Correlational networking guides the discovery of cryptic proteases for lanthipeptide maturation

Dan Xue[1,8], Ethan A. Older[1,8], Zheng Zhong[2,8], Zhuo Shang[1], Nanzhu Chen[2], Michael D. Walla[3], Shi-Hui Dong[4], Xiaoyu Tang[5], Prakash Nagarkatti[6], Mitzi Nagarkatti[6], Yong-Xin Li [2,7*] and Jie Li[1*]

[1]Department of Chemistry and Biochemistry, University of South Carolina, Columbia, South Carolina, USA. [2]Department of Chemistry, The University of Hong Kong, Pokfulam Road, Hong Kong, China. [3]The Mass Spectrometry Center, Department of Chemistry and Biochemistry, University of South Carolina, Columbia, South Carolina, USA. [4]State Key Laboratory of Applied Organic Chemistry, College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, China. [5]Institute of Chemical Biology, Shenzhen Bay Laboratory, Shenzhen, China. [6]Department of Pathology, Microbiology and Immunology, School of Medicine, University of South Carolina, Columbia, South Carolina, USA. [7]The Swire Institute of Marine Science and Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), The University of Hong Kong, Pokfulam Road, Hong Kong, China

[8]These authors contributed equally to this work.

[*]To whom correspondence may be addressed.
Email: yxpli@hku.hk; li439@mailbox.sc.edu

Abstract

Proteases required for lanthipeptide maturation are not encoded in many of their respective biosynthetic gene clusters. These cryptic proteases hinder the study and application of lanthipeptides as promising drug candidates. Here, we establish a global correlation network bridging the gap between lanthipeptide precursors and cryptic proteases. Applying our analysis to 161,954 bacterial genomes, we establish 6,041 correlations between precursors and cryptic proteases, with 91 prioritized. We use network predictions and co-expression analysis to reveal a previously missing protease for the maturation of class I lanthipeptide paenilan. We further discover widely distributed bacterial M16B metallopeptidases of previously unclear biological function as a new family of lanthipeptide proteases. We show the involvement of a pair of bifunctional M16B proteases in the production of new class III lanthipeptides with high substrate specificity. Together, these results demonstrate the strength of our correlational networking approach to the discovery of cryptic lanthipeptide proteases.

Introduction

The increasing resistance of bacteria to conventional antimicrobial therapy has necessitated the urgent requirement for the discovery of novel antibiotics, including lantibiotics, a type of lanthipeptides, that have exhibited antimicrobial activity against a range of multi-drug-resistant (MDR) bacteria[1,2]. Lanthipeptides also possess anti-fungal[3], anti-HIV[4], and antinociceptive activities[5], attracting increasingly more research interests for discovery and application of lanthipeptides.

As one of the most common ribosomally synthesized and post-translationally modified peptides (RiPPs)[6,7], the biosynthesis of all four different classes of lanthipeptides involves a crucial protease-mediated cleavage between the leader and core peptides for final maturation. However, only two types

of proteases have been relatively well studied. The first is the subtilisin-like serine protease LanP, employed by class I and II lanthipeptides[8–10]. The second is the papain-like cysteine protease domain of the LanT transporter protein, involved exclusively in the biosynthesis of class II lanthipeptides[10–13]. Due to the absence of protease-encoding genes in most characterized class III and IV lanthipeptide biosynthetic gene clusters (BGCs), the maturation of these two classes is barely understood, with FlaP[14] and AplP[15] being recently reported as potential proteases for class III lanthipeptides.

With exponentially increasing microbial genome sequences available, increasingly more lanthipeptide BGCs are being identified as lacking BGC-associated genes to encode proteases[16–18]. Thus, a missing link between these BGCs and their cryptic proteases hinders the discovery, biosynthetic study, heterologous production, and bioengineering of these potentially bioactive lanthipeptides. We hypothesize that lanthipeptide BGCs without any colocalized protease-encoding genes may rely on proteases encoded elsewhere in the genome. Thus, in this study, we developed a genome mining workflow and used correlation analysis complemented by co-expression analysis to establish the first global correlation network between lanthipeptide precursor peptides and proteases from 161,954 bacterial genome sequences. This correlation network provides guidance for targeted discovery of cryptic lanthipeptide proteases encoded by genes outside of the BGCs. As a proof of principle, we selected two representative correlations from the network for study, leading to a simultaneous discovery of new lanthipeptides and responsible cryptic proteases. Particularly, a family of bacterial M16B metallopeptidases with previously unclear biological functions was identified as being responsible for maturation of several new class III lanthipeptides.

Results

Establishment of a global correlation network between lanthipeptide precursor peptides and proteases. We established a global precursor-protease network for all four classes of lanthipeptides using 161,954 bacterial genomes obtained from the NCBI RefSeq database. Analyzing this large number of genomes with antiSMASH 5.0[19], we identified 21,225 putative lanthipeptide BGCs widely distributed across all bacterial taxa. These BGCs harbor 29,489 highly diverse precursors (Supplementary **Fig. 1**). We observed that approximately one third of lanthipeptide BGCs do not harbor any protease genes, especially class III system (Supplementary **Fig. 2**). BGCs without any colocalized protease-encoding genes, may rely on proteases encoded outside of those BGCs for leader peptide removal[6,12,20]. In this scenario, we hypothesized that the specificity between precursors and corresponding proteases still exists, at least to some extent, based on two observations: (i) only certain homologs of a protease in a genome have proteolytic activity against a specific precursor[21,22], and (ii) the proteolytic activity is affected by core peptide modifications[14].

Thus, based on the specificity between precursors and proteases, we performed a global correlation analysis, aiming to identify potential lanthipeptide maturation proteases, regardless of their gene location inside or outside of a BGC. Due to the large number of proteases contained in the genomes performing general functions, it was not rational or practical to directly correlate lanthipeptide precursors to all the proteases in the genomes. Thus, we started with pathway-specific proteases encoded by lanthipeptide BGCs. We hypothesized that pathway-specific proteases likely evolved from general proteases encoded elsewhere in the genome and, at the large scale of the dataset, pathway-specific proteases could collectively represent all the functional domains of cryptic proteases encoded outside of the BGCs. Indeed, by searching proteases from 21,225 putative lanthipeptide BGCs, we generated a library of 44,260 prospective lanthipeptide proteases, representing 120 unique Pfam domains (Supplementary **Table 1**). In contrast, previously characterized proteases associated with lanthipeptides only encompass 6 Pfam domains. We used these 120 Pfam domains to search for

proteases from the full set of 161,954 bacterial genomes, resulting in 23,777,967 putative lanthipeptide related proteases. Grouping these proteases based on their sequence similarity using MMseqs2[23] led to the formation of 288,416 groups of proteases. Among these, 10,263 groups each containing 100 or more members were selected for downstream analysis. We also applied the same clustering approach to 29,489 lanthipeptide precursors, forming 4,527 groups. Among these, 263 groups each containing ten or more precursors were selected for downstream analysis (**Fig. 1a**). We then sought to identify links between the selected groups of proteases and precursors. To reduce the effect of phylogenetic relatedness, we performed Spearman rank correlation analysis individually for each genus, leading to identification of 6,041 significant correlations (ρ>0.3, pAdj<1E-5, **Fig. 1b**) between precursors and proteases. These significant correlations ranged from 6 phyla and 114 genera, suggesting that widely distributed proteases, even encoded by genes outside of the BGCs, may function against specific lanthipeptide precursors.

We next focused on class III lanthipeptides due to their elusive maturation process[17]. We identified 1,833 putative precursors encoding class III lanthipeptides. These precursor peptides were classified into 217 groups based on sequence similarity, including 18 groups that each contained at least 10 members. We selected the correlations identified in at least ten genomes (I>=10) for further analysis, leading to prioritization of 91 significant correlations (ρ>0.3, pAdj<1E-5, I>=10) between 8 groups of precursors and 87 groups of proteases. These significant correlations were distributed in two phyla, including 80 correlations in Actinobacteria and 11 correlations in Firmicutes. The core information of these prioritized significant correlations is summarized in **Fig. 1c**, Supplementary **Fig. 3**, and Supplementary **Table 2**, with some representative discoveries described below.

Among the 91 significant correlations, metallopeptidases appeared in many significant correlations, suggesting that the metallopeptidase superfamily may play an important role in the maturation of class III lanthipeptides. In contrast, families of serine protease and cysteine protease have been reported for the maturation of class I and II lanthipeptides[13,20,24,25]. Remarkably, among the 91 significant correlations representing 864 precursors, 758 precursors (88%) were strongly correlated with only one or two groups of proteases. For example, the precursors of groups Pre_5, Pre_24, and Pre_115 were only correlated to two, two, and one group(s) of proteases (ρ>0.3, pAdj<1E-5, I>=10; **Fig. 1c** and Supplementary **Table 2**) at the genus level of *Streptomyces*, *Paenibacillus*, and *Lentzea*, respectively. None of these 888 candidate proteases had been characterized and only 19 of them are encoded by genes within the corresponding BGCs. Thus, this result exhibited the potential of our correlation network in identifying cryptic proteases for maturation of class III lanthipeptides.

On the other hand, five groups of precursors, Pre_77, Pre_117, Pre_134, Pre_181, and Pre_228, were each correlated to multiple groups of proteases, forming five multiple-correlation clusters at the genus level of *Amycolatopsis*, *Streptomyces*, *Alkalihalobacillus*, *Lactobacillus*, and *Rhodoccocus*, respectively (**Fig. 1c**, Supplementary **Fig. 3**). Together, these five groups only represent the remaining 12% of the precursors from the 91 significant correlations. Some groups of proteases, e.g. Prot_1169, Prot_2308, and Prot_9513, were observed to share similar functional domains, which may partially account for their simultaneous correlations with the same precursor group. At first glance, this multiple-correlation pattern presented a challenge to identify a responsible protease. However, the correlation strength between different groups of proteases and the same group of precursors appeared to be different based on Spearman's correlation coefficient (ρ), likely suggesting different strength in the specificity between the precursors and proteases involved. In addition, we noticed that multiple-correlation clusters could be distinguished by the classes of their correlated proteases. For example, proteases associated with Pre_117 belong primarily to the α/ß hydrolase class while those associated with Pre_134 mainly fall into the metallopeptidase class (Supplementary **Table 2**).

