Global identification of peptidase specificity by multiplex substrate profiling

Anthony J O'Donoghue[1], A Alegra Eroy-Reveles[1,2], Giselle M Knudsen[1], Jessica Ingram[3], Min Zhou[1], Jacob B Statnekov[1], Alexander L Greninger[4,5], Daniel R Hostetter[1], Gang Qu[6], David A Maltby[1], Marc O Anderson[2], Joseph L DeRisi[4,5], James H McKerrow[3], Alma L Burlingame[1] & Charles S Craik[1]

We developed a simple and rapid multiplex substrate-profiling method to reveal the substrate specificity of any endo- or exopeptidase using liquid chromatography–tandem mass spectrometry sequencing. We generated a physicochemically diverse library of peptides by incorporating all combinations of neighbor and near-neighbor amino acid pairs into decapeptide sequences that are flanked by unique dipeptides at each terminus. Addition of a panel of evolutionarily diverse peptidases to a mixture of these tetradecapeptides generated information on prime and nonprime sites as well as on substrate specificity that matched or expanded upon known substrate motifs. This method biochemically confirmed the activity of the klassevirus 3C protein responsible for polypeptide processing and allowed granzyme B substrates to be ranked by enzymatic turnover efficiency using label-free quantitation of precursor-ion abundance. Additionally, the proteolytic secretions from schistosome parasitic flatworm larvae and a pancreatic cancer cell line were deconvoluted in a subtractive strategy using class-specific peptidase inhibitors.

Peptidases represent the largest class of post-translational modifying enzymes in the human proteome. An estimated 2% of human genes encode 687 peptidases or peptidase-like homolog transcripts that result in ~550 predicted active enzymes[1]. The human proteome represents the potential substrate library for these enzymes, and all proteins are proteolytically modified, either by limited proteolysis or final degradation. Uncovering the substrate specificity of these peptidases is central to understanding their physiological role in homeostasis and disease.

Although it is still common practice to use generic protein or peptide substrates to test for proteolytic activity, the field has been continuously developing biological and chemical tools to characterize substrate specificity in greater detail[2]. Diverse peptide sequences expressed on the surface of phages or bacteria have revealed the substrate specificity of factor Xa[3] and caspase-3 (ref. 4), and N-terminal sequencing of peptide mixtures has been used to profile matrix metallopeptidases[5]. In addition, positionally arranged fluorogenic substrate libraries have been used to rapidly uncover the nonprime-side (N-terminal to the scissile bond) specificity of many enzymes[6,7]. More recently, researchers have combined proteome-derived substrate libraries with mass spectrometry to identify cleavage products within a protein extract[8]. These 'degradomic' methods require complex labeling strategies to differentiate between specific cleavages produced by the peptidase of interest and those produced by other peptidases within the sample[9,10].

There remains a need for a rapid, quantitative and highly reproducible assay that can provide specificity profiles and kinetic constants for any endo- or exo-acting peptidase. Presented here is a direct cleavage assay that uses mass spectrometry–based peptide sequencing for detection of degradation products in a mixture of synthetic peptides. We based the design of the sequences on our hypothesis that cleavage by a peptidase frequently requires no more than two amino acids suitably positioned within a peptide substrate. Therefore, the library contains all combinations of neighbor and near-neighbor amino acid pairs. A peptide length of 14 residues was selected to allow sufficient substrate length for binding of endopeptidases and to minimize tertiary structure that could limit side-chain accessibility. In this study, we examined whether the depth of information obtained from a chemically defined pool of peptides could be sufficient to determine a specificity signature of a peptidase. We demonstrate that multiplex substrate profiling by mass spectrometry (MSP-MS) can generate prime- and nonprime-side substrate specificity data for representative endo- and exopeptidases from multiple families. Because of the high information content relative to low background, this method requires no additional labeling or sample fractionation and is therefore highly reproducible. Peptide degradation kinetics can be monitored by label-free quantitation of parent-ion mass spectrometry peaks, and we deconvolute biological samples containing multiple peptidases through the use of class-specific inhibitors.

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco (UCSF), San Francisco, California, USA. [2]Department of Chemistry and Biochemistry, San Francisco State University, San Francisco, California, USA. [3]Sandler Center for Drug Discovery, UCSF, San Francisco, California, USA. [4]Howard Hughes Medical Institute, UCSF, San Francisco, California, USA. [5]Department of Biochemistry and Biophysics, UCSF, San Francisco, California, USA. [6]Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland , USA. Correspondence should be addressed to C.S.C. (Charles.Craik@ucsf.edu).

RECEIVED 22 MAY; ACCEPTED 29 AUGUST; PUBLISHED ONLINE 30 SEPTEMBER 2012; DOI:10.1038/NMETH.2182

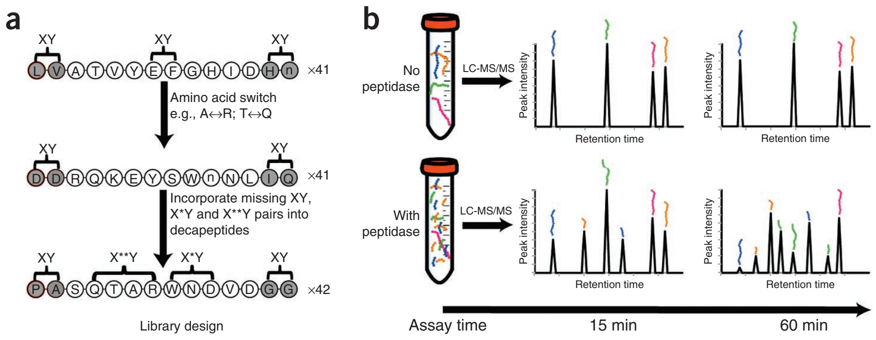

Figure 1 | Design of a physiochemically diverse peptide library and development of a multiplex substrate assay. (**a**) Design of a library of 14-mer peptides by accommodating all neighbor (XY) and near-neighbor pairs (X*Y and X**Y) into a core decapeptide (unshaded residues). X and Y correspond to defined amino acids and * to a random amino acid. The termini (shaded residues) were generated using amino acid pairs (XY) selected from 11 pools of amino acids. n, norleucine. (**b**) Illustration of the multiplex substrate-profiling assay. A peptidase is added to the peptides, and aliquots are removed at multiple time intervals and quenched with acid or inhibitor. Each time point is injected into an LC-MS/MS system to detect time-dependent appearance of cleavage products. Cleavage sites are identified by comparison with a control assay that lacks a peptidase.

