Identification of Protease Specificity by Combining Proteome-Derived Peptide Libraries and Quantitative Proteomics

Martin L. Biniossek[1], Melanie Niemer[2], Ken Maksimchuk[3], Bettina Mayer[1], Julian Fuchs[4], Pitter F. Huesgen[5], Dewey G. McCafferty[3], Boris Turk[7], Guenther Fritz[8], Jens Mayer[9], Georg Haecker[10], Lukas Mach[2], Oliver Schilling[1,11,12]

This is the peer reviewed version of the following article published in *Molecular & Cellular Proteomics*. 2016; Jul;15(7):2515-24. doi: 10.1074/mcp.O115.056671.

http://www.mcponline.org/content/15/7/2515

1. Institute of Molecular Medicine and Cell Research, University of Freiburg, Freiburg, Germany
2. Department of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences, Vienna, Austria
3. Department of Biochemistry, Duke University Medical Center, Durham, NC 27708, United States of America
4. Institute of General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innsbruck, Austria
5. Central Institute for Engineering, Electronics and Analytics, ZEA-3 Analytics, Forschungszentrum Jülich, 52425 Jülich, Germany
6. Department of Chemistry, Duke University, B120 Levine Science Research Center Box 90346, 450 Research Drive, Durham, NC 27708-0346, United States of America

7. Department of Biochemistry and Molecular and Structural Biology, Jozef Stefan Institute, Ljubljana, Slovenia

8. Institute of Neuropathology, University of Freiburg, Breisacherstrasse 64, 79106 Freiburg

9. Department of Human Genetics and Center of Human and Molecular Biology, Medical Faculty, University of Saarland, Homburg, Germany

10. Institute for Medical Microbiology and Hygiene, University Medical Center Freiburg, Freiburg, Germany

11. BIOSS Centre for Biological Signaling Studies, University of Freiburg, D-79104 Freiburg, Germany

12. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

Corresponding Author:

Dr. Oliver Schilling, Stefan-Meier-Strasse 17, 79104 Freiburg, Germany. Tel: +49 761 203 9615 Email: oliver.schilling@mol-med.uni-freiburg.de

Running title: Tag-Free Identification of Protease Specificity

Abbreviations

C-terminal, carboxy-terminal; cDNA, complementary DNA; CPAF, chlamydial protease-like activity factor; HERV, human endogenous retrovirus; LC–MS/MS, liquid chromatography–tandem mass spectrometry; N-terminal, amino-terminal; P, non-prime side; P', prime-side; PICS, proteomic identification of protease cleavage sites

Summary

We present protease specificity profiling based on quantitative proteomics in combination with proteome-derived peptide libraries. Peptide libraries are generated by endoproteolytic digestion of proteomes without chemical modification of primary amines before exposure to a protease under investigation. After incubation with a test protease, treated and control libraries are differentially isotope-labeled using cost-effective reductive dimethylation. Upon analysis by liquid chromatograph – mass spectrometry, cleavage products of the test protease appear as semi-specific peptides that are enriched for the corresponding isotope label. We validate our workflow with two proteases with well-characterized specificity profiles: trypsin and caspase-3. We provide the first specificity profile of a protease encoded by a human endogenous retrovirus and for chlamydial protease-like activity factor (CPAF). For CPAF, we also highlight the structural basis of negative subsite cooperativity between subsites S1 and S2'. For "a disintegrin and metalloproteinase with thrombospondin motifs" (ADAMTS) -4, -5, and -15 we show a canonical preference profile, including glutamate in P1 and glycine in P3'. In total, we report nearly 4000 cleavage sites for seven proteases. Our protocol is fast, avoids enrichment or synthesis steps, and enables probing for lysine selectivity as well as subsite cooperativity. Due to its simplicity, we anticipate usability by most proteomic laboratories. (198 words)

Keywords

Degradome, peptide library, protease, specificity

INTRODUCTION

Proteolysis is an irreversible post-translational modification, ubiquitously shaping every proteome. Degradative proteolysis regulates proteome composition, mediates protein turnover, and ensures proteome integrity by removing misfolded proteins. At the same time, limited proteolysis yields stable cleavage products; thereby controlling enzyme or chemokine activity, assembly of structural proteins, and release of bioactive proteins from the cell surface ("shedding") or from larger precursors (1-3). Dysregulated proteolysis is a hallmark of numerous diseases with examples including tumor biology, neurodegeneration or hereditary skin diseases. In short, proteolytic processing is a key event in health and disease.

Proteases typically recognize substrates in an extended active-site cleft, involving multiple interactions between substrate side-chains and corresponding binding pockets of the protease. In many cases, these interactions occur both N- and C–terminally to the scissile peptide bond. This relation has been formalized in 1968 with the Schechter and Berger nomenclature (4). Here, substrate residues N-terminal to the scissile peptide bond are denoted as P1, P2, P3,… and residues C-terminal to the scissile peptide bond are denoted as P1', P2', P3',…; the corresponding subsites of the protease are denoted as S1, S2, S3 and S1', S2', S3', correspondingly.

Many proteases have strict subsite specificities. Prominent examples include trypsin (lysine or arginine in P1) or GluC (glutamate or aspartate in P1). On the other hand, numerous proteases are characterized by comparably broad subsite preferences, often encompassing multiple subsites. Examples include cysteine cathepsins or

matrix metalloproteases (5-7). Determination of protease specificity is an important step in biochemical protease characterization. It enables deorphanizing of previously uncharacterized proteases and probing protease activity *in vitro* (8). Knowledge of active-site specificity also serves to guide protease inhibitor development (9) and allows to identify structural origins of protease specificity and promiscuity (10, 11).

Several complementary experimental strategies exist for protease specificity profiling (12). Phage and bacterial display techniques employ genetic approaches and iterative enrichment to screen peptide pools with a large sequence variety for preferred cleavage motifs (13, 14). Positional scanning synthetic combinatorial libraries (PS-SCLs) represent a powerful biochemical approach for the determination of subsite selectivity (15). Synthetic mixture-based oriented peptide libraries represent a two-step strategy to characterize proteases with specificity profiles that include both prime- and non-prime sites (7). In peptide nucleic acid arrays, peptidic substrates are coupled to defined nucleic acid sequences, allowing spatial deconvolution of complex mixtures on complementary microarrays (16). In combination with *in vitro* translation, nucleic acid sequencing has been used to determine protease cleavage sites (17). Such a set-up has recently enabled the family-wide portrayal of MMP specificity determinants (18). Cysteine cathepsin specificity has also been investigated by a non-peptide approach termed fast profiling of protease specificity (FPPS) (19).