To completement correlational analysis at the genomic level, we integrated co-expression analysis

using whole genome transcriptomic data. We rationalized that even if protease-encoding genes are not colocalized with lanthipeptide BGCs, they are likely to co-express with the BGCs in the same period. By integrating co-expression analysis, we aimed to simplify the multiple-correlation pattern. For example, two groups of proteases, Prot_241 (metal-dependent hydrolase; PF10118) and Prot_1365 (carbon-nitrogen hydrolase; PF00795), were both strongly correlated with one group of precursors, Pre_5, at the genomic level (ρ>0.3, pAdj<1E-5, I>=10; **Fig. 1c**). Taking advantage of the wide distribution of Pre_5 in the genus *Streptomyces*, we compiled publicly available transcriptomic data consisting of 80 samples from three strains which expressed Pre_5. Co-expression analysis using these data led to a reduction of the original two strong correlations to only one group, Prot_24 (ρ>0.4, pAdj<0.05). Thus, we demonstrate here establishment of a global correlation network, integrating co-expression data to fortify lanthipeptide precursor-protease association predictions.

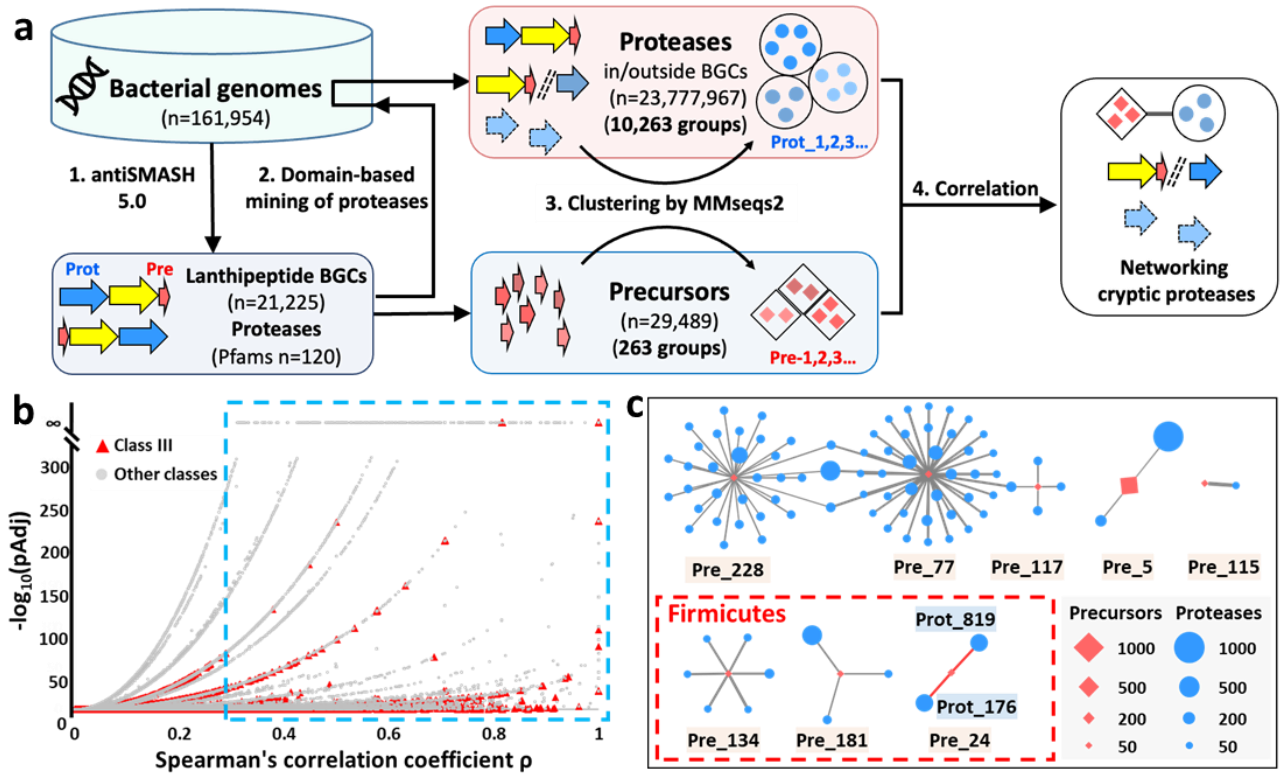

Fig. 1 Overview of the precursor-protease correlational networking for discovery of cryptic lanthipeptide proteases. **a**, A genome-mining workflow for discovering potential cryptic lanthipeptide proteases, including four steps: (1). antiSMASH 5.0 was used to analyze 161,954 bacterial genomes for identification of lanthipeptide BGCs and BGC-associated proteases; (2). A Pfam domain-based mining of proteases (either inside or outside of BGCs) was performed against entire bacterial genomes based on the 120 Pfams of BGC-associated proteases identified in step (1); (3). Proteases and precursors were clustered by sequence similarity using MMseqs2; and (4). Correlation networks at both genomic and transcriptomic levels were constructed to link precursors and cryptic proteases that are not encoded by the BGCs. Pre: precursor peptides; Prot: proteases. **b**, Volcano plot of 81,323 correlations at the genus level. The blue box encompasses 6,041 significant correlations. Red triangles indicate correlations for class III lanthipeptides. **c**, Prioritized significant correlations (ρ>0.3, pAdj<1E-5) for class III lanthipeptides from eight genera, showing only correlations identified in at least ten genomes (I>=10). Red diamonds represent groups of precursors and blue circles represent groups of proteases. Shape sizes are proportional to the number of precursors or proteases identified in a group at the genus level. Edges connecting diamonds and circles indicate a significant correlation, with the increasing thickness of the edge representing the increasing strength of the correlation (Spearman's ρ). The 11 correlations in Firmicutes (indicated in red box) were formed between 3 groups of class III precursors (Pre_24, 134, and 181, totally 68

precursors) and 11 groups of peptidases. The remaining 80 correlations found in Actinobacteria were formed between 5 groups of class III precursors (Pre_5, 77, 115, 117, and 228, totally 796 precursors) and 76 groups of proteases.

Validation of the precursor-protease correlation network. We used previously characterized representative lanthipeptide proteases to validate the correlation network by first examining better understood class I and II lanthipeptide systems. In our network, we located the previously studied class I lanthipeptides epidermin[26] and gallidermin[26] in one group, Pre_1, due to their precursor similarity, while class II lanthipeptides thuricin[27], cerecidin[28], and cytolysin[29] were found in four groups of precursors, Pre_11, Pre_12, Pre_10, and Pre_13 (cytolysin is comprised of two distinct precursors resulting in its representation in two precursor groups) (Supplementary **Fig. 4**). The BGCs producing these six lanthipeptides also encode pathway-specific LanP and LanT homologs, two types of proteases known to be involved in class I and II lanthipeptide biosynthesis[10,12,30] (Supplementary **Fig. 5)**. Indeed, these LanP and LanT homologs each were shown in our network to correlate with Pre_1, Pre_11, Pre_12, Pre_10, and Pre_13, respectively (Supplementary **Table 3)**.

Additionally, we used the previously characterized class III lanthipeptide protease FlaP for validation. Besides the genera of *Kribbella* and *Streptomyces* that were reported to harbor a FlaP-like protease colocalized with a class III precursor[14], we identified FlaA homologs in additional genera, mainly *Amycolatopsis* and *Jiangella* (Supplementary **Fig. 6)**. FlaP-like proteases were simultaneously identified in our network to significantly correlate with FlaA in *Amycolatopsis* and *Jiangella*, with ρ=0.77, pAdj=3E-16 and ρ=1, p=8E-78, respectively (**Supplementary File 1**). Such strong correlations supported the effectiveness of applying our correlation network to class III lanthipeptides.

Identification of a cryptic protease for maturation of class I lanthipeptide paenilan. To identify lanthipeptide proteases that are not encoded by their respective BGCs, we first selected a group of class I lanthipeptide precursors, Pre_49 (**Fig. 2a**), which presumably encodes paenilan as a representative product based on the precursor sequences[31] (**Fig. 2b**). All 44 BGCs of this group do not harbor any protease-encoding genes within the BGCs and the responsible proteases for maturation of this group of lanthipeptides have not been reported[31], consistent with the recent finding that increasingly more class I lanthipeptide BGCs have been noticed to lack a pathway-specific protease[12,16,18]. In our network, 34 groups of proteases were shown to be strongly correlated with Pre_49 (ρ>0.3, pAdj<1E-5, I>=10). Additionally, using transcriptomic data of 18 samples from *Paenibacillus polymyxa* ATCC 842 expressing the precursor of Pre_49 (**Supplementary File 2**), we reduced the putative correlations from 34 groups to four groups, Prot_686, Prot_1570, Prot_2941, and Prot_8704, that showed strong co-expression with Pre_49 (ρ>0.4, pAdj<0.05; **Fig. 2c**). Thus, we expressed the corresponding proteases of these four groups from *P. polymyxa* ATCC 842 as recombinant proteins and tested their proteolytic activity in vitro against the PllB/PllC-modified precursor peptide from the same strain. LC-MS analysis revealed that paenilan, the same product as the native strain, was produced in the assay with the member of Prot_686 (**Fig. 2d**). Prot_686 belongs to the S8 family of peptidases. This family has been previously reported for the maturation of class I lanthipeptides[8,20]. Thus, we used the correlation of Prot_686 with a paenilan-producing precursor to demonstrate that proteases predicted by our network analysis could in fact be responsible for lanthipeptide maturation. This motivated us to further exploit the power of our network in the less understood class III lanthipeptide system.

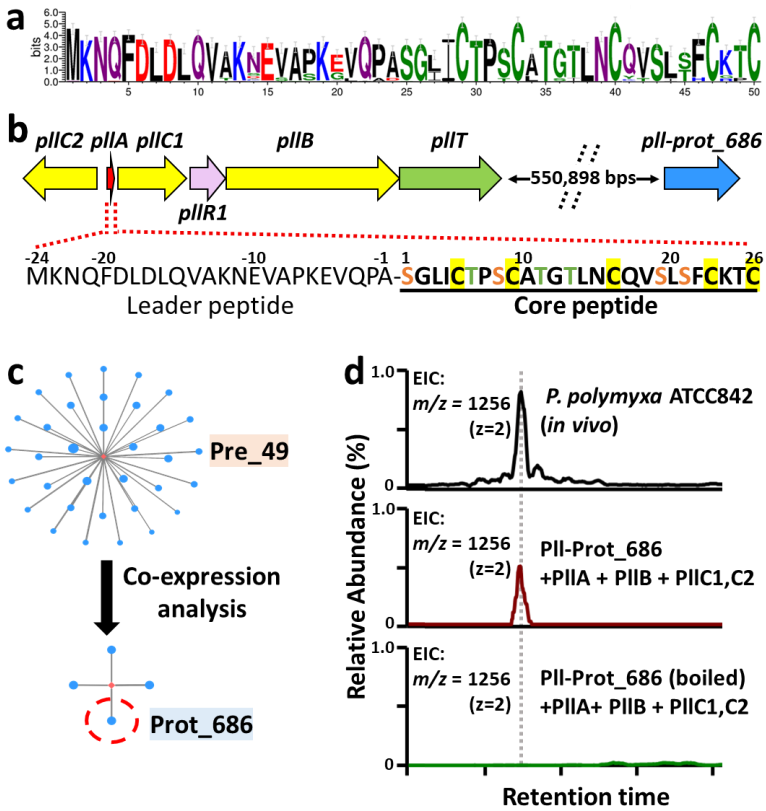

Fig. 2 Identification of a cryptic protease from Prot_686 for maturation of paenilan. a, Sequence logo of the group Pre_49. **b**, A representative BGC of Pre_49 presumably producing paenilan based on the precursor sequence. All 44 BGCs of this group do not harbor any protease-encoding genes, suggesting cryptic proteases encoded elsewhere in the genome. Dashed lines separating *pllT* and *prot_686* indicate a gap in their respective genetic locations. This type of notation is also used in other figures below. The distance of this gap, 550,898 base pairs, is represented here to convey magnitude. **c**, Precursor-protease correlation complemented by co-expression analysis suggested candidate proteases(s). **d**, LC-MS analysis indicated the proteolytic activity of Prot_686 for the maturation of paenilan.

