts using iceLogos. There are 2144 unique proteolytic events in healthy human cells, and the cleavage patterns are predominantly tryptic or chymotryptic-like with corresponding preference for lysine, arginine, or large hydrophobic P1 side chains, and small side-chains at P1' (18). The iceLogo changes dramatically for the apoptotic dataset. From the nearly 7000 unique proteolytic events detected, there is an emergence of aspartate as the predominant P1 residue. Focusing on the 1706 unique aspartate cut sites (Figure 1a), the DEVD↓ represents the highest cleavage consensus motif and matches the apoptotic caspases-3 and -7 consensus motif based on synthetic substrates (28, 29). Remarkably, the iceLogo for the 253 unique P1 glutamate cleaved sites (DEVE)↓ (Figure 1b) is virtually identical to that seen for the P1 aspartate cleaved sites from P4 to P2'. The mouse data set shows essentially the same patterns (Supplementary Figure 2)(27).

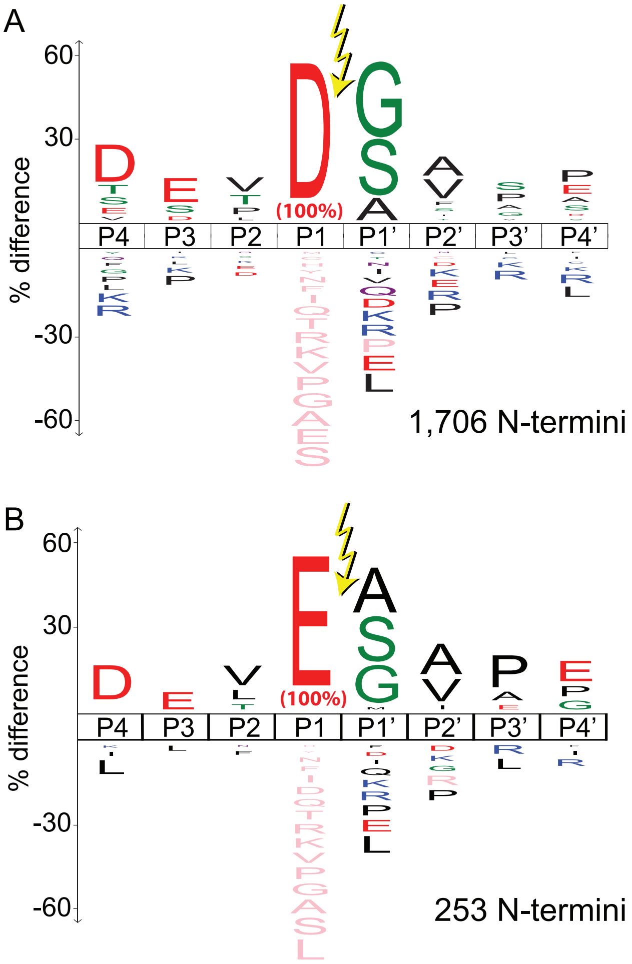


Figure 1. There is strong similarity between iceLogos for substrates cleaved after **(a)** P1 aspartate and **(b)** glutamate in apoptotic human samples. The iceLogos for P1 aspartate and P1 glutamate are virtually identical in the extended P4-P4’ sequence around the cleavage site. **(a)** reproduced from Crawford et al. 2014(18), with permission.

These data strongly suggest the same protease(s) are responsible for cleaving after aspartate and glutamate, likely the executioner caspases-3 and/or -7. Therefore we quantified the rates that these two caspases cleave standard synthetic substrates containing the fluorogenic 7-amino-4-timethylfluoro-coumarin (*AFC*) reporter attached to the peptide, Ac-DEVx-*AFC*, where x is either aspartate or glutamate (Supplementary Figure 3a-b). These data allowed calculation of the kinetic constants (k_{cat} and K_M) for hydrolysis (Table 3). Remarkably, the catalytic efficiency values (k_{cat}/K_{M}) for cleaving after aspartate are only about two-fold higher than for cleaving after glutamate by either caspase-3 or -7.

						D/E Ratio		
Caspase	Substrate	V_{max} (mM/s)	K_M (mM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))	K_M (mM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))
3	Asp	5.3 (0.16)	57.9 (5.6)	3.6E+02	6.1E+06	3.55	8.15	2.29
3	Glu	0.65 (0.04)	16.3 (3.8)	4.4E+01	2.7E+06			
7	Asp	2.6 (0.05)	49.4 (3.1)	1.7E+02	3.5E+06	0.37	0.79	2.14
7	Glu	3.3 (0.06)	133.8 (6.3)	2.2E+02	1.6E+06			

Table 3. Summary of kinetic values for cleavage of P1 aspartate and glutamate by caspases-3 and -7. The values for V_{max} *and* K_M with standard errors are reported from triplicate measurements on three biological replicates. The ratio of k_{cat}/K_M for P1 aspartate/P1 glutamate is just over two-fold for both caspases-3 and -7.

There are, however, counterbalancing differences seen in the specific k_{cat} and K_M values between the caspases. For example, the k_{cat} values for caspase-3 are eight-fold higher for cleaving aspartate, but the K_M values are nearly four-fold higher. This suggests binding is actually weaker for P1 aspartate, but the catalytic step is faster. Caspase-7 shows the opposite trend, where the k_{cat} is about the same for the P1 aspartate and glutamate, but the K_M value is about two-fold lower for the P1 aspartate. Commercial preparations of caspases-3 and -7 also had rate constants virtually identical to our recombinant forms (Supplementary Figure 3c-d). The rate constants we obtained with the P1 aspartate substrate are within 3-fold of published values for the (28, 30, 31). We extended the studies to mouse caspase-3 and -7 (Supplementary Figure 3e-f). The catalytic efficiency values for the P1 aspartate substrate were consistently only two-fold above the P1 glutamate substrate, and mouse caspase-3 and -7 (Supplementary Table 1) similarly differed in their relative k_{cat} and K_M values as seen in the human homologs.

Since Ac-DEVE-*AFC* is cleaved only two-fold slower than Ac-DEVD-*AFC* by caspase-3 and -7, we wondered if it is possible to substitute a P1 glutamate site for a known P1 aspartate cut site in cells and observe it being cleaved during apoptosis. PARP is a classic caspase substrate cut in apoptosis at a DEVD↓ sequence ending at aspartate-214 (32). We introduced three human expression constructs into HEK293 cells: PARP harboring the wild-type P1 aspartate site (DEVD), a P1 glutamate site (DEVE), or a P1 alanine control site (DEVA) (Supplementary Figure 4). We induced apoptosis by treatment with staurosporine, and monitored cleavage of PARP by Western blot. As expected, the native DEVD construct was rapidly cleaved, the DEVE was cut but more slowly, while the DEVA construct remained uncleaved. Thus, a P1 glutamate can substitute for P1 aspartate for cleavage in apoptotic cells, albeit more slowly.

Crystal structures show minimal differences in binding for P1 aspartate and glutamate peptides. To understand the structural basis for the interactions of the P1 glutamate with caspases-3 and -7, we determined the crystal structures of the two caspases covalently labeled with DEVE-cmk (chloromethylketone) each at ~2.7 Å resolution (Figure 2, Supplementary Table 2, Supplementary Figure 5). The asymmetric unit of the caspase-3-DEVE complex (PDB 5IC4) contains two structurally similar copies of the biological dimer that align well with a caspase-3-DEVD structure (Cα RMSD 0.4, PDB 2DKO(31)). The caspase-7-DEVE complex (PDB 5IC6) contains one biological dimer in the asymmetric unit, which is virtually identical to a caspase-7-DEVD structure (Cα RMSD 0.4, PDB 3H1P(33)). Importantly, the backbone and side-chains of the DEVE and DEVD peptide inhibitors as well as the positions of the caspase peptide binding residues align closely for both caspases-3 and -7 (Figure 2c and 2f). Although the P1 glutamate contains an additional methylene group compared to aspartate, both P1 acidic side-chains interact with caspases-3 and -7 in virtually identical binding modes.

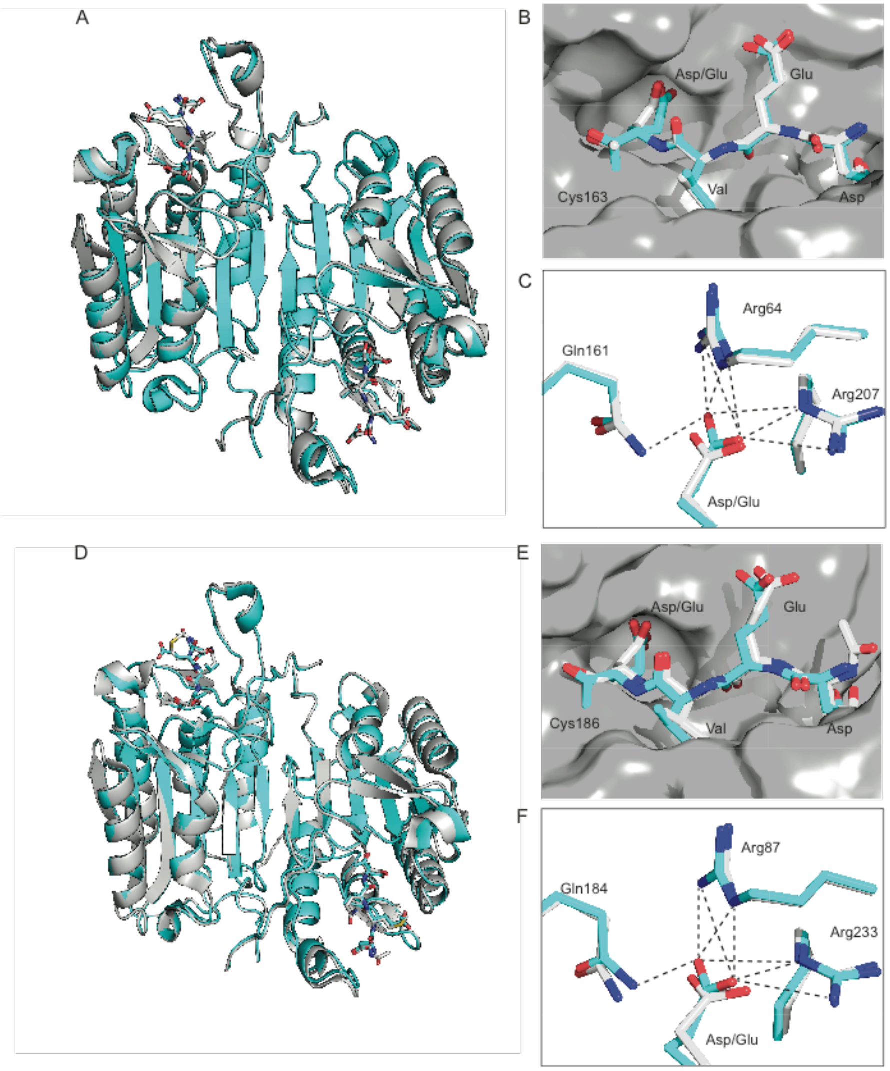


Figure 2. Overlap of P1 glutamate and aspartate substrates in the caspase-3 and caspase-7 pockets show minimal structural rearrangements to accommodate the extra methylene in P1 glutamate. **(a-c)**. Human caspase-3 was labeled with Ac-DEVE-cmk, crystallized and X-ray structure solved to 2.65 Å (PDB: 5IC4). **(d-f).** Human caspase-7 was labeled with Ac-DEVE-cmk, crystallize and X-ray structure solved to 2.7 Å (PDB: 5IC6). The P1 glutamate labeled protein (Blue) has minimal changes when overlayed with the P1 aspartate labeled protein caspase-3 (2DKO, Gray) or caspase-7 (3H1P, Gray). The differences between the glutamate and aspartate structures are localized to the binding pocket and substrate itself.

Caspase-3 can cleave after phosphorylated serine. Given the significant promiscuity for apoptotic caspases to cleave after glutamate, we wondered if it was possible that they cleave directly after phosphorylated residues. It was possible to model a P1 phosphoserine into both caspases-3 and -7 with only small perturbations in the active site, though the bulkier phosphothreonine and phosphotyrosine predicted steric clashes. We explored our Degrabase to see if cleavages occur after serine, threonine and tyrosine residues known to be phosphorylated from the PhosphoSite Database (Table 4) (34). Remarkably, we found 260 cuts after annotated phosphorylated sites during apoptosis compared to only 69 in healthy cells. A list of the annotated P1 phosphorylation substrates is shown in Supplementary Table 3. We were intrigued to find that P1 serine and threonine annotated phosphosites are among the most abundantly enriched in the apoptotic set by 2.2 and 1.8 fold, respectively. In contrast, the much larger phosphotyrosine is significantly de-enriched in the apoptotic set. We analyzed the iceLogos for the phosphosites as well as the specific phosphorylated residues (Figure 3). There is little consensus from the iceLogo when grouping all phosphosites together or for the phosphotyrosine set. There is somewhat more defined consensus sequences for cleavages at phosphothreonine and phosphoserine. GO term analysis did not reveal any remarkably unique biology for the phosphoserine and phosphothreonine groups relative to each other or other P1 acid cleaved substrates.

P1	Apoptotic	%	Healthy	%	Apoptotic/ Healthy
all Phospho	260	3.72%	69	3.22%	1.16
p-Tyr	84	1.20%	43	2.01%	0.60
p-Thr	54	0.77%	9	0.42%	1.84
p-Ser	121	1.73%	17	0.79%	2.18

Table 4. Enrichment of annotated phosphorylation sites observed cleaved at the P1 position in apoptotic samples compared to healthy cells. The fold enrichment of each residue is calculated from the ratio of the percent of peptides found with each P1 residue in apoptosis to the percent found in healthy cells compared to the total peptides (6990 apoptotic, 2144 healthy). The fold enrichment values (final column) for phospho-serine and phospho-threonine are highly enriched in the apoptotic dataset whereas phospho-tyrosine is de-enriched.

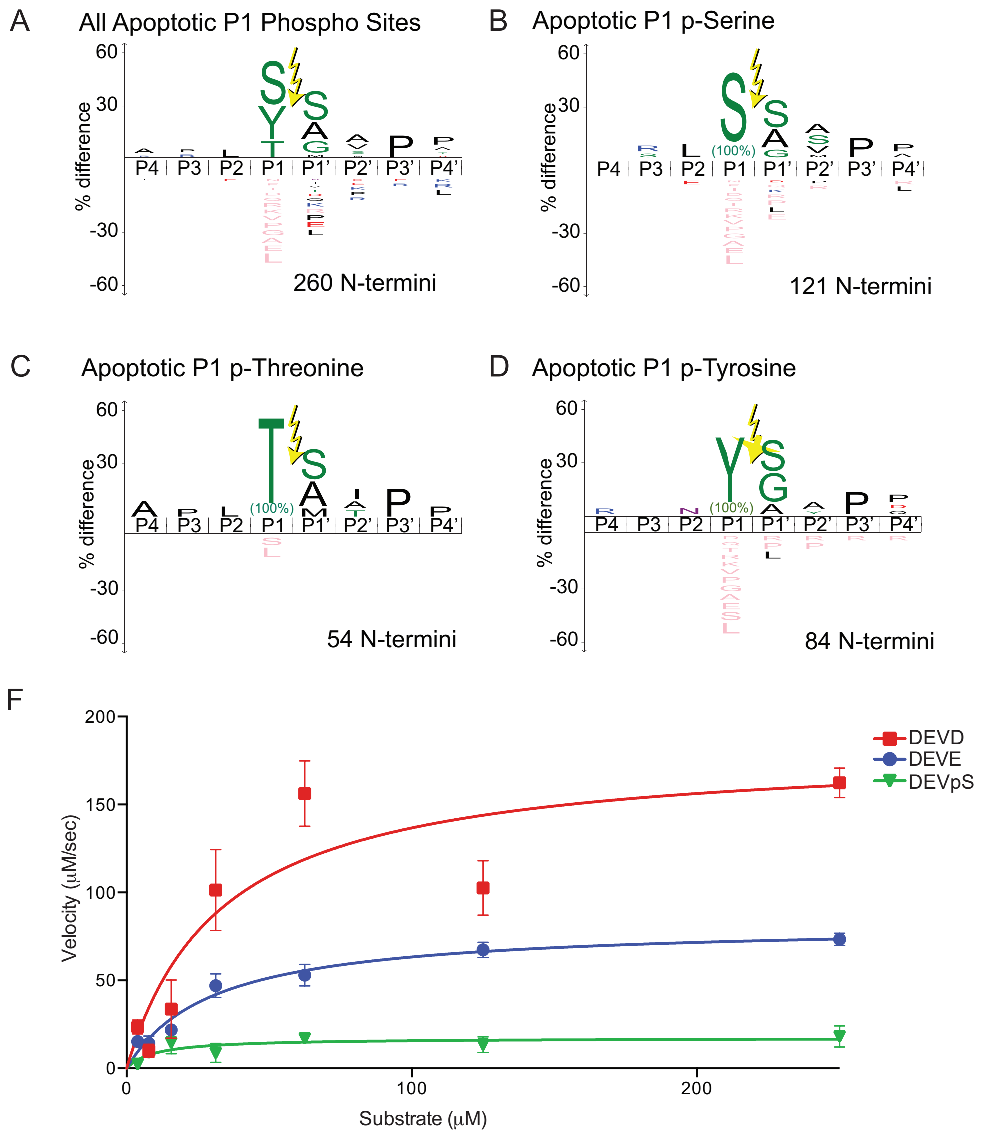


Figure 3. Cleavage of P1-phosphoserine sites are frequently found in the Degrabase and caspase-3 can cleave synthetic substrate containing a P1 phosphoserine, biotin-WDEV(pS)↓SGVDEK(DNP). **(a-d)** The iceLogos for the combined group of three phospho-sites (a), phospho-serine (b), phospho-threonine (c) and phospho-tyrosine (d), do not match the typical caspase cleavage motif. **(e)** Caspase-3 can cleave the **biotin-WDEV(pS)↓SGVDEK(DNP)** substrate only 3-fold slower than the **biotin-WDEV(D)↓SGVDEK(DNP)** substrate. Error bars represent standard error from mean.

