



OPEN

Identification and characterization of the proteolytic flagellin from the common freshwater bacterium *Hylemonella gracilis*

Ulrich Eckhard^{1,2}✉, Constantin Blöchl¹, Benjamin G. L. Jenkins³, Michael J. Mansfield^{3,5}, Christian G. Huber^{1,4}, Andrew C. Doxey^{3,6}✉ & Hans Brandstetter^{1,4,6}

Flagellins are the protein components of bacterial flagella and assemble in up to 20,000 copies to form extracellular flagellar filaments. An unusual family of flagellins was recently discovered that contains a unique metalloprotease domain within its surface-exposed hypervariable region. To date, these proteolytic flagellins (also termed flagellinolysins) have only been characterized in the Gram-positive organism *Clostridium haemolyticum*, where flagellinolysin was shown to be proteolytically active and capable of cleaving extracellular protein substrates. The biological function of flagellinolysin and its activity in other organisms, however, remain unclear. Here, using molecular biochemistry and proteomics, we have performed an initial characterization of a novel flagellinolysin identified from *Hylemonella gracilis*, a Gram-negative organism originally isolated from pond water. We demonstrate that *H. gracilis* flagellinolysin (HgrFlaMP) is an active calcium-dependent zinc metallopeptidase and characterize its cleavage specificity profile using both trypsin and GluC-derived peptide libraries and protein substrates. Based on high-throughput degradomic assays, HgrFlaMP cleaved 784 unique peptides and displayed a cleavage site specificity similar to flagellinolysin from *C. haemolyticum*. Additionally, by using a set of six protein substrates, we identified 206 protein-embedded cleavage sites, further refining the substrate preference of HgrFlaMP, which is dominated by large hydrophobic amino acids in P1', and small hydrophobic or medium-sized polar residues on the amino-terminal side of the scissile bond. Intriguingly, recombinant HgrFlaMP was also capable of cleaving full-length flagellins from another species, suggesting its potential involvement in interbacterial interactions. Our study reports the first experimentally characterized proteolytic flagellin in a Gram-negative organism, and provides new insights into flagellum-mediated enzymatic activity.

Proteolytic flagellins, also known as flagellinolysins, are a recently discovered, enzymatically active class of bacterial flagellins that contain a unique zinc metallopeptidase domain within their hypervariable region (HVR). As demonstrated in the animal pathogen *Clostridium haemolyticum*¹, flagellinolysins are produced as surface-exposed components of flagellar filaments, and provide flagella with proteolytic activity against external substrates. Currently, 265 putative flagellinolysins (FlaMPs) have been bioinformatically identified across 194 species from the NCBI database, spanning both Gram-positive and Gram-negative bacteria (<https://doxey.uwaterloo.ca/flagellinolysin-DB/>; accessed 27-05-2020), and form a distinct family (M101) within the metallopeptidases in the MEROPS database². To date, the only studies of flagellinolysin have been in the Gram-positive organism *C. haemolyticum*. Thus, the enzymatic activities and physiological functions of flagellinolysins are still unclear.

The bacterial flagellum is generated by the successive addition of flagellin monomers, which translocate through the central channel of the growing flagellum by the flagellar type III secretion system, generating a

¹Department of Biosciences, University of Salzburg, Hellbrunner Straße 34, 5020 Salzburg, Austria. ²Proteolysis Lab, Department of Structural Biology, Molecular Biology Institute of Barcelona, CSIC, Barcelona Science Park, Baldiri Reixac, 15-21, 08028 Barcelona, Catalonia, Spain. ³Department of Biology, University of Waterloo, 200 University Ave. West, Waterloo, ON N2L 3G1, Canada. ⁴Christian Doppler Laboratory for Innovative Tools for Biosimilar Characterization, University of Salzburg, Hellbrunner Straße 34, 5020 Salzburg, Austria. ⁵Present address: Genomics and Regulatory Systems Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan. ⁶These authors jointly supervised this work: Andrew C. Doxey and Hans Brandstetter ✉email: ueccri@ibmb.csic.es; acdoxey@uwaterloo.ca

gigantic helical assembly of approximately 20,000 protein subunits^{3,4}. The prototypic structural flagellin FliC from *Salmonella typhimurium* harbors two non-enzymatic HVR domains of ill-defined function⁵. Notably, removal of the corresponding domains in the *Escherichia coli* flagellin has no significant effect on filament formation⁶, suggesting that HVR-embedded domains may modulate surface properties of flagellar filaments without affecting core filament structure. This is possible due to the three-dimensional structure of flagellins where the conserved N and C terminal regions form together the D0 and D1 domains of the filament core, while the central HVR domain is surface exposed⁵. Consistent with this idea, the filament-forming N- and C-terminal domains of flagellins are extremely conserved across bacteria, but there is considerable variation in the sequence and length of flagellins HVRs, and clear evidence of diversifying selection at the HVR sequence level^{7,8}. For instance, a recent study identified a new class of “giant flagellins”, encoding large HVR regions comprising over 1000 residues. These giant flagellins were shown to be abundant in marine Gammaproteobacteria, in which they produce thick flagellar filaments and may represent a mechanical adaptation to their environment⁹.

Importantly, the genomes of about 45% of all flagellated species encode more than one flagellin gene¹⁰. The production of multiple flagellin variants may diversify the structural and biophysical properties of flagellar filaments, which is consistent with findings that different flagellins show distinct patterns of filament localization^{10–15}. *C. haemolyticum* is one such organism that produces multiple types of flagellin, including a structural flagellin (FliC, UniProt ID A0A0A0IUA0), and flagellinolysin (UniProt ID A0A0A0IT63). In purified flagellar filaments of *C. haemolyticum*, it was shown that FliC forms the major filament protein, while flagellinolysin is a less abundant component that is non-specifically distributed throughout the filament¹.

Flagella not only function in motility, but also play critical roles in a growing variety of processes including biofilm development^{16,17}, mechanosensing, host adhesion and invasion during bacterial pathogenesis, and secretion of virulence factors^{18–21}. Considering the wide diversity of functions associated with secreted bacterial proteases, this opens a wide avenue of potential roles for flagella-embedded proteolysis. Flagellinolysins may provide additional nutrition by degrading extracellular proteins into peptides, which are subsequently taken up by peptide permeases²². Since the permease DppA represents the sensor for the chemotactic-signal transducer TAP in peptide chemotaxis²³, it is possible that flagellinolysins may generate extracellular peptide gradients to guide the bacterium towards favorable protein-rich environments. Alternatively, flagellinolysins may represent potent toxins in bacterial warfare by attacking cell components of competing species²⁴, and/or the extracellular polymeric substances of biofilms^{25,26}. Similarly, proteolytic flagellins could aid host invasion, colonization and toxin diffusion by degrading the extracellular matrix, similar to what has been shown for bacterial collagenases^{27–29}, either embedded in the filament or after secretion, as seen for example in *Campylobacter jejuni* for the non-enzymatic flagellin FlaC³⁰.

To gain further insights into flagellinolysin function, we investigated the predicted flagellinolysin gene product from the Gram-negative bacterium *Hylemonella gracilis*, a common microbe in freshwater ecosystems throughout the world. *H. gracilis* forms motile spirilla of small diameter (0.2 to 0.3 μm), and is notable for its ability to pass through 0.1 μm sterilizing grade filters³¹. It produces bipolar tufts of flagella, and is classified within the Burkholderiales family of Betaproteobacteria^{32,33}. Cultures are urease positive, a prominent virulence factor in *Helicobacter pylori*³⁴, but cannot form spores or liquefy gelatin³³. Interestingly, it has been reported that *H. gracilis* can prevent the long-term persistence of the plague-causing bacteria *Yersinia pestis* in filtered fresh water, suggesting bactericidal activity. And moreover, *Y. pestis* appears to condition its competitor for outgrowth, and allows *H. gracilis* to thrive on otherwise unsuitable media³⁵.

Here, using molecular biochemistry and proteomics, we performed an in-depth proteolytic profiling of the flagellinolysin (flagellin metallopeptidase, FlaMP) from *H. gracilis*, herein referred to as HgrFlaMP. We demonstrate that HgrFlaMP is an active metalloprotease, and report cleavage specificity profiles against both peptide libraries and protein substrates. Our work establishes HgrFlaMP as the second biochemically described proteolytic flagellin to date, expanding our knowledge of this intriguing family of enzymes.