Networking analysis reveals a new family of lanthipeptide proteases and links it to the production of new class III lanthipeptides. Class III lanthipeptide BGCs are abundant in Firmicutes[7]. However, the lack of a protease-encoding gene in most class III BGCs has been a challenge toward exploiting these lanthipeptides[17]. Therefore, we focused on Firmicutes for a simultaneous discovery of new class III lanthipeptides and their associated cryptic proteases. Our attention was drawn to two groups of proteases designated as Prot_176 and Prot_819, both of which had not been studied as lanthipeptide proteases and showed strong correlations (ρ=0.69, pAdj=4E-64) to a group of uncharacterized precursors, Pre_24 (**Fig. 1c**). Pre_24, consisting of 115 precursors, is solely distributed in Firmicutes and represents the largest precursor group in Firmicutes (**Fig. 3a**). These 115 precursors are distributed across 48 class III lanthipeptide BGCs from three genera, *Alkalihalobacillus*, *Bacillus*, and *Paenibacillus*. The strains harboring *pre_24* all possess one pair of *prot_176* and *prot_819* present side by side outside of the 48 BGCs (**Fig. 3b**). Among these 48 BGCs, an extra pair of *prot_176* and *prot_819* was identified within 17 BGCs (**Fig. 3b**). In addition, many other strains from *Alkalihalobacillus*, *Bacillus*, and *Paenibacillus* that do not encode Pre_24 also contain a pair of *prot_176* and *prot_819* in their genomes (**Fig. 3b**).

To investigate Prot_176 and Prot_819 as potential proteases for Pre_24, we first produced and characterized representative class III lanthipeptides from Pre_24. From the network, we selected two

Pre_24-encoding BGCs for study. The first was from *Paenibacillus thiaminolyticus* NRRL B-4156, designated as *ptt*. The gene organization of the *ptt* BGC is shown in **Figure 3c**. Genes *pttA1*, *pttP1*, and *pttP2* belong to *pre_24*, *prot_819*, and *prot_176*, respectively (**Fig. 1c**). Outside of the *ptt* BGC, there is an additional pair of protease-encoding genes, *bcn-genomeP1* and *bcn-genomeP2*, that also belong to *prot_819* and *prot_176,* respectively. The *ptt* BGC was heterologously expressed in the host *Bacillus subtilis* 168, leading to the detection of five lanthipeptides named paenithopeptins A-E (**Fig. 3d**). These compounds were also produced by the native strain. We isolated paenithopeptin A, the most abundant product, and established its chemical structure via high-resolution mass spectrometry (HRMS), MS/MS fragmentation, Marfey's analysis, and 1D and 2D NMR (Supplementary **Figs. 7-19** and Supplementary **Table 4**). Detailed structure elucidation is described in the Supplementary Information. Paenithopeptins B-E identified by MS/MS fragmentation analysis have different overhangs of amino acids at the N-terminus compared to paenithopeptin A.

The second BGC selected was from *Bacillus nakamurai* NRRL B-41092, designated as *bcn*. The gene organization of the *bcn* BGC is shown in **Figure 3e**. The genes *bcnA1*, *bcn-genomeP1*, and *bcn-genomeP2* correspond to *pre_24*, *prot_819*, and *prot_176*, respectively (**Fig. 1c**), with *bcn-genomeP1* and *bcn*-genomeP2 located outside of the *bcn* BGC. Heterologous expression of *bcn* in the host *Bacillus subtilis* 168 led to the detection of two lanthipeptides named bacinapeptins A and B, which are dimethylated at the N-terminus and do not bear a disulfide bond due to the lack of the second cysteine in the respective core peptides (**Fig. 3f**).

Notably, paenithopeptins A-E and bacinapeptins A and B represent new class III lanthipeptides. The tricyclic ring system (labionin in a larger disulfide-bridged ring) in paenithopeptins A-E has not been previously reported in Firmicutes.

Next, we used PttP1/PttP2 and Bcn-genomeP1/P2 as examples to confirm the proteolytic activity of Prot_819/Prot_176 against Pre_24 as the precursor peptide. We began our investigation with both in vivo and in vitro characterization of the *ptt* BGC. The chemical structures of paenithopeptins A-E suggest that they were all derived from the precursor PttA1. Thus, for in vivo assays, different combinations of four genes, *pttA1*, *pttKC*, *pttP1*, and *pttP2*, were constructed into a series of pDR111-based integrative plasmids (Supplementary **Table 5**) and expressed in an engineered host *B. subtilis* 168 with homologs of *pttP1* and *pttP2* deleted by us. For in vitro assays, we individually expressed and purified recombinant His-tagged PttA1, PttKC, PttP1, or PttP2 (Supplementary **Fig. 20**). We then incubated PttA1, with or without PttKC, followed by adding PttP1, PttP2, or both, respectively, for an incubation of 12 hours. For both in vivo and in vitro assays, we observed that (i) PttP1 and PttP2 were simultaneously required to produce the highest yield of the major product, paenithopeptin A (in vivo) and paenithopeptin A′ (in vitro, lacking a disulfide bond compared to paenithopeptin A presumably due to absence of a disulfide bond-forming oxygenase from the host *B. subtilis* 168[32,33]. Hereafter, Paenithopeptin A′-E′ represent the counterparts of paenithopeptin A-E lacking the disulfide bond) and (ii) the PttP1/PttP2 complex has specificity against PttKC-modified precursor bearing a labionin ring (**Fig. 4a** and Supplementary **Fig. 21**). Likewise, we performed the in vitro characterization of Bcn-genomeP1/P2 against BcnA1, with or without BcnKC, using the same assay conditions mentioned above. We observed similar results as described for PttP1/PttP2. Since we did not add BcnMT, the predicted methyltransferase, in the assay, the observed product was demethylated bacinapeptin A, designated as bacinapeptin A′ (**Fig. 4b**). Taken together, these results indicated Prot_819/Prot_176 as previously unknown lanthipeptide proteases and established their role in the maturation of several new class III lanthipeptides represented by paenithopeptins and bacinapeptins.

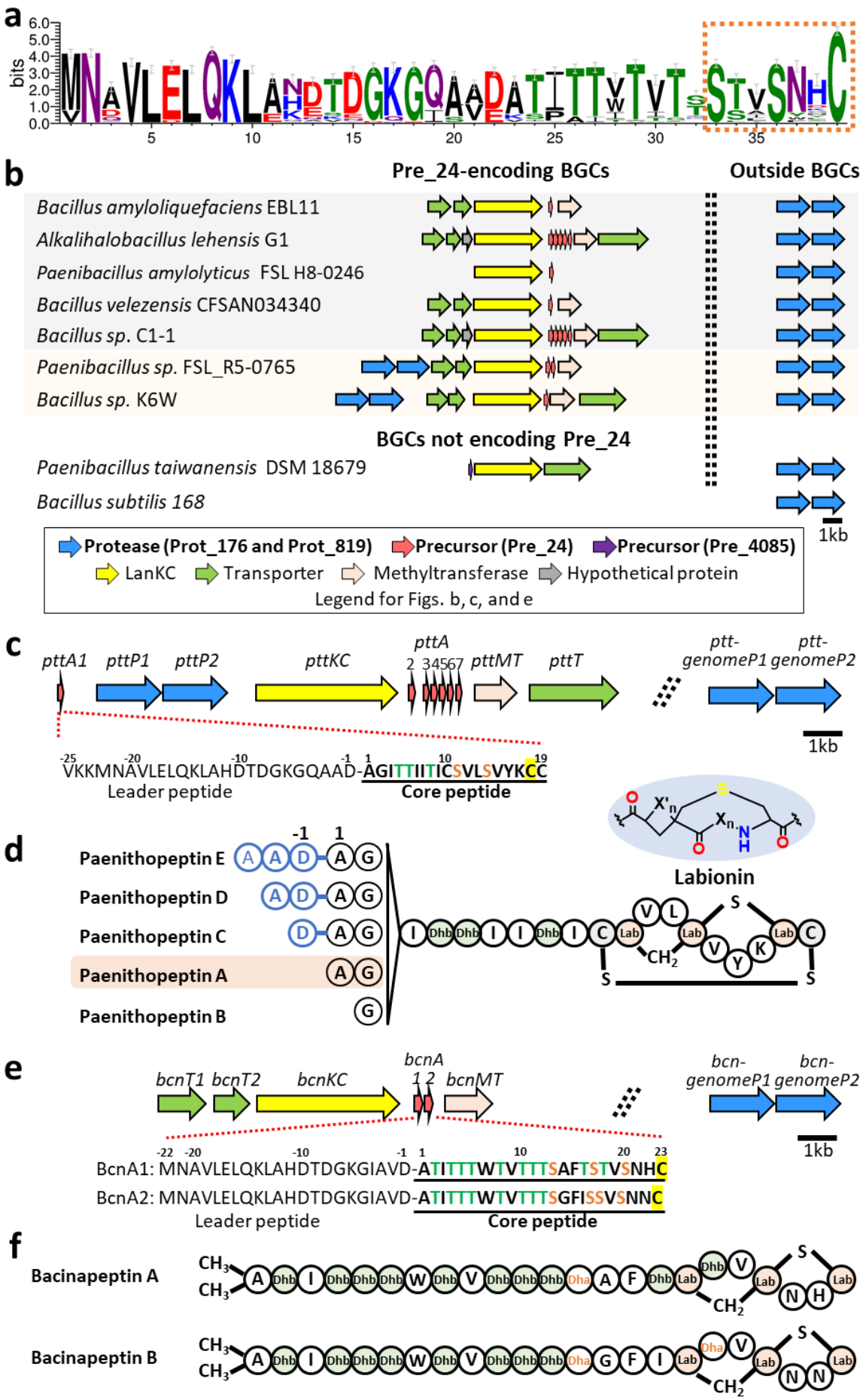

Fig. 3 Discovery of a family of previously unknown lanthipeptide proteases linked to the production of new class III lanthipeptides. **a**, Sequence logo of Pre_24. The putative core peptides contain only one S-$(X)_2$-