RESULTS

To profile the substrate specificity of all peptidase families, we synthesized a defined library of 124 peptides with extensive physicochemical diversity. Within the central decapeptide region of each sequence, two copies of every amino acid pair (XY) and one copy of every X*Y and X**Y pair were accommodated, where X and Y represent defined amino acids and * indicates a random amino acid. To create diversity at each terminus for exo-acting enzymes, a unique dipeptide sequence was placed at both the N terminus and C terminus of each decapeptide core to produce a library of 124 tetradecapeptides resulting in 1,612 potential cleavage sites (**Fig. 1a**). The abundance of each amino acid within the library ranges from 4.2% to 6.8%. By comparison, vertebrate protein sequences in the SwissProt database show bias in amino acid usage ranging from 1.3% (tryptophan) to 8.1% (leucine) (**Supplementary Fig. 1**). For the MSP-MS assay, peptides were pooled at equimolar concentration and combined with a peptidase in assay buffer. At defined time intervals, samples were quenched and analyzed by liquid chromatography–tandem mass spectrometry (LC-MS/MS) (**Fig. 1b**).

Validation of the MSP-MS assay using cathepsin E

The aspartyl peptidase cathepsin E was selected to validate the MSP-MS assay because it has been thoroughly characterized using proteome-derived peptide libraries[11]. Cathepsin E was incubated with the peptide library, and samples were removed at eight time intervals and quenched with pepstatin. After 5 min of incubation, LC-MS/MS sequencing uncovered 114 cathepsin E cleavage sites, and by 1,200 min, 14.5% of all peptide bonds in the library were hydrolyzed (**Fig. 2a**). For each cleaved bond observed at 5 min, the residues in P4 to P4′ (labeled here using peptidase substrate nomenclature) were identified and a substrate signature was generated using iceLogo[12] (**Fig. 2b**). Cathepsin E favored phenylalanine and norleucine at the P1 position and norleucine and valine at P1′, whereas glycine at P1 and histidine at P1′ were disfavored. After incubation for 1,200 min, the specificity at both P1 and P1′ broadened to include leucine and tryptophan at P1 and isoleucine and tyrosine at P1′ (**Fig. 2c**), which strongly correlated (Pearson score > 0.75) with cathepsin E specificity obtained from proteome-derived sequences[11] (**Supplementary Table 1**).

The MSP-MS assay uncovered a previously uncharacterized carboxypeptidase function of cathepsin E, as 12.9% of cleavage sites occurred at the carboxy terminus of substrates. Furthermore, in 98.3% of cleavage sites, the S3 to S1′ subsites were occupied, which suggests these subsites were most important for substrate recognition. To examine the neighboring effects within the P3 to P1′ sites, substrates containing with the **F↓* motif were analyzed in detail, as phenylalanine in the P1 position is the single most

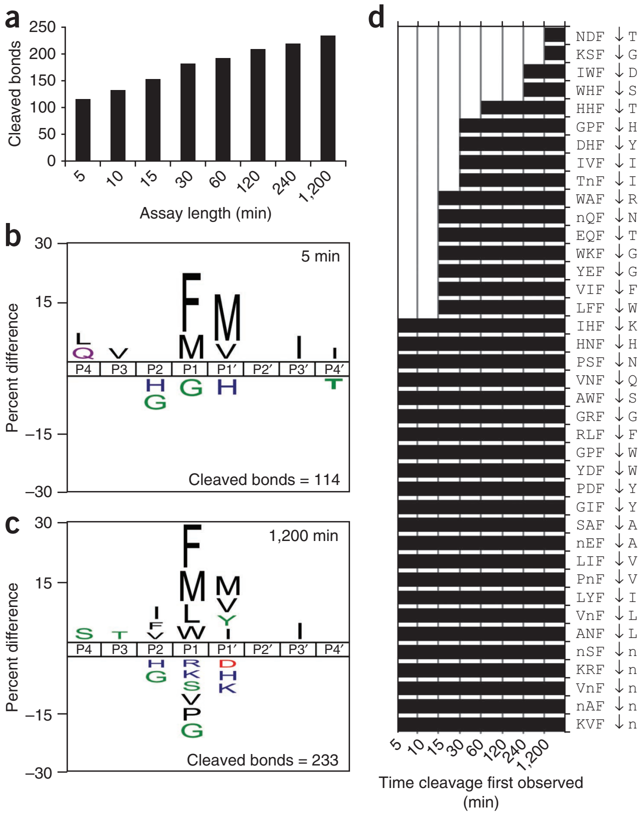

Figure 2 | Validation of the multiplex substrate-profiling technique using the aspartyl peptidase cathepsin E. (**a**) Quantitative assessment of bonds cleaved by cathepsin E at increasing time intervals. (**b**,**c**) iceLogos generated from amino acids that are enriched or de-enriched in the P4 through P4′ positions of cathepsin E cleavage sites after incubation for 5 min (**b**) and 20 h (**c**). Percent difference corresponds to the difference of amino acid frequency surrounding the cathepsin E cleavage sites relative to the frequency of amino acids surrounding all peptide bonds in the library (*n* = 1,612). In the iceLogo, M corresponds to norleucine. (**d**) Bar chart representing tetrapeptide sequences containing phenylalanine in the third position (**F↓*) and the time that cleavage is first observed. All **F* motifs in the library are listed in **Supplementary Table 2**.

preferred residue in the cathepsin E substrate-binding pocket (**Fig. 2d** and **Supplementary Table 2**). When phenylalanine was paired with a hydrophobic residue in the P1′ position, cleavage was generally observed within 30 min, whereas proline and polar residues were disfavored. In certain peptide sequences, P1 phenylalanine is surrounded by nonfavorable residues; however, cleavage still occurs rapidly, indicating flexibility in the substrate binding pocket.

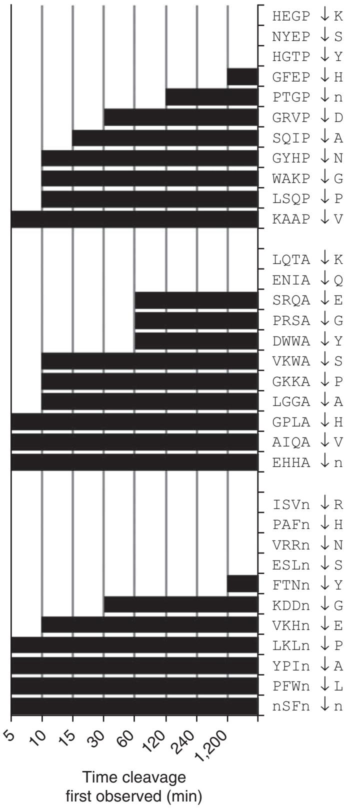

Figure 3 | Substrate profiling of an exopeptidase PRCP. PRCP releases amino acids from the C terminus of peptides when proline, alanine or norleucine are in the P1 position. The time, if any, that cleavage was first observed is illustrated by the length of each bar.