In recent years, proteome-derived peptide libraries have been introduced for the specificity profiling of proteases (5), including cysteine, serine, and metalloproteases

(5, 20, 21) with adaptions to enable multiplexed stable isotope tagging for kinetic investigations (22) as well as investigating the specificity of carboxypeptidases and Nα-acetyltransferases (23, 24). Using proteome-derived peptide libraries, Eckhard *et al.* recently reported thousands of matrix metalloprotease cleavage sites (25).

Here we present a tag–free, straightforward strategy to employ proteome-derived peptide libraries for protease specificity profiling. The technique is based on the comparison of differentially stable isotope labeled protease-treated peptide libraries and control samples. LC-MS/MS analysis allows relative quantitation of each isotope-labeled peptide and thereby specific selection of peptides that occur only in the protease-treated sample, indicating cleaved sequences. Notably, affinity enrichment or modification of primary amines prior to the actual profiling step is no longer required. The protocol enables identification of specificity towards unmodified lysine residues as well as probing of subsite cooperativity. Presently, 3502 of the 3988 (88 %) proteases annotated in the MEROPS database (version 9.12) (26) have less than 10 known substrates. This highlights that deorphanizing remains an important goal in protease research and underlines the need for straightforward approaches for protease cleavage site identification.

Experimental Procedures

Experimental Design and Statistical Rationale

A total of 12 specificity experiments were conducted. Semi-specific peptides that are enriched more than 8-fold (log2 “fold-change” = 3) are considered to unambiguously represent cleavage products of the proteases under investigation. This is validated

by the control experiments presented below, which accurately depict the specificity for trypsin and caspase-3.

Peptide Library Preparation

Escherichia coli MG1655 was grown in Luria–Bertani (LB) medium. Cells were lysed and lysates were digested by either trypsin or GluC as described elsewhere (27). GluC digests were performed in the presence of 1.0 μM tosyl phenylalanyl chloromethyl ketone and 1.0 μM tosyl-L-lysine chloromethyl ketone hydrochloride. LysC digestion was performed similarly to tryptic digestion but using LysC instead. Following digestion, primary amines were not modified. The peptide digest was further purified by C18 solid phase extraction (Sep-Pak, Waters) according to the manufacturer's instructions. Aliquots of the peptide library were stored in water at -80 ° C.

Proteases

Trypsin (Worthington, sequencing grade), GluC (Sigma, sequencing grade), LysC (Sigma, sequencing grade), and ADAMTS-4, -5, -15 (RnD Systems) were purchased commercially. Caspase-3 and chlamydial protease-like activity factor (CPAF) were prepared as described elsewhere (28, 29) as was a protease encoded by a human endogenous retrovirus belonging to the HERV-K(HML-2) group (30, 31).

Specificity Determination

Protease:library ratios, incubation times and temperature are stated in the corresponding sections below. Between 100 and 300 µg of peptide library were used per assay. Incubation buffers were 100 mM HEPES (pH 8.0) for trypsin; 50 mM HEPES (pH 7.4), 100 mM NaCl, 1 mM EDTA, 10 mM DTT, 10% (wt/vol) sucrose for caspase-3; 50 mM MES pH 5.0, 1M NaCl, 1mM EDTA for HERV-K(HML-2) protease; 50 mM HEPES (pH 7.5), 150 mM NaCl, 5 mM DTT for chlamydial protease-like activity factor; 50 mM HEPES (pH 7.5), 100 mM NaCl, 5 mM $CaCl_2$ for ADAMTS-4, -5, and -15. After incubation, the test protease was heat-inactivated. Protease-treated and control samples were isotopically labeled by reductive dimethylation using H_2CO or $D_2C^{13}O$, respectively, as described elsewhere (32, 33). Following labeling, protease-treated and control samples were mixed at equal ratios and desalted by C18 solid phase extraction (Sep-Pak, Waters) according to manufacturer's instructions.

Nanoflow−HPLC−MS/MS

Samples were analyzed on a Q-Exactive plus (Thermo Scientific) mass spectrometer coupled to an Easy nanoLC 1000 (Thermo Scientific) with a flow rate of 300 nl / min. Buffer A was 0.5 % formic acid, and buffer B was 0.5 % formic acid in acetonitrile (water and acetonitrile were at least HPLC gradient grade quality). A gradient of increasing organic proportion was used for peptide separation (5 – 40 % acetonitrile in 80 min). The analytical column was an Acclaim PepMap column (Thermo Scientific), 2 µm particle size, 100 Å pore size, length 150 mm, inner diameter 50 µm. The mass spectrometer operated in data dependent mode with a top 10 method at a mass range of 300 − 2000.

Data analysis

Raw LC-MS/MS data was converted to the mzXML format (34), using msconvert (35) with centroiding of MS1 and MS2 data and deisotoping of MS2 data. For spectrum to sequence assignment X! Tandem (Version 2013.09.01) was used (36). The *E. coli* proteome database (strain K12, reference proteome) was used as described previously (37), consisting of 4304 protein entries and 8608 randomized sequences, derived from the original *E. coli* proteome entries. The decoy sequences were generated with the software DB toolkit (38). X! Tandem parameters included: precursor mass error of ± 10 ppm, fragment ion mass tolerance of 20 ppm, semi-tryptic, semi-GluC or semi-LysC specificity with up to one missed cleavage, static residue modifications: cysteine carboxyamidomethylation (+57.02 Da); lysine and N-terminal dimethylation (light formaldehyde 28.03 Da; heavy formaldehyde 34.06 Da or 36.08 Da, respectively); no variable modifications. X!Tandem results were further validated by PeptideProphet (39) at a confidence level of > 95 %. For relative peptide quantification, XPRESS (40) was used. Mass tolerance for quantification was 0.015 Da. An in-house PERL computer script (downloadable at http://www.mol-med.uni-freiburg.de/mom/schilling/TAILS_v21_xpress-only/at_download/file) then filters for semi-specific peptides, with their N- or C-terminus generated by the initial digestion protease and the respective other terminus derived from cleavage by the test protease. Similar to the original PICS procedure (5), the script then determines bioinformatically, through database lookup, the corresponding prime- or non-prime sequence. Web-PICS was used to generate heat-map style representation of protease specificity (41).

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (42) via the PRIDE partner repository with the dataset identifier PXD002785 (Username: reviewer44222@ebi.ac.uk, Password: e0vfaRaU). The identifier for the HERV-K(HML-2) dataset is PXD003833 (Username: reviewer41375@ebi.ac.uk, Password: iZv3YslS). In order to provide access to annotated spectra, pepXML files, along with mzXML files have been deposited.