S-$(X)_{3\text{-}5}$-C motif (X stands for random amino acid) as opposed to two or three such motifs commonly seen in previously known class III lanthipeptides. **b**, Representative class III lanthipeptide BGCs encoding Pre_24, showing distribution of a pair of protease-encoding genes, *prot_176* and *prot_819*, outside of the BGCs and an extra pair in a minority of the BGCs. In addition, a pair of *prot_176* and *prot_819* is also present in the genomes of many strains that do not harbor Pre_24-encoding BGCs. **c**, A Pre_24-encoding BGC, *ptt*, selected for study. The *ptt* BGC putatively encodes a characteristic class III lanthipeptide synthetase (PttKC) and seven precursors, PttA1-PttA7. While *pttA2-pttA7* are together, *pttA1* is located separately. The putative protease-encoding genes *pttP1* and *pttP2* are adjacent to each other downstream of *pttA1*. An extra pair of *pttP1* and *pttP2* homologs is located outside of the *ptt* BGC. *pttMT* and *pttT* are predicted to encode a methyltransferase and a transporter, respectively. **d**, Chemical structures of paenithopeptins A-E produced by the *ptt* BGC. **e**, Another Pre_24-encoding BGC, *bcn*, selected for study. The *bcn* BGC putatively encodes a class III lanthipeptide synthetase (BcnKC), two precursors (BcnA1 and BcnA2), a methyltransferase (BcnMT), and a transporter (BcnT). A pair of protease-encoding genes is absent in the *bcn* BGC, but present elsewhere in the genome, designated as *bcn-genomeP1* and *bcn-genomeP2*. **f**, Chemical structure of bacinapeptins A and B produced by the *bcn* BGC.

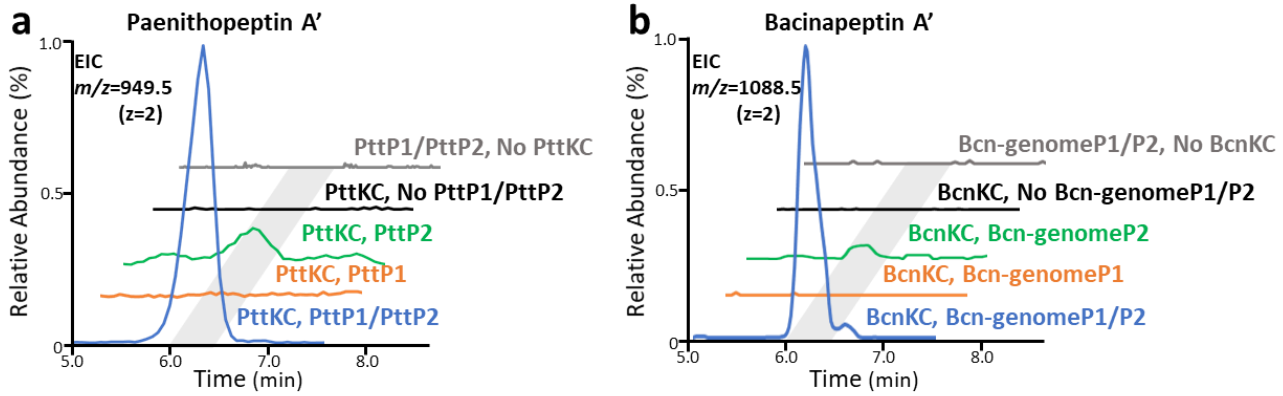

Fig. 4 Characterization of the proteolytic activity of PttP1/PttP2 and Bcn-genomeP1/P2. a, In vitro characterization of the proteolytic activity of PttP1/PttP2. PttP1 and PttP2 were required at the same time to produce the highest yield of paenithopeptin A′. In addition, the lack of PttKC in the assay completely abolished the production, suggesting that PttP1/PttP2 shows specificity towards PttKC-modified precursor bearing a labionin ring. **b**, In vitro characterization of the proteolytic activity of Bcn-genomeP1/P2, showing that Bcn-genomeP1 and Bcn-genomeP2 were required at the same time for the highest production yield and that the lack of BcnKC abolished the production.

Different efficiencies of Prot_819/Prot_176 proteases suggest potentially evolved protease specificity for class III lanthipeptides. Due to the wide distribution of *prot_819*/*prot_176* members in bacterial genomes, we investigated the potential activity of different Prot_819/Prot_176 members for leader removal. The distribution of *prot_819*/*prot_176* pairs were classified into four scenarios for our comparison (**Fig. 3b**): (i) within the *ptt* BGC, i.e. *pttP1*/*pttP2*, (ii) outside of the *ptt* BGC, designated as *ptt-genomeP1*/*P2*, (iii) outside of other Pre_24-encoding class III lanthipeptide BGCs, e.g. *bcn-genomeP1*/*P2* from *Bacillus nakamurai* NRRL B-41092, and (iv) in the genomes that do not harbor any Pre_24-encoding class III lanthipeptide BGCs, e.g. *ymfF*/*ymfH* from *B. subtilis* 168. We compared the in vitro proteolytic activity of these Prot_819/Prot_176 members using the same assay conditions described above. The results show that YmfF/YmfH, Bcn-genomeP1/P2, and Ptt-genomeP1/P2 were able to cleave PttKC-modified PttA1 and produce paenithopeptins, but the efficiency, reflected as the production yield under the same conditions, was ~4%, ~13%, and ~15% compared to that of PttP1/PttP2, respectively (**Fig. 5**).

We further evaluated the activity of these Prot_819/Prot_176 proteases against an additional

member of Pre_24, i.e. BcnA1 from *B. nakamurai* NRRL B-41092. As expected, we observed that YmfF/YmfH, Bcn-genomeP1/P2, Ptt-genomeP1/P2, and PttP1/PttP2 were all active against BcnKC-modified BcnA1, with an efficiency trend similar to that against PttKC-modified PttA1 (**Fig. 5**, Supplementary **Fig. 22**).

On the other hand, we tested whether the Prot_819/Prot_176 members could act on precursors from groups other than Pre_24. The precursor selected from a class III lanthipeptide BGC of *Paenibacillus taiwanensis* DSM 18679 was designated as PbtA (belonging to Pre_4085) (**Fig. 3b**). We expressed PbtA and Pbt-genomeP1/P2 as recombinant proteins for in vitro proteolytic assays. Despite a variety of assay conditions used, we did not observe any obvious maturation products when PbtA, with or without PbtKC, was incubated with Pbt-genomeP1/P2, YmfF/YmfH, Bcn-genomeP1/P2, Ptt-genomeP1/P2, or PttP1/PttP2.

Taken together, the specificity between Pre_24 and Prot_819/Prot_176 supports our precursor-protease correlation network. Furthermore, considering the wide distribution of Prot_819/Prot_176, the different efficiencies of Prot_819/Prot_176 members appeared supportive of our hypothesis that certain proteases with general functions in the genome might have evolved more or less specific activity against class III lanthipeptides, and extra copies of them might have further evolved into pathway-specific proteases for enhanced activity and specificity. We therefore performed a phylogenetic analysis of Prot_176 members, the catalytic component of the Prot_819/Prot_176 pair. A phylogenetic tree was built at the level of genus *Paenibacillus* where paenithopeptins were isolated. Indeed, pathway-specific Prot_176 appeared in closely related lineages in the phylogenetic tree, forming an independent branch from other Prot_176 members encoded by genes outside of the BGCs, implying a process of gene divergence (Supplementary **Fig. 23**). Thus, the substrate specificity of Prot_819/Prot_176 likely gained during evolution provides evidence for our correlational networking that uses substrate specificity to look for cryptic proteases.

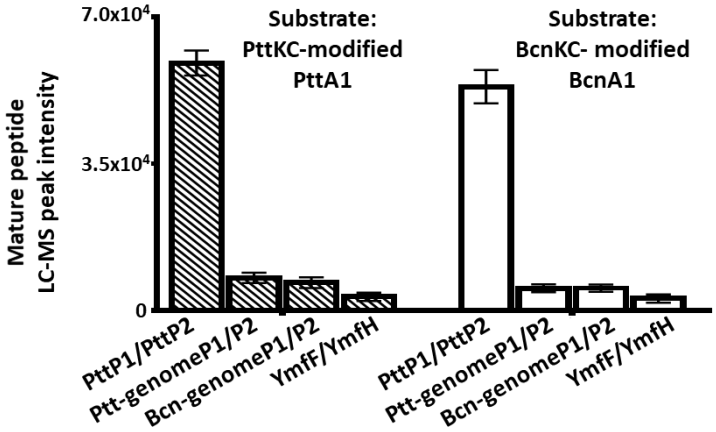

Fig. 5 Different efficiencies of Prot_819/Prot_176 members. PttP1/PttP2, Ptt-genomeP1/P2, Bcn-genomeP1/P2, and YmfF/YmfH against PttKC-modified PttA1 and BcnKC-modified BcnA1, respectively, with PttP1/PttP2 showing the highest efficiency in both cases. Error bars indicate standard deviation across triplicate experiments.

PttP1/PttP2 functions as a heterodimeric metallopeptidase and shows specificity and unique activity for leader peptide removal. Due to the highest activity of PttP1/PttP2 shown above, we further characterized this protease pair in detail as a representative member of Prot_819/Prot_176. We first performed a bioinformatics analysis of PttP1 and PttP2, which suggested that both belong to the family of zinc-dependent M16 peptidases. Specifically, PttP2 possesses an HXXEH motif essential for Zn^{2+} binding and catalytic activity, while PttP1 contains an R/Y pair in the C-terminal domain for

substrate binding (**Fig. 6a**). The characteristic sequences of PttP1 and PttP2 are highly reminiscent of two known M16B peptidases from *Sphingomonas* sp. A1, Sph2681 and Sph2682, that form a heterodimer[34,35]. To investigate this, we performed homology modeling of PttP1 and PttP2, using the heterodimeric crystal structure of Sph2681/Sph2682 (PDB code 3amj) as the template. The best model of the PttP1 and PttP2 heterodimer (hereafter called PttP1/PttP2) displayed a high degree of structural similarity toward the template structure (Supplementary **Fig. 24**). We performed a pulldown assay and confirmed the heterodimeric interaction between PttP1 and PttP2 (**Fig. 6b**).