Results

Identification of flagellinolysin from the *Hylemonella gracilis* genome. Using the sequence of flagellinolysin from *C. haemolyticum* in a BLAST query against the NCBI and UniProtKB reference database, we identified a putative flagellinolysin protein from the organism, *H. gracilis* ATCC 19624 (UniProt F3KXG2, NCBI accession WP_006299301). The *H. gracilis* and *C. haemolyticum* homologs share 37% amino acid identity and are well-aligned over the zinc-binding HExxH catalytic motif (residues 249–253 and 233–237 respectively) of the metallopeptidase active site. According to the domain annotations from the Conserved Domain Database (CDD)³⁶, the *H. gracilis* protein contains the characteristic flagellinolysin architecture consisting of conserved flagellin N-terminal and C-terminal domains, which form the filament core, and a metallopeptidase domain within the central hypervariable region (Fig. 1A). In addition to *H. gracilis* strain ATCC 19624, highly similar proteins (97% and 92% amino acid identity) were also identified in two other sequenced *H. gracilis* strains within the NCBI, which clustered together in a phylogenetic analysis (Fig. 1B). *H. gracilis* flagellinolysins are most closely related to flagellinolysin homologs from related taxa including *Rhodospirillum rubrum* sp. (68% amino acid identity), *Limnospira* sp. (66% amino acid identity) and *Comamonas* sp. (55–56% amino acid identity), which are also in the Comamonadaceae family of bacteria. This suggests that HgrFlaMP has not evolved through a unique domain fusion event within *Hylemonella*, but instead flagellinolysin may be a conserved protein within the Comamonadaceae lineage and that these proteins share common ancestry, which may have originated earlier through lateral gene transfer from a more distantly related organism. Interestingly, while flagellinolysin is encoded within a flagellar gene cluster in the *C. haemolyticum* genome, this is not the case for the gene from *H. gracilis* (Fig. 1C). In all three *H. gracilis* strains, flagellinolysin is encoded nearby non-flagellar genes encod-

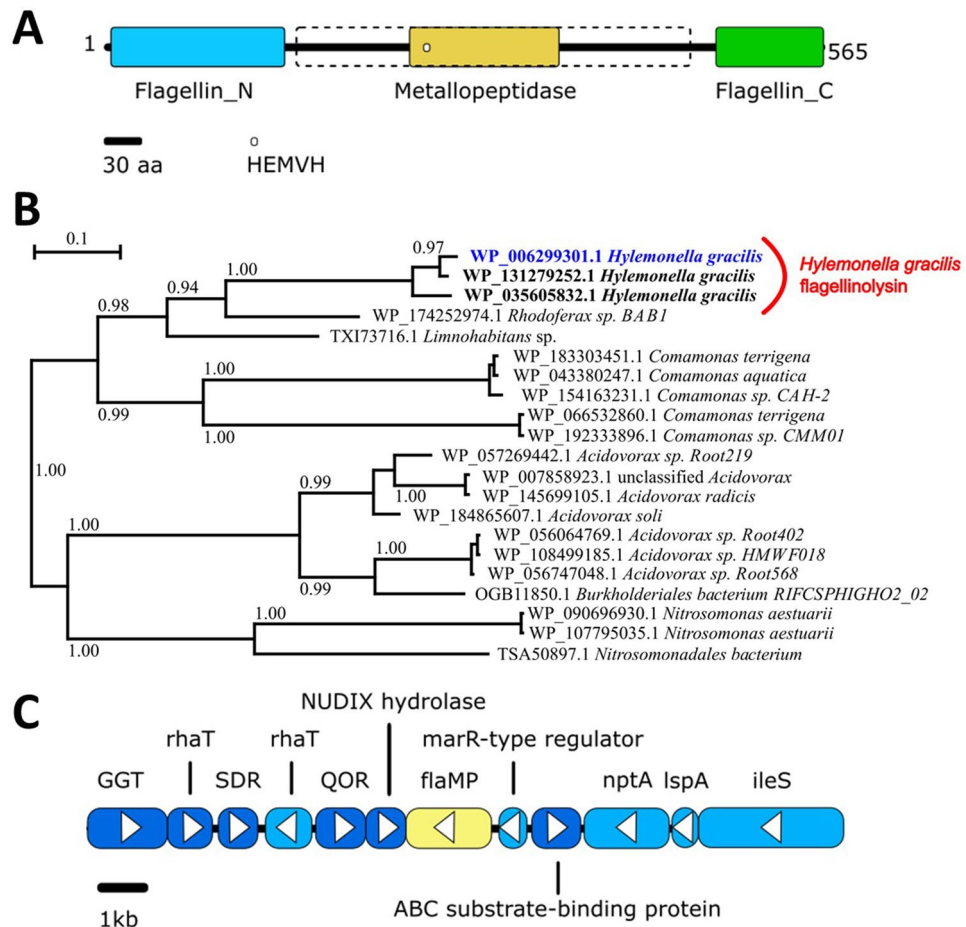


Figure 1. Identification of a putative proteolytic flagellin (flagellinolsin) from *Hylemonella gracilis*. (A) Bioinformatically annotated domain architecture of a predicted flagellinolsin from *H. gracilis* ATCC 19624 (UniProt F3KXG2, NCBI accession WP_006299301). The catalytic HEXXH motif of zinc-metallopeptidases is indicated as a white circle and is found within the central hypervariable region, which is flanked by the N- and C-terminal flagellin domains. The dashed box indicates the domain boundary used for the biochemical characterization of the HVR-only protein (Leu163 to Ile450). (B) Phylogenetic tree of *H. gracilis* flagellinolsin and its top identified homologs by BLASTp from the NCBI database. The phylogenetic tree was generated from a MUSCLE-generated multiple alignment using PhyML with default parameters and an LG model, as implemented within SeaView. (C) Genomic context of the *H. gracilis* flagellinolsin gene (labeled as *flaMP* and highlighted in yellow), displaying its surrounding gene neighborhood.

ing NUDIX hydrolase³⁷, a MarR-type transcriptional regulator³⁸, and others. The functional relevance of this genomic context is unclear.

Protein expression and purification. To examine the biological activity of the putative flagellinolsin from *H. gracilis*, we cloned and expressed the full-length (Met1-Arg565) and HVR-only (Leu163-Ile450) proteins. Both variants yielded inclusion bodies in *Escherichia coli*, and were refolded on-column by decreasing concentrations of urea. A size-exclusion chromatography step right after refolding and before tag-removal with SUMO-protease was crucial to separate folded protein from soluble aggregates, as the latter hindered separation of tagged and untagged protein via NiNTA re-chromatography, likely due to assemblies of mixed species. This was especially a problem for the full-length protein, likely due to the presence of the terminal flagellin D0 domain, which represents the major driver of self-polymerization³⁹. However, we have previously demonstrated that both flagella-embedded full-length flagellinolsin from *C. haemolyticum* and its recombinantly-produced HVR-protein show similar proteolytic behavior¹. Thus, after initial trials with full-length HgrFlaMP, we focused on the HVR-protein as it showed better stability and was better suited for the full purification protocol. The purification process for the HVR protein is shown in Fig. 2, and representative size-exclusion chromatograms are included in Supplementary Figure 1. The partial purification of the full-length protein prior to purification-tag removal is shown in Supplementary Figure 2AB. Importantly, despite solubilization in 8 M urea, not all the recombinant HVR protein was available for NiNTA purification, probably due to soluble protein aggregates and thus inaccessibility of the His6-tag (Fig. 2B). However, reloading of the flow-through after another 24 h of incubation in 8 M urea—but not immediately after the first round of NiNTA—rescued approximately 50%

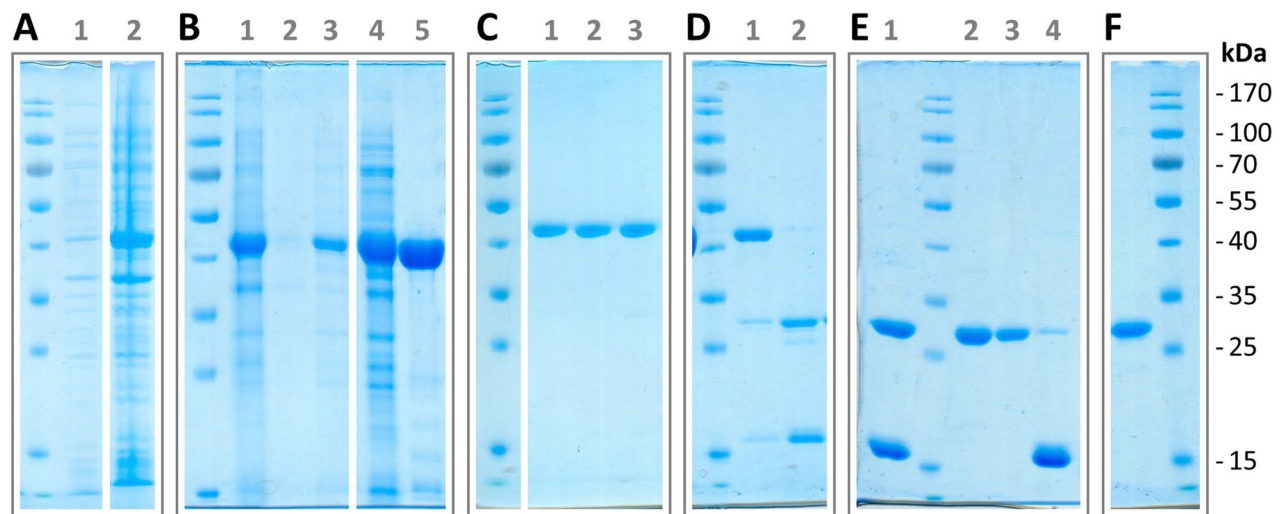


Figure 2. Expression and purification of the HVR-domain of HgrFlaMP. (A) Protein expression samples of His6-SUMO tagged recombinant protein before (1) and after (2) induction with IPTG (expected molecular weight: 44 kDa). (B) Cell lysis and subsequent NiNTA purification including refolding. 1 total lysate; 2 soluble protein after lysis; 3 insoluble protein in 8 M urea; 4 NiNTA flow-through; 5 NiNTA elution. (C) Protein after size-exclusion chromatography; peak fraction and the two surrounding SEC fractions are shown. (D) Removal of the N-terminal purification tag (expected molecular weight: 14 kDa) by SUMO protease. Samples right after protease addition (1) and 1 h of incubation (2) are shown. (E) NiNTA re-chromatography. 1 recombinant protein after tag removal (expected molecular weight: 30 kDa); 2 NiNTA flow-through and 3 wash containing the target protein; 4 NiNTA elution dominated by the His6-SUMO tag. (F) Recombinant tag-depleted HgrFlaMP-HVR after size-exclusion chromatography. Figure panels with lanes from different gels show white separation lines, and all original scans can be found in the supplementary materials.

of the remaining target protein. As protein purifications started from rather pure inclusion body preparations, one round of NiNTA and size-exclusion chromatography yielded highly pure and monodisperse target protein, albeit still His6-SUMO-tagged (Fig. 2C, Supplementary Figure 1A). The purification tag was subsequently removed by incubation with His6-tagged SUMO protease, and the tag-depleted protein was purified by final NiNTA and size-exclusion chromatography steps (Fig. 2D–F).