Molecular modeling of CPAF peptide complex

The structure of mature CPAF (pdb code 3DOR) and active site mutant CPAF-S499A mutant in complex with a peptide (pdb code 3DPN) (43) served as templates for molecular modeling. The corresponding residues P1, P2, P3, P1', P2' and P3' were changed with Coot (44) and figures were generated with PyMol (45). The PyMOL Molecular Graphics System, Version 1.3r1.)

Results and Discussion

Overview and Proof of Concept

Our procedure is outlined in **Fig. 1**. Briefly, cellular proteomes, such as cellular or bacterial lysates are harvested and digested by specific endopeptidases ("digestion protease"), such as trypsin, GluC or LysC, yielding a proteome-derived peptide library. After inactivation of the digestion protease, the peptide library is divided into a control sample and a sample for incubation with a protease under investigation ("test protease"). Following this incubation step, the samples are differentially labeled, for example by reductive dimethylation using different stable isotopes of formaldehyde

(46). The samples are then mixed, desalted, and analyzed by LC-MS/MS analysis. In data analysis, the procedure focuses on semi-specific peptides that are more abundant in the sample treated with the test protease. Semi-specific refers to peptides for which only the N- or C-terminus represents cleavage by the digestion protease, while the other terminus is generated by a protease with a different specificity. By focusing on semi-specific peptides, which are more abundant in the sample treated with the test protease, the procedure investigates the specificity of this enzyme. If only the C-terminus of such a peptide is derived from treatment with the digestion protease, the N-terminus was likely generated by the test protease and the peptide thus constitutes the prime-site cleavage sequence. The requirement that the semi-specific peptide needs to be present only in the test protease treated sample, but not in the control, excludes artifacts arising from unspecific background proteolysis in the proteome. Similar to the original PICS procedure, the corresponding non-prime-site cleavage sequences are bioinformatically retrieved through a database lookup. We have previously shown that this combination of experimentally and bioinformatically derived sequences is suitable for profiling of protease specificities (5). We have now expanded the strategy to enable bioinformatic retrieval of prime-site cleavage sequences based on experimentally determined non-prime-site cleavage sequences (i.e. N-terminus generated by the digestion protease, C-terminus produced by the test protease). In a last step, all cleaved sequences are aligned, optionally normalized, and the aggregate specificity motifs visualized, e.g. as heatmaps (41) or sequence logos (47).

As an initial proof-of-concept study we generated a proteome-derived peptide library from *Escherichia coli* lysates by GluC digestion. We then incubated this library with

trypsin at a protease:library ratio of 1:500 for either 1 h or 16 h at 37 ° C. Upon LC-MS/MS data analysis, we used an 8-fold increase in abundance to discriminate semi-specific peptides that are generated by trypsin incubation from background signals. 786 (1 h trypsin incubation, sequences in Sup. Tables 1a and 1b) and 819 (16 h trypsin incubation, sequences in Sup. Tables 2a and 2b) semi-specific peptides fulfilled this criterion and were further analyzed. As shown in **Fig. 2a**, both incubation times clearly result in specificity profiles that reflect the prototypical trypsin selectivity for arginine and lysine in P1. This result nicely validates our experimental strategy and highlights its suitability for assessing the capacity of proteases to act on lysine-containing substrates.

We further chose human caspase-3 for a second proof-of-concept experiment. Caspase-3 features subsites with differing grades of specificity, including strict selectivity (P1D), moderate selectivity (P4D), and partial preference (P2V). Using our modified PICS strategy, we profiled recombinant human caspase-3 with tryptic *E. coli* peptide libraries. We identified 63 caspase-3 cleavage sites (sequences in Sup. Tables 3a and 3b). The resulting specificity profile (**Fig. 2b**) matches prototypical caspase specificity (48, 49). In P3, methionine rather than glutamate appears to be preferred. This is corroborated by a phage-display study on caspase-3, which identified DLVD rather than DEVD as the preferred sequence motif for caspase-3 (50).

Specificity Profiling of HERV-K(HML-2) Protease

The human genome contains numerous sequences that originate from germ line insertions of exogenous retroviruses, so-called human endogenous retroviruses (HERVs). HERV sequences, that is, “fossilized” retroviral sequences, account for approximately 8% of the human genome (51). Similar to present-day exogenous retroviruses, a relatively small number of HERV sequences in the human genome comprises ORFs for typical retroviral proteins, among them a retroviral protease. The protease of an evolutionarily young HERV group, named HERV-K(HML-2), bears similarity to the aspartyl protease of viruses such as the human immunodeficiency virus (HIV) (52) and it has been suggested that the HERV-K(HML-2) protease may functionally complement HIV-1 protease (53).

Despite its interesting biology, little is known about the specificity of the HERV-K(HML-2) protease. Here, we sought to elucidate its preferred cleavage site motif using proteome-derived peptide libraries. Incubation of a tryptic peptide library with recombinant, purified HERV-K(HML-2) protease (protease:library ratio 1:100) yielded 95 cleavage sites (specificity profile in **Fig. 2c,** sequences in Sup. Tables 4a and 4b). Aromatic residues (Phe, Trp, Tyr) in P1 constitute the primary specificity determinant of HERV-K(HML-2) protease. This finding is corroborated by its ability to process HERV-K(HML-2) Gag protein at Tyr134 and Phe252 (51, 54, 55). At the same time, we notice that there is a substantial proportion (21 %) of cleavage sites with a P1 Gly. In line with our observation, HERV-K(HML-2) protease also cleaves the HERV-K(HML-2) Gag protein at Gly532 *in vitro* (51, 54, 55). Further minor specificity determinants are represented by a preference for either aliphatic residues in P2 and for aromatic or aliphatic residues in P1’. The acidic residues Asp and Glu are found in position P2 in 32 % of all cleavage sites. However, due to the low pH of the reaction condition (pH 5.0), they were likely present in their protonated, non-charged form.

Overall, the specificity profile of HERV-K(HML-2) protease bears similarity to HIV-1 protease, which also displays a preference for aliphatic residues in P2 as well as for aromatic residues in P1 and P1' (5). HIV-1 protease is a prototypical case for subsite cooperativity (5, 56), In particular, bulky residues in P1 (e.g. Phe) result in a preference for small residues in P3 (e.g. Ala) and *vice-versa*. This behavior is also found for HERV-K(HML-2) protease (data not shown). Interestingly, the similar specificity profiles of HERV-K(HML-2) and HIV-1 proteases persist despite a rather limited sequence homology of the two enzymes (53).

Using the HERV-K(HML-2) dataset, we also investigated whether the appearance of cleavage products coincides with the depletion of the original tryptic peptides. For 22 of the 95 cleavage sequences, we also identified and quantified the original tryptic peptide. As expected, protease treatment resulted in depletion of the substrate peptides; for HERV-K(HML-2) the average depletion was more than 4-fold. This observation further corroborates our strategy.