Next, we confirmed the importance of the Zn-binding HXXEH motif and the R/Y pair in the proteolytic activity of PttP1/PttP2 against PttKC-modified PttA1 precursor. We individually mutated H67, E70, and H71 of the HXXEH motif and R298 and Y305 of the R/Y pair to Ala, leading to decreased activity of PttP1/PttP2 in vitro for each mutation (Supplementary **Fig. 25**). Simultaneous mutation of all these five residues into Ala residues completely abolished the production of paenithopeptins A′-E′ (Supplementary **Fig. 25**). In addition, the metal-chelating compound, *o*-phenanthroline, significantly inhibited the activity of PttP1/PttP2 (Supplementary **Fig. 26**).

We further investigated how PttP1/PttP2 led to production of a series of compounds with different N-terminal overhangs from PttA1 (**Fig. 6c**). We first performed an in vitro enzymatic assay with an incubation of PttA1 and PttKC for four hours, followed by adding both PttP1 and PttP2 for different incubation lengths, ranging from 15 minutes to 36 hours. We observed that at the 15 minute-point, paenithipeptin E′ with the longest leader overhang was formed as the major product (**Fig. 6d**), suggesting an endopeptidase activity of PttP1/PttP2. Over time, paenithipeptin E′ diminished while paenithipeptin A′ (with no leader overhang) was accumulated (**Fig. 6d**), followed by removal of the N-terminal alanine of paenithipeptin A′ to form paenithipeptin B′ after an extended incubation of 36 hours (**Fig. 6d**). This transformation process suggested an aminopeptidase activity of PttP1/PttP2. Next, we directly used paenithopeptin E′ as a substrate for incubation with PttP1/PttP2, leading to the production of paenithipeptin A′ (**Fig. 6e**). We also incubated paenithopeptin A with PttP1/PttP2, resulting in the generation of paenithopeptin B (**Fig. 6f**). These experiments confirmed the aminopeptidase activity of PttP1 and PttP2. Taken together, our results show that the QAAD motif in the leader of PttA1 appeared to be specifically cleaved by PttP1/PttP2, through an initial endopeptidase activity to partially remove the leader followed by an aminopeptidase activity to successively remove the remaining overhangs.

We then investigated whether PttP1 and PttP2 have proteolytic activity against PttA2-PttA7. Although *pttA2-pttA7* are located separately from *pttA1* in the *ptt* BGC (**Fig. 3c**), they also belong to the Pre_24 group in our network. As expected, in vitro enzymatic assays using conditions described above revealed corresponding products cleaved from the conserved Q-A-(A/I)-(D/E) motif of PttA2-PttA7 (Supplementary **Fig. 27**). The products also showed different overhangs of amino acids at the N-terminus, suggesting again both endo- and aminopeptidase activities of PttP1/PttP2. In addition, formation of these products was also dependent on the precursor modification by PttKC (Supplementary **Fig. 27**).

We further studied substrate specificity of PttP1/PttP2 in comparison with AplP-like proteases recently discovered from Actinobacteria for the maturation of class III lanthipeptides[15]. Since the native producer of paenithopeptins A-E also harbors an AplP-like protease, we expressed this protease as a recombinant protein and tested whether it could cleave PttA1 (belonging to Pre_24), with or without modification by PttKC. However, although AplP-like proteases were hypothesized as a universal strategy for leader peptide removal of class III lanthipeptides[15], we could not detect any products, suggesting that the most abundant group of class III precursors in Firmicutes was not recognizable by Actinobacteria-derived AplP-like proteases (**Fig. 6g**). On the other hand, we cloned AciA (an experimentally validated precursor for AplP-like proteases; belonging to Pre_2495) from

Catenulispora acidiphila DSM 44928 for a comparison experiment. When AciA, with or without modification by AciKC, was incubated with a pair of PttP1/PttP2 homolog from *C. acidiphila* DSM 44928, no products were detected (**Fig. 6g**). These results indicated that PttP1/PttP2 and AplP-like proteases have respective specificity against different groups of precursors and supported our protease-precursor correlation network.

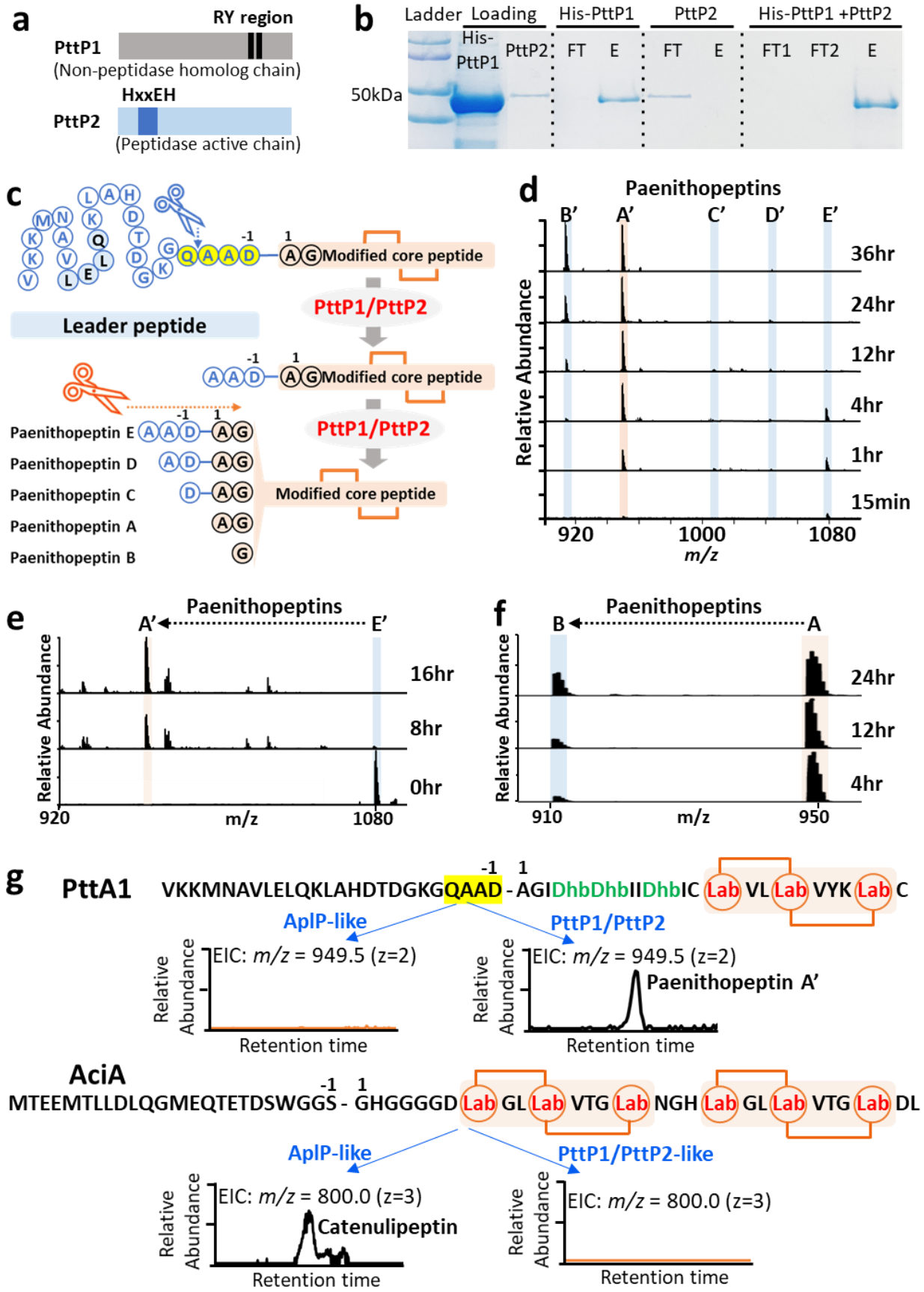

Fig. 6 PttP1/PttP2 is a heterodimeric metallopeptidase showing substrate specificity for leader peptide removal. a, Representation of the Zn^{2+}-binding HXXEH motif of PttP2 and the substrate-binding R/Y region of PttP1. **b**, A pull-down assay showing protein-protein interaction between PttP1 and PttP2. His_6-tag-free PttP2 was immobilized in the nickel affinity column only when His_6-tagged PttP1 was present, as reflected by examining the flowthrough using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). FT: flowthrough; E: elute. **c**, Representation of the leader peptide processing by PttP1/PttP2, through an initial endopeptidase activity to partially remove the leader followed by an aminopeptidase activity to successively remove the remaining overhangs. **d**, LC-MS analysis of an in vitro assay of PttP1/PttP2, showing production of paenithopeptins A′-E′ at different time points. **e**, LC-MS spectrum showing transformation of paenithopeptin E′ to paenithopeptin A′ by aminopeptidase activity of PttP1/PttP2. **f**, LC-MS spectrum showing transformation of paenithopeptin A to paenithopeptin B by aminopeptidase activity of PttP1/PttP2. **g**, In vitro assay of PttP1/PttP2-like proteases in comparison with AplP-like proteases recently reported for class III lanthipeptide maturation, indicating that PttP1/PttP2-like and AplP-like proteases have respective substrate specificity. The conserved Q-A-(A/V)-(D/E) motif of Pre-24, absent in the precursor peptides for AplP-like proteases, was specifically cleaved by PttP1/PttP2-like proteases.

Discussion

Using a correlation-guided genome-mining strategy, we established a global network bridging the gap between bacterial lanthipeptide precursors and cryptic proteases which are not encoded by their respective BGCs. This proof-of-principle study showcases the promise of precursor-protease correlational networking in targeted discovery of both missing proteases for previously known lanthipeptides (e.g., Prot_686 for the maturation of class I lanthipeptide paenilan) and new families of lanthipeptide proteases involved in the maturation of novel products (e.g., M16B metallopeptidases leading to the production of new class III lanthipeptides paenithopeptins and bacinapeptins).

Many natural product BGCs are found not to harbor a full set of necessary genes, especially tailoring enzymes, and must rely on enzymes encoded elsewhere in the genome for biosynthesis. These cryptic enzymes are particularly important when the missing biosynthetic steps are responsible for installation of critical moieties. For example, the antibiotic gentamicin C complex and the antitumor reagent geldanamycin both rely on remote genes located apart from their BGCs for methylation steps that are essential for their biological activities[36,37]. In lanthipeptides, a final protease performs an indispensable maturation step. Notably, with the advancement of sequencing and genome-mining technologies, increasingly more lanthipeptide BGCs lacking protease-encoding genes are being identified, consistent with our data that about one third of lanthipeptide BGCs do not harbor protease genes. Particularly, most reported class III and IV lanthipeptide BGCs lack protease-encoding genes [17]. Computational workflows have been developed to predict lanthipeptide BGCs with high accuracy by following canonical co-localized biosynthetic gene rules[19,38]. However, these approaches cannot account for cryptic proteases located far outside of the BGC. Focusing on protein domain architecture rather than global sequence similarity, we used functional domain-based genome mining to search for cryptic proteases in 161,954 bacterial genomes, expanding the pool of lanthipeptide protease Pfam domains to 120 as opposed to the 6 domains of previously reported lanthipeptide proteases LanP, LanT, FlaP, and AplP. Furthermore, using Spearman rank correlation analysis within each bacterial genus, our networking strategy linked cryptic proteases to lanthipeptide precursors.