Biochemical characterization. Initial biochemical characterizations were performed using the His6-SUMO-tagged protein. As initial candidate substrates, we selected bovine casein, given its standard use as a general protease target, as well as flagellins from another bacterial species, given our hypothesis that flagellinolysin may target flagella from other competing cells. While only minor activity was detected for the full-length flagellinolysin against bovine casein (Supplementary Figure 2C), full degradation of both casein and the structural flagellin from *Pseudoalteromonas tunicata* could be observed when using the tagged HVR protein (Fig. 3A,B). The proteolytic activity of HgrFlaMP could be further confirmed by casein zymography (Fig. 3C). Interestingly, the recombinant protein displayed a time- and substrate-dependent tendency of self-digestion. While substrate degradation clearly dominated over self-cleavage in the presence of casein, cleavage of the structural flagellin appeared to be only slightly favored over autolysis. Intriguingly, the purified tag-depleted HgrFlaMP-HVR protein revealed a highly similar—but not identical—proteolytic behavior, probably due to the presence of soluble aggregates next to correctly folded protein in the His-SUMO tagged preparations (Fig. 4). Both bovine casein and the *P. tunicata* structural flagellin were readily digested by the tag-depleted HgrFlaMP-HVR protein, but instead of self-degradation, a shift in molecular weight by 1–2 kDa was observed, consistent with the removal of 10–20 amino acids from the N- or C-terminus. Additionally, a preferential degradation of beta-casein over alpha- and kappa-casein⁴⁰ could be observed (Fig. 4A), and which we later confirmed by mass spectrometry (Supplementary Figures 3–6). We then additionally included inactive full-length flagellinolysin from *P. tunicata* into the proteinaceous substrate mix, which was likewise degraded by HgrFlaMP (Fig. 4B). Both flagellins from *P. tunicata* were expressed and purified as tag-depleted variants as described for HgrFlaMP. Representative size-exclusion chromatograms of the final purification step are shown in Supplementary Figure 7. Importantly, four hours of incubation with HgrFlaMP-HVR at a protease to substrate ratio of 1:25 at 22 °C were sufficient to ensure nearly complete degradation of the provided substrates (Fig. 4B, lanes 6–7). Consistent with our finding for flagellinolysin of *C. haemolyticum*¹, addition of 10 mM CaCl₂ and 10 μM ZnCl₂ were needed to reach full enzymatic activity, while 10 mM EDTA abrogated proteolysis. This further supports the notion of flagellinolysins as zinc- and calcium-dependent metalloproteases, as bioinformatically predicted due to the remote homology between the flagellinolysin HVR domain and the peptidase domain of clostridial collagenases^{1,41}. A multiple sequence alignment of 10 selected flagellinolysins around their active site, and sequence logos of the conserved regions surrounding the zinc-binding HEXXH motif and the putative calcium binding aspartate are shown in Supplementary Figure 8.

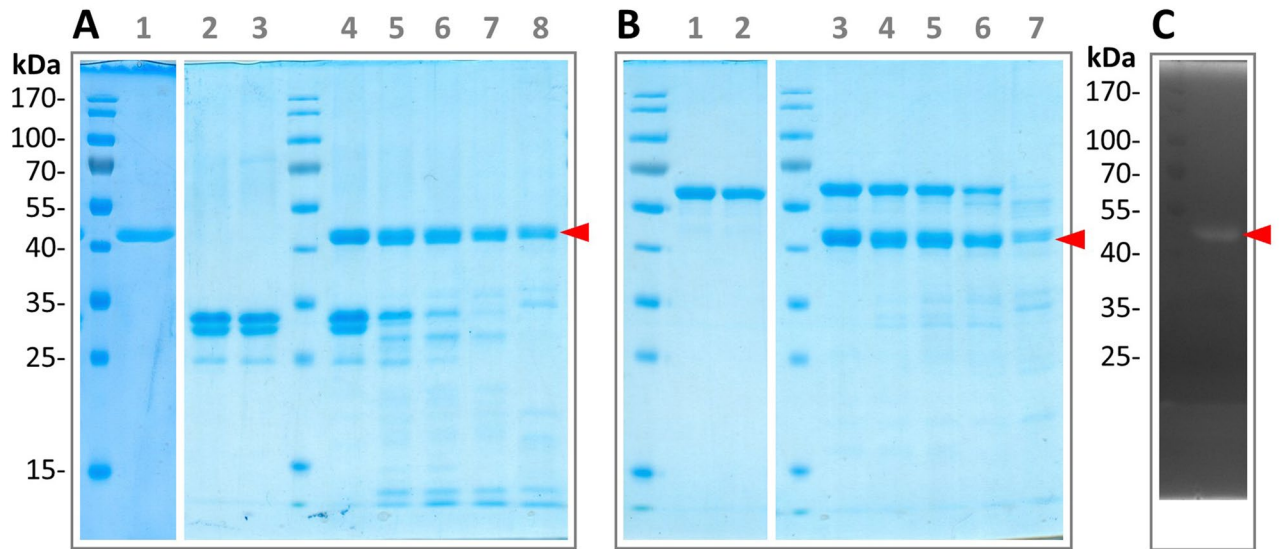


Figure 3. Degradation of bovine casein and the structural flagellin from *P. tunicata* by His6-SUMO-tagged HgrFlaMP-HVR. (A) 1 tagged HgrFlaMP after NiNTA and SEC purification (expected molecular weight: 44 kDa). 2–3 bovine casein after 0 and 16 h of incubation at 16 °C (control). 4–8 bovine casein after incubation with HgrFlaMP-HVR for 0, 1, 2, 4, and 16 h, respectively. (B) 1–2 structural flagellin from *P. tunicata* (expected molecular weight 61 kDa) after 0 and 16 h of incubation at 16 °C. 3–7 structural flagellin after incubation with HgrFlaMP-HVR for 0, 1, 2, 4, and 16 h, respectively. (C) Casein zymography of His6-SUMO-tagged HgrFlaMP-HVR. For clarity, the flagellinolysin band is indicated by a red triangle. Figure panels with lanes from different gels show white separation lines, and all original scans can be found in the supplementary materials.

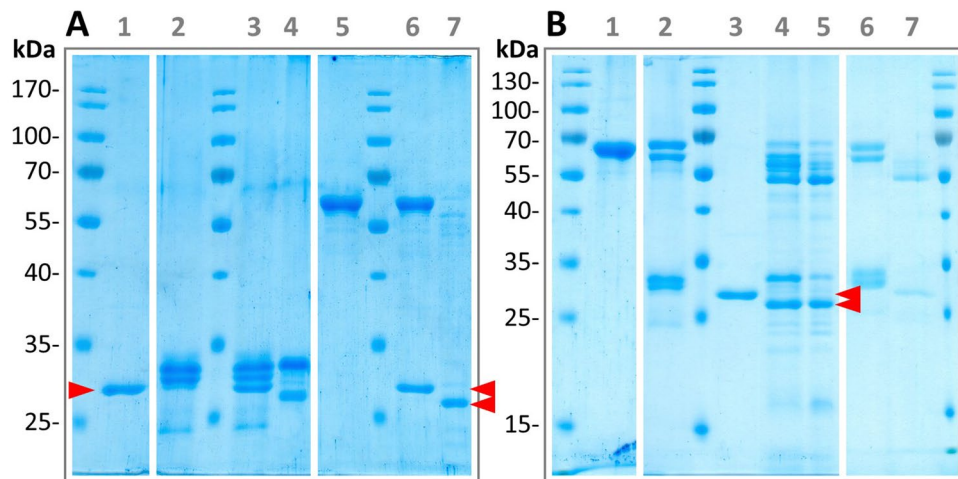


Figure 4. Degradation of bovine casein and the structural and proteolytic flagellin from *P. tunicata* by recombinant HgrFlaMP-HVR. (A) 1 tag-depleted HgrFlaMP-HVR (expected molecular weight: 30 kDa). 2 bovine casein. 3–4 bovine casein incubated with HgrFlaMP-HVR in presence of EDTA or $\text{Ca}^{2+}/\text{Zn}^{2+}$, respectively. 5 structural flagellin from *P. tunicata* (expected molecular weight: 61 kDa). 6–7 structural flagellin incubated with flagellinolysin HVR in presence of EDTA or $\text{Ca}^{2+}/\text{Zn}^{2+}$, respectively. (B) 1 inactivated HAXXH variant of the full-length proteolytic flagellin from *P. tunicata* (expected molecular weight: 65 kDa). 2 substrate mix consisting of bovine casein and both the structural and proteolytic flagellin from *P. tunicata*. 3 tag-depleted HgrFlaMP-HVR. 4–5 substrate mix after 2 and 4 h of incubation with HgrFlaMP-HVR at 16 °C. 6–7 incubation of substrate mix with flagellinolysin HVR in presence of either 10 mM EDTA or 10 mM Ca^{2+} and 10 μM Zn^{2+} after 2 h at 20 °C. HgrFlaMP-HVR is indicated by a red triangle. Figure panels with lanes from different gels show white separation lines, and all original scans can be found in the supplementary materials.