We next directly measured the kinetics of proteolysis of P1 phosphoserine or phosphothreonine substrates by caspases-3 and -7. We designed FRET substrates based on the PARP caspase cleavage sequence, DEV**(x)**↓GVDE containing a P1 aspartate, glutamate, phosphoserine or phosphothreonine and control substrates of P1 alanine and serine. As expected, both the P1 aspartate and glutamate FRET peptides were cleaved rapidly (Figure 3e). We observed quantitative cleavage of P1 phosphoserine by caspase-3 but not caspase-7, and no cleavage of P1 phosphothreonine by either caspase-3 or -7. There was no observable activity for the control alanine or serine substrates. The caspase-3 catalytic efficiency of the FRET peptides for aspartate was two-fold more active than glutamate and only three-fold more active than phosphoserine (Table 5). The caspase-3 phosphoserine kinetic constants k_{cat} and K_{M} are both lower than glutamate, and further reduced compared to aspartate. Thus, caspase-3 is more promiscuous than caspase-7 and able to cleave all three acidic functionalities at rates that are within an order of magnitude of each other.

Caspase	Substrate	V_{max} (mM/s)	K_M (mM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))	k_{cat}/K_M ratio	
3	Asp	182 (21.4)	33.4 (11.9)	1.8E-01	5.5E+03	Asp/Glu	2.00
3	Glu	82.4 (3.9)	30.2 (4.6)	8.2E-02	2.7E+03	Asp/pSer	3.14
3	phospho-Ser	17.3 (2.4)	9.95 (5.6)	1.7E-02	1.7E+03	Glu/pSet	1.57

Table 5. Summary of kinetic values for cleavage after aspartate, glutamate or phospho-serine FRET peptides and comparison of the ratios of k_{cat}/K_M for three FRET peptides are shown in the last column.

P1 aspartate and P1 glutamate sites are well conserved between mouse and human, but not for distant metazoans. We took the 1706 observed P1 aspartate cleaved human sites and determined the frequency these were conserved in homologs by sequence alignments and compared this with the frequency they were found altered to glutamate. There is a weak tendency for the 1706 aspartate sites to switch to glutamate (Figure 4a), and a similarly weak tendency for the 226 human glutamate sites to switch to aspartate (Figure 4b). However, both residues were more likely to remain unchanged than switch to the other, and almost as likely to switch to another amino acid as to its acidic relative.

We analyzed the relationship between evolutionary conservation of the acidic cleavage sites relative to recently measured rates of proteolysis of protein substrates from a subset of rate data compiled from 246 substrates cleaved by caspases-2, -3, -6, -7 and -8 (21). We find a very weak trend where the faster the substrate is cleaved the stronger the acidic residue conservation, but only for caspases-3 and -7, and not caspases-2, -6 and -8 (Supplementary Figure 6). Additionally, there is no significant relationship between conservation of acidic sites among the 84 substrates that are cleaved by more than one caspase within our kinetic dataset of 246 substrates (Supplementary Figure 7). It is likely that these acidic cleavage sites evolve so quickly that only very mild trends are apparent in the evolutionary record.

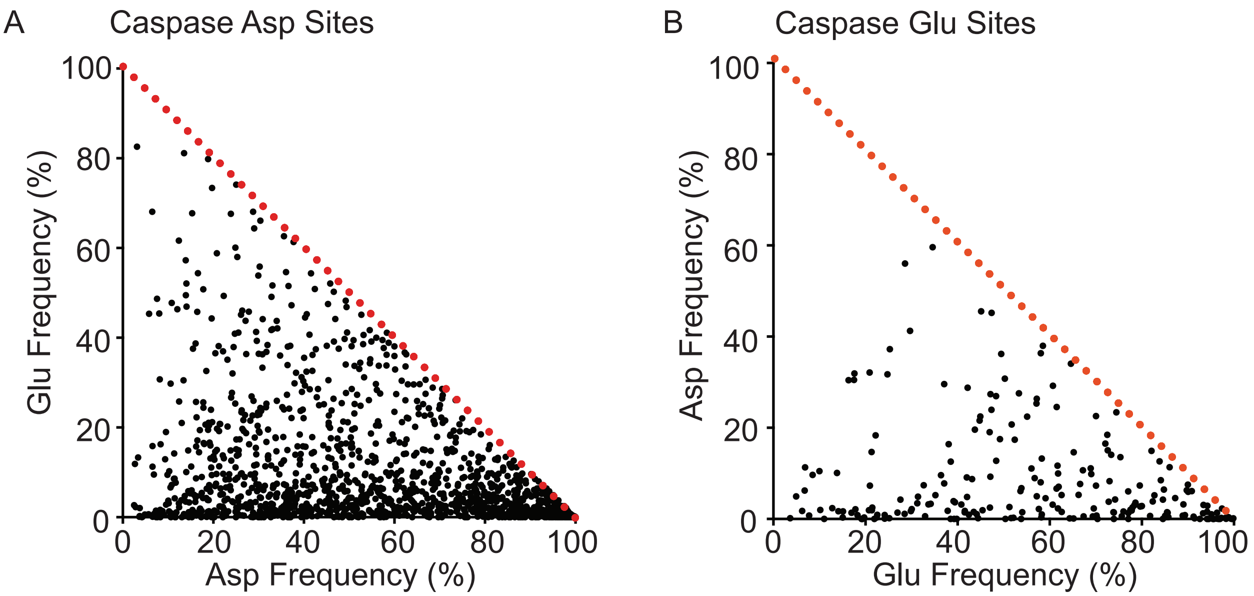


Figure 4. P1 aspartate and glutamate cleavage sites are not strongly conserved throughout the broader metazoan evolutionary record. Human caspase substrates with a P1 aspartate (**a**) and glutamate sites (**b**) are compared to all homologs for site conservation of a P1 aspartate or P1 glutamate in the sequence. The dashed red line indicates maximal conservation of 100% acidic residue (Asp + Glu). For both aspartate and glutamate sites, there is a tendency to remain itself as the data is skewed to the bottom left of the chart and minimal activity switching to the other acidic residue.

Discussion

These studies highlight the use of proteomics coupled with detailed kinetic and structural analysis to reveal a greater prevalence and promiscuity for cleaving after glutamate residues in apoptosis. We found no remarkable differences in target proteins cleaved after aspartate or glutamate based upon iceLogo, secondary structure, evolution, or cellular functions. The structural studies show that P1 glutamate can be accommodated in the substrate binding sites of both caspases-3 and -7 with minimal perturbations. Moreover, detailed kinetic analysis for matched synthetic substrates showed that aspartate is cleaved only two-fold faster than glutamate by either caspases-3 or -7; this difference is roughly the same in cells undergoing apoptosis for the protein PARP containing the wild type DEVD or replacement DEVE site. The two-fold kinetic difference we see does not fully account for the eight-fold difference from N-terminomics of aspartate over glutamate cleavages but prevalence and intrinsic rate measurements may not scale linearly. It is also possible that cellular cleavage activity of P1 glutamate substrate may rely on more optimized residues than is required for cleavage of P1 aspartate substrates. The rates of individual caspase substrates vary over 500-fold so this two- to eight-fold rate differences between aspartate and glutamate cleavages may be almost biologically indistiguishable(14, 21). Our studies indicate that P1 glutamate cleavages are biologically relevant and isofunctional with P1 aspartate cleavages.

There have been isolated reports of single human proteins being cleaved after glutamate by human caspases-1, -3, -5 and -9 (17, 22-26) and a P1 glutamate synthetic substrate cleaved by caspase-14 (35). Indeed, initial *in vitro* tests of human and other species' apoptotic and non-apoptotic caspase activities for Ac-DEVE-*AFC* have indicated broad glutamic acid activity (Supplementary Table 4) (36). Additionally, *in silico* caspase substrate algorithms have predicted, but not empirically confirmed, the ability for caspases to cleave P1 glutamate substrates (37, 38). Probably the best-characterized functional example is for DRONC, a drosophila caspase, that cleaves itself during activation after glutamate (39-41). The active-site sequence around the catalytic cysteine of DRONC is quite different (PF**C**R versus QA**C**Q, where C is the catalytic cysteine) from that of other fly caspases, such as DCP-1, which is shown to prefer DEVD (42). Hawkins and co-workers proposed this difference could facilitate more promiscuity for glutamate binding (39). However, the structures we present for inhibitors bound to caspases-3 and -7 show the P1 glutamate can be readily accommodated in the binding site.

Co-regulation of between kinases and caspases has long been recognized(43). Most caspases are known to be phosphorylated, and nearly half of the 518 human protein kinases can be cut during apoptosis (12, 18, 44). Phosphorylation on caspase substrates can also directly inhibit or activate cleavage rates by caspases (12). Our data here show that caspase cleavage may also occur directly following a phosphorylated serine site during apoptosis, suggesting kinase directed targeting of proteolysis. While we infer the P1 residue from the gene sequence due to the method capturing the N-terminus of the P1' cleavage fragment, we do not know the explicit phosphylation state of the residue. Nonetheless we know these P1 serines can be phosphorylated at these sites, and have shown caspase-3 can directly cleave after a synthetic phosphoserine substrate. The synthetic substrates were not optimized to reflect the (R/P)LpS motif from apoptotic substrates. If these phosphoserine substrates were further optimized, it is possible these enzymes may indeed have higher activity for other P1 phosphoserine substrates and activity for phosphothreonine. Further experiments will be needed to pursue optimization and kinetic determination of other phosphopeptides. To our knowledge this is the first example of any protease that cleaves immediately after a phosphorylated residue.

It is well-known that substitution of acidic residues at sites of phosphorylation can function as constitutive mimics (45), so perhaps it is not surprising we see capsase-3 cleave after phosphoserine. Ferrell and co-workers have systematically studied the evolutionary interplay between acidic and phosphorylation sites in kinases and estimate that upwards of 5% of the phosphorylation sites are derived from acidic precursors (46). They suggest phosphorylation sites tend to develop later in evolution as regulatory complexity develops. Phosphorylation sites are known to evolve rapidly in the sequence record, but they noted that phosphoserine sites most commonly mutate to threonine, glutamate and aspartate by enrichment factors of 1.5, 1.3 and 1.2, respectively. We have similarly found that caspase sites evolve very rapidly, with target choice being more conserved and pathway choice most conserved (27). With the direct interplay between kinases and caspases in apoptosis, we wondered if there was a common evolutionary pattern that may mimic or connect the two pathways. Indeed there is a trend for caspase sites to remain as an acidic amino acid, but it is not as strong as phosphorylation and is only mildly evident for caspase-3 and -7 substrates.

The data set presented here on nearly 2000 natural caspase substrates suggest human caspases have a much greater promiscuity for cleaving acidic functionalities than has been reported. Why was this overlooked? Before the application of LC-MS, proteomic studies only focused on tens of candidate-based protein substrate identifications using SDS-PAGE to estimate where a candidate protein was cleaved, and for only some of these did protein sequencing and mutagenesis identify the precise sites of cleavage. With such a small data set, it is not surprising P1 glutamate cuts would be missed systematically. Furthermore, specificity studies using positional-scanning libraries of synthetic substrates were focused on optimizing assays and finding strong inhibitors. These studies fixed on P1 aspartate as the sole P1 substrate for greatest activity levels and then thoroughly analyzed preferences in P4-P2 to identify optimal assay substrates and inhibitor sequences based on P1 aspartate (3). Subsequent proteomic studies, including our own, only focused on P1 aspartate cleavages as a traditional caspase signature. The more in-depth and unbiased bioinfomatic analysis of the DegraBase revealed the P1 glutamate and P1 phosphoserine signatures. We suggest P1 acidic residue promiscuity is general throughout the caspase family in all metazoans, and the term “cacidases” aptly reflects this broader specificity.

Materials and Methods

Database Analysis:

The DegraBase (http://wellslab.ucsf.edu/degrabase/) compiles 8090 unique N-termini from 3206 proteins identified in subtiligase-based positive enrichment mass spectrometry experiments in 11 healthy and apoptotic human cell lines (18). The database was split into multiple subsets for analysis: N-termini from healthy cells, N-termini from apoptotic cells, apoptotic N-termini following an aspartate cleavage (Apop Asp), and apoptotic N-termini following a glutamate cleavage (Apop Glu). Sequence logos were made using iceLogo (47) with the whole human SwissProt library as background control. All cleavage iceLogo images were made with the percent difference scoring system and significance was determined by chi-square analyses. The secondary structure was calculated using the NetSurfP server for the whole protein(48). The P4-P4' iceLogo was made using the filled logo option in iceLogo.

GO term enrichment was determined using the GO::TermFinder software(49). A list of unique proteins for each dataset was created and uploaded to the database and tested for enrichment against the human SwissProt background using all evidence codes except ND (No biological Data available) and IEA (Inferred from Electronic Annotation). Enriched terms were defined using a corrected p-value cutoff of less than 0.01. To compare terms between datasets, a pairwise chi-square test was performed using the Benjamini-Hochberg multiple testing correction procedure.

Enzyme purification and kinetics

Caspase-3 and -7 were purified using a C-terminal 6xHis tag in *E. coli* as previously described (50). Enzyme activity was assayed using a 7-Amino-4-trifluoromethylcoumarin (*AFC*) substrate (excitation 395nm, emission 505nm) or a Tryptophan/lysine-DNP FRET peptide substrate (excitation 280nm, emission 355nm) in caspase activity buffer (50mM KCl, 50mM HEPES pH 7.4, 0.1mM EDTA, 0.1% CHAPS, 10mM DTT) at room temperature. The *AFC* substrates, Ac-DEVD-*AFC* and Ac-DEVE-*AFC* (SMBiochemicals, Anaheim, CA), were used with final caspase concentrations ranging from 10-50nM. The FRET substrates, biotin-WDEV(x)GVDEK(dinitrophenol) where x is D, E, S, T, A, pS (phosphoserine) (Chinapeptides, Shanghai, China) or pT (phosphothreonine) (Shengnuo Peptide, Menlo Park, CA), were used with enzyme concentrations ranging from 500nM-2μM for optimized signal timing and intensity. FRET substrates were used instead of *AFC* reporter substrates for ease of synthesis of specified phospho-P1 residues. All enzymes were assayed at specified concentrations as determined by active site titrations using z-VAD-fmk (Bachem Americas, Torrance, CA). The final kinetics determined using triplicate biological replicates and each substrate concentration was assayed in triplicate (nine assays per substrate concentration total). Commercial sources of human caspase-3 (BD Biosciences, San Jose, CA) and caspase-7 (G-Biosciences, St. Louis, MO) were tested with the same kinetic protocols. All substrates showed regular parabolic Michaelis-Menten kinetic curves consistent with normal steady state kinetics and no evidence of product inhibition over the initial rate measurements.

Cellular Assays

HEK293T (ATCC CLR-3216, Manassas, VA) cells were cultured in DMEM high-glucose medium supplemented with sodium pyruvate, nonessential amino acids, penicillin/streptomycin, and 10% FBS at 37°C. The cells were transiently transfected with pcDNA3.1 containing PARP (Uniprot P09874) with endogenous sequence (Asp214), or mutated to glutamate (Asp214Glu) or alanine (Asp214Ala), using Lipofectamine 2000 (ThermoFisher Scientific, Waltham, MA) in

OptiMEM media. After 6 hours, the cells were supplemented with full DMEM media. The cells were treated with 500nM staurosporine 24 hours after transfections and harvested at various time-points up to 24 hours later in M-PER Mammalian Protein Extract Reagent (Pierce, Rockford, IL) with protease inhibitors (cOmplete, Mini, EDTA-free, Roche, Indianapolis, IN). Extracts were run on SDS-PAGE gels and PVDF western blots. The antibodies used were anti-PARP (9542, Cell Signaling, Danvers, MA) and anti-V5 (V8012, Sigma-Aldrich, St. Louis, MO).

Crystallization, data collection, structure determination, and refinement

Purified caspases-3 and -7 were treated with 1:8 and 1:32 ratio excess of z-DEVE-cmk (chloromethylketone) (American Peptide Company, Sunnyvale, CA), respectively, in 50mM HEPES pH 7.4, 50mM KCl, 0.1mM EDTA and 1mM DTT for two hours at 37 ^{o}C. Complete and single site covalent labeling was confirmed by mass spectrometry. The covalent caspase-DEVE complexes were dialyzed into 10mM Tris pH 8.0, 50mM NaCl, 10 mM DTT (caspase-3) or 2mM DTT (caspase-7) and concentrated to 10mg/mL. The caspase-3-DEVE and caspase-7-DEVE complexes were initially screened for crystallization using the Index Suite (Hampton Research, Aliso Viejo, CA), PACT Suite, PEGs Suites and PEGs 2 suite (Qiagen) screens. Caspase-3-DEVE crystal hits were optimized by mixing 1 μl of protein (10 mg/ml) with 1 μl of mother liquor (60% (v/v) tacsimate pH 7.0) at room temperature and crystals appeared in 14 days. Caspase-7-DEVE crystal hits were optimized by mixing 1 μl of protein (10 mg/ml) with 1 μl of mother liquor (0.2 M ammonium fluoride, 0.1 M sodium acetate pH 4.6, 20% (w/v) PEG 10000) at room temperature and crystals appeared in 14 days. The crystals were cryoprotected with their respective mother liquors supplemented with 10% glycerol, flash cooled, and stored in liquid nitrogen until data collection. Diffraction data were collected on the 8.3.1 beamline at the Advanced Light Source (ALS) and were processed in space groups F222 and $P2_12_12_1$ for the caspase-3-DEVE and caspase-7-DEVE complexes, respectively, using the XDS package (51). The structures were determined by molecular replacement with Phaser (52) using PDB codes 2DKO (31) for caspase-3 and PDB code 2QL5(53), chains A and B, for caspase-7. The search models were prepared by removing the active site peptide ligands. The complexes were iteratively built using Coot (54) and refined in PHENIX (55). The covalent DEVE ligand was manually built into electron density. The restraint parameter file for the DEVE-cmk labeled protein was prepared with JLigand (56). The caspase-peptide linkages were modeled as a covalent bond between the caspase catalytic cysteine residue and the ketomethylene group of the DEVE-cmk peptides. Refinement parameters included an initial round of simulated annealing, rigid body refinement, and restrained refinement including TLS, NCS, and weight optimization. Ramachandran statistics were calculated using MolProbity (57). Coordinates and structure factors have been deposited in the RCSB Protein Data Bank under accession codes 5IC4 (caspase-3-DEVE) and 5IC6 (caspase-7-DEVE).

Evolutionary Analysis

All human apoptotic proteins cleaved after aspartate and glutamate were used as the input list. The identification of homologs (52 to 14,000 per protein substrate) were generated using a stand-alone protein BLAST (blastp) (58) against the non redundant “nr” database with a E-value cutoff $\leq 10^{-16}$, as described previously(46). For each set of proteins and homologs, multiple sequence alignments were generated using Clustal Omega (clustalo) with default options(59). The output alignments were run through Perl codes to isolate the experimentally observed cleavage site conservation and all other non-cleaved background amino acids of the same type using only the human sequence as reference.