As mentioned above, multiple-correlation clusters represent a challenge in determining candidate proteases. Additionally, the 6,041 precursor-protease correlations identified in this study may contain weak correlations which would require more data to refine. Addressing this, we integrated co-

expression analysis to complement our approach. We envision that with increasing availability of bacterial whole genome transcriptome data, we can continuously increase the power of our correlation network.

The biological function of widely distributed M16B metallopeptidases has been unclear in bacteria [39]. Our network guided the discovery of a new function of these proteins as lanthipeptide proteases. We characterized PttP1/PttP2-like proteases of the M16B family and showed their high substrate specificity toward lanthipeptide precursors, despite being encoded outside of the respective BGCs. This result supports the basis of the correlational networking approach that cryptic proteases exhibit a measurable degree of substrate specificity. Furthermore, varying efficiencies of PttP1/PttP2-like proteases from different genetic loci supports a process of functional evolution and provides a method to identify higher efficiency proteases for more productive bioengineering of lanthipeptides.

Consistent with the literature[7,40], our study revealed Firmicutes as a prolific source for discovery of the underexplored class III lanthipeptides that often lack pathway-specific proteases. Exploring this feature in the network, we identified new lanthipeptides paenithopeptins and bacinapeptins which represent the largest precursor group in Firmicutes. Our network further reveals the wide distribution of uncharacterized proteases in Firmicutes as exemplified by PttP1/PttP2-like proteases. Their discovery and characterization showcase the potential of our networking approach to link cryptic proteases to corresponding BGCs.

Taken together, our correlational networking approach takes advantage of readily available data and easily accessible computational platforms which can be developed into a streamlined process to discover cryptic proteases for the exploitation of lanthipeptides. Furthermore, our approach is envisioned to be applicable to other natural product classes that lack necessary genes in their respective BGCs.

Online content

Any methods, additional references, source data, supplementary information, acknowledgements, author contributions, and competing interests are available.

References

1. Gomes, K. M., Duarte, R. S. & Bastos, M. do C. de F. Lantibiotics produced by Actinobacteria and their potential applications (A review). *Microbiol. (United Kingdom)* **163**, 109–121 (2017).
2. D., J. *et al.* Efficacy of the new lantibiotic NAI-107 in experimental infections induced by multidrug-resistant gram-positive pathogens. *Antimicrob. Agents Chemother.* **55**, 1671–1676 (2011).
3. Mohr, K. I. *et al.* Pinensins: The first antifungal lantibiotics. *Angew. Chemie. Int. Ed.* **54**, 11254–11258 (2015).
4. Férir, G. *et al.* The lantibiotic peptide labyrinthopeptin A1 demonstrates broad anti-HIV and anti-HSV activity with potential for microbicidal applications. *PLoS One* **8**, e64010 (2013).
5. Iorio, M. *et al.* A glycosylated, labionin-containing lanthipeptide with marked antinociceptive activity. *ACS Chem. Biol.* **9**, 398–404 (2014).
6. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
7. Walker, M. C. *et al.* Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genomics* **21**, 387 (2020).
8. Xu, Y. *et al.* Structure of the nisin leader peptidase NisP revealing a C-terminal autocleavage

activity. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 1499–1505 (2014).
9. Montalbán-lópez, M., Deng, J., Heel, A. J. Van & Kuipers, O. P. Specificity and application of the lantibiotic protease NisP. **9**, 1–16 (2018).
10. Tang, W. *et al.* Applications of the class II lanthipeptide protease LicP for sequence-specific, traceless peptide bond cleavage. *Chem. Sci.* **6**, 6270–6279 (2015).
11. Caetano, T., van der Donk, W. & Mendo, S. Bacteroidetes can be a rich source of novel lanthipeptides: The case study of *Pedobacter lusitanus*. *Microbiol. Res.* **235**, 126441 (2020).
12. Repka, L. M., Chekan, J. R., Nair, S. K. & van der Donk, W. A. Mechanistic understanding of lanthipeptide biosynthetic enzymes. *Chem. Rev.* **117**, 5457–5520 (2017).
13. Wang, J., Ge, X., Zhang, L., Teng, K. & Zhong, J. One-pot synthesis of class II lanthipeptide bovicin HJ50 via an engineered lanthipeptide synthetase. *Sci. Rep.* **6**, 38630 (2016).
14. Völler, G. H., Krawczyk, B., Ensle, P. & Süssmuth, R. D. Involvement and unusual substrate specificity of a prolyl oligopeptidase in class III lanthipeptide maturation. *J. Am. Chem. Soc.* **135**, 7426–7429 (2013).
15. Chen, S. *et al.* Zn-dependent bifunctional proteases are responsible for leader peptide processing of class III lanthipeptides. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2533–2538 (2019).
16. Barbosa, J., Caetano, T. & Mendo, S. Class I and class II lanthipeptides produced by *Bacillus spp*. *J. Nat. Prod.* **78**, 2850–2866 (2015).
17. Hegemann, J. D. & Süssmuth, R. D. Matters of class: coming of age of class III and IV lanthipeptides. *RSC Chem. Biol.* 110–127 (2020).
18. Zhang, Q., Doroghazi, J. R., Zhao, X., Walker, M. C. & van der Donk, W. A. Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in *Actinobacteria*. *Appl. Environ. Microbiol.* **81**, 4339–4350 (2015).
19. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
20. Corvey, C., Stein, T., Düsterhus, S., Karas, M. & Entian, K. D. Activation of subtilin precursors by *Bacillus subtilis* extracellular serine proteases subtilisin (AprE), WprA, and Vpr. *Biochem. Biophys. Res. Commun.* **304**, 48–54 (2003).
21. Ren, H. *et al.* Discovery and characterization of a class IV lanthipeptide with a nonoverlapping ring pattern. *ACS Chem. Biol.* **15**, 1642–1649 (2020).
22. Wiebach, V. *et al.* An amphipathic alpha-helix guides maturation of the ribosomally-synthesized lipolanthines. *Angew. Chemie - Int. Ed.* **59**, 16777–16785 (2020).
23. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
24. Tang, W. *et al.* Applications of the class II lanthipeptide protease LicP for sequence-specific, traceless peptide bond cleavage. *Chem. Sci.* **6**, 6270–6279 (2015).
25. Ortega, M. A. *et al.* Substrate specificity of the lanthipeptide peptidase ElxP and the oxidoreductase ElxO. *ACS Chem. Biol.* **9**, 1718–1725 (2014).
26. Bierbaum, G. The biosynthesis of the lantibiotics epidermin, gallidermin, Pep5 and epilancin K7. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **69**, 119–127 (1996).
27. Favret, M. E. & Yousten, A. A. Thuricin: The bacteriocin produced by *Bacillus thuringiensis*. *J. Invertebr. Pathol.* **53**, 206–216 (1989).
28. Wang, J. *et al.* Cerecidins, novel lantibiotics from *Bacillus cereus* with potent antimicrobial activity. *Appl. Environ. Microbiol.* **80**, 2633–2643 (2014).
29. Tang, W. & Van Der Donk, W. A. The sequence of the enterococcal cytolysin imparts unusual lanthionine stereochemistry. *Nat. Chem. Biol.* **9**, 157–159 (2013).
30. Skaugen, M., Andersen, E. L., Christie, V. H. & Nes, I. F. Identification, characterization, and

expression of a second, bicistronic, operon involved in the production of lactocin S in *Lactobacillus sakei* L45. *Appl. Environ. Microbiol.* **68**, 720–727 (2002).

31. Park, J. E., Kim, H. R., Park, S. Y., Choi, S. K. & Park, S. H. Identification of the biosynthesis gene cluster for the novel lantibiotic paenilan from *Paenibacillus polymyxa* E681 and characterization of its product. *J. Appl. Microbiol.* **123**, 1133–1147 (2017).
32. Meindl, K. *et al.* Labyrinthopeptins: A new class of carbacyclic lantibiotics. *Angew. Chemie - Int. Ed.* **49**, 1151–1154 (2010).
33. Dorenbos, R. *et al.* Thiol-disulfide oxidoreductases are essential for the production of the lantibiotic sublancin 168. *J. Biol. Chem.* **277**, 16682–16688 (2002).
34. Martins, A. M. *et al.* A two-component protease in Methylorubrum extorquens with high activity toward the peptide precursor of the redox cofactor pyrroloquinoline quinone. *J. Biol. Chem.* **294**, 15025–15036 (2019).
35. Maruyama, Y., Chuma, A., Mikami, B., Hashimoto, W. & Murata, K. Heterosubunit composition and crystal structures of a novel bacterial M16B metallopeptidase. *J. Mol. Biol.* **407**, 180–192 (2011).
36. Li, S. *et al.* Methyltransferases of gentamicin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1340–1345 (2018).
37. Yin, M. *et al.* The missing C-17 O-methyltransferase in geldanamycin biosynthesis. *Org. Lett.* **13**, 3726–3729 (2011).
38. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
39. Dabonné, S. *et al.* Cloning, expression and characterization of a 46.5-kDa metallopeptidase from *Bacillus halodurans* H4 sharing properties with the pitrilysin family. *Biochim. Biophys. Acta - Gen. Subj.* **1725**, 136–143 (2005).
40. Grigoreva, A. *et al.* Identification and characterization of andalusicin: N-terminally dimethylated class III lantibiotic from *Bacillus thuringiensis* sv. *andalousiensis*. *iScience* **24**, 102480 (2021).

Methods

BGCs, precursors, and proteases detection. Publicly available bacterial genomes were downloaded from NCBI RefSeq database (accessed in Aug. 2019) and analyzed by antiSMASH 5.0 with default parameters. Lanthipeptide BGCs and precursors were detected automatically by antiSMASH 5.0. Lanthipeptide BGCs with a protein containing both LANC_Like domain (PF05147) and the Pkinase domain (PF00069) were classified as class III or class IV BGCs and further distinguished by profile-Hidden Markov models (pHMMs) developed previously.