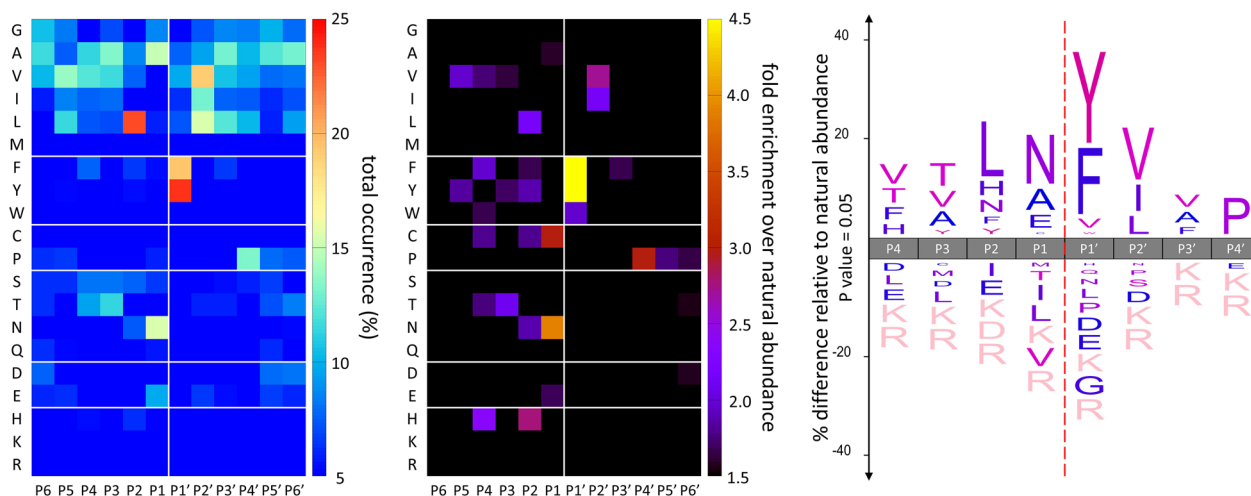
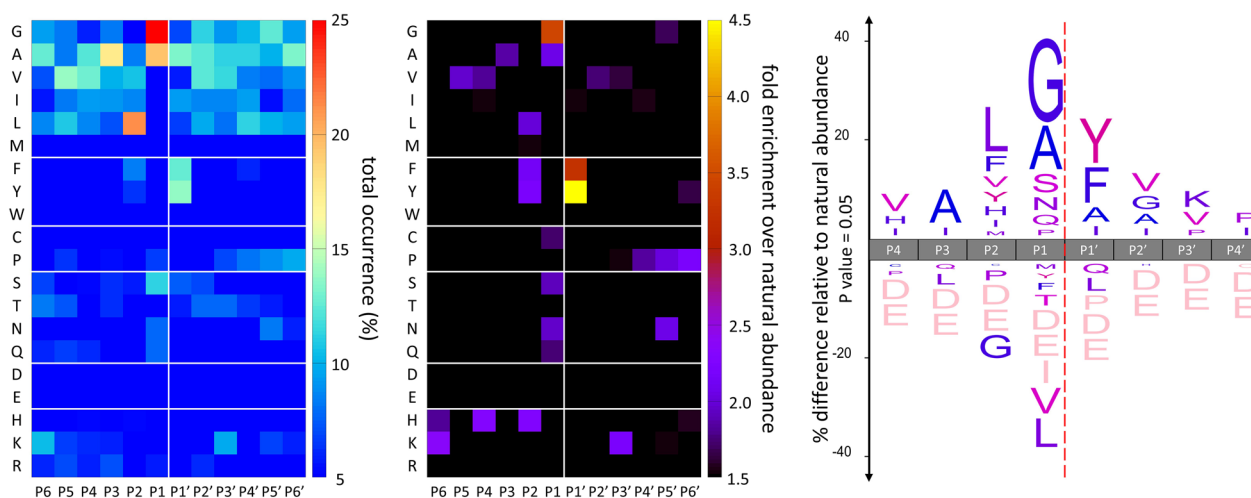
A Cleavage site analysis of HgrFlaMP in a trypsin-generated *E. coli* K12 peptide library (n=434)**B** Cleavage site analysis of HgrFlaMP in a GluC-generated *E. coli* K12 peptide library (n=350)

Figure 5. PICS specificity profiling of the proteolytic flagellin from *Hylemonella gracilis*. The alignment of 434 and 350 unique cleavage sites that were determined in a (A) trypsin- and (B) GluC-derived peptide library, respectively, is summarized as heat maps showing the relative occurrence (left) and fold-enrichment over natural abundance (middle) for the individual amino acids at the respective cleavage subsites. P6 to P6' subsite positions are shown on the x-axis, and the single letter code for amino acid residues is plotted on the y-axis. In iceLogo⁴⁴ (right), significantly over- and under-represented amino acids (p value 0.05) are shown above and below the x-axis, respectively, and amino acids that have not been identified are depicted in pink. Heat maps were created using gnuplot (www.gnuplot.info).

Specificity profiling by proteomic identification of protease cleavage sites (PICS). As a first step, we generated both trypsin- and GluC-generated peptide libraries from purified *E. coli* proteomes, which thereby provide complementary peptide substrates of biological amino acid diversity. The purified HgrFlaMP-HVR was then incubated with library aliquots in the presence of either EDTA (control) or calcium plus zinc, and cleavage products were identified together with non-cleaved library peptides by quantitative proteomics using light and heavy dimethylation for control and protease-treated samples, respectively. Following peptide-spectrum matching at a false discovery rate (FDR) of 1% using X!Tandem⁴² and PeptideShaker⁴³, complete cleavage sites were bioinformatically inferred from the *E. coli* proteome (Supplemental Table 1), from which we obtained 434 and 350 unique P6 to P6' cleavage sites for the trypsin- and GluC-generated peptide library, respectively (Fig. 5). Amino acid occurrences were counted for each sub-site, normalized to the natural amino acid abundance, and summarized as heat maps. Over- and under-representation of each amino acid was further analyzed using iceLogo⁴⁴ to evaluate individual subsite preferences. Importantly, we used both trypsin- and GluC-generated peptide libraries to account for the amino acid bias introduced by the library enzyme; *i.e.* trypsin-generated peptide libraries lack internal basic residues while GluC libraries are devoid of acidic residues⁴⁵. Consequently, a preference or dislike of acidic residues can only be seen in a tryptic library, as GluC-derived peptides have these amino acids only at their C-termini. Intriguingly, the two peptide libraries yielded similar, but not identi-

cal specificity profiles, but which is not unexpected due to the different peptide composition of the two libraries. While the substrate preferences at the P1' and P2 positions—substrate and active-site cleft subsite nomenclature according to Schechter and Berger 1967⁴⁶—were particularly conserved between the two PICS library preparations, a preference for glycine emerged only in the GluC-derived peptide library (Fig. 5B). Large hydrophobic and aromatic residues, such as phenylalanine and tyrosine, dominated P1' in both libraries, and were also readily found in P2, where they were accompanied by other large amino acids such as leucine and histidine. Overall, these specificity results are largely similar to our specificity findings for *C. haemolyticum* flagellinolysin¹. However, in *H. gracilis*, small (e.g. Ala, Gly) or medium-sized polar amino acids (e.g. Asn, Gln) dominated with the negatively charged glutamate in subsite P1, which contrasts the previously identified strong preference for both large hydrophobic and polar residues observed in *C. haemolyticum*. Moreover, the amino acid preference in the positions flanking the scissile bond, namely P1 and P1', closely resembles the specificity profiles of mammalian matrix metalloproteinases (MMPs), where Gly/Ala and Asn/Glu are frequently found in P1 next to medium-to-large hydrophobic amino acids in P1'^{45,47}. However, while MMPs additionally enrich for proline in P3, this feature is completely absent in both flagellinolysins, pointing to a comparably open S3 subsite, *i.e.* without the subsite roof typical for MMPs. This supports the aforementioned inability of *H. gracilis* to liquefy gelatin³³ as gelatin contains around 24% of proline and hydroxyproline. Overall, the identified specificity profiles were rather broad, highlighting the cleavage promiscuity of flagellinolysins. Anchored by a mostly hydrophobic P2 residue and particularly by the P1' position, which has a clear preference for an aromatic residue such as tyrosine or phenylalanine, the main determinants span only the central P2 to P2' region, and only a few amino acids are significantly underrepresented relative to their natural abundance throughout the profile.