Profiling exopeptidase substrate specificity with MSP-MS

Exopeptidases, such as prolylcarboxypeptidase (PRCP), process peptide hormones by removing residues from their carboxy terminus. Known PRCP substrates have a strict specificity for Pro-Xaa bonds[13]; however, the MSP-MS assay revealed that the enzyme readily accepted alanine or norleucine in the S1 pocket (**Fig. 3**). Cleavage occurred rapidly when P1 proline, alanine or norleucine were paired with hydrophobic residues such as norleucine (n), valine, leucine, alanine and proline in the P1′ position. PRCP had no tolerance for arginine or lysine in the S1′ pocket but accepted histidine, aspartic acid and glutamic acid in certain cases. In several substrates, time-dependent trimming was observed. In one example, cleavage at Pro-Val, Ala-Pro, Ala-Ala and Lys-Ala bonds in the substrate RnENYnVLTKAAPV was evident at 5, 10, 60 and 1,200 min, respectively.

Quantification of cleavage efficiency for substrates

Granzyme B and human rhinovirus 14 (HRV14) 3C are highly specific peptidases that are involved in apoptosis and viral polypeptide processing, respectively. Unlike with other enzymes with broader specificity, cleavage using these peptidases occurred in the MSP-MS assay at a single site within a substrate, producing products that were not subsequently degraded at secondary sites. This allowed for the quantification of the extracted ion chromatogram for each product as it accumulated over time and for the calculation of catalytic efficiency. The best substrate for granzyme B (KHPLETVYAD↓SSEW) had a catalytic efficiency of 127,000 ± 13,000 M^{-1} s^{-1} (**Fig. 4**), which closely matched the results of previous studies (116,000 M^{-1} s^{-1}) using a fluorescent substrate containing the sequence VVAD↓SSMES[14]. HRV14 3C cleaved at only one site (YnDSIRHQ↓GPFWnL); this substrate was therefore pooled with other peptides containing Gln-Gly and Gln-Ser pairs and a selection of peptides of known or putative viral polypeptide processing sites (**Supplementary Table 3**). The catalytic efficiency of cleavage in YnDSIRHQ↓GPFWnL was comparable with that of the HRV14 2C-3A polypeptide release site and >100-fold superior to that of the HRV14 3C-3D release site (**Supplementary Fig. 2**), as has been observed previously[15].

Screening peptidase gene products for activity

We have confirmed that the peptide library is broadly applicable for profiling purified peptidases from multiple families that include cruzain, matriptase, DPP-IV, MMP2, eqolisin, aspergillopepsin, and HIV-1 and HIV-2 proteases (**Supplementary Fig. 3**). Furthermore, we have previously used the MSP-MS assay to profile the specificity of a gut-associated hemoglobinase from the Lyme disease tick vector *Ixodes ricinus*[16]. However, isolation of pure, stable peptidases from native or recombinant sources is time consuming and labor intensive; therefore, we investigated whether sufficient enzymatic activity could be generated from an *in vitro* transcription-translation system to perform an MSP-MS assay. Using a bacterial *in vitro* transcription-translation system, a selection of highly specific 3C cysteine peptidases from picornaviruses were expressed, partially purified (**Supplementary Fig. 4**) and assayed with the same pool of peptides as outlined above for commercial-grade HRV14 3C (**Supplementary Table 3**). Cleavage of YnDSIRHQ↓GPFWnL was evident by 3C peptidases from HRV14, poliovirus, enterovirus 71, hepatitis A virus and a previously uncharacterized klassevirus 3C peptidase. The HRV14 2C-3A substrate was cleaved by only the related enterovirus peptidases. Taken together, these data indicate that rapid expression and partial purification of peptidases (<2 h) can be combined with a highly sensitive peptidase assay to distinguish active from inactive enzymes.

Identifying the proteolytic signature of complex samples

Schistosomiasis affects over 200 million people worldwide and is ranked second only to malaria in overall morbidity caused by a parasitic organism. Secreted proteases are used by *Schistosoma mansoni* infective larvae, termed cercariae, for transmission from their intermediate snail host and penetration of human skin. The process of cercariae being shed from snails can be recapitulated *in vitro*[17]. Using the MSP-MS technique, the background activity in conditioned water from noninfected snails contained P1-arginine specificity consistent with snail tryptases[18] (**Fig. 5a**). The peptidase activity of *S. mansoni* cercarial secretions had a preference for tyrosine, phenylalanine and norleucine in the

P1 position, proline in P2 and norleucine in P1′ (**Fig. 5b**). Proteomic analysis determined that *S. mansoni* secretions contained serine and cysteine peptidases and metallopeptidases[18,19]; therefore, to dissect the contribution of each activity, secretions were incubated with either the metal chelator EDTA, the cysteine peptidase inhibitors E-64 and CAO74, or an elastase-specific chloromethyl ketone inhibitor[17]. No major changes were observed following treatment with EDTA or a mixture of E-64 and CAO74 (**Supplementary Fig. 5**), whereas the elastase inhibitor resulted in a 36% reduction in cleaved bonds and a de-enrichment of tyrosine, phenylalanine and norleucine at the P1 position (**Fig. 5c**). MSP-MS profiling of an elastase-enriched fraction from an *S. mansoni* cercarial extract confirmed the source of the major peptidase activity in the parasite secretions to be elastase (**Fig. 5d**).