Proteases in lanthipeptide BGCs were obtained by searching for keywords: "peptidase", "proteinase", "protease", "hydrolase", "beta-lactamase" from antiSMASH generated annotations. These proteases were further subjected to Pfam domain analysis by hmmsearch[1] (HMMER v3.3) with default parameters against the Pfam-A database[2] (v33.1). Pfam domains with hit score >0 with at least 5 occurrences in all lanthipeptide BGCs were selected and manually curated to generate a pool of 120 Pfam domains. These domains were considered as potential lanthipeptide protease Pfams (Supplementary File 1). Proteases in bacterial genomes were obtained by searching for proteins in bacterial genomes and selecting proteases which contained at least one Pfam domain from the pool with hit score >0.

Protease and precursor clustering. Proteases in bacterial genomes were clustered using MMseqs2 (v11.e1a1c) using the easy-cluster workflow with the following parameters: easy-cluster --min-seq-id 0.45 --single-step-clustering --cluster-mode 1 (using "connected component" mode to cover more remote homologs). Lanthipeptide precursor sequences were clustered by MMseqs2 with the same parameters except for a larger similarity cutoff: easy-cluster --min-seq-id 0.6 --single-step-clustering --cluster-mode 1.

Correlation network construction. Number of occurrences of protease groups and precursor groups were counted in each bacterial genome. Spearman correlation coefficient and corresponding p-value were calculated for each protease group to precursor group pairs in each genus. P-values were further adjusted by Benjamini–Hochberg procedure[3] (false-discovery rate). All statistics were calculated in R (v.3.6). Calculated correlations were filtered by Spearman correlation coefficient (ρ), adjusted p-values (pAdj) and number of genomes containing the precursor-protease group pair (I). Networks were visualized by Cytoscape[4] (v.3.8.2), in which the size of node indicates the number of proteases or precursors in a genus, width of edge indicates the correlation strength (Spearman coefficient, ρ).

Transcriptomic analysis. Transcriptomic data from *Streptomyces albidoflavus* J1074, *Streptomyces coelicolor* A3(2) and *Streptomyces davaonensis* JCM 4913 were downloaded from the NCBI SRA database. SRA runs were filtered by paired library layout, random or cDNA library selection, and average length >=100. A full list of SRA runs used in this study are shown in Supplementary File 1. BBMap[5] (v38.90) was used to remove base pairs with low quality as well as adapter sequences and PhiX sequences with the following parameters: qtrim=rl ktrim=r mink=11 trimq=10 minlen=45 corresponding to read quality cutoff of 10 and read length cutoff of 45 bp. rRNA sequences were filtered out using SortMeRNA[6] (v.4.3.1) with default parameters with the smr_v4.3_sensitive_db_rfam_seeds database. Quality controlled transcriptomic data were mapped to corresponding genome sequences by BWA mem[7] (v0.7.17) with default parameters. Mapped reads were counted by featureCounts[8] (v2.0.1) with the following parameters: -M -O --fraction -p --primary --minOverlap 40 -Q 10 (minimum mapping quality score of 10, also counted multi-mapping reads). Transcripts per million (TPM) were calculated for each gene.

To calculate proteases that co-expressed with precursors, TPM of genes that belong to the same protease group (or precursor group) were summed to generate a protease and precursor TPM matrix. The matrix was further trimmed to contain only precursor groups and protease groups that exist in the correlation network calculated by genomic data. Spearman correlation coefficient between precursor groups and protease groups

were calculated and p-values were adjusted by Benjamini–Hochberg procedure. Correlations presented in results of all 3 strains were taken.

For self-sequenced transcriptomic data from *Paenibacillus polymyxa* ATCC 842, the same quality control and statistical analysis workflow was performed.

Phylogenetic analysis. For proteases inside or outside BGCs at the genus *Paenibacillus*, protein sequences of all Prot_176 or Prot_819 within the Pre_24-containing BGCs, or in Pre_24-containing genomes (but outside BGCs) were selected. For proteases in genomes that didn't encode Lanthipeptide BGCs, protein sequences of all these members in Prot_176 or Prot_819 were clustered by CD-HIT[9] (v4.8.1) with default parameters to remove redundant sequences sharing >=90% sequence similarity, and representative sequences were randomly selected.

Sequences from proteases inside BGCs, outside BGCs, random representatives, as well as a manually added outgroup (WP_023988223.1) were aligned together using MAFFT[10] (v7.480) with the high accuracy linsi method. The resulting alignment was subjected to phylogenetic tree construction by IQ-TREE[11] (v2.1.3) using substitution model LG+G4. Bootstrap was calculated using the Ultrafast bootstrap[12] algorithm implemented in IQ-TREE with 1000 bootstrap replicates. The resulted tree was re-rooted at the manually added outgroup and visualized with Interactive Tree of Life[13].

Sequence logos. Full precursor peptide sequences were aligned by MAFFT (v7.480) using the linsi method and trimmed by trimAl[14] (v1.4.rev22) with the -automated1 option. The resulted alignment was used to generate sequence logos using WebLogo[15] (v3.7.4).

General materials, strains, and culture conditions. Biochemicals and media components for bacterial cultures were purchased from Thermo Fisher Scientific Co. Ltd. (USA) unless otherwise stated. Restriction endonucleases were purchased from New England Biolabs, Inc. (USA). Chemical reagents were purchased from standard commercial sources.

All strains used in this study are listed in Supplementary Table 4. *Paenibacillus thiaminolyticus* NRRL B4156, *Bacillus nakamurai* NRRL B41092, *P. taiwanensis* DSM 18679*, P. polymyxa* ATCC 842 and *B. subtilis* 168 strains were cultured in Tryptic Soy Broth (TSB) at 30°C. *Catenulispora acidiphila* DSM44928 was grown in ISP2 medium at 28°C. *E. coli* DH10B strains were grown in Luria-Bertani broth (LB) at 37°C. *E. coli* BL21(DE3) strains were cultured in Terrific Broth (TB) at 37°C for general growth and 22°C for protein expression.

General genetic manipulation methods. Genomic DNA was isolated according to previously described methods[16]. PCR amplifications were carried out on an Eppendorf® Mastercycler® Nexus X2 Thermal Cycler (Eppendorf Co., Ltd. Germany) using PrimeSTAR HS DNA polymerase (Takara Biotechnology Co., Ltd. Japan). DNA sequencing was performed by Eton Bioscience, Inc. (USA).

All plasmids used in this study are listed in Supplementary Table 5. General plasmid construction procedure is described as follows. Target genes were amplified by PCR from genomic DNA of corresponding strain, using corresponding primers listed in Supplementary Table 6. Amplifications were confirmed by gel electrophoresis, and PCR products were purified using an E.Z.N.A.® Gel Extraction Kit (Omega Bio-tek, Inc., USA). Target DNA fragment(s) and selected vector were assembled by Gibson Assembly[17] to generate corresponding plasmids, using NEBuilder® HiFi DNA Assembly master mix (New England Biolabs, Inc., USA). Sequences of the resulting plasmid products were confirmed by DNA sequencing. In all heterologous expression plasmids, native promoters for genes were included in initial PCR amplifications.

Mutagenesis was performed by PCR-based site-directed mutagenesis[18]. Briefly, primers incorporating the desired base changes were designed and applied through PCR to amplify target genes containing desired

mutations. Mutated PCR products were used for Gibson assembly with corresponding vectors to generate plasmids containing mutations. Sequences of all resulting plasmids were confirmed by DNA sequencing as described above.

The PttP1 and PttP2 homologs in the genome of *B. subtilis* 168, *ymfF* and *ymfH*, were knocked out using the CRISPR-Cas9 system[19]. The sgRNA Designer tool provided by the Broad Institute[20] was used to check the *ymfF* and *ymfH* genes and a high-scoring 20-nucleotide (nt) sequence was identified. The candidate sgRNA sequence was synthesized as two complementary oligonucleotides and inserted between the BsaI sites of pJOE8999 to construct pJOE8999.1. Then two 700-bp fragments flanking the target region were amplified by PCR and inserted between the SfiI sites of pJOE8999.1 to generate pJOE8999.2. *B. subtilis* 168 was transformed with pJOE8999.2 and grown on LB agar plates containing 5 μg/mL kanamycin and 0.2% mannose. After incubation at 30 °C for 1 day, colonies were placed on LB plates without antibiotics and incubated at 50 °C. On the next day, they were streaked on LB plates to obtain single colonies at 42°C. Finally, the colonies were tested again for plasmid loss by transferring single colonies to LB agar plates with kanamycin. Knock-out of the *ymfF-ymfH* sequence in selected colonies was confirmed by colony PCR using the outer primers from the 700-bp homology templates.

pDR111 constructs were integrated into the chromosome of *B. subtilis* 168 or *B. subtilis* 168_Δ*ymfFH* as previously described[21] for heterologous expression of the *ptt* cluster. Plasmid constructed with pHis8 or pET-28(a)+ were transferred into *E. coli* BL21(DE3) by electroporation for protein expression.

Paenithopeptin Isolation. *P. thiaminolyticus* NRRL B4156 was inoculated in a 150 ml Ultra Yield flask (Thomson Scientific) containing 30 mL of TSB broth and shaken at 220 rpm, 30 °C, overnight, as seed culture. The seed culture was used to inoculate 500 mL of TSB in 2.5 L Ultra Yield flasks (20×) for a total of 10 L TSB, which were subsequently shaken at 220 rpm at 30 °C for 3 days. The bacterial broth was extracted with an equal volume of butanol, and the combined organic phase was concentrated in vacuo to yield the extract, which was partitioned on a reversed-phase C18 open column with a 25% stepwise gradient elution from 50% H_2O/MeOH to 100 % MeOH. Pure MeOH with 1% formic acid was applied for the final elution. The MeOH/formic acid fraction was further purified by HPLC (Phenomenex Luna RP-C18, 250 mm × 10 mm, 5 μm, 100 Å, 3.5 mL/min isocratic elution at 32% H_2O/MeCN over 30 min with constant 0.1% formic acid) to yield paenithopeptin A. The paenithopeptin A characterization data can be found in the Supplementary Information.

High-resolution ESI-MS, MSn, and NMR characterization of paenithopeptins and bacinapeptins. High-resolution ESI-MS spectra and MSn analysis of paenithopeptins A-E and bacinapeptins A and B were recorded on a Thermo Q-Exactive mass spectrometer. ^{1}H, ^{13}C, ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, ^{1}H-^{1}H NOESY, ^{1}H-^{13}C HSQC, and ^{1}H-^{13}C HMBC NMR spectra for paenithopeptin A in DMSO-d6 were acquired on a Bruker Avance III HD 400 MHz spectrometer with a 5 mm BBO ^{1}H/^{19}F-BB-Z-Gradient prodigy cryoprobe, a Bruker Avance III HD 500 MHz spectrometer with a PA BBO 500S2 BBF-H-D_05 Z SP probe, or a Bruker Avance III HD Ascend 700 MHz equipped with 5 mm triple-resonance Observe (TXO) cryoprobe with Z-gradients. Data were collected and reported as follows: chemical shift, integration multiplicity (s, singlet; d, doublet; t, triplet; m, multiplet), coupling constant. Chemical shifts were reported using the DMSO-d6 resonance as the internal standard for ^{1}H-NMR DMSO-d6: δ = 2.50 p.p.m. and ^{13}C-NMR DMSO-d6: δ = 39.6 p.p.m.