Refinement of the peptide-derived specificity profile by using a set of proteinaceous substrates. To bolster the peptide-based specificity profiles obtained from the PICS experiments, we repeated the aforementioned substrate mix digests and analyzed them also by quantitative mass spectrometry. To do so, a protein substrate mix consisting of the bovine caseins alpha-1, alpha-2, beta, and kappa, and the structural and proteolytic flagellins from *Pseudoalteromonas tunicata* (PtuFlaMP), the latter as a mutationally inactivated HAXXH variant, was incubated with recombinant HgrFlaMP at a protease to substrate ratio of 1:25. To induce activity, 10 mM Ca²⁺ and 10 μM Zn²⁺ were added, while 10 mM EDTA was added to the negative control. After incubation for two hours at 22 °C, reactions were stopped by adding 100 mM EDTA, and the sample was halved. While one part was directly acidified and purified using C18 peptide clean-up tips, the other was incubated for another hour at 37 °C with trypsin to alter the physicochemical properties of HgrFlaMP-generated cleavage peptides that are otherwise inadequate for mass spectrometry analysis. The trypsin digests were stopped by addition of 0.1% formic acid, and the derived peptides purified as above. Without any further optimization of the cleavage conditions and by using HgrFlaMP alone, we obtained sequence coverages of 66%, 38%, 71%, and 45% for the mature bovine caseins alpha-S1, alpha-S2, beta, and kappa, respectively, (Supplementary Figures 3–6), and a total of 95 cleavage sites were identified in these four proteins. The corresponding peptide-spectrum matches are documented in Supplementary Table 2. Importantly, none of the identified peptides matched the respective signal peptides annotated in the UniProt database, validating our data-analysis pipeline as these are missing from mature caseins purified from bovine milk. Taking a closer look at alpha-S1-casein (Supplementary Figure 3), the uncovered part mainly corresponds to the amino acid stretch between Leu55 and Gly108. Importantly, this region has only one of the protein's 18 tyrosine and phenylalanine residues, which are, however, the primary specificity determinants in P1' based on our specificity analysis, and both Ser and Asp are enriched more than threefold, which are two rather disfavored residues by HgrFlaMP. Additionally, its only tyrosine residue is located in position 105, and cleavage would generate a tripeptide inaccessible for detection with our applied mass spectrometry settings.

While we obtained a sequence coverage of 61% for the structural flagellin from *P. tunicata* (Supplementary Figure 9), we could only identify cleavage products accounting for 39% of its proteolytic counterpart despite similar molecular weight (Supplementary Figure 10). Intriguingly, the major difference is attributed to the central HVR region of the flagellinolysin, spanning approximately Leu155 to Gly593, which appears to be hardly cleaved by HgrFlaMP, while the N-terminal and C-terminal coiled-coil domains were readily degraded. And, contrary to the above-mentioned example of alpha-S1-casein, two thirds of the protein's 32 tyrosine/phenylalanine residues reside within the HVR-region. Strikingly, the specificity logos obtained this way (Fig. 6) agreed well with our profiling results using PICS, especially with the one from the trypsin-derived peptide library (Fig. 5A). In total, we identified 206 unique cleavage sites using this approach, 168 after HgrFlaMP digest alone (Fig. 6A), 99 by the sequential digestion protocol, and hence 61 in both experimental set-ups (Fig. 6B). In P1', tyrosine and phenylalanine were again the dominating residues, followed by glutamine, isoleucine, leucine and methionine. In P2, leucine was clearly preferred over asparagine and histidine, while in P1, glutamine, asparagine, and alanine were again favored. However, it is important to note that the repertoire of potential cleavage sites is less balanced in such a small protein-based cleavage assay than in a proteome-derived peptide library assay such as PICS. Finally, due to the sequence similarity between structural flagellins, especially in their N- and C-terminal domains (Supplementary Figure 11), we analogously expect potential cleavage of the *H. gracilis* structural flagellin by HgrFlaMP, and of other structural and proteolytic flagellins.

Discussion

We successfully cloned, expressed, and purified HgrFlaMP, the putative flagellinolysin from *Hylemonella gracilis* (Figs. 1, 2, Supplementary Figure 1), and experimentally assessed its predicted function as an active protease capable of cleaving both peptide and protein substrates (Figs. 3, 4, 5, 6). Importantly, flagellinolysins were

Cleavage specificity analysis of HgrFlaMP against a set of proteinaceous substrates

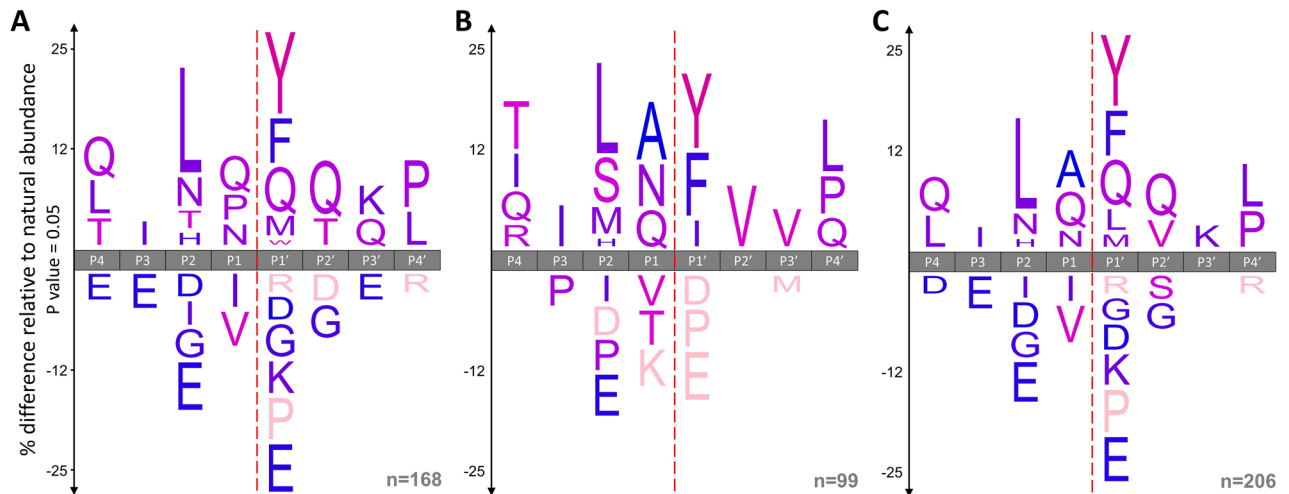


Figure 6. iceLogo representation of the proteolytic cleavage preference of HgrFlaMP using a set of six protein substrates. **(A)** HgrFlaMP specificity profile derived from 168 cleavage sites identified in bovine caseins and the structural and proteolytic flagellin from *P. tunicata*. **(B)** A total of 99 HgrFlaMP cleavage sites, 38 of which were intangible in the previous setting, were identified by a two-step digestion protocol utilizing trypsin as orthogonal cleavage enzyme. **(C)** Combined specificity logo based on 206 unique cleavage events, 95 within bovine caseins, and 111 in the flagellins from *P. tunicata*.

discovered only recently in 2017¹, and so biochemical information about these enzymes is still scarce. Thus, this report represents the second characterization of a proteolytic flagellin, but the first from a Gram-negative organism.

Using proteomics and both proteome-derived peptide libraries and a set of intact protein substrates, we showed that HgrFlaMP is indeed a protease and that it exhibits a similar cleavage specificity to its previously characterized homolog from *C. haemolyticum*¹. Substrate preference is dominated by large, predominantly hydrophobic residues in P1', such as the aromatic amino acids phenylalanine and tyrosine, which are also favored in P2 together with leucine and histidine. Sequence specificity profiles obtained from the peptide-based PICS assay and the protein-cleavage assay are in good agreement, and perfectly explain the difference in sequence coverage for alpha-1 and alpha-2 casein by HgrFlaMP, which were 66 and 38% respectively. For beta-casein, we obtained a peptide coverage of over 71% in a single experiment without any optimization, and overall, we could identify more than 200 cleavage sites in a set of only six proteins, which pinpoints a rather broad sequence specificity combined with a high substrate promiscuity for proteolytic flagellins. Of note, the identified cleavage site preferences against both peptide libraries and a set of protein substrates must be interpreted collectively, and individual amino acid occurrences, fold enrichment, and relative difference to natural abundance have to be considered side-by-side for comprehensive interpretation, as each experiment and type of analysis highlights a slightly different aspect of the identified specificity profiles. Notably, the His6-SUMO-tagged protein exhibited lower stability than the tag-depleted enzyme after the addition of calcium and zinc, and the presence of substrate appears to stabilize the protease, probably by competing with self-cleavage. However, in the biological context of the bacterial flagellum, this may not be relevant as the N- and C-terminal coiled-coil domains form the inner backbone of the filament and are thus largely inaccessible for the HVR-domain, and the proteolytic units would be spread out over the filament with structural flagellins as additional spacer molecules.

While we did not succeed in purifying full-length HgrFlaMP because of its high propensity to self-aggregation during purification and autolysis, we did obtain crystallization-grade protein of our HVR variant. However, despite extensive trials, no crystallization conditions could be established at this point, but we are confident that our cleavage assays will help us to narrow down the minimal domain boundaries for a proteolytically active and stable variant. This may be of great importance, as with the current preparation we observe partial to complete self-cleavage, as indicated by an approx. 1–2 kDa shift in SDS-PAGE upon addition of calcium and zinc (Fig. 4). Intriguingly, when taking a closer look at our mass spectrometry cleavage data, we observed a preference for the N- and C-terminal coiled-coil regions in the proteolytic flagellin of *P. tunicata*, while the central HVR region remained nearly intact (Supplementary Figure 10). On the other hand, despite a slight preference for the terminal D0/D1 domains, complete degradation of the recombinant structural flagellin is observed in our degradomic experiment (Supplementary Figure 9), which supports the idea of flagellinolysins being able to target flagella from other organisms. Further work is needed to fully substantiate this finding.