Acknowledgements:

We thank Guy Salvesen and Scott Snipas at the Sanford Burnham Medical Research Institute for their helpful discussions and repeating some of kinetics with our *AFC* subtrates. We thank Emily Crawford, Zachary Hill (University of California, San Francisco), Dennis Wolan and Robyn Stanfield (The Scripps Research Institute) for helpful discussions and the staff of the ALS beamline 8.3.1 for support. This project was supported in part by Grants from National Institutes of Health R21 CA186077 (C.S.C.), R01 GM081051, R01 GM097316, and RO1 CA154802 (J.A.W.). It was also supported in part by NIH Training Grant T32 GM007175 (J.E.S.), the Global Healthy Living Foundation (J.E.S), NIH F31 CA180378 (T.J.R.), and the postdoctoral fellowships of a Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research, Government of Canada (O.J.) and the Damon Runyon Cancer Research Foundation 2082-11 (N.D.T.).

Conflict of Interest:

The authors and no conflicts of interests that affect the presentation of these results and nor do we have or advise companies whose value would directly benefit from the results of these experiments.

References:

1. Alnemri ES, Livingston DJ, Nicholson DW, Salvesen G, Thornberry NA, Wong WW, et al. Human ICE/CED-3 protease nomenclature. Cell. 1996;87(2):171.

2. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun. 1967;27(2):157-62.

3. Thornberry NA, Rano TA, Peterson EP, Rasper DM, Timkey T, Garcia-Calvo M, et al. 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.

4. Poreba M, Strozyk A, Salvesen GS, Drag M. Caspase substrates and inhibitors. Cold Spring Harb Perspect Biol. 2013;5(8):a008680.

5. Rano TA, Timkey T, Peterson EP, Rotonda J, Nicholson DW, Becker JW, et al. A combinatorial approach for determining protease specificities: application to interleukin-1beta converting enzyme (ICE). Chem Biol. 1997;4(2):149-55.

6. Pham VC, Anania VG, Phung QT, Lill JR. Complementary methods for the identification of substrates of proteolysis. Methods Enzymol. 2014;544:359-80.

7. van den Berg BH, Tholey A. Mass spectrometry-based proteomics strategies for protease cleavage site identification. Proteomics. 2012;12(4-5):516-29.

8. Staes A, Impens F, Van Damme P, Ruttens B, Goethals M, Demol H, et al. Selecting protein N-terminal peptides by combined fractional diagonal chromatography. Nat Protoc. 2011;6(8):1130-41.

9. Impens F, Colaert N, Helsens K, Ghesquiere B, Timmerman E, De Bock PJ, et al. A quantitative proteomics design for systematic identification of protease cleavage events. Mol Cell Proteomics. 2010;9(10):2327-33.

10. Drag M, Bogyo M, Ellman JA, Salvesen GS. Aminopeptidase fingerprints, an integrated approach for identification of good substrates and optimal inhibitors. J Biol Chem. 2010;285(5):3310-8.

11. Wejda M, Impens F, Takahashi N, Van Damme P, Gevaert K, Vandenabeele P. Degradomics reveals that cleavage specificity profiles of caspase-2 and effector caspases are alike. J Biol Chem. 2012;287(41):33983-95.

12. Turowec JP, Zukowski SA, Knight JD, Smalley DM, Graves LM, Johnson GL, et al. An unbiased proteomic screen reveals caspase cleavage is positively and negatively regulated by substrate phosphorylation. Mol Cell Proteomics. 2014;13(5):1184-97.

13. Dix MM, Simon GM, Cravatt BF. Global identification of caspase substrates using PROTOMAP (protein topography and migration analysis platform). Methods Mol Biol. 2014;1133:61-70.

14. Julien O, Zhuang M, Wiita AP, O'Donoghue AJ, Knudsen GM, Craik CS, et al. Quantitative MS-based enzymology of caspases reveals distinct protein substrate specificities, hierarchies, and cellular roles. Proc Natl Acad Sci U S A. 2016.

15. Stoehr G, Schaab C, Graumann J, Mann M. A SILAC-based approach identifies substrates of caspase-dependent cleavage upon TRAIL-induced apoptosis. Mol Cell Proteomics. 2013;12(5):1436-50.

16. Shimbo K, Hsu GW, Nguyen H, Mahrus S, Trinidad JC, Burlingame AL, et al. Quantitative profiling of caspase-cleaved substrates reveals different drug-induced and cell-type patterns in apoptosis. Proc Natl Acad Sci U S A. 2012;109(31):12432-7.

17. Lamkanfi M, Kanneganti TD, Van Damme P, Vanden Berghe T, Vanoverberghe I, Vandekerckhove J, et al. Targeted peptidecentric proteomics reveals caspase-7 as a substrate of the caspase-1 inflammasomes. Mol Cell P

roteomics. 2008;7(12):2350-63.
18. Crawford ED, Seaman JE, Agard N, Hsu GW, Julien O, Mahrus S, et al. The DegraBase: a database of proteolysis in healthy and apoptotic human cells. Mol Cell Proteomics. 2013;12(3):813-24.
19. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell. 2008;134(5):866-76.
20. Timmer JC, Zhu W, Pop C, Regan T, Snipas SJ, Eroshkin AM, et al. Structural and kinetic determinants of protease substrates. Nat Struct Mol Biol. 2009;16(10):1101-8.
21. Agard NJ, Mahrus S, Trinidad JC, Lynn A, Burlingame AL, Wells JA. Global kinetic analysis of proteolysis via quantitative targeted proteomics. Proc Natl Acad Sci U S A. 2012;109(6):1913-8.
22. Srinivasula SM, Hegde R, Saleh A, Datta P, Shiozaki E, Chai J, et al. A conserved XIAP-interaction motif in caspase-9 and Smac/DIABLO regulates caspase activity and apoptosis. Nature. 2001;410(6824):112-6.
23. Soares J, Lowe MM, Jarstfer MB. The catalytic subunit of human telomerase is a unique caspase-6 and caspase-7 substrate. Biochemistry. 2011;50(42):9046-55.
24. Moretti A, Weig HJ, Ott T, Seyfarth M, Holthoff HP, Grewe D, et al. Essential myosin light chain as a target for caspase-3 in failing myocardium. Proc Natl Acad Sci U S A. 2002;99(18):11860-5.
25. Krippner-Heidenreich A, Talanian RV, Sekul R, Kraft R, Thole H, Ottleben H, et al. Targeting of the transcription factor Max during apoptosis: phosphorylation-regulated cleavage by caspase-5 at an unusual glutamic acid residue in position P1. Biochem J. 2001;358(Pt 3):705-15.
26. Checinska A, Giaccone G, Rodriguez JA, Kruyt FA, Jimenez CR. Comparative proteomics analysis of caspase-9-protein complexes in untreated and cytochrome c/dATP stimulated lysates of NSCLC cells. J Proteomics. 2009;72(4):575-85.
27. Crawford ED, Seaman JE, Barber AE, 2nd, David DC, Babbitt PC, Burlingame AL, et al. Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis. Cell Death Differ. 2012;19(12):2040-8.
28. Talanian RV, Quinlan C, Trautz S, Hackett MC, Mankovich JA, Banach D, et al. Substrate specificities of caspase family proteases. J Biol Chem. 1997;272(15):9677-82.
29. Pop C, Salvesen GS. Human caspases: activation, specificity, and regulation. J Biol Chem. 2009;284(33):21777-81.
30. Thomsen ND, Koerber JT, Wells JA. Structural snapshots reveal distinct mechanisms of procaspase-3 and -7 activation. Proc Natl Acad Sci U S A. 2013;110(21):8477-82.
31. Ganesan R, Mittl PR, Jelakovic S, Grutter MG. Extended substrate recognition in caspase-3 revealed by high resolution X-ray structure analysis. J Mol Biol. 2006;359(5):1378-88.
32. Lazebnik YA, Kaufmann SH, Desnoyers S, Poirier GG, Earnshaw WC. Cleavage of poly(ADP-ribose) polymerase by a proteinase with properties like ICE. Nature. 1994;371(6495):346-7.

33. Witkowski WA, Hardy JA. L2' loop is critical for caspase-7 active site formation. Protein Sci. 2009;18(7):1459-68.

34. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics. 2004;4(6):1551-61.

35. Park K, Kuechle MK, Choe Y, Craik CS, Lawrence OT, Presland RB. Expression and characterization of constitutively active human caspase-14. Biochem Biophys Res Commun. 2006;347(4):941-8.

36. Quistad SD, Stotland A, Barott KL, Smurthwaite CA, Hilton BJ, Grasis JA, et al. Evolution of TNF-induced apoptosis reveals 550 My of functional conservation. Proc Natl Acad Sci U S A. 2014;111(26):9567-72.

37. Kumar S, van Raam BJ, Salvesen GS, Cieplak P. Caspase cleavage sites in the human proteome: CaspDB, a database of predicted substrates. PLoS One. 2014;9(10):e110539.

38. Wee LJ, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. Bioinformatics. 2007;23(23):3241-3.

39. Hawkins CJ, Yoo SJ, Peterson EP, Wang SL, Vernooy SY, Hay BA. The Drosophila caspase DRONC cleaves following glutamate or aspartate and is regulated by DIAP1, HID, and GRIM. J Biol Chem. 2000;275(35):27084-93.

40. Kumar S, Doumanis J. The fly caspases. Cell Death Differ. 2000;7(11):1039-44.

41. Snipas SJ, Drag M, Stennicke HR, Salvesen GS. Activation mechanism and substrate specificity of the Drosophila initiator caspase DRONC. Cell Death Differ. 2008;15(5):938-45.

42. Song Z, Guan B, Bergman A, Nicholson DW, Thornberry NA, Peterson EP, et al. Biochemical and genetic interactions between Drosophila caspases and the proapoptotic genes rpr, hid, and grim. Mol Cell Biol. 2000;20(8):2907-14.

43. Kurokawa M, Kornbluth S. Caspases and kinases in a death grip. Cell. 2009;138(5):838-54.

44. Dix MM, Simon GM, Wang C, Okerberg E, Patricelli MP, Cravatt BF. Functional interplay between caspase cleavage and phosphorylation sculpts the apoptotic proteome. Cell. 2012;150(2):426-40.

45. Thorsness PE, Koshland DE, Jr. Inactivation of isocitrate dehydrogenase by phosphorylation is mediated by the negative charge of the phosphate. J Biol Chem. 1987;262(22):10422-5.

46. Pearlman SM, Serber Z, Ferrell JE, Jr. A mechanism for the evolution of phosphorylation sites. Cell. 2011;147(4):934-46.

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. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol. 2009;9:51.

49. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710-5.

50. Wolan DW, Zorn JA, Gray DC, Wells JA. Small-molecule activators of a proenzyme. Science. 2009;326(5954):853-8.

51. Kabsch W. Xds. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):125-32.

52. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. J Appl Crystallogr. 2007;40(Pt 4):658-74.

53. Agniswamy J, Fang B, Weber IT. Plasticity of S2-S4 specificity pockets of executioner caspase-7 revealed by structural and kinetic analysis. FEBS J. 2007;274(18):4752-65.

54. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 4):486-501.

55. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 2):213-21.

56. Lebedev AA, Young P, Isupov MN, Moroz OV, Vagin AA, Murshudov GN. JLigand: a graphical tool for the CCP4 template-restraint library. Acta Crystallogr D Biol Crystallogr. 2012;68(Pt 4):431-40.

57. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 1):12-21.

58. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36(Web Server issue):W5-9.

59. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

Supplementary Tables

A. Mouse

Caspase	Substrate	V_{max} (uM/s)	K_M (uM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))	D/E Ratio		
						K_M (uM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))
3	D	6.48 (0.28)	126.5 (14.0)	431.67	3.41E+06	1.68	6.65	3.96
3	E	0.97 (0.03)	75.43 (6.0)	64.92	8.61E+05			
7	D	0.51 (0.01)	36.77 (3.3)	33.99	9.24E+05	0.16	0.30	1.85
7	E	1.71 (0.12)	228.2 (34.2)	114.00	5.00E+05			

B. Commercial

Caspase	Substrate	V_{max} (uM/s)	K_M (uM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))	D/E Ratio		
						K_M (uM)	k_{cat} (1/s)	k_{cat}/K_M (1/(Ms))
3	D	1.37 (0.02)	36.15 (2.2)	91.27	2.52E+06	1.78	8.21	4.61
3	E	0.17 (0.004)	20.3 (1.7)	11.12	5.48E+05			
7	D	1.25 (0.06)	50.34 (7.6)	83.40	1.66E+06	0.38	0.96	2.52
7	E	1.30 (0.08)	131.7 (19.4)	86.60	6.58E+05			

Supplementary Table 1. Summary of kinetic values for cleavage of P1 aspartate and glutamate by mouse and commercial caspases-3 and -7. The values for V_{max} *and* K_M with standard errors are reported from triplicates measurements on three biological replicates.

Supplementary Table 2

Data collection and refinement statistics of caspase-DEVE complexes

Data Collection	**Caspase-3-DEVE**	**Caspase-7-DEVE**
Beamline	ALS 8.3.1	ALS 8.3.1
Wavelength (Å)	1.116	1.116
Space group	F222	$P2_12_12_1$
Unit cell dimensions		
a, *b*, *c* (Å)	133.8, 177.8, 193.4	66.1, 88.5, 103.4
α, *β*, *γ* (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution	50 – 2.65 (2.78 – 2.65)[a]	50 – 2.70 (2.83 – 2.70)
Observations	140,636	125,195
Unique reflections	33,509 (4,401)	17,178 (2,211)
R_{merge} (%)[b]	17.4 (83.5)	16.9 (89.0)
R_{pim} (%)[b]	9.7 (46.3)	6.7 (34.7)
I/sigma	8.5 (2.0)	12.4 (2.7)
Completeness (%)	99.8 (99.3)	99.6 (98.2)
Multiplicity	4.2 (4.2)	7.3 (7.4)
Refinement statistics		
Resolution (Å)	48.4 – 2.65	47.2 – 2.70
Reflections (total)	33,447	17,139
Reflections (test)	1,696	876
R_{cryst} (%)[c]	20.7	19.7
R_{free} (%)[d]	23.9	24.1
Protein atoms	7,467	3,682
Ligand atoms	136	68
Waters	53	0
Average *B*-value (Å^2)		
Protein	41	40
Ligand	47	49
Wilson	41	38
RMSD from ideal geometry		
Bond length (Å)	0.002	0.005
Bond angles (°)	0.696	0.814
Ramachandran statistics (%)[e]		
Favored	98.8	98.5
Outliers	0	0
PDB code	5IC4	5IC6

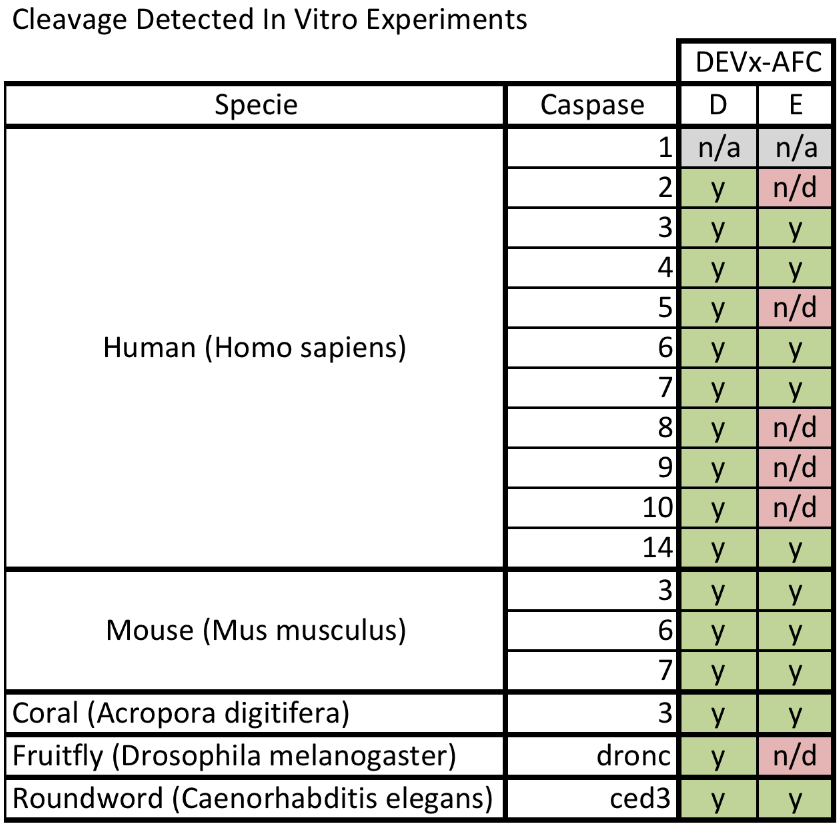
[a] Numbers in parenthesis refer to the highest resolution shell.
[b] $R_{\text{merge}} = \Sigma_{hkl} \Sigma_i \mid I_{hkl,i} - <I_{hkl}> \mid / \Sigma_{hkl} \Sigma_i I_{hkl,I}$ and $R_{\text{pim}} = \Sigma_{hkl}[1/(N\text{-}1)]^{1/2}\Sigma_i \left|I_{hkl,i} - <I_{hkl}>\right| /\Sigma_{hkl}\Sigma_i I_{hkl,i}$, where $I_{hkl,i}$ is the scaled intensity of the i$^{\text{th}}$ measurement of reflection *h*, *k*, *l*, $< I_{hkl}>$ is the average intensity for that reflection, and *N* is the redundancy.
[c] $R_{\text{cryst}} = \Sigma \mid F_o - F_c \mid / \Sigma \mid F_o \mid$ x 100, where F_o and F_c are the observed and calculated structure factors, respectively.
[d] R_{free} was calculated as for R_{cryst}, but on a random test set comprising 5% of the data excluded from refinement.
[e] Calculated using MolProbity.

Supplementary Table 2. Data collection and refinement statistics of caspase-DEVE complexes.

Supplementary Table 3

For Supplementary Table 3, see
http://www.nature.com/search?journal=cdd&q=J%20E%20Seaman

Supplementary Table 3 – List of annotated P1 phosphorylation sites.