Specificity and Subsite Cooperativity of Chlamydial Protease-like Activity Factor

Chlamydial protease-like activity factor (CPAF, MEROPS ID S41.011) is a serine protease secreted by the obligate intracellular pathogen *Chlamydia trachomatis*. MEROPS, the peptidase database, does not yet list any peptidic or proteinaceous CPAF substrates. However, autocatalytic cleavage of CPAF has been suggested (43), occurring between Met^{242} and Arg^{243}, Met^{264} and Val^{265}; as well as between Ser^{283} and Gly^{284}.

For the present work, we used purified recombinant CPAF. To probe CPAF specificity we generated a proteome-derived peptide library by GluC digestion of

Escherichia coli lysates. CPAF incubation (protease:library ratio of 1:500 (wt/wt)) at pH 7.5 for 1 h and 16 h, respectively, at 37 ° C resulted in the identification of 501 cleavage sequences (1 h incubation sequences in Sup. Tables 5a and 5b) and 680 cleavage sequences (16 h incubation sequences in Sup. Table 6a and 6b). Both incubation times yielded highly similar specificity profiles (**Fig. 3a**). P1 constitutes the major specificity determinant with a preference for alanine, glycine or methionine. P1M accounts for 8.4 % (1 h incubation) or 8.5 % (16 h incubation) of the CPAF cleavage sites. Due to its rare natural occurrence, this corresponds to a strong over-representation. The selectivity for P1M is corroborated by the three previously reported CPAF self-cleavage sites, two of which feature a P1M (43). Further positional preferences include tyrosine in P3, aliphatic residues (isoleucine and valine) in P2', and proline in P3'. Mixed specificity in a given subsite has been previously observed for example for cathepsin B, which prefers small or aromatic residues in P1' (20).

We further investigated whether CPAF preference for P1M is positively or negatively correlated to other positional preferences. We noticed that P1M negatively correlates to P2'I and vice-versa (**Fig. 3b, c**). In order to further characterize the substrate recognition by CPAF we analyzed the structure of the active site and created a molecular model of a complex of CPAF with two model substrates containing either the motif VM↓VA or VM↓VI (P2 – P2', "↓" indicates the cleavage site, **Fig. 3d**). The side chain of methionine in the P1 position reaches into a large hydrophobic pocket close to the active site serine (Ser499) in CPAF. This interaction places the peptide bond in optimal distance to the active site serine. The side chain of valine in position P1' points away from the protein surface. The side chain of the residue in position P2' points towards the protein surface. However, there is limited space and only residues

with a short side chain such as alanine fit well. Side chains of larger size such as isoleucine clash with the protein matrix and give a rationale for the negative correlation between isoleucine in position P2' and methionine in P1. Such cooperativity for subsites with mixed specificities has been previously observed, e.g. in the case of cathepsin B (20) and other proteases (57). Generally, our technique enabled detailed, high-content profiling of CPAF specificity and unraveled subsite cooperativity for the mixed P1 specificity. We conclude that CPAF is a protease with multiple enzyme-substrate interactions and comparably broad subsite specificities.

Canonical Specificity of ADAMTS Proteases

The "A Disintegrin And Metalloproteinase with Thrombospondin Motifs" (ADAMTS) family comprises 19 mammalian members (58). These are secreted proteases with a variety of ancillary domains. Members of the ADAMTS family are involved in a plethora of functions in health and disease. Here we focus on ADAMTS-4, -5, and -15. For ADAMTS-15, MEROPS does not yet report any cleavage sites and we present its first specificity profile. We employed tryptic and LysC libraries to profile the active site specificity of ADAMTS-4, -5 and -15, using an enzyme:library ratio of 1:50. Earlier reports on specificity profiling of related proteases of the "a disintegrin and a metalloproteinase" (ADAM) family employed enzyme:library ratios of 1:10 (59). Our own studies highlighted that specificity profiling with proteome derived peptide libraries is not prone to overdigestion or blurred specificity motifs as a result of elongated incubation times (20) or elevated enzyme:library ratios (5). For ADAMTS-4, we identified 180 tryptic and 335 LysC-based cleavage sites (Sup. Tables 7a/b and 8a/b); for ADAMTS-5, we identified 215 tryptic and 246 LysC-based cleavage

sites (Sup. Tables 9a/b and 10a/b); for ADAMTS-15 the corresponding numbers were 77 tryptic and 63 LysC-based cleavage sites (Sup. Tables 11a/b and 12a/b).

The specificity profiles are summarized in **Fig 4**. All three ADAMTS proteases yielded similar specificity profiles. Glutamate in P1 constitutes the major specificity determinant. ADAMTS-4 is unique in also displaying a minor preference for tyrosine in P1. Glycine in P2' is prominently preferred by all three ADAMTS proteases. In P2 there is a common, but less pronounced preference for glutamine and in P1' there is a shared preference for alanine. The LysC library further unraveled a preference for arginine in P2', which is shared by all three ADAMTS proteases. Generally, there is good agreement between the tryptic and LysC-based specificity profiles, which further testifies to the robustness of our approach. For ADAMTS-4 and -5, our specificity profiles are in good agreement with cleavage site data deposited in MEROPS. Interestingly, the canonical ADAMTS specificity is distinct from the specificity profiles of further metzincin proteases such as MMPs, ADAMs or meprins.

Conclusion

We present a simple and robust protocol for profiling of protease specificity, including subsite cooperativity, with proteome-derived peptide libraries. We showcase and validate our technique through characterization of the well-known trypsin and caspase-3 specificity. We demonstrate its wider applicability by deorphanizing HERV-K(HML-2) protease and CPAF, which are both proteases with previously unknown specificity. Furthermore, we determine the canonical specificity profile of ADAMTS proteases.

Acknowledgement

O.S. is supported by grants of the Deutsche Forschungsgemeinschaft (DFG) (SCHI 871/2, SCHI 871/5, SCHI 871/6, GR 1748/6, and INST 39/900-1) and the SFB850 (Project B8), a starting grant of the European Research Council (Programme “Ideas” – Call identifier: ERC-2011- StG 282111-ProteaSys), and the Excellence Initiative of the German Federal and State Governments (EXC 294, BIOSS). M.N. and L.M. acknowledge support by the Austrian Science Fund (FWF): project W1224-B09. B.T. was supported by a grant from Slovene Research Agency (P1-0140). G. H.’s work on CPAF is supported by the DFG (SPP1580). D.G.M. acknowledges support by the NIH National Institute of Allergy and Infectious Disease (research grant number 1RO1-AI107951). K.M. is the recipient of an NSF Predoctoral Fellowship. J.M. is supported by DFG (MA2298/12-1). The authors thank Franz Jehle, Esther Maldener and Birgit Herrmann for excellent technical assistance.