Overall, flagellin-mediated proteolysis opens an entire avenue of potential biological functions. Flagellinolysins may act as secreted proteases, either after extracellular export via the flagellar type III secretion system, or by liberating the proteolytic HVR domain alone while leaving the filament intact. In doing so, proteolytic flagellins could perform any of the documented roles of secreted microbial proteases. However, if flagellinolysin activity is localized to the filament, this would also provide advantages by specific localization of proteolytic activity to external substrates that are in direct physical contact with the bacterium. Furthermore, retention of

flagellinolysin within the flagellar filament avoids loss of enzymatic subunits through diffusion from the cell. The proteolytic activity of flagellinolysins could be of nutritional value, as intact proteins are more difficult to import than cleaved peptides that are generated in close proximity to the cell. Additionally, proteolytic flagellins may play a crucial role in host–pathogen interaction by degrading the extracellular matrix and thus aiding host infiltration and colonization. And in a biofilm setting, flagellinolysins may contribute to interspecies competition and microbial warfare by degrading the extracellular polymeric substance and/or proteinaceous cell wall compounds of competing bacteria, or aid biofilm formation and remodeling.

Notably, only monomeric but not filament-embedded flagellins activate Toll-like receptor 5 (TLR5), as the main TLR5 recognition motifs lie within the evolutionary conserved D0 and D1 domains^{48–50}. Based on our proteomic data with recombinant protein, these same regions are preferentially targeted by flagellinolysins. Consequently, proteolytic flagellins could contribute to immune evasion by targeting flagellin monomers after their release from the filament, and thus deterring the binding of this prominent pathogen-associated molecular pattern to its cognate detector. Equivalently, proteolytic targeting of TLR5 and/or its associated adapter and signaling proteins may also abrogate a potential inflammatory and immune response, and thus may allow a pathogen to remain undetected during infection^{51,52}.

In summary, we successfully produced a recombinant construct of the HVR-region of the flagellinolysin from *H. gracilis*, and performed a first biochemical characterization of its proteolytic activity using both classical cleavage assays and proteomics. We thereby established HgrFlaMP as the second characterized proteolytic flagellin to date and the first from a Gram-negative organism, and combined with the specificity profile of *C. haemolyticum* flagellinolysin, a common cleavage specificity profile is emerging for this protease family. Since flagellum-assisted proteolysis can be detected in hundreds of bacterial species across the tree of life throughout multiple environments, further characterization of flagellinolysin function and specificity will be important for understanding its potential roles in a wide range of possible activities, including biofilm development, interbacterial warfare, and host colonization and pathogenesis.

Methods

Bioinformatic and phylogenetic analysis. BLASTp searches were performed using the peptidase domain of the flagellinolysin from *C. haemolyticum* (NCBI accession number WP_039229452.1) to identify homologs in *Hylemonella gracilis* (UniProt ID F3KXG2), which were subsequently used to search the NCBI and UniprotKB databases for related proteins. A multiple sequence alignment of the top homologs was constructed using MUSCLE⁵³ over the full-length proteins, and a phylogenetic analysis was performed using PhyML⁵⁴ using the LG model and aLRT values for clade support, as implemented within SeaView⁵⁵. Domain architectures were predicted using the NCBI's Conserved Domain Database³⁶ with default thresholds. Multiple sequence alignments of structural and proteolytic flagellins in the Supplementary Figures were performed using ClustalO⁵⁶, and amino acid identity and similarity of protein regions calculated by BLASTp⁵⁷. The molecular representation of the structural flagellin from *Salmonella typhimurium* (PDB entry 2uxu)⁵ was created in PyMol (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC). Sequence Logo analysis of the active site HExxH motif and the putative calcium-binding site was performed using WebLogo⁵⁸ with sequencing numbering according to the proteolytic flagellin FliA(H) from *Clostridium haemolyticum* (ChaFlaMP) to allow for better comparability with our previous sequence analysis of flagellinolysins¹.

Recombinant protein production. The full-length gene of the proteolytic flagellin from *Hylemonella gracilis* (UniProt ID F3KXG2) was synthesized de novo by GeneArt (Life technologies, Thermo Fisher). Both full-length HgrFlaMP (M1 to R565; approx. 59 kDa) and a variant representing the hypervariable region (HVR; L163–I450; approx. 30 kDa), were subcloned into pET28-N-His-SUMO, sequenced at Eurofins Genomics, and transfected into *E. coli* BL21 for protein expression. Cultures were inoculated 1:500 and grown at 37 °C and 230 rpm until an OD₆₀₀ of ~1.2. Temperature was lowered to 20 °C, and cells induced with 1 mM isopropylthiogalactosid (IPTG). After 16 h induction, cells were harvested by centrifugation at 4000 RCF for 15 min at 10 °C, and cell pellets stored at –20 °C until lysis. Cells were lysed by sonication in 10 ml of ice-cold Buffer A (100 mM Tris, 300 mM NaCl, 5 mM of beta-mercaptoethanol (bME), pH 8.0) per wet gram of pellet, and 10 µg/ml DNase (Thermo Fisher Scientific) and 10 mM MgSO₄ were added for DNA removal. The mixture was incubated on ice for 1 h and soluble proteins were removed by centrifugation at 15,000 RCF and 4 °C. The pellet was resuspended using a Potter–Elvehjem Homogenizer in the original volume of chilled Buffer B (2 M Urea in 100 mM Tris, 300 mM NaCl, 5 mM bME, pH 8.0), incubated for 2 h on a rotator, and supernatant was removed by centrifugation. The remaining pellet containing insoluble proteins was resuspended in Buffer C (8 M Urea in 100 mM Tris, 5 mM bME, pH 8.0), incubated overnight, and clarified by two rounds of centrifugation prior loading the sample onto pre-equilibrated NiNTA gravity columns. After capture of the unfolded protein, a step-wise refolding protocol was performed from 6 to 0 M Urea at pH 8.0 in the presence of 100 mM Tris, 300 mM NaCl, 10 mM imidazole and 5 mM bME. For every refolding step, 10 column volumes of buffer were used. Prior elution of the target protein with 250 mM imidazole, columns were washed with 25 column volumes of 20 mM imidazole. Note, the last two buffers did not contain any bME. Elution samples were concentrated using centrifugal filtration (Amicon Ultra-15, MWCO 10 kDa, Merck), and aggregates removed via size-exclusion chromatography using a Superdex S200 16/60 GL column, and 50 mM Tris, 100 mM NaCl, pH 8.0, as running buffer. Peak fractions were pooled and incubated with home-made SUMO-protease overnight at a ratio of 1:1000 (w/w) in the presence of 5 mM bME. A second round of NiNTA ensured separation of the purification tag and the added SUMO protease from the target protein. Flow-through samples were concentrated and further purified using size-exclusion chromatography. Peak fractions were pooled and either used directly for biochemical analysis or frozen in liquid nitrogen as 50 µl aliquots at 1–5 mg/ml. All protein purification steps were performed at 4 °C in a

cold room or on ice. Of note, protein preparations of HgrFlaMP appeared inactive after purification and needed addition of calcium and zinc to restore proteolytic activity.

SDS PAGE and casein zymography. Proteins were separated using 10 or 12% SDS-PAGE under reducing or non-reducing conditions and stained using 1% Coomassie Brilliant Blue G250 (BioRad). The PageRuler Prestained Protein Ladder (Thermo Scientific) was used as molecular weight reference throughout the project. For zymography, protein samples were separated by SDS-PAGE containing 0.1% of bovine casein (Sigma Aldrich) under non-reducing conditions. To remove SDS, gels were first washed with double distilled H₂O, then rinsed twice with renaturation buffer (2.5% Triton X-100 in 50 mM Tris, 200 mM NaCl, 5 mM EDTA, pH 8.0), and incubated in the same buffer for 1 h. Next, gels were rinsed twice with developing buffer (0.02% Brij-35 in 50 mM Tris, 200 mM NaCl, 20 mM CaCl₂, 20 μM ZnCl₂, pH 8.0), and incubated for 2–4 h at room temperature, and stained using 0.5% Coomassie Brilliant Blue R250 (BioRad). Note, due to the lack of cysteine residues, the proteolytic flagellin from *H. gracilis* does not contain any disulfide bridges, and the protein behaves identical on reducing and non-reducing SDS-gels.

In vitro cleavage assays of protein substrates. For protein-based cleavage assays, various proteins were incubated with recombinant flagellinolysin from *Hylemonella gracilis* at various protease to substrate ratios. The following proteins were included in the final substrate mix: bovine caseins (Sigma Aldrich), and the structural and proteolytic flagellins of *Pseudoalteromonas tunicata* (UniProt IDs A4C6G1 and A4C6E0, respectively). The latter was produced as a HAXXH variant, where the general base glutamate of the HEXXH motif is exchanged to an alanine, rendering the enzyme inactive. The flagellin samples from *P. tunicata* were prepared as described above for *H. gracilis*. All cleavage assays were initiated by the addition of 10 mM CaCl₂ and 10 μM ZnCl₂, and negative controls were performed in the presence of 10 mM EDTA. Reactions were stopped after their respective time-points by adding 100 mM EDTA, and analyzed either by SDS-PAGE or mass spectrometry. For the proteomic analysis of the cleavage assay, the sample was split after inactivation of HgrFlaMP, and half was additionally incubated with trypsin to render otherwise too long peptides suitable for mass spectrometry. Reactions were stopped after 1 h at 37 °C by adding formic acid to a final concentration of 0.1%. All samples were purified via C18 tips (100 μl bed, Pierce, Thermo Fisher Scientific, Vienna, Austria) prior HPLC-MS/MS analysis.