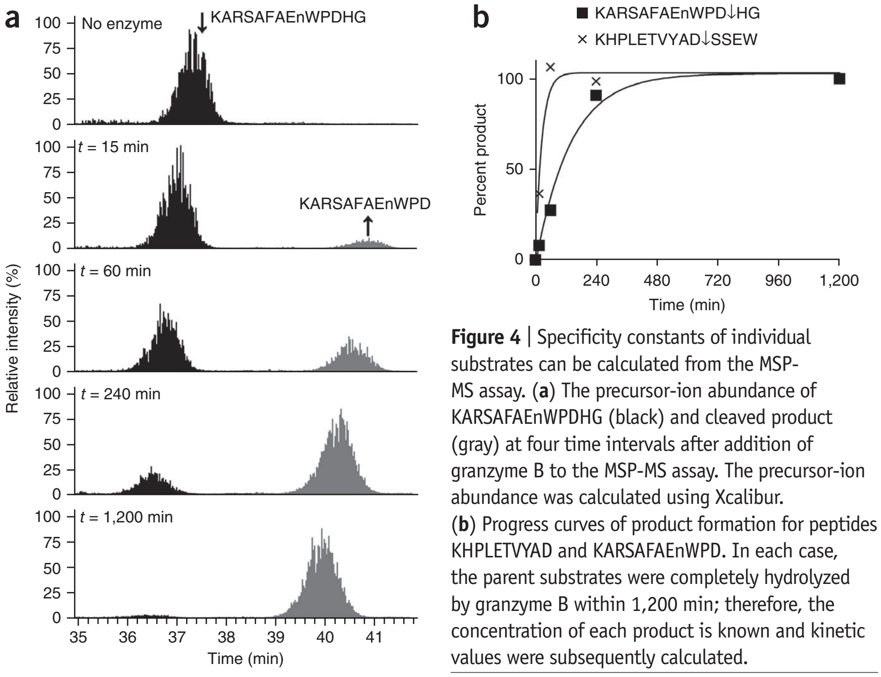

Figure 4 | Specificity constants of individual substrates can be calculated from the MSP-MS assay. (**a**) The precursor-ion abundance of KARSAFAEnWPDHG (black) and cleaved product (gray) at four time intervals after addition of granzyme B to the MSP-MS assay. The precursor-ion abundance was calculated using Xcalibur. (**b**) Progress curves of product formation for peptides KHPLETVYAD and KARSAFAEnWPD. In each case, the parent substrates were completely hydrolyzed by granzyme B within 1,200 min; therefore, the concentration of each product is known and kinetic values were subsequently calculated.

In many human cancers, dysregulation of protease activity can lead to degradation of extracellular matrices, thereby facilitating neoplastic progression[20]. Pancreatic ductal adenocarcinoma (PDAC) is an aggressive form of cancer with limited response to treatment leading to an average survival time after diagnosis of 6 months. To interrogate the role of extracellular proteases in PDAC, proteomic analysis and MSP-MS was performed on conditioned medium from a primary mouse PDAC cell line. The proteomic study identified six peptidases in the conditioned medium (**Supplementary Table 4**), most of which were optimally active between pH 4.5 and 6 (refs. 21–24). The extracellular environment within pancreatic tumors is known to be acidic, so the MSP-MS assay was performed at pH 5.2. Under these conditions, a total of 98 unique cleavage sites were identified, and the substrate signature revealed a preference for hydrophobic residues in the P1 and P1′ positions (**Fig. 5e**). When conditioned medium was pretreated with either E-64 or the metallopeptidase inhibitor 1,10-phenanthroline, the majority of cleavage sites remained unchanged (**Fig. 5f,g**). However, treatment with pepstatin completely altered the cleavage signature (**Fig. 5h**). As cathepsin E was the only pepstatin-sensitive peptidase detected in the medium and the substrate signature was similar to that obtained from the recombinant enzyme (**Fig. 2c**), we concluded that cathepsin E is the major proteolytic activity secreted by PDAC cells.

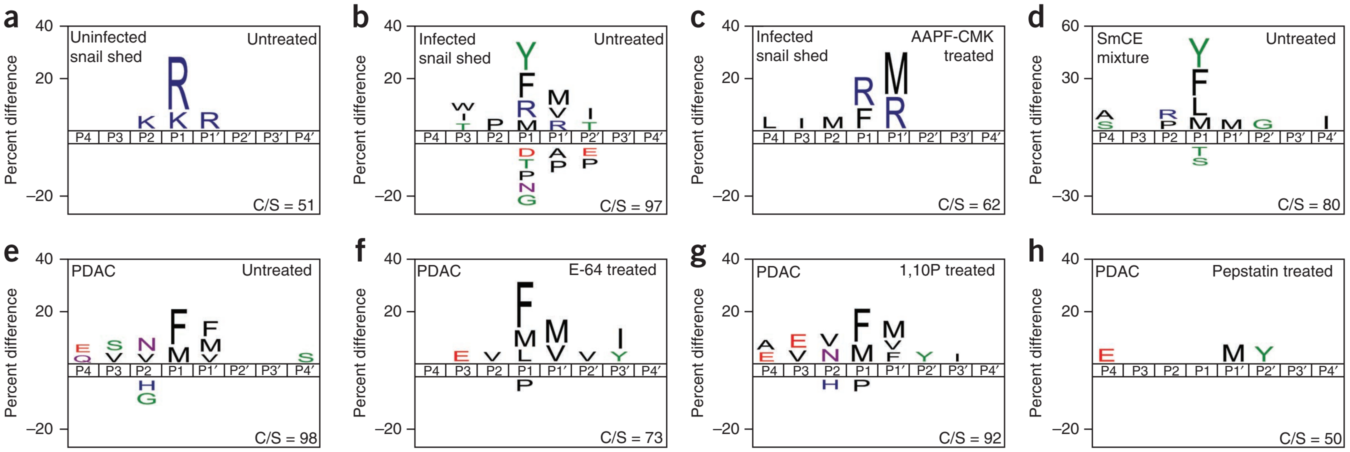

Figure 5 | Class-specific peptidase inhibitors can dissect the proteolytic signatures of biological samples. (**a**,**b**) Substrate signature of material from uninfected (**a**) and *S. mansoni*–infected (**b**) freshwater snails after 240 min of incubation with the substrate library. (**c**) Infected snail material was pretreated with Ala-Ala-Pro-Phe–chloromethyl ketone (AAPF-CMK) inhibitor before the multiplex substrate profile assay. (**d**) Cleavage-site data set of two semi-pure preparations of *S. mansoni* cercarial elastases (SmCEs) were combined to generate a substrate signature for this peptidase group. (**e**) Substrate signature of all cleavage sites observed within 1,200 min using conditioned medium from pancreatic ductal adenocarcinoma (PDAC) cells as a source of proteases. Medium was pretreated with E-64 (**f**), 1,10-phenanthroline (1,10P) (**g**) or pepstatin (**h**), and a substrate signature was generated of the remaining cleavage events. In all cases, only amino acids that were enriched or de-enriched are illustrated in the substrate signature; M corresponds to norleucine. C/S, cleavage sites.

DISCUSSION

Fluorescent and colorimetric substrates have been the standard reagents for detecting and characterizing peptidases for decades. However, with the recent advances in mass spectrometry–driven degradomics, substrates containing reporter groups are no longer required to obtain subsite specificity[25]. We hypothesized that substrate recognition by peptidases requires no more than two amino acids suitably positioned within a peptide. This hypothesis was generated from the wealth of data derived from synthetic and proteome-derived peptide libraries[26], and examples include P2 and P1 of cathepsin L, K, S and F[7]; P4 and P1 of granzyme B[27]; and P3 and P1′ of MMP2 and MMP9 (ref. 28). Using a synthetic peptide library containing all combinations of amino acid pairs, we generated comprehensive substrate signatures for enzymes representing five families that matched or expanded upon known substrate motifs.