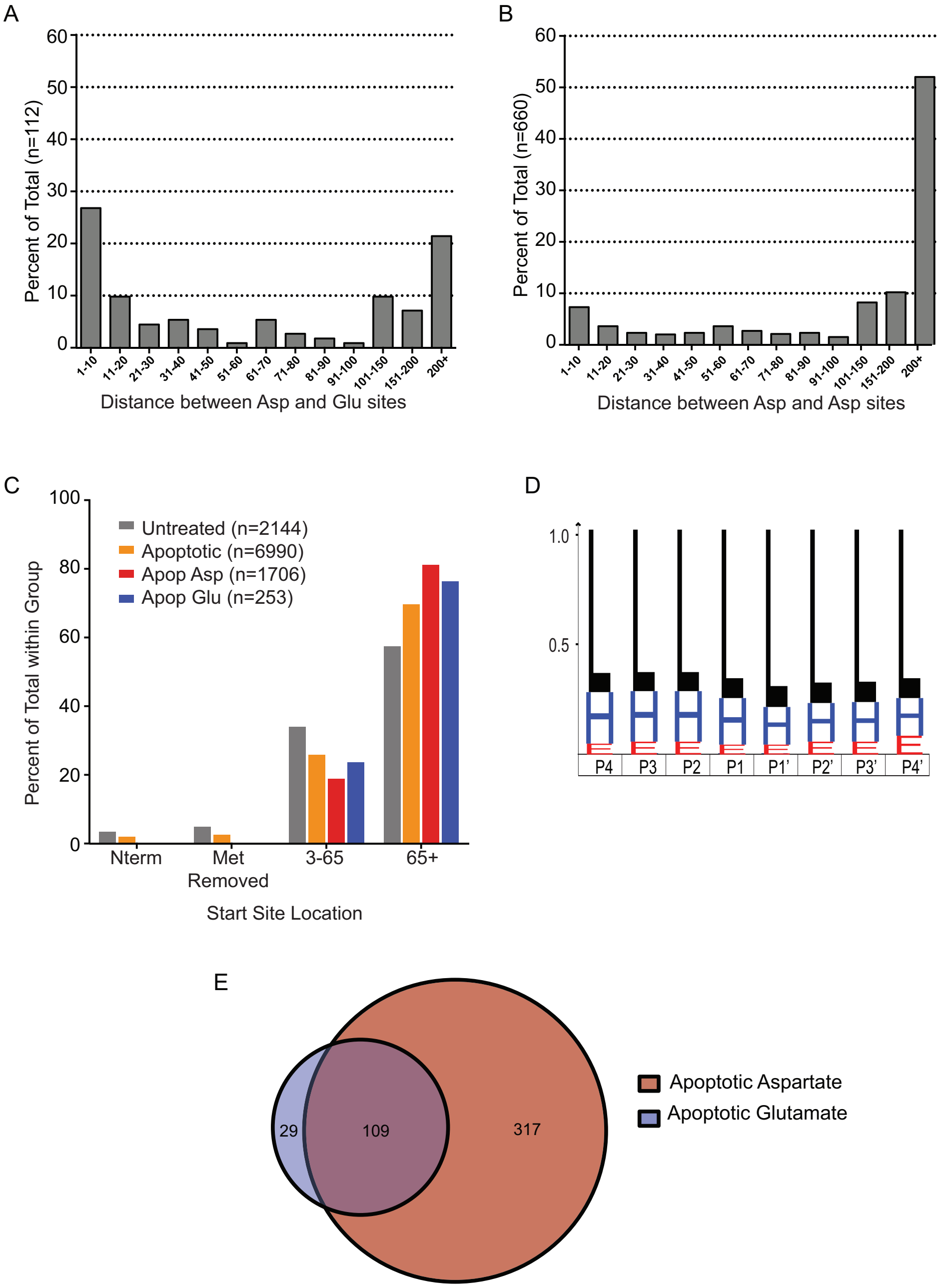


Cleavage Detected In Vitro Experiments

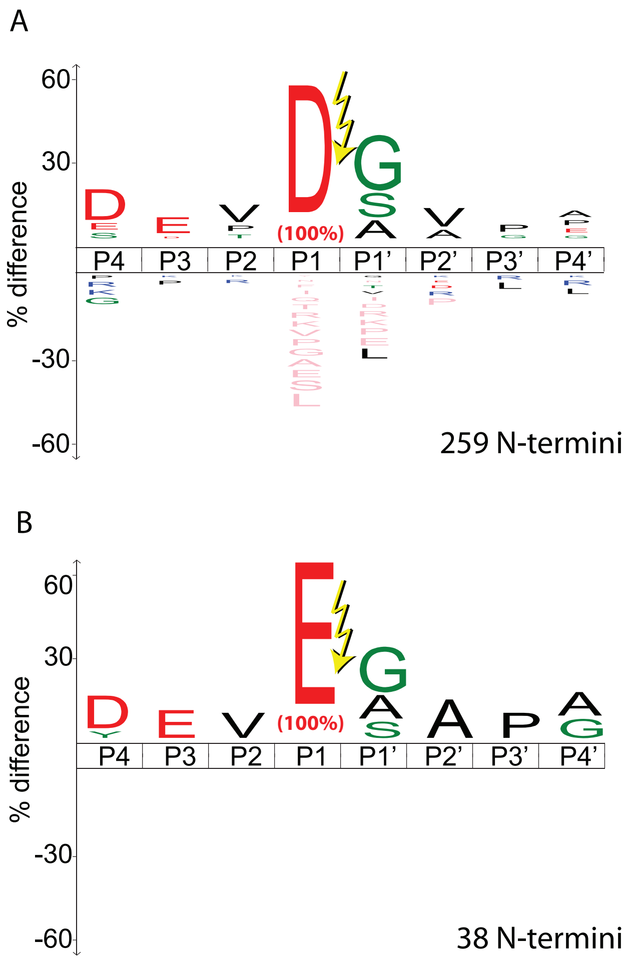
		DEVx-AFC	
Specie	Caspase	D	E
Human (Homo sapiens)	1	n/a	n/a
	2	y	n/d
	3	y	y
	4	y	y
	5	y	n/d
	6	y	y
	7	y	y
	8	y	n/d
	9	y	n/d
	10	y	n/d
	14	y	y
Mouse (Mus musculus)	3	y	y
	6	y	y
	7	y	y
Coral (Acropora digitifera)	3	y	y
Fruitfly (Drosophila melanogaster)	dronc	y	n/d
Roundword (Caenorhabditis elegans)	ced3	y	y

Supplementary Table 4. *In vitro* cleavage activity for glutamic acid was tested for human, mouse, coral, fruitfly and worm caspases. If activity was found, it is marked with a "y." No activity detected is marked as "n/d." Human caspase 1 was not tested, "n/a."

Supplementary Figures

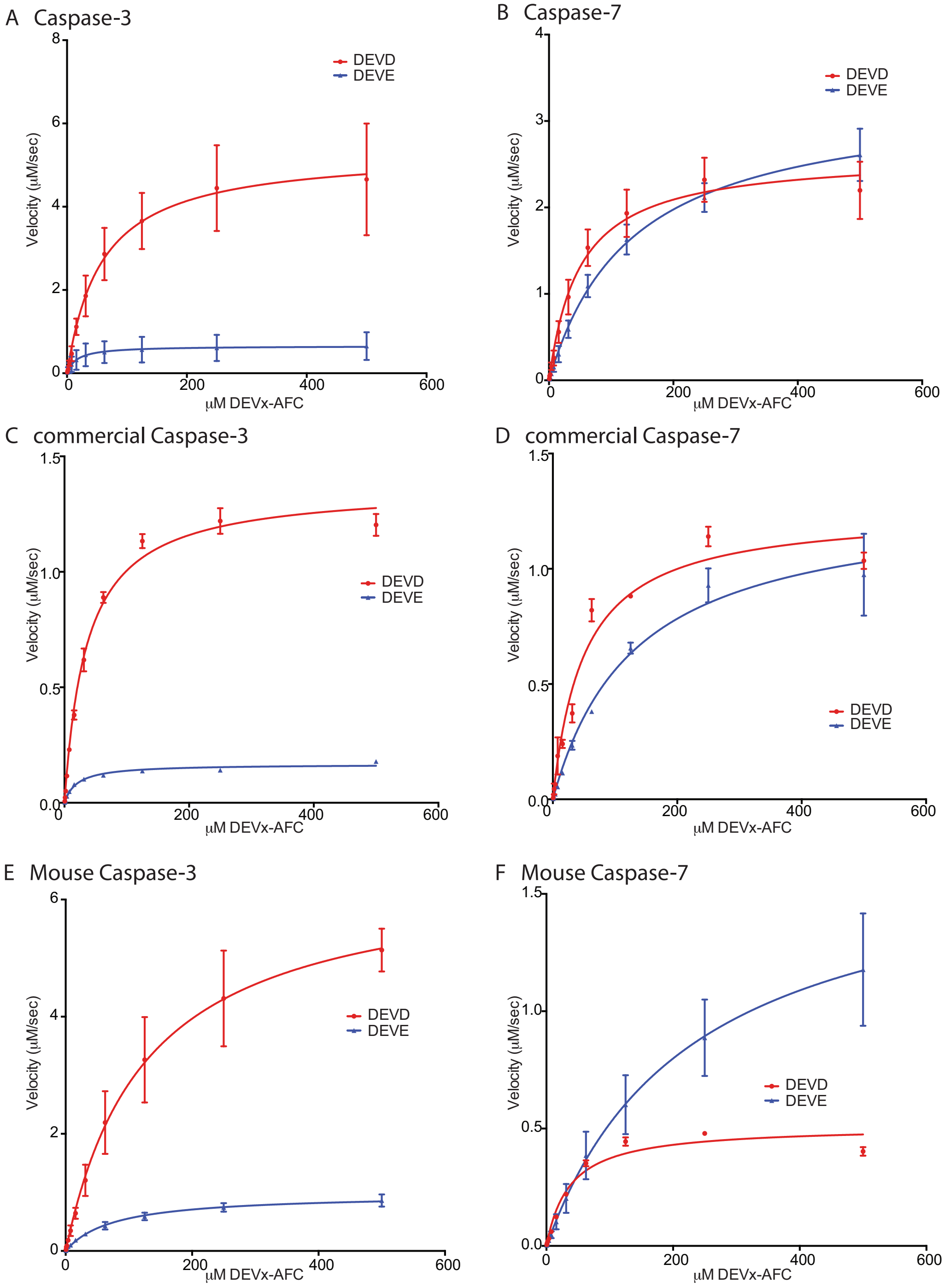


Supplementary Figure 1. (a-b) Distance between cut sites within a protein for Glu/Asp (a) and Asp/Asp (b). **(c)** Site of cleavage in proteins is very similar for apoptotic Asp and Glu**. (d)** Filled logo of predicted structure for P4-P4’ of apopotic Glu sites (L=loop, H=alpha helix, E= Beta sheet). (**e**) The overlap of significant GO terms for proteins cleaved after Asp and Glu during apoptosis.

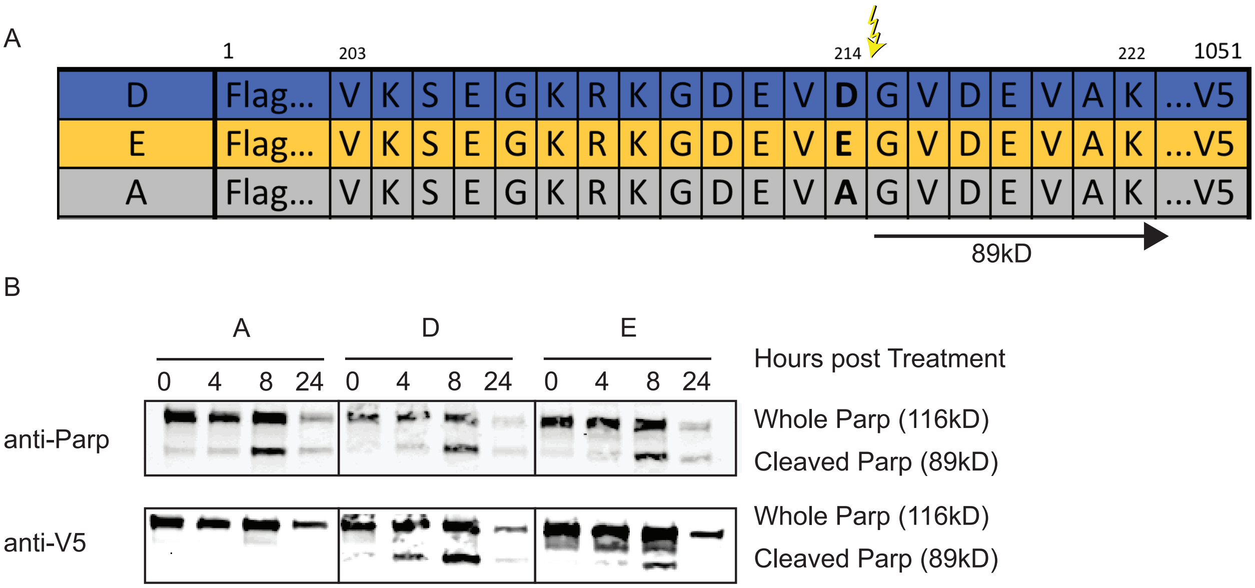


Supplementary Figure 2. iceLogos of apoptotic mouse P1 aspartate and P1 glutamate cut sites.

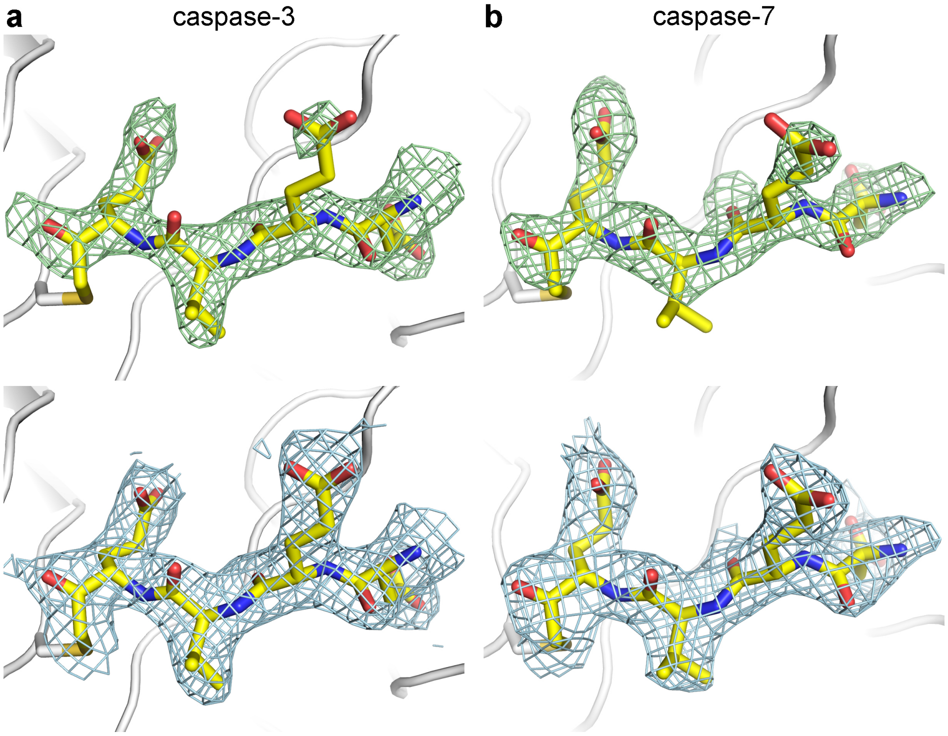
Apoptotic mouse P1 aspartate iceLogo reproduced with permission from Crawford, et al (27).



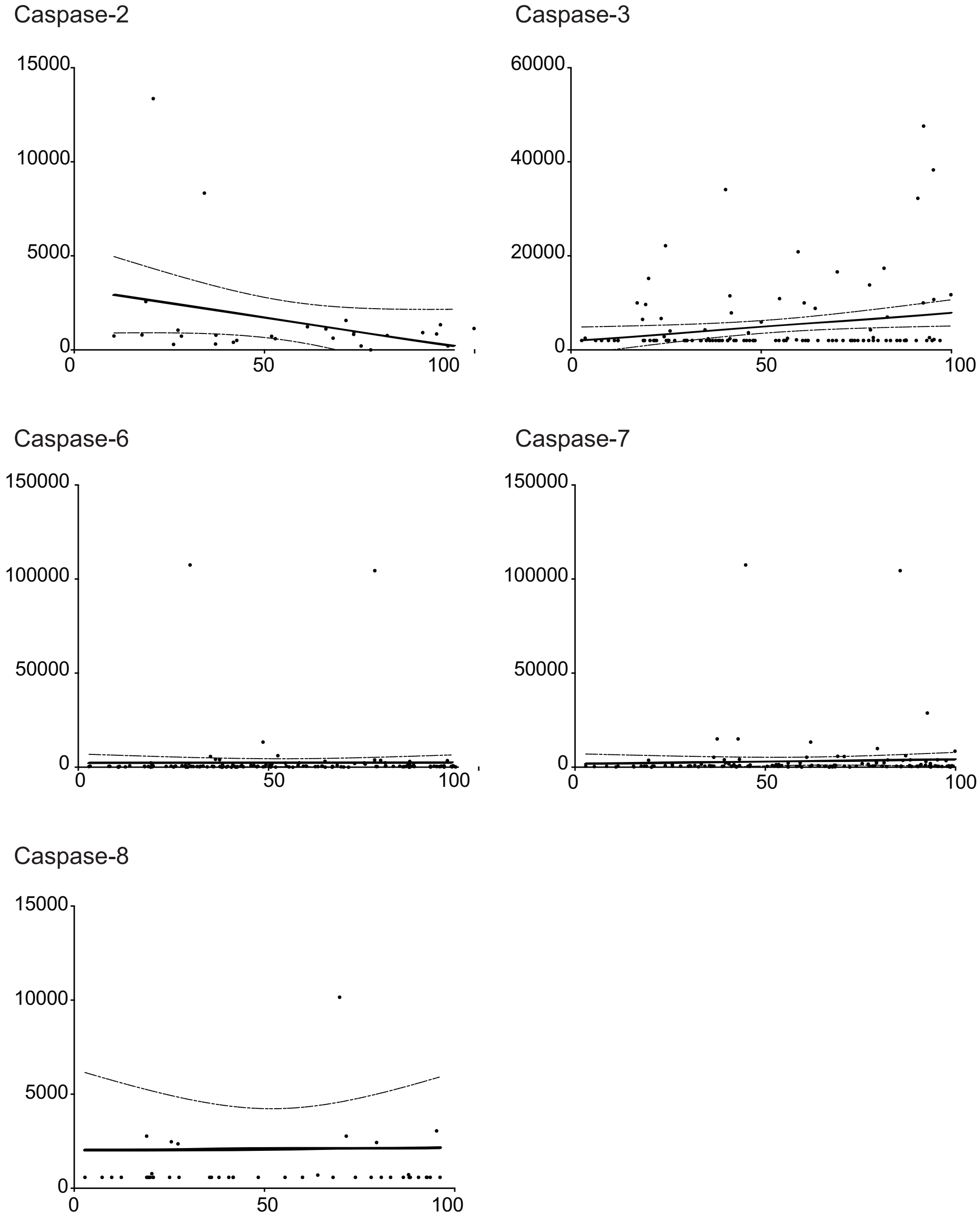
Supplementary Figure 3. Michaelis-Menten plots for caspases-3 and -7. Human (**a-b**), commercial human (**c-d**) and mouse (**e-f**) varieties' of caspase-3 and caspase-7 cleavage of Ac-DEVD-afc and Ac-DEVE-afc show similar activity differences for P1 aspartate and P1 glutamate substrates.