COMPETING INTERESTS

The authors declare no competing interests.

References

1. Lai, Z. W., Petrera, A., and Schilling, O. (2014) The emerging role of the peptidome in biomarker discovery and degradome profiling. *Biol Chem* 0
2. Petrera, A., Lai, Z. W., and Schilling, O. (2014) Carboxyterminal protein processing in health and disease: key actors and emerging technologies. *J Proteome Res* 13, 4497-4504
3. Shahinian, H., Tholen, S., and Schilling, O. (2013) Proteomic identification of protease cleavage sites: cell-biological and biomedical applications. *Expert Rev Proteomics* 10, 421-433

4. Schechter, I., and Berger, A. (1968) On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem Biophys Res Commun* 32, 898-902
5. Schilling, O., and Overall, C. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol* 26, 685-694
6. Ratnikov, B. I., Cieplak, P., Gramatikoff, K., Pierce, J., Eroshkin, A., Igarashi, Y., Kazanov, M., Sun, Q., Godzik, A., Osterman, A., Stec, B., Strongin, A., and Smith, J. W. (2014) Basis for substrate recognition and distinction by matrix metalloproteinases. *Proc Natl Acad Sci USA* 111, E4148-4155
7. Turk, B. E., Huang, L. L., Piro, E. T., and Cantley, L. C. (2001) Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat Biotechnol* 19, 661-667
8. Perera, N. C., Schilling, O., Kittel, H., Back, W., Kremmer, E., and Jenne, D. E. (2012) NSP4, an elastase-related protease in human neutrophils with arginine specificity. *Proc Natl Acad Sci USA* 109, 6229-6234
9. Drag, M., and Salvesen, G. S. (2010) Emerging principles in protease-based drug discovery. *Nat Rev Drug Discov* 9, 690-701
10. Fuchs, J. E., von Grafenstein, S., Huber, R. G., Margreiter, M. A., Spitzer, G. M., Wallnoefer, H. G., and Liedl, K. R. (2013) Cleavage entropy as quantitative measure of protease specificity. *PLoS Comput Biol* 9, e1003007
11. Fuchs, J. E., von Grafenstein, S., Huber, R. G., Kramer, C., and Liedl, K. R. (2013) Substrate-driven mapping of the degradome by comparison of sequence logos. *PLoS Comput Biol* 9, e1003353
12. Turk, B., Turk, D., and Turk, V. (2012) Protease signalling: the cutting edge. *EMBO J* 31, 1630-1643
13. Matthews, D. J., Goodman, L. J., Gorman, C. M., and Wells, J. A. (1994) A survey of furin substrate specificity using substrate phage display. *Protein Sci* 3, 1197-1205
14. Boulware, K. T., and Daugherty, P. S. (2006) Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proc Natl Acad Sci U S A* 103, 7583-7588
15. Rano, T. A., Timkey, T., Peterson, E. P., Rotonda, J., Nicholson, D. W., Becker, J. W., Chapman, K. T., and Thornberry, N. A. (1997) A combinatorial approach for determining protease specificities: application to interleukin-1beta converting enzyme (ICE). *Chem Biol* 4, 149-155
16. Winssinger, N., Damoiseaux, R., Tully, D. C., Geierstanger, B. H., Burdick, K., and Harris, J. L. (2004) PNA-encoded protease substrate microarrays. *Chem Biol* 11, 1351-1360
17. Kozlov, I. A., Thomsen, E. R., Munchel, S. E., Villegas, P., Capek, P., Gower, A. J., Pond, S. J., Chudin, E., and Chee, M. S. (2012) A highly scalable peptide-based assay system for proteomics. *PLoS One* 7, e37441
18. Kukreja, M., Shiryaev, S. A., Cieplak, P., Muranaka, N., Routenberg, D. A., Chernov, A. V., Kumar, S., Remacle, A. G., Smith, J. W., Kozlov, I. A., and Strongin, A. Y. (2015) High-Throughput Multiplexed Peptide-Centric Profiling Illustrates Both Substrate Cleavage Redundancy and Specificity in the MMP Family. *Chem Biol* 22, 1122-1133
19. Vizovisek, M., Vidmar, R., Van Quickelberghe, E., Impens, F., Andjelkovic, U., Sobotic, B., Stoka, V., Gevaert, K., Turk, B., and Fonovic, M. (2015) Fast profiling of protease specificity reveals similar substrate specificities for cathepsins K, L and S. *Proteomics* 15, 2479-2490

20. Biniossek, M. L., Nägler, D. K., Becker-Pauly, C., and Schilling, O. (2011) Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S. *J Proteome Res* 10, 5363-5373
21. Becker-Pauly, C., Barré, O., Schilling, O., Auf dem Keller, U., Ohler, A., Broder, C., Schütte, A., Kappelhoff, R., Stöcker, W., and Overall, C. (2011) Proteomic analyses reveal an acidic prime side specificity for the astacin metalloprotease family reflected by physiological substrates. *Mol Cell Proteomics* 10, M111.009233-M009111.009233
22. Jakoby, T., van den Berg, B. H., and Tholey, A. (2012) Quantitative protease cleavage site profiling using tandem-mass-tag labeling and LC-MALDI-TOF/TOF MS/MS analysis. *J Proteome Res* 11, 1812-1820
23. Van Damme, P., Evjenth, R., Foyn, H., Demeyer, K., De Bock, P. J., Lillehaug, J. R., Vandekerckhove, J., Arnesen, T., and Gevaert, K. (2011) Proteome-derived peptide libraries allow detailed analysis of the substrate specificities of N(alpha)-acetyltransferases and point to hNaa10p as the post-translational actin N(alpha)-acetyltransferase. *Mol Cell Proteomics* 10, M110 004580
24. Tanco, S., Lorenzo, J., Garcia-Pardo, J., Degroeve, S., Martens, L., Aviles, F. X., Gevaert, K., and Van Damme, P. (2013) Proteome-derived peptide libraries to study the substrate specificity profiles of carboxypeptidases. *Mol Cell Proteomics* 12, 2096-2110
25. Eckhard, U., Huesgen, P. F., Schilling, O., Bellac, C. L., Butler, G. S., Cox, J. H., Dufour, A., Goebeler, V., Kappelhoff, R., Keller, U. A., Klein, T., Lange, P. F., Marino, G., Morrison, C. J., Prudova, A., Rodriguez, D., Starr, A. E., Wang, Y., and Overall, C. M. (2016) Active site specificity profiling of the matrix metalloproteinase family: Proteomic identification of 4300 cleavage sites by nine MMPs explored with structural and synthetic peptide cleavage analyses. *Matrix Biol* 49, 37-60
26. Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 42, D503-509
27. Schilling, O., Huesgen, P. F., Barré, O., Auf dem Keller, U., and Overall, C. (2011) Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. *Nat Protoc* 6, 111-120
28. Stennicke, H. R., and Salvesen, G. S. (1999) Caspases: preparation and characterization. *Methods* 17, 313-319
29. Bednar, M. M., Jorgensen, I., Valdivia, R. H., and McCafferty, D. G. (2011) Chlamydia protease-like activity factor (CPAF): characterization of proteolysis activity in vitro and development of a nanomolar affinity CPAF zymogen-derived inhibitor. *Biochemistry* 50, 7441-7443
30. Kuhelj, R., Rizzo, C. J., Chang, C. H., Jadhav, P. K., Towler, E. M., and Korant, B. D. (2001) Inhibition of human endogenous retrovirus-K10 protease in cell-free and cell-based assays. *J Biol Chem* 276, 16674-16682
31. Mayer, J., Sauter, M., Racz, A., Scherer, D., Mueller-Lantzsch, N., and Meese, E. (1999) An almost-intact human endogenous retrovirus K on human chromosome 7. *Nat Genet* 21, 257-258
32. Tholen, S., Biniossek, M. L., Gansz, M., Ahrens, T. D., Schlimpert, M., Kizhakkedathu, J. N., Reinheckel, T., and Schilling, O. (2014) Double deficiency of cathepsins B and L results in massive secretome alterations and suggests a degradative cathepsin-MMP axis. *Cell Mol Life Sci* 71, 899-916