PRCP releases amino acids from the C terminus of substrates such as α-melanocyte-stimulating hormone and angiotensins II and III, all of which have proline in the penultimate position[29]. Using MSP-MS, PRCP cleaved single amino acids from the C terminus but not exclusively after proline as its name would suggest. In fact, cleavage after alanine or norleucine was often identified at earlier time points than was proline cleavage. Inhibition of PRCP in mice causes a reduction in body weight, making the enzyme a target for treatment of obesity; however, until now the substrate specificity has not been thoroughly investigated. Norleucine is an isosteric analog of methionine, and therefore, proteins or peptides containing either methionine or alanine in the penultimate position should now be assessed as potential physiological substrates for PRCP.

Functional characterization of peptidases in a biological system has traditionally involved identifying a candidate protein, determining the substrate specificity and generating a specific probe or inhibitor. Although the candidate approach has a proven track record, we sought to characterize proteolysis using an unbiased global approach. In this study, peptidase substrate profiling using conditioned water from healthy snails revealed trypsin-like specificity, whereas conditioned water containing secretions from *S. mansoni* larvae had a mixture of trypsin and elastase-like activity. A previous study has shown that topical application of Ala-Ala-Pro-Phe–chloromethyl ketone on human skin prevents invasion by the parasite[30]. Here we demonstrate that this inhibitor reduces the overall proteolytic activity in parasite secretions by targeting the elastase-like peptidases of these larvae.

The secretome of a primary PDAC revealed multiple secreted peptidases in the medium. The same preparation demonstrated robust cleavage activity between pairs of hydrophobic residues that could be diminished with an aspartic protease inhibitor. The sole aspartic peptidase, cathepsin E, was a lower-abundance protease but represented the major proteolytic activity, indicating that proteomic studies alone cannot be used to predict the role of active proteolytic enzymes. Abundant cathepsin E activity is partially explained by the presence of endogenous inhibitors for the cysteine and metalloproteinases and the absence of endogenous aspartic peptidase inhibitors in the PDAC-secreted proteome. These data provide functional support for a recent study that uses a fluorescent cathepsin E substrate to detect pancreatic cancer in a mouse model[31]. The cleavage site data generated for cathepsin E can be used in future studies to design improved fluorescent substrate probes to selectively image cathepsin E activity in pancreatic tumors.

Successful substrate prediction tools require an experimentally determined list of protein substrates to identify sequence and structural features important for peptidase recognition[32]. Using the MSP-MS assay, individual substrates in a peptide mixture can be ranked using kinetic values extracted from progress curves. Any peptide that is not cleaved is considered a true negative substrate. The predictive power for identifying natural substrates can be improved by incorporating negative and ranked positive substrates into the bioinformatic analysis. In certain cases, the preferred peptide substrate can be a powerful probe for identifying endogenous substrates[33]. Although the current set of peptides is sufficient to identify cleavage site specificity, we anticipate that greater sequence resolution can be achieved by synthesizing secondary libraries that iteratively explore the preferred sequence space for a peptidase. However, determination of substrates *in vitro* must take into account the complex biology associated with the enzyme. Detailed knowledge of the substrate specificities of individual peptidases and those in complex biological systems affords new opportunities to understand their role in homeostasis and disease and will aid in the development of chemical tools for detection or inhibition.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank F.A. Lewis at the US National Institutes of Health–National Institute of Allergy and Infectious Disease Schistosomiasis Resource Center for providing *S. mansoni*–infected *Biomphalaria glabrata* and K.C. Lim of the Sandler Center for Drug Discovery at UCSF for maintenance of the *S. mansoni* life cycle. We thank C. Tajon, C. Brown, G. Lee and S. Clarke from UCSF, M. Tuohy from National University of Ireland, Galway, and W. Geissler from Merck & Co. for providing purified peptidases and D. Hanahan from UCSF for the mouse PDAC cell line. This project was supported by grants from the US National Institutes of Health: P50 GM082250 (C.S.C.), R01 CA128765 (C.S.C.), P41 RR001614 (A.L.B.), P41 GM103481 (A.L.B.) and K12 GM081266 (A.A.E.-R.) and the Sandler Center (C.S.C.).

AUTHOR CONTRIBUTIONS

A.J.O., M.Z. and C.S.C. conceptualized the MSP-MS library and assay. C.S.C. directed and coordinated the project. M.O.A. and G.Q. wrote the pair-fitting script and J.B.S. wrote the MSP-MS extractor script. A.J.O., A.A.E.-R. and M.Z. synthesized and purified the peptides. A.J.O., A.A.E.-R. and G.M.K. performed the MSP-MS assays and analyzed the data. G.M.K., D.A.M. and A.L.B. developed the mass spectrometry protocol. J.I. and J.H.M. generated *S. mansoni*–infected snail samples. A.J.O. and D.R.H. generated conditioned PDAC medium. A.L.G. and J.L.D. provided viral peptidases. A.J.O., G.M.K., A.A.E.-R. and C.S.C. wrote the manuscript, and all authors participated in editing it.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2182. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Puente, X.S., Gutiérrez-Fernández, A., Velasco, G. & López-Otín, C. A genomic view of the complexity of mammalian proteolytic systems. *Biochem. Soc. Trans.* **33**, 331–334 (2005).
2. Van Damme, P., Vandekerckhove, J. & Gevaert, K. Disentanglement of protease substrate repertoires. *Biol. Chem.* **389**, 371–381 (2008).