A
1
203
214
222
1051
D Flag... V K S E G K R K G D E V D G V D E V A K ...V5
E Flag... V K S E G K R K G D E V E G V D E V A K ...V5
A Flag... V K S E G K R K G D E V A G V D E V A K ...V5
89kD
B
A
D
E
0 4 8 24 0 4 8 24 0 4 8 24
Hours post Treatment
anti-Parp
Whole Parp (116kD)
Cleaved Parp (89kD)
anti-V5
Whole Parp (116kD)
Cleaved Parp (89kD)

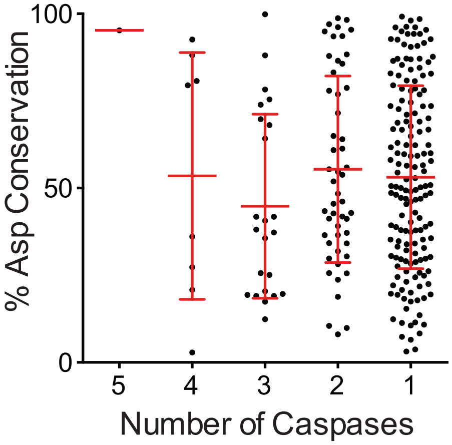
Supplementary Figure 4. Cellular apoptotic cleavage of overexpressed V5-tagged PARP at Asp214, Asp214Glu and Asp214Ala in HEK293T cells treated with 500nM staurosporine**. (a)** Cells were overexpressed with PARP protein tagged C-terminally with V5 without knockdown of endogenous levels. The endogenous sequence contains a caspase cleavage at Asp214 ("D") and tested cleavages after Asp214Glu ("E") or Asp214Ala ("A"). **(b)** Apoptosis was induced with 500nM treatment of staurosporine and lystes were collected at the indicated times. Cellular lysates were blotted for Parp (endogenous and mutant) and V5 (mutant only). Whole Parp appears around 116kD while the apoptotic fragment from cleavage after residue 214 is around 89kD.



Supplementary Figure 5 Electron density of the DEVE ligands for (**a**) caspase-3 and (**b**) caspase-7. The simulated annealing omit $F_o - F_c$ densities are colored green and contoured at 3σ. The $2F_o - F_c$ electron density maps are colored blue and contoured at 1σ. The catalytic cysteine side chain is shown as sticks.



Supplementary Figure 6. Conservation rate of aspartate at caspase cleavage sites compared to kinetic rate of cleavages for substrates of caspase-2, -3, -6, -7 and -8 with regression line (bold line) and 95% confidence interval (curved line). The known substrate rates for individual caspase are not well correlated to evolutionary conservation of aspartate at the human cleavage site positions. However, there is a slight positive trend for caspase-3 and -7 for higher rates having higher conservation. There is limit of detection for rates within some of the datasets.



Supplementary Figure 7. Conservation rate of aspartate at caspase cleavage sites compared to number of identified caspases that cleave the substrate. Proteins are grouped by how many caspases (-2, -3, -6, -7, or -8) are able to cleave it. There is no trend of Asp conservation at human site of cleavage and the number of caspases that can cleave the protein. The single protein cut by all 5 caspases is DGCR8 (Uniprot Q8WYQ5) and has 93% conservation of the aspartate.

Chapter 5

Initial results and development of tools to identify apoptotic as biomarkers of chemotherapeutic efficacy

This work was done in collaboration with Charlie Morgan.

Introduction

The 2010 US cancer prevalence was estimated at almost 14 million invasive cancer sites, or almost 5% of the total US population, and predicted to grow to over 18 million cancer sites by 2020. The total cost for cancer care was estimated at $124.57 billion dollars in 2010 and predicted to almost double in the next 10 years (1). Recent research has helped improve treatment for the growing patient populations through greatly increasing our knowledge of cancer biology in classification of subtypes, genetic susceptibility and appropriate treatment regimens, and there are now over 300 therapeutics commonly prescribed during treatment (2). However, the efficacy of drug treatments may not be known until weeks into a drug course or even months after treatment has finished. An ineffective drug course postpones possibly more productive treatment, thereby allowing the disease to worsen or develop resistance, and exposing the patient to unnecessary side effects, tests and costs. Current assays of disease response are often expensive, unique to the disease, may be invasive, or not very accurate (3, 4). However, there are many biochemical changes that occur at the tumor site within hours or even minutes after drug exposure that can reveal a response much sooner than traditional monitoring methods. Multiple studies have detected chemotherapeutic-induced release of cytochrome c, nucleosomes, tumor DNA, caspase-cleaved cytokeratin 18, and apoptotic cellular debris in the blood (5-8). Although many of these markers are stable, some detectable up to weeks in serum, there is no successful biomarker panel model yet for treatment-induced apoptotic response.

A common mode of action of chemotherapy is to induce apoptosis, a form of programmed cell death. Apoptosis is a universal molecular mechanism in the metazoan world and the pathway structure is highly conserved throughout phylogeny (9). Apoptosis occurs during normal healthy homeostasis and plays a crucial function in organisms' development, maturation and diseases (10). Caspases, a family of cysteine proteases with aspartic acid specificity, are responsible for initiating the hallmarks of apoptosis: nuclear condensation, DNA laddering, and membrane blebbing. The apoptotic pathway can be initiated from external or internal endogenous signals, and drugs target multiple levels upstream and directly in the apoptotic pathway. Caspases-3 and -7 are the apoptotic effector caspases and are responsible for cleaving a large swath of cellular proteins that ultimately create an efficient pathway to destroy the cell. Multiple studies, including many performed in the Wells lab, have revealed 2200 caspase cleavage sites, though only 784 have been labeled in any one experiment(11-16). The Wells lab has also discovered 257 mouse, 130 fly and 50 worm caspase cleavage sites(17). Although not as comprehensive, comparisons of these non-human sites to the human dataset reveal a strong evolutionary conservation at the pathway, protein, and motif site between homologous proteins. Through these experiments, we have also discovered that caspase substrates make up only a quarter of the total proteolysis identified in apoptotic cells, revealing almost 5000 non-caspase apoptotic sites. Within these sites, only 10 out of the 2900 non-caspase, non-tryptic sites have physiological relevant annotations in MEROPS. There is a lot of unexamined apoptotic proteolytic activity, both caspase and other, and the downstream pathway ultimately leading to cell death is not clear.

As many chemotherapeutics induce apoptosis, efforts have been made to find biomarkers in patient blood that are released from apoptotic tumors. Plasma-based biomarkers are appealing as collection is minimally invasive, as it may be part of routine blood draws, and tumor markers in samples appear highly stable and resistant to degradation (18). The plasma proteome background has been analyzed in-depth as analytical techniques improve, revealing the known background proteome and N-terminome for plasma biomarkers (19, 20). Apoptotic proteolytic peptides, especially caspase cleavage products, have a direct biological connection to apoptosis making them ideal targets for a biomarker study. Recent work has demonstrated the potential of using caspase cleaved cytokeratin 18 (ccCK18) as a biomarker for treatment response in gastrointestinal cancer and survivor outcome for breast cancer (21, 22). However, ccCK18's correlation to treatment and disease outcome appears more promising as diagnostic for liver diseases than cancer biomarkers (23). The ccCK18 research does demonstrate that caspase cleavage events can be easily detectable days after treatment in patient blood samples and may have a correlation with successful treatments, supporting the proposed project to discover apoptotic peptides biomarkers (5). In fact, preliminary experiments in two post-chemotherapeutic patient plasma samples label ~100 apoptotic related potential biomarker peptides, many of them highly conserved and with physiologically relevant kinetic rates.

A major step to achieve biomarker identification and confidence measurements will be creating a positive control system. There are many hurdles for adapting apoptotic labeling from cell cultures into organisms. First and foremost is the increased complexity of cellular types within the sample and those being exposed and reacting to the chemotherapeutic. This leads to signals within the sampling from normal, healthy apoptotic signals and off-target chemotherapeutic effects. Additionally, the timescale and dosing methodology may not scale up from cell culture to a mouse or human patient easily. Therefore, this study intends create an engineered system to induce cell death independent of chemical inducers. While several strategies for a "death switch" have been previously reported, we sought to harness our vast experience in protein and cellular engineering and combine it with our novel N-terminomics workflow for discovering a comprehensive inclusion list of apoptotic biomarkers. Previous inducible death switches predominately focus on caspase activation, relying on either dimerization of initiator caspase zymogens, stabilizing of a degron-caspase, or inducible caspase expression (24-26).

Financial and ethical hurdles make the acquisition of human samples difficult, preventing large-scale biomarker studies in human patients. Mouse models are used to increase sample size, reduce confounding variables, and discover potential biomarkers (27-29). As apoptosis is highly conserved between mouse and human, potential apoptotic biomarkers and sample methodology from a murine system may be applicable for human patients. This study intends to develop a "death switch" positive control system with the goal to discover and validate murine apoptotic related peptides, to classify them using conservation, kinetics and cell-/drug- induction pattern to elucidate more of the apoptotic pathway, and to create an accurate biomarker panel to predict treatment response **(Figure 1)**.

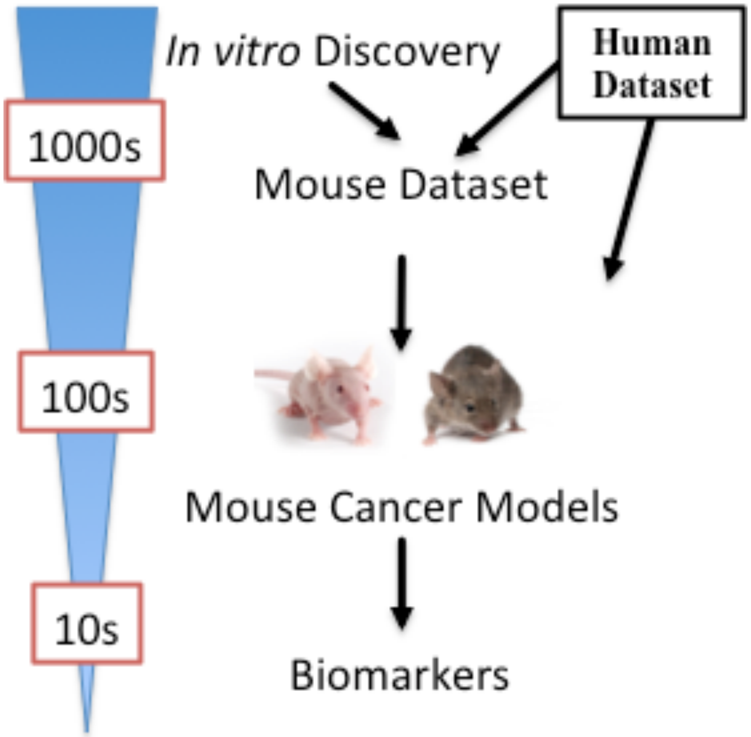


Figure 1. An overview schematic of identifying biomarkers. The *in vitro* discovery datasets, combining mouse and human apoptotic cellular samples, contains thousands of peptides. These will be used to narrow down and identify apoptotic-derived peptides from *in vivo* mouse cancer model studies. The peptides of interest will be furthered narrowed down to the top individual peptides that correlate qualitatively (what peptides are they) and quantitatively (how much of it is there) to treatment progression.

Results

Initial mouse studies reveal numerous potential apoptotic biomarkers but also highlight study design setbacks.

The mouse experimental design utilizes human T cell lymphocyte (Jurkat) or multiple myeloma (MM1S) cell lines as xenografts on a mouse background (Figure 2). As the model will ultimately be used for biomarker identification and detection, the distinction in species between human and mouse-based peptides will allow for greater confidence in determining a peptide's origin from the tumor (human) or off-target and background (mouse). The mice will be treated with clinically relevant, apoptosis-inducing therapies of bortezomib or its next generation carfilzomib, with or without lenalidomide.

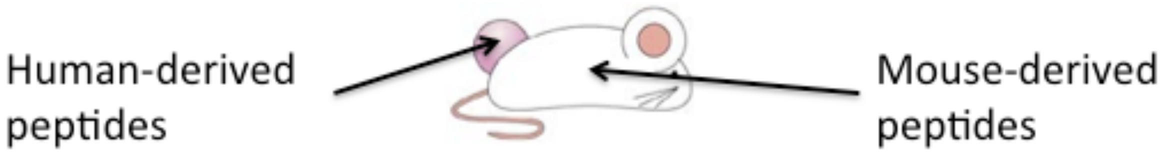


Figure 2. The mouse models will have human tumor cell xenografts. Therefore, peptides originating from the tumor itself will consist of primarily human-derived sequences, while peptides from the background tissues will be mouse-derived sequences.

Study I: RPMI 8826 hindleg xenograft on Balb/c SCID female mice**.**

Design: As an initial experiment, 13 mice previously implanted for a different study were used to test the dosing, bleeding and labeling protocols. Of the 13 mice, 5 mice had tumors that failed to grow and were used as no-tumor, drug-treated control group (Group 1). The remaining tumor-bearing 8 mice were split with 5 mice into the drug-treated test group (Group 2) and 3 mice into the vehicle-treated control group (Group 3). All mice were given 0.2 mg/kg bortezomib or equivalent volume of saline i.p. twice a week. Before treatment started and the days after treatment, a saphenous blood draw took up to 200uL of whole blood. After 4 weeks, all mice were euthanized with a terminal bleed, and tumor, kidney and liver were extracted and saved.

Results: Overall there were little phenotypic differences in response between drug and vehicle treatment groups. There is no significant difference in tumor size between Group 1 and 2 (Figure 3a). However, there was a slight difference in body weight between the three groups. The drug treated groups both maintained initial body weight, while the vehicle group began to lose weight, an indication of poor health (Figure 3b). However, overall, there was a lot of individual variation in tumor volume and body weight for Groups 1 and 2.

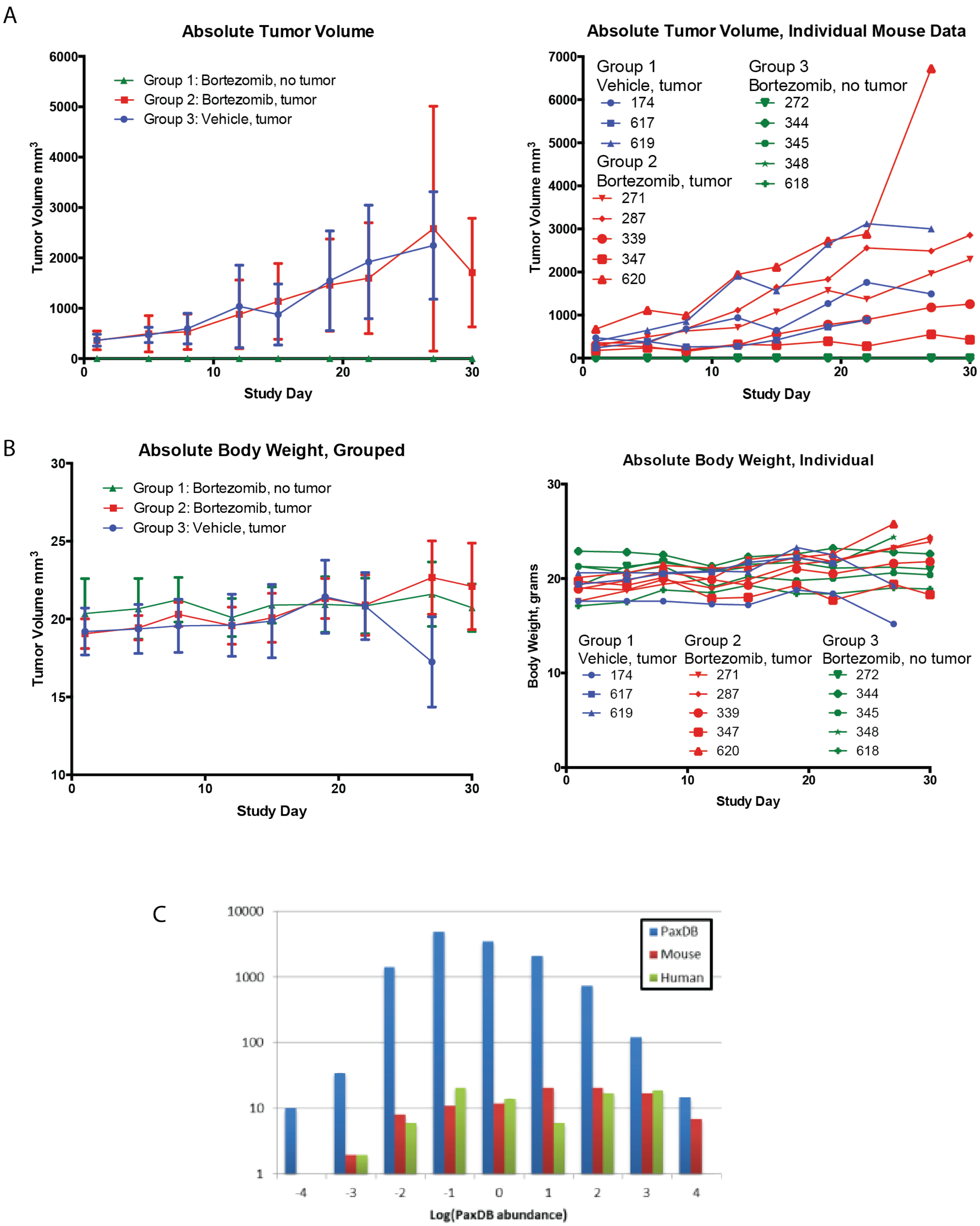


Figure 3: Results from Study I for tumor growth (A) and body weight (B). The peptides from Study I mice cover a wide range of abundance values for both human and mice (C).