Protease specificity profiling using proteome-derived peptide libraries. Proteomic identification of protease cleavage sites (PICS) experiments were performed with trypsin and GluC-derived peptide libraries from whole proteome extracts of *Escherichia coli* as described in great detail in the literature^{45,59,60}. PICS cleavage assays were performed by incubation of 50 to 100 μg of peptide library with recombinant flagellinolysin at a protease to peptide library ratio of 1:50 (w/w) in 50 mM HEPES, 150 mM NaCl, 10 mM CaCl₂, 10 μM ZnCl₂, pH 7.4. For negative controls, 10 mM EDTA was added instead of zinc and calcium ions, and reactions were stopped by the addition of 100 mM EDTA after overnight incubation at 20 °C. Peptides were stable isotope labeled by reductive dimethylation with CH₂O and ¹³CD₂O for controls and protease-treated samples, respectively, cysteines were carbamidomethylated, and reactions stopped by addition of 100 mM Tris-HCl (pH 7.5). Protease-treated samples and controls were subsequently mixed and purified via C18 tips.

High-performance liquid chromatography and mass spectrometry. Peptide separations were carried out on a nanoHPLC instrument (UltiMate U3000 RSLCnano, Thermo Fisher Scientific, Germering, Germany) equipped with an Acclaim PepMap 100 C18 column (500 × 0.075 mm i.d., 3 μm particle size, 100 Å pore size, Thermo Fisher Scientific, Sunnyvale, CA, USA) at a flow rate of 350 nL min⁻¹ and a column oven temperature of 50 °C. Eluent A (H₂O (MilliQ Integral 3 instrument, Millipore, Billerica, MA, USA) + 0.1% formic acid (Sigma-Aldrich)) and eluent B (acetonitrile (VWR International, Vienna, Austria) + 0.1% formic acid) were employed in a stepped linear gradient as follows: 1.0–22% B in 200 min, 22–40% in 40 min, 80% B for 20 min and 1.0% B for 40 min. Samples were injected in full-loop mode (loop volume 1.0 μl) at a peptide concentration of 2.0 mg mL⁻¹. Mass spectrometric data was acquired on a quadrupole-Orbitrap hybrid mass spectrometer (Thermo Scientific QExact Plus benchtop quadrupole-Orbitrap mass spectrometer) hyphenated to the HPLC system via a Nanospray Flex ion source (Thermo Fisher Scientific, Bremen, Germany). A SilicaTip emitter with 360 μm o.d., 20 μm i.d. and a tip i.d. of 10 μm (New Objective, Woburn, MA, USA) was used for electrospray ionization (ESI). Spray voltage was set to 1.5 kV, S-lens RF level to 55.0, and capillary temperature to 320 °C. A MS1 AGC target of 3e6 was employed in an *m/z* range of 400–2000 with a maximum injection time of 100 ms. Each MS1 scans at a resolution setting of 70,000 at 200 *m/z* was followed by 15 data-dependent MS2 scans for PICS experiments and 10 data-dependent MS2 scans for cleavage experiments on protein substrates, both at a resolution of 17,500 at *m/z* 200. HCD fragmentation was carried out at 28.0 NCE in a 2.0 *m/z* isolation window with an AGC target of 1e5 and a maximum injection time of 100 ms. Fragmented peptides were excluded from fragmentation for 30 s. The mass spectrometer was calibrated using the Pierce LTQ Velos ESI Positive Ion Calibration Solution from Life Technologies (Vienna, Austria), and the MS raw data associated with the present paper are available upon request.

Proteomic data analysis. Thermo raw files were converted to .mgf peaklists using ProteoWizard msConvert (v3.0.20075)^{61,62}, and data were analyzed using X!Tandem⁴² within SearchGUI (v3.3.19)⁶³ for peptide-spectrum matching and PeptideShaker (v1.16.45) for statistical analysis⁴³. PICS cleavage data were searched non-specific against a decoy database containing the UniProt annotated *E. coli* K12 proteome (downloaded December 2019)⁶⁴, common contaminants such as human keratins, and proteases under investigation. For the substrate-based cleavage assay, the amino acid sequences of bovine caseins alpha-1, alpha-2, beta, and kappa

were added to the database together with the flagellins from *P. tunicata*. Search parameters included a mass tolerance of 20 ppm for both parent and fragment ions, and methionine oxidation (+ 15.99 Da) and cysteine carboxyamidomethylation (+ 57.02 Da) were set as variable. Search settings allowed for free and dimethylated lysines and N-termini (light dm: + 28.05 Da; heavy dm: + 34.06 Da), and the “quick-pyrrolidone” (− 17.03 Da from Q; − 18.02 Da from E) and “quick-acetyl” (+ 42.01 Da) options were enabled. Results were imported into PeptideShaker, analyzed to a false discovery rate of 1%, and exported as csv-files for further analysis. The complete cleavage sites were reconstructed similar to the web-based program WebPICS⁶⁵, and the generated non-redundant cleavage sites summarized as heat maps using the open source plotting tool Gnuplot (www.gnuplot.info) and iceLogos⁴⁴. Heat maps show either relative occurrence or fold-change over natural abundance of the 20 proteinaceous amino acids at the P6 to P6' cleavage position, while iceLogos⁴⁴ were used to display the percent difference of occurrence compared to natural abundance. Of note, for both the heat maps and iceLogo analysis, the *E. coli* K12 reference proteome was used to ensure comparability with our previous specificity profiling data on clostridial collagenases⁶⁶, the human matrix metalloproteinase family^{45,67}, and the proteolytic flagellin FliA(H) from *Clostridium haemolyticum*¹. Please see Supplementary Tables 1 and 2 for all identified peptide-spectrum matches and cleavage sites. For the protein coverage maps shown in Supplementary Figures 3 to 6 and 9 to 10, the MS Tools web application *Draw Map* (<https://www.hxms.com/mstools/>) was used⁶⁸. Importantly, peptide-based specificity profiling is not designed for the discovery of folded substrates; thus, the provided UniProt identifiers are for reference only.

Received: 24 June 2020; Accepted: 22 October 2020

Published online: 04 November 2020

References

- Eckhard, U. *et al.* Discovery of a proteolytic flagellin family in diverse bacterial phyla that assembles enzymatically active flagella. *Nat. Commun.* **8**, 521 (2017).
- Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **46**, D624–D632 (2018).
- Samatey, F. A. *et al.* Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature* **410**, 331–337 (2001).
- Imada, K. Bacterial flagellar axial structure and its construction. *Biophys. Rev.* **10**, 559–570 (2018).
- Yonekura, K., Maki-Yonekura, S. & Namba, K. Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy. *Nature* **424**, 643–650 (2003).
- Kuwajima, G. Construction of a minimum-size functional flagellin of *Escherichia coli*. *J. Bacteriol.* **170**, 3305–3309 (1988).
- Hu, D. & Reeves, P. R. The remarkable dual-level diversity of prokaryotic flagellins. *mSystems* **5**, e00705-19 (2020).
- Reid, S. D., Selander, R. K. & Whittam, T. S. Sequence diversity of flagellin (fliC) alleles in pathogenic *Escherichia coli*. *J. Bacteriol.* **181**, 153–160 (1999).
- Thomson, N. M., Ferreira, J. L., Matthews-Palmer, T. R., Beeby, M. & Pallen, M. J. Giant flagellins form thick flagellar filaments in two species of marine γ -proteobacteria. *PLoS ONE* **13**, e0206544 (2018).
- Faulds-Pain, A. *et al.* Flagellin redundancy in *Caulobacter crescentus* and its implications for flagellar filament assembly. *J. Bacteriol.* **193**, 2695–2707 (2011).
- Kühn, M. J. *et al.* Spatial arrangement of several flagellins within bacterial flagella improves motility in different environments. *Nat. Commun.* **9**, 5369 (2018).
- Bardy, S. L., Mori, T., Komoriya, K., Aizawa, S.-I. & Jarrell, K. F. Identification and localization of flagellins FlaA and FlaB3 within flagella of *Methanococcus voltae*. *J. Bacteriol.* **184**, 5223–5233 (2002).
- Lövgren, A., Zhang, M. Y., Engström, A. & Landén, R. Identification of two expressed flagellin genes in the insect pathogen *Bacillus thuringiensis* subsp. *alesti*. *J. Gen. Microbiol.* **139**, 21–30 (1993).
- Mohari, B., Thompson, M. A., Trinidad, J. C., Setayeshgar, S. & Fuqua, C. Multiple flagellin proteins have distinct and synergistic roles in *Agrobacterium tumefaciens* motility. *J. Bacteriol.* **200**, e00327-18 (2018).
- Kostrzynska, M., Betts, J. D., Austin, J. W. & Trust, T. J. Identification, characterization, and spatial localization of two flagellin species in *Helicobacter pylori* flagella. *J. Bacteriol.* **173**, 937–946 (1991).
- Wood, T. K., González Barrios, A. F., Herzberg, M. & Lee, J. Motility influences biofilm architecture in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **72**, 361–367 (2006).
- Lemon, K. P., Higgins, D. E. & Kolter, R. Flagellar motility is critical for *Listeria monocytogenes* biofilm formation. *J. Bacteriol.* **189**, 4418–4424 (2007).
- Young, G. M., Schmiel, D. H. & Miller, V. L. A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6456–6461 (1999).
- Chaban, B., Hughes, H. V. & Beeby, M. The flagellum in bacterial pathogens: for motility and a whole lot more. *Semin. Cell Dev. Biol.* **46**, 91–103 (2015).
- Haiko, J. & Westerlund-Wikström, B. The role of the bacterial flagellum in adhesion and virulence. *Biology (Basel)* **2**, 1242–1267 (2013).
- Duan, Q., Zhou, M., Zhu, L. & Zhu, G. Flagella and bacterial pathogenicity. *J. Basic Microbiol.* **53**, 1–8 (2013).
- Payne, J. W. & Smith, M. W. Peptide transport by micro-organisms. *Adv. Microb. Physiol.* **36**, 1–80 (1994).
- Manson, M. D., Blank, V., Brade, G. & Higgins, C. F. Peptide chemotaxis in *E. coli* involves the Tap signal transducer and the dipeptide permease. *Nature* **321**, 253–256 (1986).
- Granato, E. T., Meiller-Legrand, T. A. & Foster, K. R. The evolution and ecology of bacterial warfare. *Curr. Biol.* **29**, R521–R537 (2019).
- Saggu, S. K., Jha, G. & Mishra, P. C. Enzymatic degradation of biofilm by metalloprotease from *Microbacterium* sp. SKS10. *Front. Bioeng. Biotechnol.* **7**, 192 (2019).
- Fleming, D. & Rumbaugh, K. P. Approaches to dispersing medical biofilms. *Microorganisms* **5**, 15 (2017).
- Hatheway, C. L. Toxigenic clostridia. *Clin. Microbiol. Rev.* **3**, 66–98 (1990).
- Kassegne, K. *et al.* Identification of collagenase as a critical virulence factor for invasiveness and transmission of pathogenic *Leptospira* species. *J. Infect. Dis.* **209**, 1105–1115 (2014).
- Eckhard, U., Schönauer, E., Nüss, D. & Brandstetter, H. Structure of collagenase G reveals a chew-and-digest mechanism of bacterial collagenolysis. *Nat. Struct. Mol. Biol.* **18**, 1109–1114 (2011).