33. Shahinian, H., Loessner, D., Biniossek, M. L., Kizhakkedathu, J. N., Clements, J. A., Magdolen, V., and Schilling, O. (2014) Secretome and degradome profiling shows that Kallikrein-related peptidases 4, 5, 6, and 7 induce TGFbeta-1 signaling in ovarian cancer cells. *Mol Oncol* 8, 68-82
34. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 22, 1459-1466
35. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534-2536
36. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467
37. Biniossek, M. L., and Schilling, O. (2012) Enhanced identification of peptides lacking basic residues by LC-ESI-MS/MS analysis of singly charged peptides. *Proteomics* 12, 1303-1309
38. Martens, L., Vandekerckhove, J., and Gevaert, K. (2005) DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* 21, 3584-3585
39. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74, 5383-5392
40. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* 19, 946-951
41. Schilling, O., Auf dem Keller, U., and Overall, C. (2011) Factor Xa subsite mapping by proteome-derived peptide libraries improved using WebPICS, a resource for proteomic identification of cleavage sites. *Biol Chem* 392, 1031-1037
42. Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P. A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H. J., Albar, J. P., Martinez-Bartolome, S., Apweiler, R., Omenn, G. S., Martens, L., Jones, A. R., and Hermjakob, H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32, 223-226
43. Huang, Z., Feng, Y., Chen, D., Wu, X., Huang, S., Wang, X., Xiao, X., Li, W., Huang, N., Gu, L., Zhong, G., and Chai, J. (2008) Structural basis for activation and inhibition of the secreted chlamydia protease CPAF. *Cell Host Microbe* 4, 529-542
44. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66, 486-501
45. Schrodinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8.
46. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S., and Heck, A. J. (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* 4, 484-494
47. 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

48. 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
49. Thornberry, N. A., Rano, T. A., Peterson, E. P., Rasper, D. M., Timkey, T., Garcia-Calvo, M., Houtzager, V. M., Nordstrom, P. A., Roy, S., Vaillancourt, J. P., Chapman, K. T., and Nicholson, D. W. (1997) 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* 272, 17907-17911
50. Lien, S., Pastor, R., Sutherlin, D., and Lowman, H. B. (2004) A substrate-phage approach for investigating caspase specificity. *Protein J* 23, 413-425
51. Mayer, J., and Meese, E. (2005) Human endogenous retroviruses in the primate lineage and their influence on host genomes. *Cytogenet Genome Res* 110, 448-456
52. Schommer, S., Sauter, M., Krausslich, H. G., Best, B., and Mueller-Lantzsch, N. (1996) Characterization of the human endogenous retrovirus K proteinase. *J Gen Virol* 77 (Pt 2), 375-379
53. Towler, E. M., Gulnik, S. V., Bhat, T. N., Xie, D., Gustschina, E., Sumpter, T. R., Robertson, N., Jones, C., Sauter, M., Mueller-Lantzsch, N., Debouck, C., and Erickson, J. W. (1998) Functional characterization of the protease of human endogenous retrovirus, K10: can it complement HIV-1 protease? *Biochemistry* 37, 17137-17144
54. George, M., Schwecke, T., Beimforde, N., Hohn, O., Chudak, C., Zimmermann, A., Kurth, R., Naumann, D., and Bannert, N. (2011) Identification of the protease cleavage sites in a reconstituted Gag polyprotein of an HERV-K(HML-2) element. *Retrovirology* 8, 30
55. Kraus, B., Boller, K., Reuter, A., and Schnierle, B. S. (2011) Characterization of the human endogenous retrovirus K Gag protein: identification of protease cleavage sites. *Retrovirology* 8, 21
56. Ridky, T. W., Cameron, C. E., Cameron, J., Leis, J., Copeland, T., Wlodawer, A., Weber, I. T., and Harrison, R. W. (1996) Human immunodeficiency virus, type 1 protease substrate specificity is limited by interactions between substrate amino acids bound in adjacent enzyme subsites. *J Biol Chem* 271, 4709-4717
57. Ng, N. M., Pike, R. N., and Boyd, S. E. (2009) Subsite cooperativity in protease specificity. *Biol Chem* 390, 401-407
58. Kelwick, R., Desanlis, I., Wheeler, G. N., and Edwards, D. R. (2015) The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. *Genome Biol* 16, 113
59. Tucher, J., Linke, D., Koudelka, T., Cassidy, L., Tredup, C., Wichert, R., Pietrzik, C., Becker-Pauly, C., and Tholey, A. (2014) LC-MS based cleavage site profiling of the proteases ADAM10 and ADAM17 using proteome-derived peptide libraries. *J Proteome Res* 13, 2205-2214

Figure Legends

Figure 1: Overview of workflow. Peptide libraries are prepared by specific endoproteolytic digestion of complex proteomes, such as cell lysates. Following inactivation of the digestion protease and clean-up, the peptide library is split in two parts. One aliquot is incubated with the protease under investigation, the second aliquot serves as a control. After incubation with the test protease, the peptide library aliquots are differentially labeled using stable isotope compounds, such as dimethylation of primary amines with either light ($C^{12}H_2O$) or heavy ($C^{13}D_2O$) formaldehyde. Equal amounts of the labeled protease-treated and control library are mixed and analyzed by LC-MS/MS. Cleavage events lead to semi-specific peptides enriched in the sample treated with the test protease. The matching prime- or non-prime sequences are derived bioinformatically by database searches, similar to the original PICS strategy (5).