The plasma samples were N-terminally labeled and identified by mass spectrometry. Due to the low volume of some individual samples, some plasma samples were pooled together from the same mouse. The pre-treatment samples were pooled for Groups 1 and 2 and within Group 3. All mass spectrometry results were searched for both human and mice peptide sequences. For all plasma peptides, there were 220 unique mouse-based N-termini and 102 unique human-based N-termini, and both species covered a similar abundance profile (Figure 3c). Every sample identified two- to ten-fold as many mouse peptides compared to human peptides, which is expected in the mouse plasma background. Additionally, 62% of the mouse peptides identified are peptides and proteins commonly found in healthy serum. Although there was limited tumor regression, 25% of the human peptides were caspase-derived peptides, defined as cleavage after aspartate or glutamate, compared to 10% of the mouse peptides.

From this sample, there are multiple peptides of interest as potential biomarkers. These were defined as peptides from Group 1, not found in Group 2 or 3 and not from healthy plasma proteins for both human and mouse proteins (Table 1). Of these, there are some of particular interest. Mouse Abcc12 protein is a plasma membrane transporter with no known substrates. However it has been shown to be upregulated in bortezomib treatment (Mol Pharmacol. 2012 Dec;82(6):1008-21.). Human FAIM3 protein is an anti-apoptotic protein that protects cells. The protein was found with a endorproteolytic cleavage which may be inhibitory to its protective function. Human LY6H protein is a lymphoblastic specific protein that indicates the Jurkat cell type. These three peptides indicate that there are drug (Abcc12), apoptotic (FAIM3) and cell specific (LY6H) potential biomarkers within this study.

Entry	MOUSE Protein names	Gene
P02089	Hemoglobin subunit beta-2	Hbb-b2
P07759	Serine protease inhibitor A3K	Serpina3k
P09036	Serine protease inhibitor Kazal-type 3	Spink3
P20065	Thymosin beta-4	Tmsb4x
P31695	Neurogenic locus notch homolog protein 4	Notch4
P42703	Leukemia inhibitory factor receptor	Lifr
Q61114	BPI fold-containing family B member 1	Bpifb1
Q61838	Alpha-2-macroglobulin	A2m
Q62523	Zyxin	Zyx
Q68SA9	A disintegrin and metalloproteinase 7	Adamts7
Q80WJ6	Multidrug resistance-associated protein 9	Abcc12
Q8BYK4	Retinol dehydrogenase 12	Rdh12
Q8K3Z9	Nuclear envelope pore membrane protein POM 12	Pom121
Q9D289	Trafficking protein particle complex subunit	Trappc6b
Q9JLZ8	Single Ig IL-1-related receptor	Sigirr
Q9Z1R3	Apolipoprotein M	Apom

Entry	HUMAN Protein names	Gene
O60667	Fas apoptotic inhibitory molecule 3	FAIM3
O94772	Lymphocyte antigen 6H	LY6H
P21589	5'-nucleotidase	NT5E
P29322	Ephrin type-A receptor 8	EPHA8
P56159	GDNF family receptor alpha-1	GFRA1
Q14765	Signal transducer and activator of transcript	STAT4
Q86YW0	1-phosphatidylinositol 4,5-bisphosphate phosp	PLCZ1
Q8N3V7	Synaptopodin	SYNPO
Q8N6Y0	Usher syndrome type-1C protein-binding protein	USHBP1
Q8NEZ3	WD repeat-containing protein 19	WDR19
Q8WYP3	Ras and Rab interactor 2	RIN2
Q96KR4	Leishmanolysin-like peptidase	LMLN
Q96MY7	Protein FAM161B	FAM161B
Q9HBA0	Transient receptor potential cation channel	TRPV4
Q9HCG1	Zinc finger protein 160	ZNF160
Q9UGQ3	Solute carrier family 2, facilitated glucose	SLC2A6

Table 1: Proteins with peptides that were enriched in the treated mice not found in healthy plasma for Study I.

There were a few issues identified with this study. First, the drug treatment did not show any tumor regression, like due to too low of a dose and too late initial dosage. The next study should have quicker treatment after tumor growth onset and adjusted dosages. There is also a big issue with individual variation in response and identified peptides. This may be helped with more mice and better sampling. There are a small but not insignificant number of peptides that cannot be confidently identified as human or mouse. From all the labeled peptides, 32 were unable to be distinguished based on the mass spectrometry sequence fingerprinting. Indeed, the average protein sequence identify for these 32 proteins is 92%, and many of the protein regions have identical sequences. Some proteins may be distinguished based on function or cell type, or most sensitive mass spectrometry for slight differences, but in general, these will require extra confidence before using them as hits of interest. Finally, there is limited sample volume and total identified peptides in the current design due to limited blood sampling before the mice become anemic. Many of these issues may be resolved with greater number of mice and analysis with the highest sensitivity mass spectrometry available.

Study II: RPMI 8826 hindleg xenograft on Balb/c SCID female mice**.**

Design: As an followup experiment, 11 mice previously implanted for a different study were used to test the dosing, bleeding and labeling protocols. The tumor-bearing were split with 3 mice into the vehicle-treated control group (Group 1) and 8 mice into the drug-treated test group (Group 2). All mice were given 1 mg/kg bortezomib or equivalent volume of saline i.p. every 72 hours. Before treatment started and the days after treatment, a saphenous blood draw took up to 200uL of whole blood. After 1 week, all remaining mice were euthanized with a terminal bleed, and tumor, kidney and liver were extracted and saved.

Results: There was a significant difference in response, with the treated group showing minimal tumor growth or tumor recession (Figure 4a). However, the more aggressive drug treatment led to much greater body weight loss for the treated group (Figure 4b). There was a lot less individual variation in response for the treatment group than during Study I. However, the control group was still varied in both tumor group and body weight over the study period.

The plasma samples were N-terminally labeled and identified by mass spectrometry. The treatment group samples were pooled by response rate based on tumor volume and compared to the pooled vehicle treatment group. For all N-termini, there were 133 unique mouse-based N-termini and 100 human-based N-termini, and both species covered similar abundance profiles (Figure 4c). Every sample identified around 30 human peptides with 10-20% caspase-derived and around 45 mouse peptides with 5-10% caspase-derived. This is a lower total number of peptides and total mouse peptides than the previous study. Indeed, there were a lot fewer healthy plasma proteins in this study, with only 11% of the mouse proteins. The human and mouse peptides contained only 22 mass spectrometry fingerprinting that could not distinguish between the species. These peptides had an average sequence identity of 91%.

There are 11 human and 5 mouse peptides that are enriched in the moderate to high responding mouse serum and not found in normal plasma (Table 2). Of interest are multiple caspase-derived peptides that may indicate the treatment-induced apoptosis. One peptide from PI3 kinase subunit alpha is identical for human and mouse. However, unlike Study I, there are no drug or cell-specific peptides that are potential biomarkers. Additionally, there was no overlap in the potential biomarker lists between the two studies.

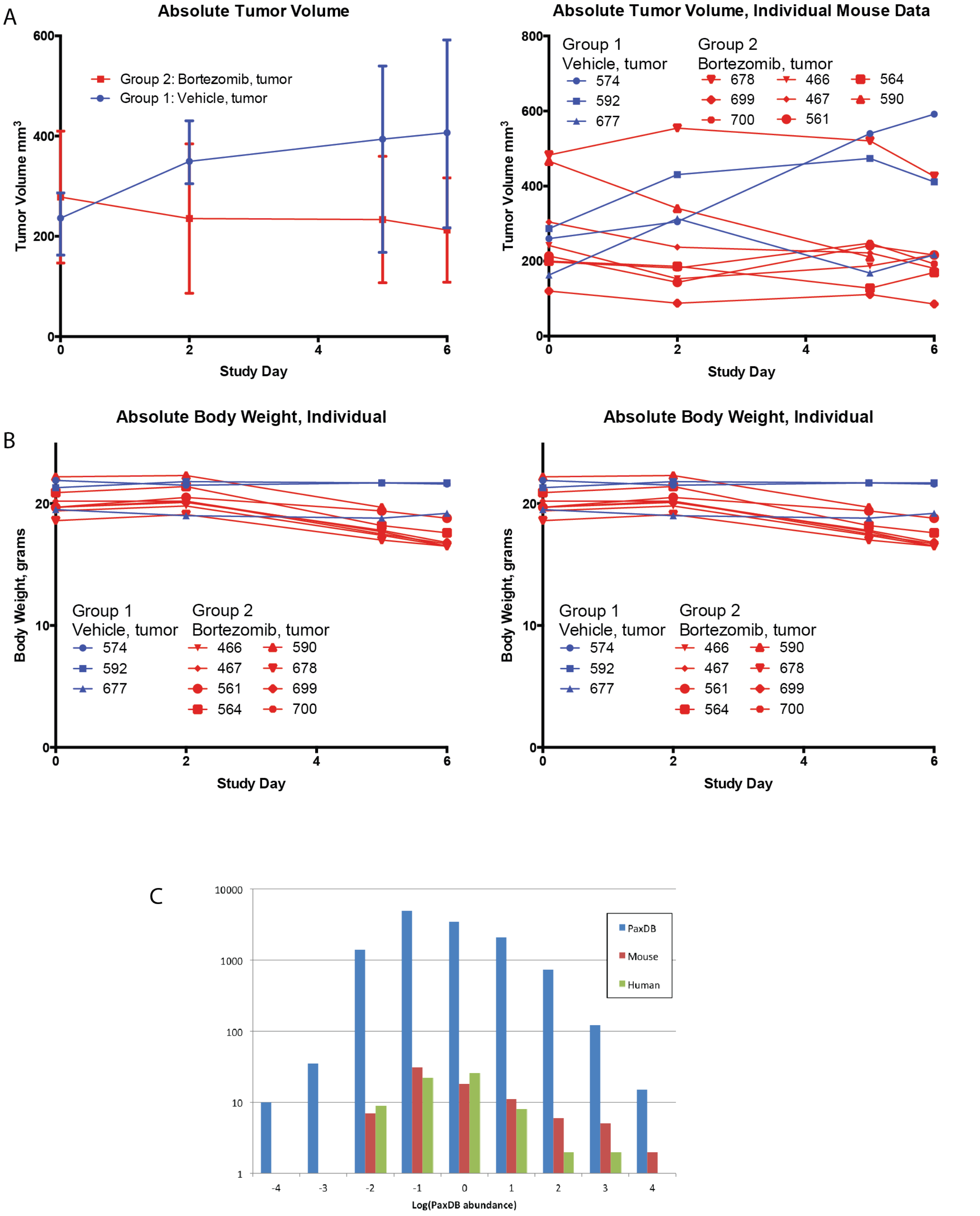


Figure 4: Results from Study II for tumor growth (A) and body weight (B). The peptides from Study I mice cover a wide range of abundance values for both human and mice (C).

HUMAN

Entry	Name	Length	Caspase?	Tumor?	P1'
O14525	Astrotactin-1	1302			576
P02679	Fibrinogen gamma chain	453			316
P42336	PI3-kinase subunit alpha*	1068	yes	yes	892
P51854	Transketolase 2	596			81
Q13488	V-ATPase 116 kDa isoform a3	830	yes		357
Q5JVG2	Zinc finger protein 484	852	yes	yes	101
Q5VTL7	Fibronectin type III domain-containing protein 7	733		yes	440
Q5VTL7	Fibronectin type III domain-containing protein 7	733	yes	yes	442
Q6UWU4	Bombesin receptor-activated protein C6orf89 (Amfion)	347	yes		33
Q9UPS6	Histone-lysine N-methyltransferase SETD1B	1923			146
Q9Y2Q0	Probable phospholipid-transporting ATPase IA	1164			692

MOUSE

Entry	Name	Length	Caspase?	Tumor?	P1'
P39087	Glutamate receptor 6	908		yes	113
P42337	PI3-kinase subunit alpha*	1068	yes	yes	892
Q61704	Inter-alpha-trypsin inhibitor heavy chain H3	889			34
Q8C3B8	Protein RFT1 homolog	541	yes		157
Q9ET38	Claudin-19	224			93

Table 2: Peptides enriched in treated mice and not found in normal plasma. The highlighted peptide cannot distinguish by sequence between the human and mouse protein.

Study III: RPMI 8826 hindleg xenograft on Balb/c SCID female mice**.**

Design: To test further drug response, a small third study was started. Fourteen tumor-bearing mice were split into four treatment groups: vehicle-control (3 mice, Group 1), carfilzomib i.v. 3 mg/kg twice a week (3 mice, Group 2), bortezomib i.p.1 mg/kg every 72 hours (4 mice, Group 3), and bortezomib i.p. 1 mg/kg every 72 hours plus lenalidomide p.o. 15 mg/kg daily (4 mice, Group 4). Before treatment started and the days after treatment, a saphenous blood draw took up to 200uL of whole blood. After 3 weeks, all remaining mice were euthanized with a terminal bleed, and tumor, kidney and liver were extracted and saved.

Results: There was no difference in tumor volume for any for the treatment groups (Figure 5). However, one mouse (505) in Group 3 showed significant tumor regression. Samples from this study except mouse 505 were not further analyzed.

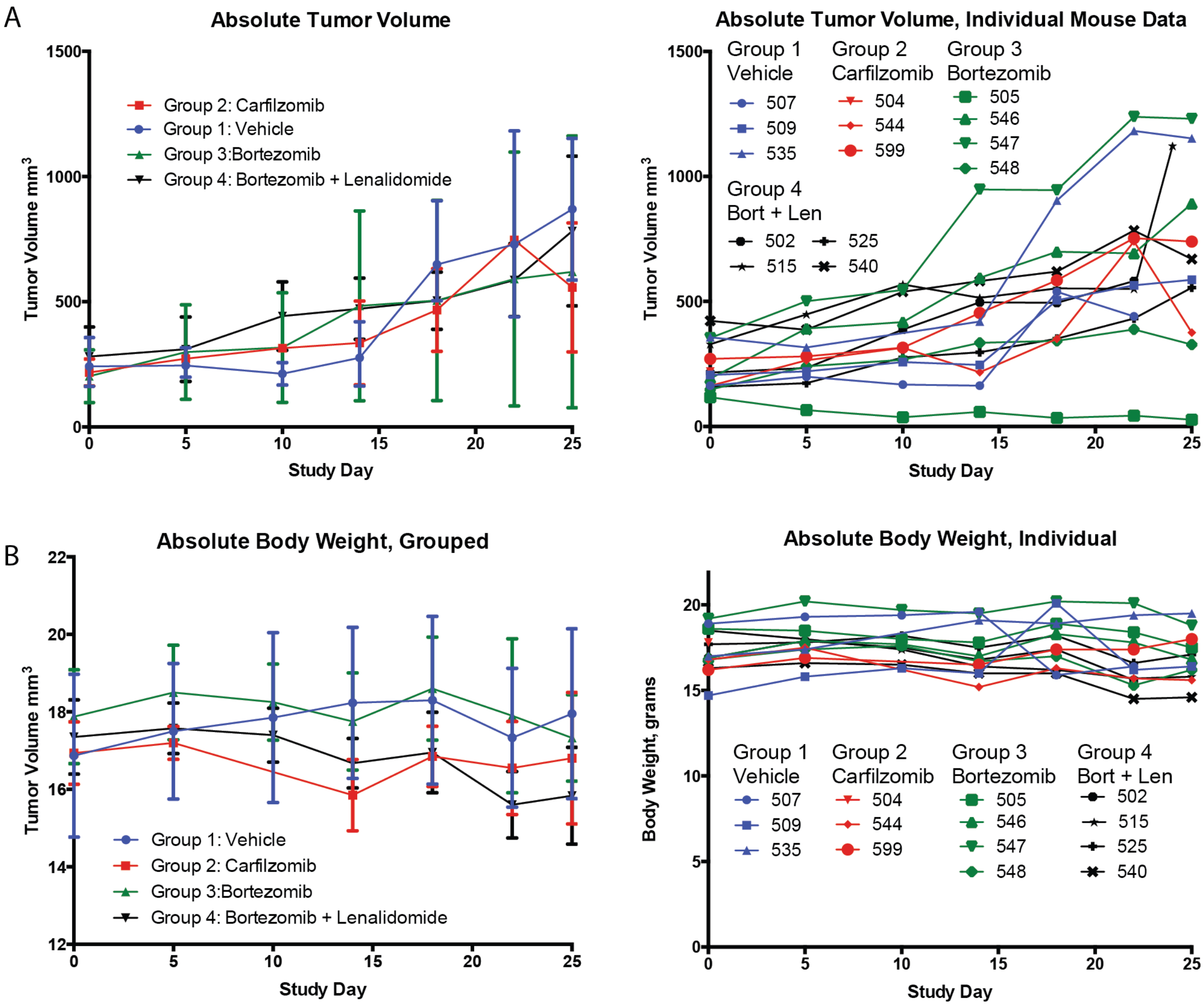


Figure 5: Results from Study III for tumor volume (A) and body weight (B).

Study IV: MM1S disseminated xenograft on Balb/c SCID female mice.

Design: To test a disseminated mouse model that demonstrates an orthotopic disease model, 9 mice were injected with MM1S cells and allowed to develop a tumor burden (30). Because the tumor is no longer visible, the MM1S cells contain luciferase for imaging. The mice were split into two groups: 1 mg/kg bortezomib (4 mice, Group 1) or equivalent volume of saline i.p. every 72 hours (5 mice, Group 2). Before treatment started and the days after treatment, a saphenous blood draw took up to 200uL of whole blood. After 2 weeks, all remaining mice were euthanized with a terminal bleed, and tumor, kidney and liver were extracted and saved. Mice were imaged on the first and last day of the study.

Results: There were significant differences in tumor growth between vehicle and treated mice (Figure 6). However, all mice had tumor progression during the study period. N-terminal labeling of plasma samples was done on individual samples pooled over the treatment period. There were 79 unique mouse and 55 unique human N-termini identified. Of those, 20 peptides could not differentiate between human and mouse due to similar sequences, with those proteins having an average of 92% sequence identity. There was a much greater overlap of peptides that had previously been seen in the apoptotic peptides in the DegraBase at the site and protein levels (Table 3). Additionally, there were two human and two mouse peptides that were in 3/4 of the treated mice, not found in healthy plasma and not present in the pre-treatment plasma (Table 4).

The study design is more relevant disease model for the biomarker study. The injected human cells reside within the bone marrow and blood compartment, replicating the multiple myeloma disease burden. Additionally, this removes the barrier between the tumor and blood from the previous hindleg xenografts studies, potentially allowing more peptides into the plasma. This model is also relatively quick to establish and responds as expected to drug treatments.

However, using this model requires the luciferase reporter engineered into the cell line. There is also significantly more animal handling. Another downside is the minimal potential for a tumor collection, and any organ or tumor samples will contain a mixture of human and mouse cells, complicating analysis. However, overall, this orthotopic design has the potential for better signal and more relevant biomarkers than the previous studies.