30. Radomska, K. A., Wösten, M. M. S. M., Ordoñez, S. R., Wagenaar, J. A. & van Putten, J. P. M. Importance of *Campylobacter jejuni* FliS and FliW in flagella biogenesis and flagellin secretion. *Front. Microbiol.* **8**, 1060 (2017).
31. Wang, Y., Hammes, F., Boon, N. & Egli, T. Quantification of the filterability of freshwater bacteria through 0.45, 0.22, and 0.1 microm pore size filters and shape-dependent enrichment of filterable bacterial communities. *Environ. Sci. Technol.* **41**, 7080–7086 (2007).
32. Spring, S., Jäckel, U., Wagner, M. & Kämpfer, P. *Ottowia thiooxydans* gen. nov., sp. nov., a novel facultatively anaerobic, N₂O-producing bacterium isolated from activated sludge, and transfer of *Aquaspirillum gracile* to *Hylemonella gracilis* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* **54**, 99–106 (2004).
33. Hylemon, P. B., Wells, J. S., Krieg, N. R. & Jannasch, H. W. The genus spirillum: a taxonomic study. *Int. J. Syst. Evol. Microbiol.* **23**, 340–380 (1973).
34. Rutherford, J. C. The emerging role of urease as a general microbial virulence factor. *PLoS Pathog.* **10**, e1004062 (2014).
35. Pawlowski, D. R., Raslawsky, A., Siebert, G., Metzger, D. J. & Karalus, R. J. Identification of *Hylemonella gracilis* as an antagonist of *Yersinia pestis* persistence. *J. Bioterror. Biodefense* **0**, 1–6 (2012).
36. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–226 (2015).
37. McLennan, A. G. The Nudix hydrolase superfamily. *Cell. Mol. Life Sci.* **63**, 123–143 (2006).
38. Grove, A. Regulation of metabolic pathways by MarR family transcription factors. *Comput. Struct. Biotechnol. J.* **15**, 366–371 (2017).
39. Ozin, A. J., Claret, L., Auvray, F. & Hughes, C. The FliS chaperone selectively binds the disordered flagellin C-terminal D0 domain central to polymerisation. *FEMS Microbiol. Lett.* **219**, 219–224 (2003).
40. Jovanovic, S., Barac, M., Macej, O., Vucic, T. & Lacnjevac, C. SDS-PAGE analysis of soluble proteins in reconstituted milk exposed to different heat treatments. *Sensors (Basel)* **7**, 371–383 (2007).
41. Eckhard, U., Schönauer, E. & Brandstetter, H. Structural basis for activity regulation and substrate preference of clostridial collagenases G, H, and T. *J. Biol. Chem.* **288**, 20184–20194 (2013).
42. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
43. Vaudel, M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33**, 22–24 (2015).
44. Colaert, N., Helsen, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 (2009).
45. Eckhard, U. et al. 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 (2016).
46. Schechter, I. & Berger, A. On the size of the active site in proteases I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162 (1967).
47. Schlage, P., Kockmann, T., Sabino, F., Kizhakkedathu, J. N. & auf dem Keller, U. Matrix metalloproteinase 10 degradomics in keratinocytes and epidermal tissue identifies bioactive substrates with pleiotropic functions. *Mol. Cell. Proteomics* **14**, 3234–3246 (2015).
48. Forstnerič, V., Ivičak-Kocjan, K., Plaper, T., Jerala, R. & Benčina, M. The role of the C-terminal D0 domain of flagellin in activation of Toll like receptor 5. *PLoS Pathog.* **13**, e1006574 (2017).
49. Andersen-Nissen, E. et al. Evasion of toll-like receptor 5 by flagellated bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9247–9252 (2005).
50. Smith, K. D. et al. Toll-like receptor 5 recognizes a conserved site on flagellin required for protofilament formation and bacterial motility. *Nat. Immunol.* **4**, 1247–1253 (2003).
51. Meylan, E. et al. Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. *Nature* **437**, 1167–1172 (2005).
52. Li, K. et al. Immune evasion by hepatitis C virus NS3/4A protease-mediated cleavage of the Toll-like receptor 3 adaptor protein TRIF. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2992–2997 (2005).
53. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
54. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
55. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
56. Madeira, F. et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
57. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
58. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
59. Biniossek, M. L. et al. Identification of Protease Specificity by Combining Proteome-Derived Peptide Libraries and Quantitative Proteomics. *Mol. Cell. Proteomics* **15**, 2515–2524 (2016).
60. Schilling, O., Huesgen, P. F., Barré, O., Auf dem Keller, U. & Overall, C. M. Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. *Nat. Protoc.* **6**, 111–120 (2011).
61. Adusumilli, R. & Mallick, P. Data conversion with ProteoWizard msConvert. *Methods Mol. Biol.* **1550**, 339–368 (2017).
62. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
63. Barsnes, H. & Vaudel, M. SearchGUI: A highly adaptable common interface for proteomics search and de novo engines. *J. Proteome Res.* **17**, 2552–2555 (2018).
64. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
65. Schilling, O., auf dem Keller, U. & Overall, C. M. 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 (2011).
66. Eckhard, U., Huesgen, P. F., Brandstetter, H. & Overall, C. M. Proteomic protease specificity profiling of clostridial collagenases reveals their intrinsic nature as dedicated degraders of collagen. *J. Proteomics* **100**, 102–114 (2014).
67. Eckhard, U. et al. Active site specificity profiling datasets of matrix metalloproteinases (MMPs) 1, 2, 3, 7, 8, 9, 12, 13 and 14. *Data Brief* **7**, 299–310 (2016).
68. Kavan, D. & Man, P. MSTools—Web based application for visualization and presentation of HXMS data. *Int. J. Mass Spectrom.* **302**, 53–58 (2011).

Acknowledgements

This study was supported by a post-doctoral fellowship from the Peter and Traudl Engelhorn Foundation (U.E.) and the Austrian Science Fund (FWF, project number: W1213; H.B., C.B. and C.G.H.). A.C.D. is supported by an Ontario Early Research Award and by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant. U.E. is currently supported by the Beatrui de Pinós Program (2020–2023), funded by the Secretary of Universities and Research (Government of Catalonia) and by the Horizon 2020 program of research and innovation of the European Union under the Marie Skłodowska-Curie COFUND actions. Michael Mansfield gratefully acknowledges funding from the Japan Society for the Promotion of Science as a JSPS International Research Fellow (Luscombe Unit, Okinawa Institute of Science and Technology Graduate University), and Michael Kohlberger is acknowledged for technical assistance during PICS library preparations. Financial support for the Christian Doppler Laboratory for Biosimilar Characterization by the

Austrian Federal Ministry of Science, Research, and Economy and by a Start-up Grant of the State of Salzburg is gratefully acknowledged.

Author contributions

U.E. and A.C.D. designed and coordinated the experiments. U.E. cloned, expressed and purified recombinant proteins, and performed biochemical studies. U.E. and C.B. performed mass spectrometry experiments. LC-MS/MS data were collected by C.B. and analyzed by U.E. B.G.L.J. performed phylogenetic analysis, and M.J.M. assisted with bioinformatic analysis. H.B., C.G.H., and A.C.D. supervised the research. All authors contributed with valuable discussions. U.E. and A.C.D. wrote the paper with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-76010-8>.

Correspondence and requests for materials should be addressed to U.E. or A.C.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020