3. Matthews, D.J. & Wells, J.A. Substrate phage: selection of protease substrates by monovalent phage display. *Science* **260**, 1113–1117 (1993).
4. Boulware, K.T. & Daugherty, P.S. Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proc. Natl. Acad. Sci. USA* **103**, 7583–7588 (2006).
5. Turk, B.E., Huang, L.L., Piro, E.T. & Cantley, L.C. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat. Biotechnol.* **19**, 661–667 (2001).
6. Harris, J.L. *et al.* Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proc. Natl. Acad. Sci. USA* **97**, 7754–7759 (2000).
7. Choe, Y. *et al.* Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.* **281**, 12824–12832 (2006).
8. auf dem Keller, U. & Schilling, O. Proteomic techniques and activity-based probes for the system-wide study of proteolysis. *Biochimie* **92**, 1705–1714 (2010).
9. Mahrus, S. *et al.* Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell* **134**, 866–876 (2008).
10. Staes, A. *et al.* Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *Proteomics* **8**, 1362–1370 (2008).
11. Impens, F. *et al.* A quantitative proteomics design for systematic identification of protease cleavage events. *Mol. Cell. Proteomics* **9**, 2327–2333 (2010).
12. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 (2009).
13. Chajkowski, S.M. *et al.* Highly selective hydrolysis of kinins by recombinant prolylcarboxypeptidase. *Biochem. Biophys. Res. Commun.* **405**, 338–343 (2011).
14. Sun, J. *et al.* Importance of the P4′ residue in human granzyme B inhibitors and substrates revealed by scanning mutagenesis of the proteinase inhibitor 9 reactive center loop. *J. Biol. Chem.* **276**, 15177–15184 (2001).
15. Cordingley, M.G., Callahan, P.L., Sardana, V.V., Garsky, V.M. & Colonno, R.J. Substrate requirements of human rhinovirus 3C protease for peptide cleavage in vitro. *J. Biol. Chem.* **265**, 9062–9065 (1990).
16. Sojka, D. *et al.* Characterization of gut-associated cathepsin D hemoglobinase from tick *Ixodes ricinus* (IrCD1). *J. Biol. Chem.* **287**, 21152–21163 (2012).
17. Ingram, J.R. *et al.* Investigation of the proteolytic functions of an expanded cercarial elastase gene family in *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* **6**, e1589 (2012).
18. Knudsen, G.M., Medzihradszky, K.F., Lim, K.-C., Hansell, E. & McKerrow, J.H. Proteomic analysis of *Schistosoma mansoni* cercarial secretions. *Mol. Cell. Proteomics* **4**, 1862–1875 (2005).
19. Curwen, R.S., Ashton, P.D., Sundaralingam, S. & Wilson, R.A. Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry. *Mol. Cell. Proteomics* **5**, 835–844 (2006).
20. Nolan-Stevaux, O. *et al.* GLI1 is regulated through Smoothened-independent mechanisms in neoplastic pancreatic ducts and mediates PDAC cell survival and transformation. *Genes Dev.* **23**, 24–36 (2009).
21. Fricker, L. Carboxypeptidase E. in *Handbook of Proteolytic Enzymes* 2nd edn. (eds. Barrett, A.J., Rawlings, N.D. & Woessner, J.F.) 840–844 (Academic, 2004).
22. Caglic, D. *et al.* Murine and human cathepsin B exhibit similar properties: possible implications for drug discovery. *Biol. Chem.* **390**, 175–179 (2009).
23. Mason, R.W., Johnson, D.A., Barrett, A.J. & Chapman, H.A. Elastinolytic activity of human cathepsin L. *Biochem. J.* **233**, 925–927 (1986).
24. Zaidi, N., Herrmann, T., Voelter, W. & Kalbacher, H. Recombinant cathepsin E has no proteolytic activity at neutral pH. *Biochem. Biophys. Res. Commun.* **360**, 51–55 (2007).
25. Impens, F. *et al.* MS-driven protease substrate degradomics. *Proteomics* **10**, 1284–1296 (2010).
26. Rawlings, N.D., Barrett, A.J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **40**, D343–D350 (2012).
27. Ruggles, S.W., Fletterick, R.J. & Craik, C.S. Characterization of structural determinants of granzyme B reveals potent mediators of extended substrate specificity. *J. Biol. Chem.* **279**, 30751–30759 (2004).
28. Prudova, A., auf dem Keller, U., Butler, G.S. & Overall, C.M. Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol. Cell. Proteomics* **9**, 894–911 (2010).
29. Zhou, C. *et al.* Design and synthesis of prolylcarboxypeptidase (PrCP) inhibitors to validate PrCP as a potential target for obesity. *J. Med. Chem.* **53**, 7251–7263 (2010).
30. Cohen, F.E. *et al.* Arresting tissue invasion of a parasite by protease inhibitors chosen with the aid of computer modeling. *Biochemistry* **30**, 11221–11229 (1991).
31. Cruz-Monserrate, Z. *et al.* Detection of pancreatic cancer tumours and precursor lesions by cathepsin E activity in mouse models. *Gut* **61**, 1315–1322 (2012).
32. Barkan, D.T. *et al.* Prediction of protease substrates using sequence and structure features. *Bioinformatics* **26**, 1714–1722 (2010).
33. Harris, J.L., Peterson, E.P., Hudig, D., Thornberry, N.A. & Craik, C.S. Definition and redesign of the extended substrate specificity of granzyme B. *J. Biol. Chem.* **273**, 27364–27373 (1998).

ONLINE METHODS

Peptide library design. For the MSP-MS assay, a library of 124 peptides with extensive physicochemical diversity was developed. Cysteine was omitted because of potential disulfide bond formation, and norleucine (Nle) replaced oxidation-prone methionine. Peptide sequence diversity was designed in a three-part strategy.

For the endopeptidase component, a novel algorithm was developed to arrange amino acids pairs into the minimal number of decapeptide sequences such that no decapeptide had more than two identical residues or more than one pair. All but one pair were incorporated into 41 sequences. To generate diversity surrounding each pair, a second set of 41 decapeptides was designed by substituting all amino acids with a physicochemically distinct counterpart: for example, all Ala residues in the first set of 41 peptides were replaced by Arg residues in the second set of peptides and vice versa. The other substitutions involved Val and Lys, Leu and Asp, Ile and Asn, Gln and Thr, Phe and Ser, Trp and Gly, Glu and Tyr, and Nle and His. Pro remained unchanged. These 82 decapeptides are defined as the 'XY decapeptide sublibrary', where X and Y represent defined amino acids. Next, all near-neighbor amino acid pairs separated by one (X*Y) or two (X**Y) random amino acids were identified in the XY decapeptide sublibrary. Any pair not present (including the missing XY pair) was manually assembled into 42 additional decapeptides to generate a final set of 124 decapeptide sequences.

For generating diversity for exopeptidases, amino acids were first combined into 11 groups having distinct physicochemical characteristics, specifically Ile/Leu/Val, Ser/Thr, Glu/Asp, Lys/Arg, Tyr/Trp/Phe, Gln/Asn, Gly, Pro, Ala, His and Nle. Next, 121 dipeptide sequences were generated by pairing a single amino acid, chosen at random, from each group. Each pair represents the amino-terminal dipeptide of the final 14-mer sequence. This procedure was repeated to generate an additional 121 pairings for positioning at the carboxy terminus.

Finally, 14-mer sequences were assembled by combining an N-terminal and C-terminal dipeptide with each core decapeptide sequence such that no more than three identical residues were present in the entire sequence. Three additional N-terminal and C-terminal pairs were generated manually to ensure that the number of exo-sequences (121) matched the number of decapeptides (124). Upon request, the full list of sequences will be made available from the corresponding author.

Peptide synthesis. Peptide synthesis was performed on an automated peptide synthesizer (Protein Systems, model 433A) using solid-phase conditions, rink amide AM resin (Novabiochem) and Fmoc main-chain protecting group chemistry. For the coupling of Fmoc-protected amino acids (Novabiochem), 10 equivalents of amino acid and a 1:1:2 molar ratio of coupling reagents HBTU/HOBt (Novabiochem)/DIEA were used. Isolation of desired peptides was achieved by trifluoroacetic acid–mediated deprotection and cleavage, ether precipitation to yield the crude product, and high-performance liquid chromatography (HPLC) (Varian ProStar) on a reversed-phase C18 column (Varian) to yield the pure compounds. Chemical composition of the pure products was confirmed by LC-MS mass spectrometry (Waters Micromass ZQ), and 5 mM stock solutions of each peptide were prepared in DMSO. The average cost of synthesis and purification was ~$10 per amino acid. A subset of peptides was synthesized and purified by AnaSpec. In all cases, the purity of each peptide was 90% or greater.