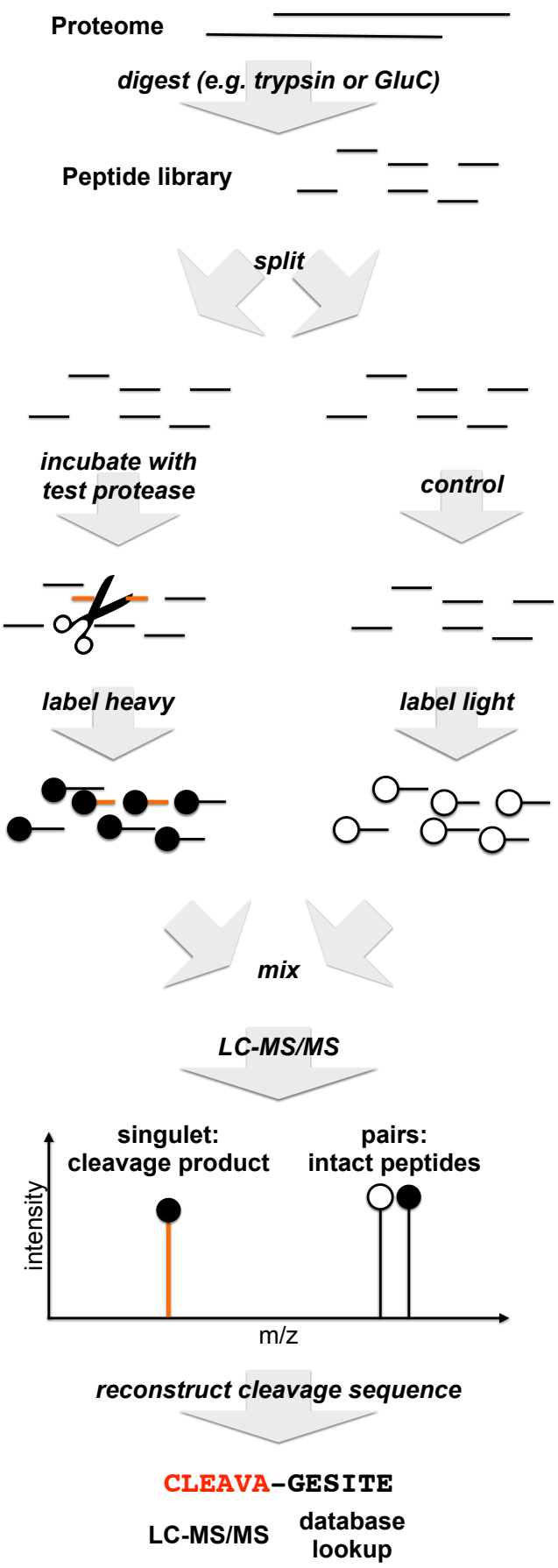
Figure 2a: Specificity profiling of trypsin using a GluC peptide library. The protease:library ratio was 1:500 (wt/wt). Incubation at pH 8.0 occurred for either 1 h or 16 h at 37 ° C. The histograms show the fold-change value distribution (log_2 of label ratios for trypsin/control) of the semi-specific peptides. Semi-specific peptides with a more than 8-fold enrichment (log2 fold-change value > 3) for the protease-treated sample (here: trypsin) are considered to represent specific cleavage events mediated by the test protease. These were used for reconstruction of the substrate cleavage sites, which were aligned and summarized as heat maps clearly showing the expected stringent trypsin specificity with arginine and lysine and P1. **2b**: Specificity profiling of human caspase-3. The protease:library ratio was 1:300 (wt/wt). Incubation at pH 7.4 occurred for 3 h at 37 ° C. The specificity heatmap is in line with canonical description of caspase-3 specificity, with the exception of the P3 position.

However, caspase-3 affinity for aliphatic residues in P3 and a preference for DLVD over the DEVD sequence present in common synthetic caspase-3 substrates have also been reported by others (50). **2d**: Specificity profiling of human endogenous retrovirus HERV-K(HML-2) protease. The protease:library ratio was 1:100 (wt/wt). Incubation at pH 5.0 occurred for 16 h at 37 °C. P1 constitutes the major specificity determinant with a preference for aromatic residues.