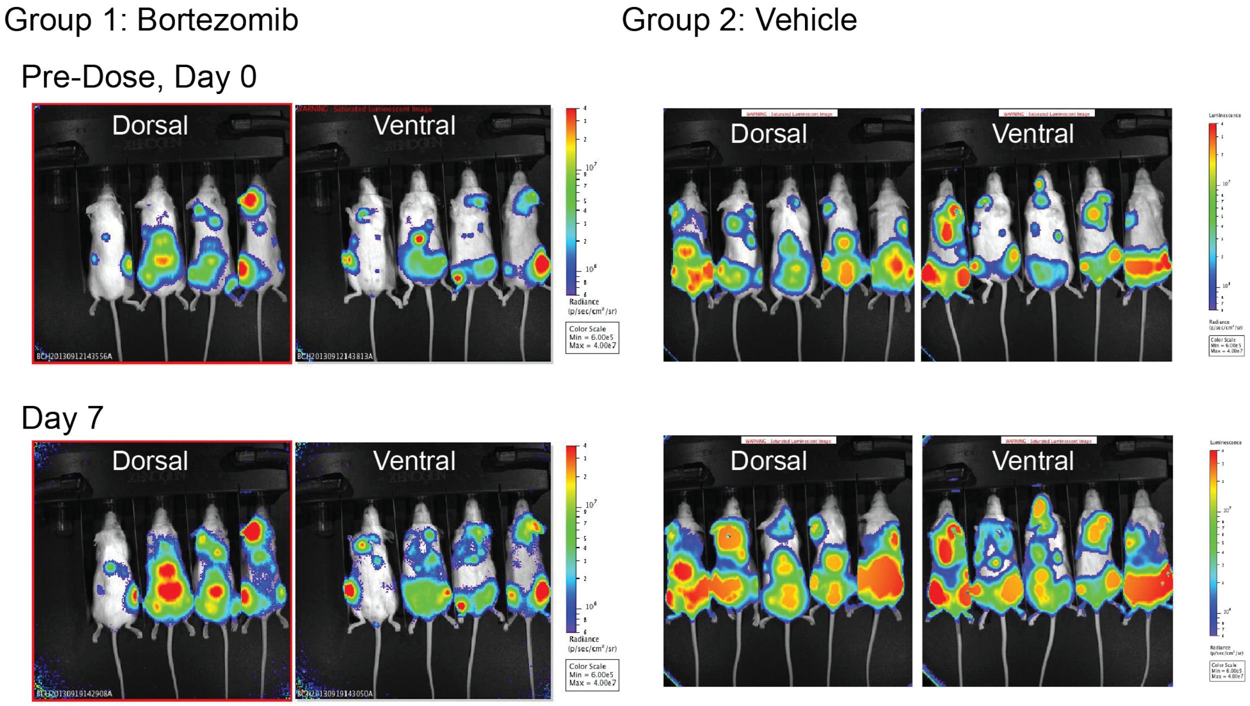


Figure 6: Luminescent imaging of mice in Study IV to compare bortezomib tumor response to vehicle treatment.

HUMAN

Entry	Name	Length	P1'
A4D1F6	Leucine-rich repeat & death domain-containing protein 1	860	850
Q9H611	ATP-dependent DNA helicase PIF1	641	73

MOUSE

Entry	Name	Length	P1'
Q00896	Alpha-1-antitrypsin 1-3	412	374
A2APY7	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex assembly factor 5	343	43

Table 3: Peptides enriched in treated mice and not found in normal plasma.

HUMAN

Entry	Name	Length	DegraBase	Caspase?	P1'
P42704	Leucine-rich PPR motif-containing protein, mitochondrial	1394	site		60
P42704	Leucine-rich PPR motif-containing protein, mitochondrial*	1394	protein		767
P54819	Adenylate kinase 2, mitochondrial	239	site		4
P62937	Peptidyl-prolyl cis-trans isomerase A	165	site		2
P62942	Peptidyl-prolyl cis-trans isomerase FKBP1A*	108	site		2
Q07021	Glycoprotein gC1qBP, Mitochondrial matrix protein p32	282	site		74
Q13177	Serine/threonine-protein kinase PAK 2 *	524	protein		321
Q14676	Mediator of DNA damage checkpoint protein 1	2089	protein		2003
Q14676	Mediator of DNA damage checkpoint protein 1	2089	protein		281
Q16602	Calcitonin gene-related peptide type 1 receptor	461	site		326
Q5JVG2	Zinc finger protein 484	852	site	yes	101

MOUSE

Entry	Name	Length	DegraBase	Caspase?	P1'
A2ASS6	Titin	35213	protein		13354
P01027	Complement C3	1663	protein		75
P26883	Peptidyl-prolyl cis-trans isomerase FKBP1A*	108	site		2
P29699	Alpha-2-HS-glycoprotein	345	protein		22
Q6PB66	Leucine-rich PPR motif-containing protein, mitochondrial*	1392	protein		766
Q8CIN4	Serine/threonine-protein kinase PAK 2*	524	protein		321

Table 4: Peptides found in Study IV that overlap peptides found in the human and mouse DegraBase. The highlighted peptides cannot distinguish by sequence between the human and mouse protein.

Analysis of tissue samples reveals apoptotic evidence and minimal organ-derived peptides in plasma.

The tumor, liver and kidney samples from mouse 505 (Study III, Group 3, good response), mouse 590 (Study II, Group 2, good response), and tumors from mouse 466 (Study II, Group 2, initial response) and mouse 546 (Study II, Group 2, initial response) were analyzed. The organ samples were split with 25% for western blot analysis and 75% for N-terminal labeling for mass spectrometry. The treated tumors show evidence of high levels of caspase-3 by western blot (Figure 7a). The mice with highest response (505 and 590) have more relative protein and more active cleaved forms. This adds evidence that the tumor regression in these mice is through drug-induced apoptosis. Additionally, there are minimal levels of whole or active caspase-3 in the kidney and liver tissues, indicating there were no significant off-target apoptosis inducing drug activity or side effects.

Analysis of the kidneys and livers by mass spectrometry reveal many tissue specific peptides. There is similar overlap between the individual organs within a mouse as there is between similar organs between mice (Figure 7b). Indeed almost a majority of peptides found in any sample were found in all the other samples. While there was great overlap within these samples, there was minimal overlap with proteins found in healthy plasma or peptides identified in earlier studies. From the tumor samples, most of the peptides did not overlap any healthy plasma proteins, but there were a minority of peptides that were found in previous treated plasma samples. There were 4 human and 2 mouse peptides of interest from Study II that were also found in mouse 590's tumor (Table 2). Therefore, peptides enriched in treated and responding mice found in treated plasma samples are likely tumor-derived, and there are minimal peptides from other organs.

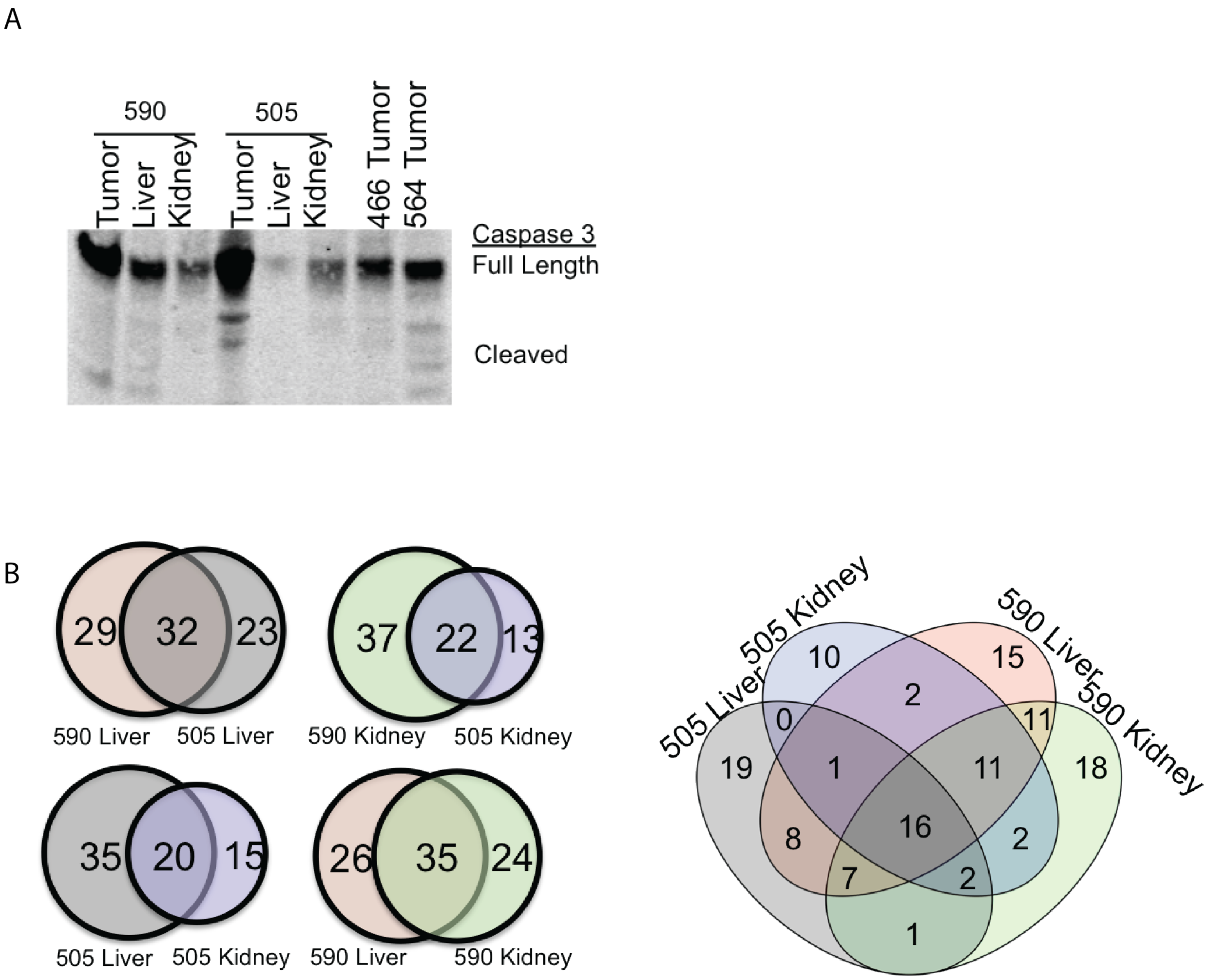


Figure 7: The tumor and liver samples from 4 mice were tested for caspase activity (A) and peptide fragments (B). The tumors that had the best response (505 and 590) had markedly more caspase-3 expression and cleaved activated forms. Additionally, there was much less caspase-3 in the non-tumor tissues. The non-tumor tissues had many unique peptides, but they almost all overlapped with the other samples, both within mouse (liver and kidney) or within organ (liver and liver, kidney and kidney).

A Death Switch Positive Control for treatment-independent apoptosis. A large issue with the previous mouse studies has been the limited sampling leading to small total number of peptides identified and potentially high noise levels. There are also many peptides that cannot be directly related to apoptosis within the peptide. For the current studies, bortezomib and carfilzomib's mechanism of action target the tumor cells and therefore apoptosis should be constrained to those cells. However, off-target and disease-related side effects may induce apoptosis in other cell types. To combat these issues, we will create a positive control of apoptosis: a "death switch." These engineered cells will undergo apoptosis independent of drug-treatment, allowing for controlled plasma sampling and revealing all potential identifiable biomarkers.

The death switch will continue four main elements: (1) cell selective and small molecule inducer of apoptosis (2) rapid implementation through genetically engineering (3) bio-orthogonal to mammalian cells and (4) dose-responsive in xenograft mouse models in order to investigate the threshold of caspase activity required for tumor cell apoptosis and downstream detection limits from processing and instrumentation. We chose to modify the SNIPer system already developed in the lab that consists of a modified human caspase-3 (31). During apoptosis, caspase-3 is activated through cleavage in its intersubunit linker. The death switch consists of caspase-3 modified to include a TEV-site in its intersubunit linker and a co-expressed TEV protease. These proteins are under the control of the small molecule tetracycline, where addition of tetracycline will begin transcription and translation of the proteins and thus inducing apoptosis (Figure 8a).

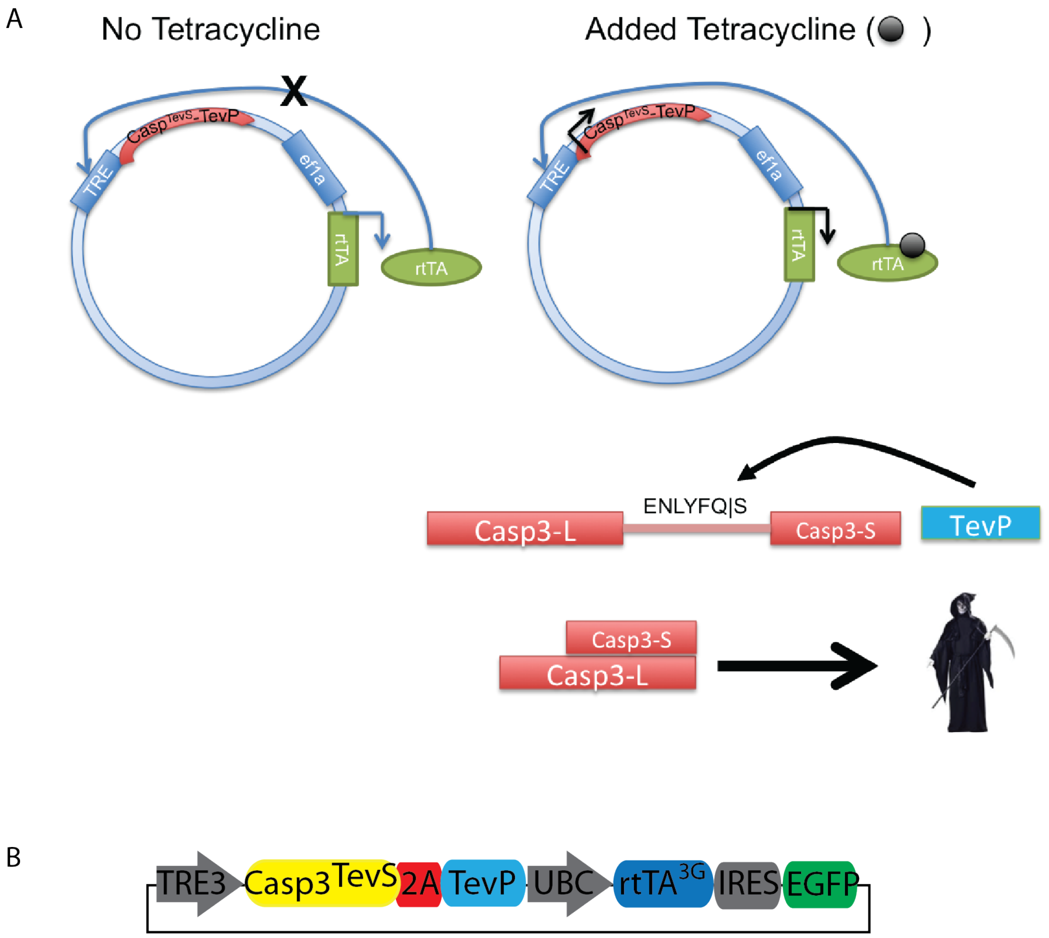


Figure 8: The schematic for the Death Switch. (A) The Death Switch is in the off state without any added tetracycline as there is no binding of rtTA to the Tet-Responsive Element (TRE). Upon addition of tetracycline, the rtTA can bind to the TRE and induce transcription of the Death Swtich Casp$^{\text{TevS}}$ and TevP. The TevP then cleaves the engineered TEV cleavable site in the intersubunit linker, activating caspase-3 and inducing apoptosis within the cell. (B) The original design for the Death Switch lentiviral vector.

Initial Death Switch Results were mixes. Early efforts to create a death switch had mixed results. The initial vector system incorporated the TEV-site caspase-3 (Casp3^{TevS}) with the TEV protease (TevP) and a GFP fluorescent marker or the TEV protease only control (Figure 8b). This system utilized a mixed backbone of second generation and third generation tet-responsive elements to minimize background levels and what was available at the time.

The vector was initially tested in HEK293 and HeLa cells. Cells were tested after two rounds of sorting on GFP for final populations of >90% positive cells. The cells were tested with a range of tetracycline to interrogate the dose responsiveness of the system. Both HEK293 and HeLa cells showed increased caspase activity and cellular death after tetracycline treatment (Figure 9). The response does appear to be concentration dependent, especially for total caspase activity. However, initiation of the caspase activity is significantly delayed and prolonged compared to drug-induced apoptosis. The peak caspase activity for staurosporine apoptosis occurs within 4-8 hours after treatment, whereas the tetracycline activity peaks 8-24 hours after treatment.

With these positive results, we next wanted to try the vector in MM1S and RPMI-8826 cell lines. Lentiviral infection of these cell lines was significantly more difficult. No protocol created any viable MM1S cell lines with these vectors. The RPMI-8826 cells were able to be sorted to 85% GFP-positive cells. However, testing these cell lines revealed inconclusive or negative results (Figure 10). Specifically, tetracycline treatment no longer induced apoptosis, though staurosporine treatment still could. Interestingly, one version of the death switch cells appeared resistant to the staurosporine treatment. Interrogating this results reveals likely alterations at the genomic DNA insertion, as the caspase DNA sequence was missing and only TEV protease myc-markers are visible after induction.