Enzymes. Proteases were either purchased or obtained through kind gifts and include mouse cathepsin E (R&D Systems), HRV 3C (EMD Chemicals), rat granzyme B (C. Tajon, UCSF), human matriptase-1 (C. Brown, UCSF), matrix metalloprotease 2 (AnaSpec), *Talaromyces emersonii* eqolisin (M. Tuohy, NUI Galway, Ireland), aspergillopepsin I (Sigma), cruzain (G. Lee, UCSF), human dipeptidyl peptidase IV (Sigma), human prolylcarboxypeptidase (W. Geissler, Merck) and HIV-1 and HIV-2 proteases (S. Clarke, UCSF).

A selection of viral proteases were amplified using the oligonucleotides listed in **Supplementary Table 5** and were cloned using InFusion Advantage (Clontech) into a pET23b vector linearized with NdeI and XhoI. One microgram of sequence-confirmed plasmid was used as input for the S30 T7 High Yield Protein Expression System (Promega) and purified using MagneHis Protein purification system (Promega) following the manufacturer's suggested protocols. Of the 50-µL eluate, 10 µL was run on a 4–12% Bis-Tris NuPage acrylamide gel (Life Technologies) and silver stained.

Isoforms of cercarial elastase were partially purified from *S. mansoni* cercariae after sonication in 300 mM sodium acetate, pH 6.5, 0.1% Triton X-100, 0.1% Tween-20, 0.05% NP40. Soluble protein was harvested by centrifugation for 15 min at 7,500*g*, and this was followed by 0.2-µm filtration. The supernatant was loaded onto an SR 16/100 column packed with Sephacryl 200 (GE Healthcare), and 4-ml fractions were collected at 4 °C. Fractions were assayed using 10 µl of sample and 100 µl of assay buffer (100 mM glycine, pH 9.0, 100 µM succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (AAPF-pNA). Proteolytically active fractions were used in the MSP-MS assay. These fractions were also size separated on a 10% Bis-Tris polyacrylamide gel (Life Technologies) and silver stained. Protein bands corresponding to the correct molecular weight of cercarial elastases were excised from the gel, digested with trypsin and analyzed by LC-MS/MS to determine the protein composition.

***S. mansoni* secretions.** Cercariae were shed from several hundred infected *B. glabrata*, and cercarial secretions were collected as previously described by Salter and colleagues[34]. Isolated cercarial secretions were lyophilized and resuspended in 50 mM Tris-HCl, pH 7.5, and sonicated for 1 min. The soluble fraction was isolated by centrifugation at 16,000*g* at 4 °C. For inhibition studies, the soluble fraction was pretreated 30 min at room temperature with either 25 mM EDTA, 500 nM succinyl Ala-Ala-Pro-Phe-chloromethyl ketone or a mixture of 250 nM CAO74 and 250 nM E-64 before addition to the MSP-MS assay. The assay was performed in 50 mM Tris-HCl, pH 7.5.

Pancreatic cancer secretions. A cell line derived from p48-Cre/+; LSL-Kras/+; Trp53F/+ mice[20] was maintained in DMEM containing 10% FBS and 1× penicillin/streptomycin and grown to ~80% confluence in triplicate T75 flasks. Medium was removed and cells were washed five times with Opti-MEM (Invitrogen) and incubated with Opti-MEM. After 20 h, the conditioned medium was removed and sterile filtered (0.22 µm). The cells were treated

doi:10.1038/nmeth.2182

with trypsin, and their viability was calculated using trypan blue staining. The conditioned medium was buffer-exchanged into PBS and concentrated 50-fold in an Amicon Ultra centrifugal filter with 10-kDa cutoff. An aliquot of each triplicate sample was digested with trypsin and subjected to LC-MS/MS sequencing (described below). The remaining conditioned medium was pooled and acidified to pH 5.2 with 200 mM ammonium acetate. The sample was split into four tubes and treated with ethanol (vehicle control), 1 μM E64, 1 μM 1,10-phenanthroline or 400 nM pepstatin for 30 min at room temperature before addition to the MSP-MS assay. The assay was performed in PBS acidified to pH 5.2 with 200 mM ammonium acetate.

Protein identification by mass spectrometry. Protein identification in pancreatic cancer secretion samples was performed using peptide sequencing by mass spectrometry. Secretion samples were digested with trypsin in solution as follows. Secretion sample (20–50 μg total protein) was incubated with urea (5 M final) and 10 mM DTT for 10 min at 56 °C. After reduction, the sample was alkylated with 15 mM iodoacetamide (45 min, dark, 21 °C) then quenched with 10 mM additional DTT, and the final volume was diluted 5× in 100 mM ammonium bicarbonate. Trypsin (sequencing grade, Promega) was added at 1:50 trypsin: total protein for digestion overnight at 37 °C. The sample was then acidified with 10% formic acid to pH 2–3 and desalted using C18 ZipTips (Millipore). Extracted peptides were sequenced using an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific) equipped with a 10,000 p.s.i. system–nanoACQUITY (Waters) UPLC instrument for reversed-phase chromatography with a C18 column (BEH130, 1.7-μm bead size, 100 μm × 100 mm). The LC was operated at 600 nL/min flow rate, and peptides were separated using a linear gradient over 42 min from 2% B to 30% B, with solvent A: 0.1% formic acid in water and solvent B: 0.1% formic acid in 70% acetonitrile. Survey scans were recorded over a 350–1,900 *m*/*z* range, and MS/MS was performed in data-dependent acquisition mode with HCD fragmentation on the seven most intense precursor ions.

Mass spectrometry peak lists were generated using in-house software called PAVA, and data were searched using Protein Prospector software v.5.10.0 (ref. 35). Database searches were performed against the SwissProt *Mus musculus* database (downloaded 21 March 2012), which contained 16,520 entries. For estimation of false discovery rate, this database was concatenated with a fully randomized set of sequence entries[36]. Data were searched with mass tolerances of 20 p.p.m. for parent ions and 30 p.p.m. for fragment ions.

For database searching, peptide sequences were matched as tryptic peptides with no missed cleavages and with carbamidomethylated cysteines as a fixed modification. Variable modifications included oxidation of methionine, N-terminal pyroglutamate from glutamine, loss of methionine and N-terminal acetylation. Protein Prospector score parameters were minimum protein score of 22, minimum peptide score of 15, and maximum expectation values of 0.01 for protein and 0.001 for peptide matches, resulting in a protein false discovery rate of 1.5%. Protein identification results from three biological replicates are reported with a spectral count as an approximation of protein abundance, along with percent sequence coverage and an expectation value for the probability of the protein identification[37,38].