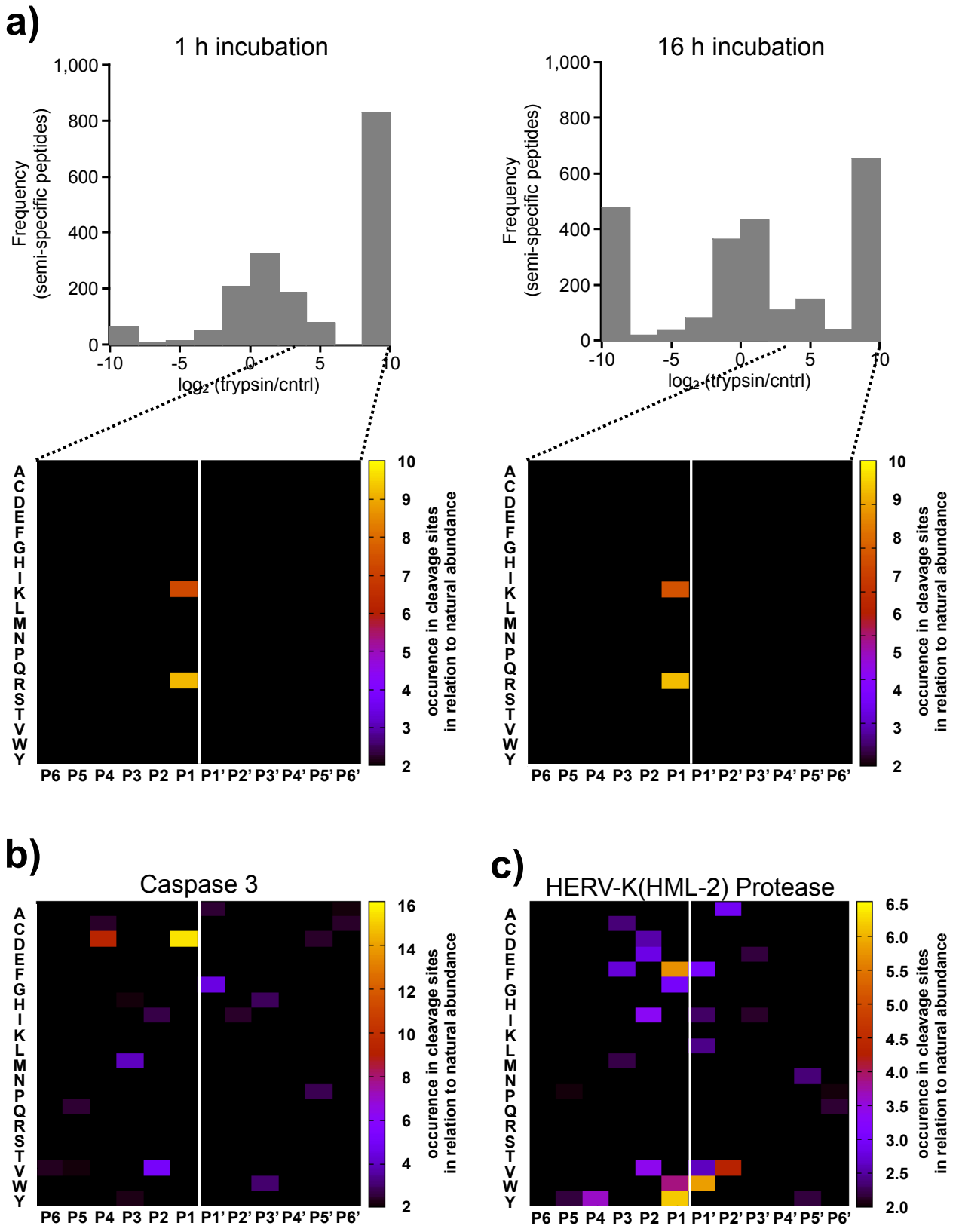
Figure 3a: Specificity profiling of chlamydial protease-like activity factor. The protease:library ratio was 1:500 (wt/wt). Incubation at pH 7.5 occurred for either 1 h or 16 h at 37 ° C. P1 constitutes the major specificity determinant with a mixed preference for either small (alanine, glycine) or aliphatic (methionine) residues. **3b/c**: Negative correlation between P1 methionine and P2' isoleucine. **3d:** Structural modeling of the peptide sequences VM↓VA or VM↓VI ("↓" indicates the cleavage site) in the active site of CPAF ("Ser" denotes active site serine 499).

Figure 4: Specificity profiling of proteases of the "A Disintegrin And Metalloproteinase with Thrombospondin Motifs" (ADAMTS) Family. The protease:library ratio was 1:50 (wt/wt). Incubation at pH 7.5 occurred for 16 h at 37 ° C. P1 constitutes the major specificity determinant with a canonical preference for glutamate.

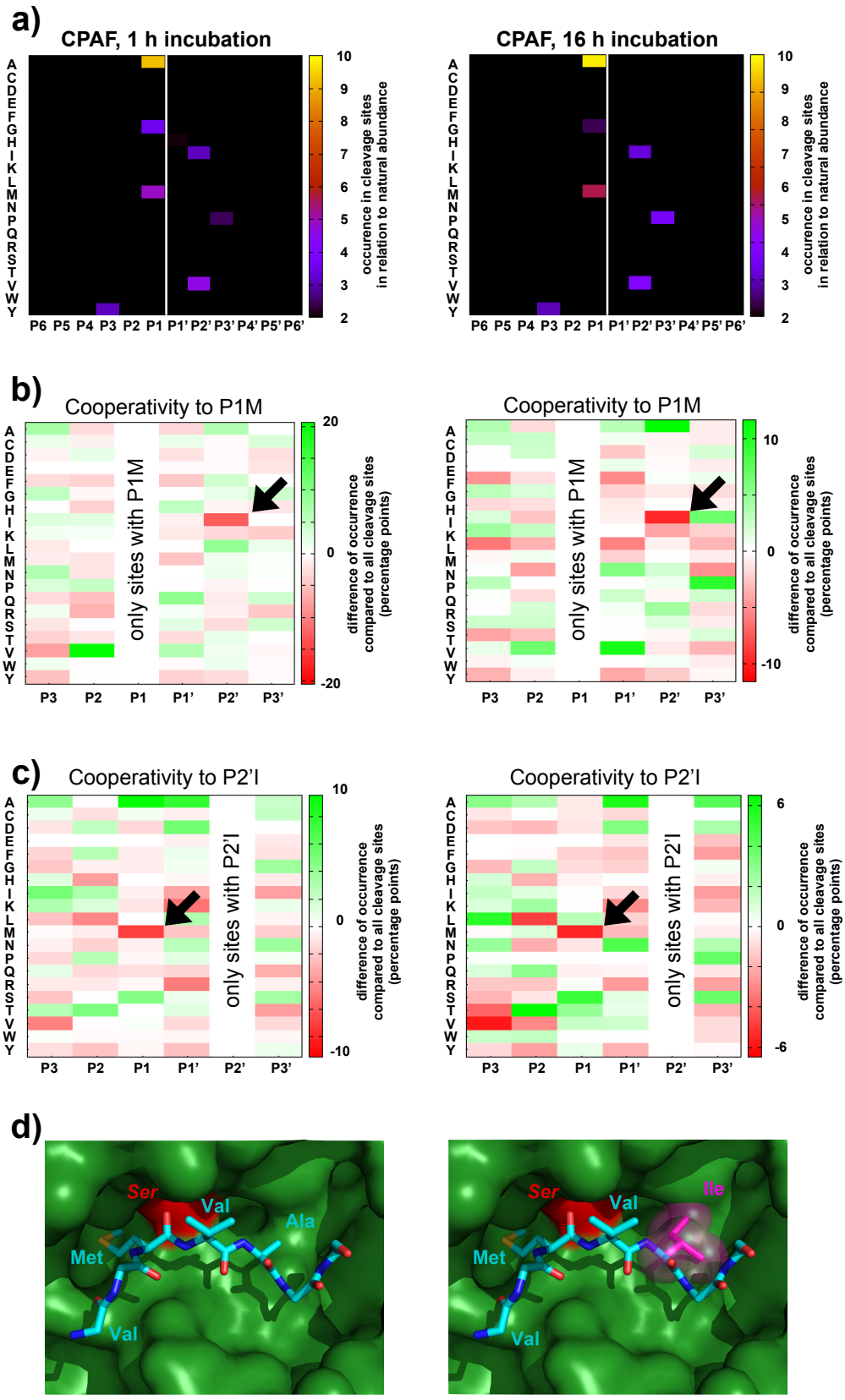
Biniossek et al, Figure 1



Biniossek et al, Figure 2



Biniossek et al, Figure 3



Biniossek et al, Figure 4