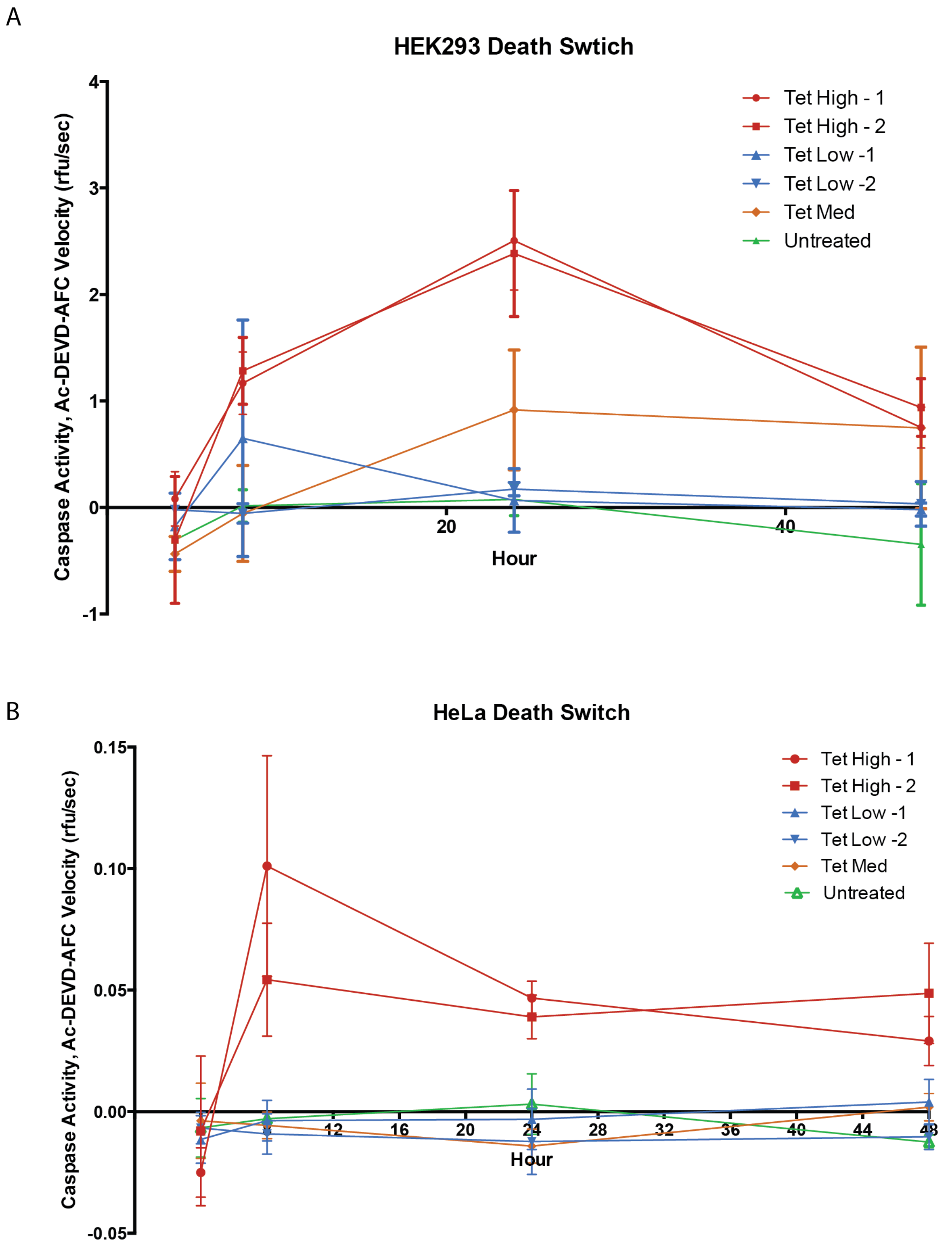


Figure 9: Testing the Death Switch in HEK293 (A) and HeLa (B) cells. Caspase activity is measured by fluorescence in cellular lysate through addition of Ac-DEVD-AFC, and the initial velocities are plotted.

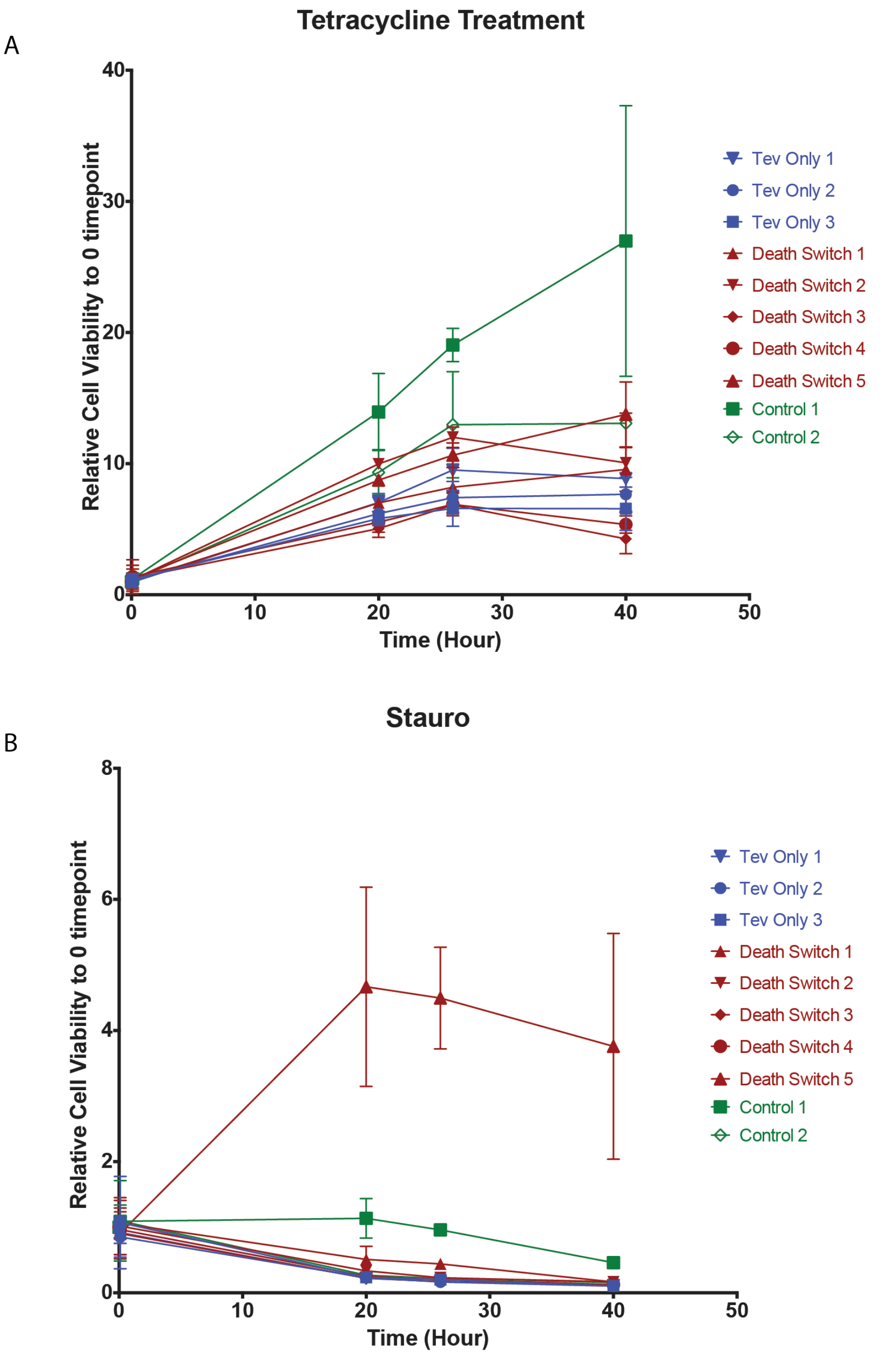


Figure 10: Relative cell viability for multiple replicates of the Death Switch, TevP-only Switch control or an empty vector control. The expected results for tetracycline treatment would be to have death in the Death Switch cells, but there is no difference between them and control cells (A). However, the positive control, staurosporine, can still induce cell death, except for one isolated Death Switch cell line (B).

Current and Future Death Switch

Because of the lack of positive results, we decided to create an optimized version of the death switch vector (Figure 11). This vector is utilizing Clontech's third generation Tet-One system as its backbone. In comparison, the previous death switch backbone was primarily second-generation and did not initially contain all of the elements. The new vector will also contain luciferase along with a mCherry fluorescent marker to allow for single vector protocols to create cells ready for xenograft into mice. This final vector will also be 1kb shorter and therefore have higher viral packaging efficiency. The design and cloning is currently ongoing.

Conclusion and Future Work

The identification of treatment-induced proteolytic biomarker fragments in plasma is still very promising. Initial mouse studies revealed numerous potential peptide fragments that were enriched in treated and responding mice. Additionally, the orthotopic disease model appears to have better relevance to clinical work, both in its tumor burden and response and its comparison to known apoptotic fragments from the DegraBase. Further work is required to optimize dosage and plasma sampling. Additionally, work is promising for a positive control for biomarker identification. The treatment-independent apoptotic tumor samples will show tumor regression and allow for maximal biomarker signal release. These work intend to identify peptide fragments that can detect drug response and used for clinical treatment decisions. Future work with these technologies can expand to distinguish apoptotic peptide fragments from specific cell types compared to background in other disease models.

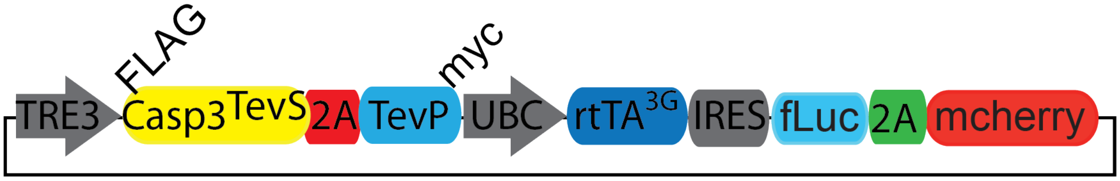


Figure 11: The second generation Death Switch vector schematic, incorporating Flag and myc tags on the Death switch and switch GFP for fLuc and mCherry.

Materials and Methods

Mouse models and sampling. All mouse models were used as excess subjects from ongoing studies using MM1S or RPMI-8826 cell lines with the Preclinical Therapeutics Core and Blake Aftab. Specifically, Female C.B17 SCID mice from Taconic were implanted with 1 x 10^6 cells through tail-vein injection or hind-leg xenograft. Xenografts were allowed to grow until evidence of tumor is visible, either through physical measurements of tumor size or luminescent reporting, generally 6 weeks. Once tumors had reached appropriate size, the drug regimen was started. Mice received either bortezomib or carfilzomib as described in protocols. Blood samples were taken from the saphenous vein up to three times a week. Whole blood was kept on ice, spun for 10 minutes at 2000 x *g* and plasma was snap frozen and stored at -80C. Upon euthanasia at the end of the study or due to ethical considerations, solid organ samples and tumor were collected and snap frozen.

N-terminal Labeling. Plasma samples were thawed on ice and prepared for labeling as previously described (32). Solid samples are crushed while snap frozen using a tissue homogenizer before following the common labeling protocol previously described (33). In brief, protease inhibitors are added to plasma samples before a 1 hour subtiligase labeling reaction. Samples are then prepped for mass spectrometry with desalting by Nap 25 columns, cysteine reduction and acylation and capture on-bead for trypsinization. The tryptic peptides are released from bead and collected for analysis.

Mass Spectrometry. All mass spectrometry data were collected in the UCSF Mass Spectrometry Facility and searched using the Protein Prospector using the most recent SwissProt database. Due to cellular sources containing both human and mouse proteins, searches were run using species chosen as “human,” “mouse,” and combined to identify potential missed peptides and those whose source is unknown due to common peptide sequences.

Transfections, Lentiviral and Cellular Assays.

Production: 293FT cells were transfected with a cocktail of 2nd generation lentiviral packaging plasmids at approximately 80% confluences. Lipofectamine 2000 (Invitrogen) was used for lipid-based transfection of the plasmids using 4 μg DNA and 2.5ul lipofectamine per well of a six well plate. Transfected DNA for viral packaging was mixed at 1:1:1 (VSV-G:pCMVdelta8.91:Lenti) ratio. Media was changed to complete DMEM 10%FBS after six hours of incubation with the lipid:DNA mixture. The supernatant was harvested and cleared by passing it through a 0.2 μm syringe at 48h and 72h post transfection. The cleared supernatant was kept at 4 °C in-between collection points. The supernatant was concentrated using Clontech LentiX Concentrator according to the manufactures supplied protocol.

Infection: Freshly adhered target HeLa cells (~50% confluent), were then incubated with the viral supernatant + 8 μg/mL polybrene for a minimum of 6 hrs before the media was changed to fresh complete DMEM. Suspension cells were spinfected with concentrated virus for 3 hours at 700 x *g*. Cells were expanded for a minimum of 48 hours before selection by fluorescent activated cell sorting.

Cellular Assays: All cells were kept in tet-free media until assayed. Tetracycline was added at concentrations of 0.1 to 100 ug/mL. Successful induction was measured through cell death assay with CellTiterGlo and western blot with multiple timepoints over 72 hours. Caspase activity was measured in cellular lysate with Ac-DEVD-AFC caspase-3 substrate.

Acknowledgements

The initial tet-responsive lentiviral vector were gifts from Martin Kappmann and the Weissman lab. The RPMI-8826 cells were donated by Diego Alcosta-Alvear and the MM1S cells were donated by Blake Aftab.

References

1. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst. 2011;103(2):117-28.
2. Cancer Drug Information: National Institute of Health; 2012 [Available from: http://www.cancer.gov/cancertopics/druginfo/alphalist.
3. NCCN Guidelines Version 3.2012 Non-Hodgkin's Lymphomas: National Comprehensive Cancer Network; 2012 [Available from: http://www.nccn.org/professionals/physician_gls/pdf/nhl.pdf.
4. NCCN Guidelines Version 3.2012 Breast Cancer: National Comprehensive Cancer Network; 2012 [Available from: http://www.nccn.org/professionals/physician_gls/pdf/breast.pdf.
5. Beachy SH, Repasky EA. Using extracellular biomarkers for monitoring efficacy of therapeutics in cancer patients: an update. Cancer Immunol Immunother. 2008;57(6):759-75.
6. Chalut KJ, Ostrander JH, Giacomelli MG, Wax A. Light scattering measurements of subcellular structure provide noninvasive early detection of chemotherapy-induced apoptosis. Cancer Res. 2009;69(3):1199-204.
7. Larson CJ, Moreno JG, Pienta KJ, Gross S, Repollet M, O'Hara S M, et al. Apoptosis of circulating tumor cells in prostate cancer patients. Cytometry A. 2004;62(1):46-53.
8. Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature. 2008;452(7187):571-9.
9. Riedl SJ, Shi Y. Molecular mechanisms of caspase regulation during apoptosis. Nat Rev Mol Cell Biol. 2004;5(11):897-907.
10. Zheng TS, Hunot S, Kuida K, Flavell RA. Caspase knockouts: matters of life and death. Cell Death Differ. 1999;6(11):1043-53.
11. Agard NJ, Mahrus S, Trinidad JC, Lynn A, Burlingame AL, Wells JA. Global kinetic analysis of proteolysis via quantitative targeted proteomics. Proc Natl Acad Sci U S A. 2012;109(6):1913-8.
12. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell. 2008;134(5):866-76.
13. Shimbo K, Hsu GW, Nguyen H, Mahrus S, Trinidad JC, Burlingame AL, et al. Quantitative profiling of caspase-cleaved substrates reveals different drug-induced and cell-type patterns in apoptosis. Proc Natl Acad Sci U S A. 2012;109(31):12432-7.
14. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2012;40(Database issue):D343-50.
15. Luthi AU, Martin SJ. The CASBAH: a searchable database of caspase substrates. Cell Death Differ. 2007;14(4):641-50.
16. Crawford ED, Seaman JE, Agard NJ, Hsu GW, Julien O, Mahrus S, et al. The DegraBase: a database of proteolysis in healthy and apoptotic human cells. Molecular and Cellular Proteomics. 2012:40.
17. Crawford ED, Seaman JE, Barber AE, 2nd, David DC, Babbitt PC, Burlingame AL, et al. Conservation of caspase substrates across metazoans suggests hierarchical importance of signaling pathways over specific targets and cleavage site motifs in apoptosis. Cell Death Differ. 2012.

18. Zimmerman LJ, Li M, Yarbrough WG, Slebos RJ, Liebler DC. Global stability of plasma proteomes for mass spectrometry-based analyses. Mol Cell Proteomics. 2012;11(6):M111 014340.

19. Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. Mol Cell Proteomics. 2011;10(9):M110 006353.

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

21. Brandt D, Volkmann X, Anstatt M, Langer F, Manns MP, Schulze-Osthoff K, et al. Serum biomarkers of cell death for monitoring therapy response of gastrointestinal carcinomas. Eur J Cancer. 2010;46(8):1464-73.

22. Olofsson MH, Ueno T, Pan Y, Xu R, Cai F, van der Kuip H, et al. Cytokeratin-18 is a useful serum biomarker for early determination of response of breast carcinomas to chemotherapy. Clin Cancer Res. 2007;13(11):3198-206.

23. Yilmaz Y. Systematic review: caspase-cleaved fragments of cytokeratin 18 - the promises and challenges of a biomarker for chronic liver disease. Aliment Pharmacol Ther. 2009;30(11-12):1103-9.

24. Chelur DS, Chalfie M. Targeted cell killing by reconstituted caspases. Proc Natl Acad Sci U S A. 2007;104(7):2283-8.

25. Di Stasi A, Tey SK, Dotti G, Fujita Y, Kennedy-Nasser A, Martinez C, et al. Inducible apoptosis as a safety switch for adoptive cell therapy. N Engl J Med. 2011;365(18):1673-83.

26. Simpson KL, Cawthorne C, Zhou C, Hodgkinson CL, Walker MJ, Trapani F, et al. A caspase-3 'death-switch' in colorectal cancer cells for induced and synchronous tumor apoptosis in vitro and in vivo facilitates the development of minimally invasive cell death biomarkers. Cell Death Dis. 2013;4:e613.

27. Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, et al. A targeted proteomics-based pipeline for verification of biomarkers in plasma. Nat Biotechnol. 2011;29(7):625-34.

28. Taguchi A, Politi K, Pitteri SJ, Lockwood WW, Faca VM, Kelly-Spratt K, et al. Lung cancer signatures in plasma based on proteome profiling of mouse tumor models. Cancer Cell. 2011;20(3):289-99.

29. Kislinger T, Gramolini AO. Proteome analysis of mouse model systems: A tool to model human disease and for the investigation of tissue-specific biology. J Proteomics. 2010;73(11):2205-18.

30. Rozemuller H, van der Spek E, Bogers-Boer LH, Zwart MC, Verweij V, Emmelot M, et al. A bioluminescence imaging based in vivo model for preclinical testing of novel cellular immunotherapy strategies to improve the graft-versus-myeloma effect. Haematologica. 2008;93(7):1049-57.

31. Gray DC, Mahrus S, Wells JA. Activation of specific apoptotic caspases with an engineered small-molecule-activated protease. Cell. 2010;142(4):637-46.

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

33. Wiita AP, Seaman JE, Wells JA. Global analysis of cellular proteolysis by selective enzymatic labeling of protein N-termini. Methods Enzymol. 2014;544:327-58.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Julia Seaman

Author Signature

10 June 2016

Date