Multiplex peptide cleavage assay. Purified peptidase concentration ranged from 10–100 nM, and biological samples were assayed at 20 μg/ml of total protein. To reproducibly detect all 124 intact peptides, three pools containing 52, 52 and 20 peptides were prepared and the concentration of each peptide in the assay was 500 nM, which is tenfold below the typical K_M for peptidases. Whenever enzymatic activity parameters such as specific activity, pH optima or cofactor requirements were known for a given peptidase, this information was used in the assay preparation. Each peptide pool was incubated at room temperature with peptidase, and aliquots were removed and acid quenched to pH 3 or less with formic acid (4% final) at defined time intervals. A control sample lacking enzyme was also prepared under identical conditions and quenched at the first and last time point of the assay to account for non-enzymatic degradation of the substrates. For enzymes with low pH optima, specific inhibitors such as pepstatin for aspartic acid peptidases and TA1 for eqolisin were used[39]. All of the peptides could be cleaved by one of four enzymes (cathepsin E, eqolisin, cruzain or matriptase), thereby suggesting that any potential secondary structure of the peptides does not appear to limit access of the sequence to proteolysis.

Peptide cleavage site identification by mass spectrometry. Cleavage site identification was performed using peptide sequencing by mass spectrometry. Samples containing 1–3 μg of total peptide (calculated as 60 μl of enzyme reaction containing peptide pools at 500 nM) were desalted using C18 ZipTips (Millipore) and rehydrated in 0.1% formic acid. Total peptide corresponding to 0.1 μg was injected on the column. For LC-MS/MS, a linear ion trap LTQ mass spectrometer (Thermo Scientific) equipped with an Ultimate HPLC, and Famos autoinjector (LC Packings) was used, with a C18 'Magic' column (Michrom Bioresources, Inc., 5-μm bead size, 0.3 × 150 mm Magic, 200 Å). The LC system was operated at 5 μL/min flow rate, and peptides were separated using a linear gradient over 42 min from 2% B to 40% B, with solvent A: 0.1% formic acid in water and solvent B: 0.1% formic acid in 70% acetonitrile. Survey scans were taken over 300–1,500 *m*/*z*, and the top three ions in the survey scan were subjected to a high-resolution MS 'zoom' scan of the precursor and then a CID fragmentation MS/MS scan. Each sample requires 1 h of mass spectrometry time, and therefore, a typical assay with four time points requires 12 h of instrument time.

Mass spectrometry peak lists were generated with PAVA software based on the Raw_Extract script from Xcalibur v.2.4 (Thermo Scientific), and data were searched using Protein Prospector software v.5.10.0 (UCSF). Database searches were performed against a defined library of 124 sequences. Data were searched with parent mass and fragment mass tolerances of 0.8 Da. For database searching, peptide sequences were matched with no enzyme specificity requirement and variable modifications including oxidation of Trp, Pro and Phe and N-terminal pyroglutamate from glutamine. For estimation of false discovery rate with this small database size, four different decoy databases containing the randomized sequences of the same 124 entries were concatenated to the original 124 entries to create a final database of 620 sequences. The data for replicate no-enzyme control samples were searched using this large random concatenated database. False discovery rate was calculated using

doi:10.1038/nmeth.2182

the formula: FDR = 0.25 × FP/(TP), where FP = false positives and TP = true positive peptides identified in the search[36]. Protein Prospector score thresholds were selected to be minimum protein score of 20, minimum peptide score of 15, and maximum expectation values were set to 0.1 for 'protein' and 0.05 for peptide matches; these settings resulted in a peptide false discovery rate of <0.17%. Cleavage-site data were extracted from Protein Prospector using the MSP-extractor software developed at UCSF. MSP-extractor identifies all cleaved bonds by matching the peptide products to the original substrate. The user can choose to output a defined number of residues at either side of the cleaved bond. If cleavage occurs close to the termini, 'X' is used to fill the void. This software is available at http://www.craiklab.ucsf.edu/extractor.html. Octapeptides corresponding to P4–P4′ residues were exported from MSP-extractor for this study and imported into iceLogo (http://code.google.com/p/icelogo/) as a positive data set. The negative set consisted of all possible cleavage sites in the library ($n = 1{,}612$). The difference in frequency of an amino acid in the positive and negative set is calculated as the percent difference. Only amino acids that are significantly over-represented or under-represented ($P \leq 0.05$) in the positive data set are illustrated in the iceLogo plots. Using this P value and the Wichura algorithm, the confidence interval was $[-1.96\sigma; 1.96\sigma]$, where σ is the s.d.[12]. Starting with a set of raw data files from multiple time points, a substrate signature can be generated for an enzyme or biological sample in ~30 min. Cleavage site sequences and a link to all RAW data files are listed in the **Supplementary Data** describing MSP-MS cleavages.

Kinetics calculations. MS acquisition, peak integration and data analysis were performed using Xcalibur software v.1.2 (Thermo Finnigan). Enzymatic progress curves were calculated from the peak areas of substrate and product species in the MS precursor scans, with percent conversion (Y) defined as 100 × [Product] / ([Product] + [Substrate]). Progress curves were modeled using the first-order kinetics formula $Y = \exp(-t \times k_{\mathrm{cat}}/K_{\mathrm{M}} \times [E_0])$, where E_0 is the total enzyme concentration, and fitted using the Marquardt method for nonlinear least-squares fitting to the enzyme kinetics model in GraphPad Prism v.5. Catalytic efficiency was solved from the overall rate by estimating total enzyme concentration and is reported as $k_{\mathrm{cat}}/K_{\mathrm{M}}$ with a standard deviation value for the quality of the data fit.

34. Salter, J.P. *et al.* Cercarial elastase is encoded by a functionally conserved gene family across multiple species of schistosomes. *J. Biol. Chem.* **277**, 24618–24624 (2002).
35. Chalkley, R.J., Baker, P.R., Medzihradszky, K.F., Lynn, A.J. & Burlingame, A.L. In-depth analysis of tandem mass spectrometry data from disparate instrument types. *Mol. Cell. Proteomics* **7**, 2386–2398 (2008).
36. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
37. Liu, H., Sadygov, R.G. & Yates, J.R. III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
38. Choi, H., Fermin, D. & Nesvizhskii, A.I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* **7**, 2373–2385 (2008).
39. O'Donoghue, A.J. *et al.* Inhibition of a secreted glutamic peptidase prevents growth of the fungus *Talaromyces emersonii*. *J. Biol. Chem.* **283**, 29186–29195 (2008).

doi:10.1038/nmeth.2182