Determination of absolute configuration of Paenithopeptin A. Paenithopeptin A (1.0 mg) was hydrolyzed in 6 M HCl (700 μL) at 115 °C for 10 h with stirring in a sealed thick-walled reaction vessel, after which the hydrolysate was concentrated to dryness under N_2 gas. The resulting hydrolysate was resuspended in distilled H_2O (700 μL) and dried again. This process was repeated three times to remove the acid completely. The hydrolysate was divided into two portions (2 × 500 μg) for chemical derivatization with 1-fluoro-2,4-dinitrophenyl-5-L-alanine amide (L-FDAA) and 1-fluoro-2,4-dinitrophenyl-5-D-alanine amide (D-FDAA). Each

hydrolysate sample was treated with 1 M $NaHCO_3$ (100 μL) and either L- or D-FDAA (100 μL, 1% solution in acetone) at 40 °C for 1 h. The reaction was then neutralized with addition of 1 M HCl (150 μL) and diluted with MeCN (150 μL). The final sample was analyzed by HPLC-DAD-ESIMS (Kinetex C_{18} HPLC column, 4.6 × 100 mm, 2.4 μm, 100 Å, 1.0 mL/min gradient elution from 95% to 38% H_2O/MeCN over 30 min with constant 0.1% formic acid; positive and negative ionization modes; UV at 340 nm). For the preparation of FDAA derivatives of amino acid standards, 50 μL of each amino acid (L-Ser, L-Asp, *N*-Me-L-Phe, *N*-Me- L-Asp, 2*S*,3*S*-MeDap, 2*S*,3*R*-MeDap, 2*S*,4*S*-hyGlu, 2*S*,4*R*-hyGlu, and 2*S*,3*R*-MeGlu) (50 mM in H_2O) reacted with L- or D-FDAA (100 μL, 1% solution in acetone) at 40 °C for 1 h in the presence of 1 M $NaHCO_3$ (20 μL). The reaction was quenched by 1 M HCl (20 μL) and diluted with MeCN (810 μL) followed by HPLC-DAD-MS analysis using the same column and elution condition as above. The L- and D-FDAA derivatives were detected by either UV or extracted ion chromatograms. Absolute configurations of amino acid residues in paenithopeptin A were established by comparing the retention times of FDAA-derivatized peptide hydrolysate with those of amino acid standards.

Protein expression and purification. Precursor peptides were produced in a form tagged with 6×His fused and a small ubiquitin-like modified (SUMO) fusion partner at the N-terminus. Modification proteins were expressed with an 8×His tag at the N-terminus. *E. coli* BL21(DE3) cells were transformed with pHis8 or pET28 derivative plasmids containing genes encoding precursor peptides or modification enzymes (Supplementary Table 5). A single colony was used to inoculate a 10 mL culture of LB supplemented with 50 mg/L kanamycin. The culture was grown at 37 °C for 8 h and used to inoculate 1 L of LB with kanamycin. Cells were grown at 37 °C to OD600 ~0.6-0.8, then cooled to 16 °C before IPTG was added to a final concentration of 0.2 mM. The culture was incubated at 20 °C for an additional 12 h. Cells were harvested by centrifugation at 5,000 ×g for 30 min at 4 °C. Cell pellets were resuspended in 30 mL of lysis buffer (20 mM Tris, pH 8.0, 300 mM NaCl, 25 mM imidazole, 5% glycerol) and the suspension was sonicated on ice for 20 min to lyse the cells. Cell debris was removed by centrifugation at 15,000 ×g for 60 min at 4 °C. The supernatant was loaded onto a 3ml His SpinTrap™ column (GE Healthcare) previously charged with Ni^{2+} and equilibrated in lysis buffer. The column was washed with 10 mL of wash buffer I (35 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl) and 10ml of wash buffer II (55 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl). The protein was eluted stepwise with elution buffer I (250 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl) and elution buffer II (500 mM imidazole, 20 mM Tris, pH 8.0, 300 mM NaCl). Resulting elution fractions were collected and analyzed by SDS-PAGE. Fractions containing target proteins were combined and concentrated using an Amicon Ultra-15 Centrifugal Filter Unit (10 kDa for precursor peptide, 30kDa for modification enzymes, MWCO, Millipore). The resulting protein sample was stored at −70 °C.

Pulldown assay. The interaction between PttP1 and PttP2 was studied by a pulldown assay using His SpinTrap™ columns (GE Healthcare) with a bed volume of 200 μl. Binding buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl, 5% (v/v) glycerol) was used to immobilize His_6-PttP1 (75 nmol) on the column then the column was washed with wash buffer (binding buffer with 30 mM imidazole added). PttP2 was applied to induce complex formation and elution buffer (binding buffer with 250 mM imidazole added) was used to elute the His_6-PttP1/PttP2 complex. A nontreated column control was also run to identify and eliminate false positives caused by nonspecific binding of PttP2, as well as a control with His_6-PttP1 but no PttP2.

In vitro enzymatic assays. Assays were based on previous reports on class I and III lanthipeptide enzymes[22–25]. The class I precursor peptide was treated as previously described for the dehydration of nisin with some modification. Briefly, precursor peptide (pllA; 20 uM) was incubated with cell extracts obtained from BL21 (DE3) expressing pllB and pllC (450 uL each) in 285 uL reaction buffer (100 mM HEPES pH 7.5, 1 mM dithiothreitol (DTT), 10 mM L-glutamate, 10 mM $MgCl_2$, 10 mM KCl, 5 mM ATP) in a final volume of 1.5

mL. The assay was incubated at 30 °C for 5h after which Tris-HCl pH 8.0 was added to 50 mM and Prot_686 to 2 uM followed by another 2h incubation at 30 °C. The reaction was quenched and extracted with butanol. The butanol phase was dried under N_2, and redissolved in 50 µL 50% methanol/H_2O for further LC-ESI-MS analysis.

Class III precursor peptides (100 µM) were incubated with corresponding modification enzymes (20 µM) in 200 µL reaction buffer (20 mM Tris, PH 8.0, 10 mM $MgCl_2$, 1 mM DTT, 5 mM ATP). After 4h incubation at 30°C, Prot_176 and/or Prot_819 protease(s) (10 µM each) and 2.5 mM $ZnSO_4$ were added into the reaction.

PttP1 and PttP2 Modeling. Homology based modeling was performed using the SWISS-MODEL platform[26].The amino acid sequences of PttP1 and PttP2 were used to search for crystal structure templates in the SWISS-MODEL repository. Templates were evaluated according to global mean quality estimate (GMQE) as calculated by SWISS-MODEL, sequence identity and similarity, as well as literature investigation to ensure a heterodimer structure was used for the modeling of putative heterodimer PttP1/PttP2. Based on these criteria, the M16B metallopeptidase heterodimer from *Sphingomonas* sp. A1 with PDB code 3amj[27] was selected as the template with a GMQE score of 0.6. The two sequences in this dimer contain the corresponding HXXEH and R/Y motifs found in PttP1 and PttP2 further supporting their use as templates. Models were then visualized and RMSD calculations were performed using PyMol.

References

1. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, (2011).
2. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
3. Yoav Benjamini & Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
4. Excoffier, L. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Nucleic Acids Res.* **13**, e149–e149 (2017).
5. Green, M. R. & Sambrook, J. BBMap: a fast, accurate, splice-aware aligner. *Cold Spring Harbor Laboratory Press* **13**, (2012).
6. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
7. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 **[q-bio.GN]** (2013).
8. Yang, L., Smyth Gordon K & Wei, S. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
9. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
10. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
11. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
12. Hoang, D. T., Chernomor, O., Haeseler, A. von, Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518-522 (2017).
13. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293-W296 (2021).
14. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
15. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

16. Green, M. R. & Sambrook, J. Molecular cloning: A laboratory manual. *Cold Spring Harbor Laboratory Press* (2012).
17. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
18. J., B. Site-directed mutagenesis. *Methods Enzym.* **529**, 241–248 (2013).
19. Altenbuchner, J. Editing of the *Bacillus subtilis* genome by the CRISPR-Cas9 system. *Appl. Environ. Microbiol.* **82**, 5421–5427 (2016).
20. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
21. Radeck, J. *et al.* The *Bacillus* BioBrick Box: Generation and evaluation of essential genetic building blocks for standardized work with *Bacillus subtilis*. *J. Biol. Eng.* **7**, 29 (2013).
22. Völler, G. H., Krawczyk, B., Ensle, P. & Süssmuth, R. D. Involvement and unusual substrate specificity of a prolyl oligopeptidase in class III lanthipeptide maturation. *J. Am. Chem. Soc.* **135**, 7426–7429 (2013).
23. Chen, S. *et al.* Zn-dependent bifunctional proteases are responsible for leader peptide processing of class III lanthipeptides. *Proc. Natl. Acad. Sci.* **116**, 2533–2538 (2019).
24. Ortega, M. A. *et al.* Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* **517**, 509–512 (2015).
25. Ortega, M. A. *et al.* Substrate specificity of the lanthipeptide peptidase ElxP and the oxidoreductase ElxO. *ACS Chem. Biol.* **9**, 1718–1725 (2014).
26. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
27. Maruyama, Y., Chuma, A., Mikami, B., Hashimoto, W. & Murata, K. Heterosubunit composition and crystal structures of a novel bacterial M16B metallopeptidase. *J. Mol. Biol.* **407**, 180–192 (2011).

Acknowledgements

This work is partially funded by a National Institutes of Health (NIH) grant P20GM103641, a National Science Foundation EPSCoR Program OIA-1655740, a Hong Kong Research Grants Council ECS grant HKU27107320, and the Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (SMSEGL20SC02). M.D. Walla acknowledges the support of NSF-MRI program (Award No. 1828059) for the acquisition of MS data with the Thermo Q-Exactive mass spectrometer. We thank L. Hou and H. Liu for help with experiments, and T. de Rond, P.J. Pellechia, T. Johnson, and W.E. Cotham for technical advice and access to equipment.

Author contributions

D.X., Z.Z., Y.-X.L., and J.L. designed research; D.X., E.A.O., and Z.Z. performed research; Z.S. characterized the structures of compounds; M.D.W performed the LC-MS analysis. All authors analyzed data and discussed the results. All authors participated in preparing the manuscript.

Competing interests

The authors declare no competing interests.