New Phytologist

Research

Substrate-binding glycine residues are major determinants for hydrolase and ligase activity of plant legumains

Xinya Hemu[1]*, Ning-Yu Chan[1]*, Heng Tai Liew[1], Side Hu[2], Xiaohong Zhang[1], Aida Serra[1,3], Julien Lescar[2], Chuan-Fa Liu[1] and James P. Tam[1,2]

[1]School of Biological Sciences, Synzymes and Natural Products Center (SYNC), Nanyang Technological University, 60 Nanyang Drive, Singapore City 637551, Singapore; [2]NTU Institute of Structural Biology, Nanyang Technological University, 59 Nanyang Drive, Singapore City 637921, Singapore; [3]Neuroscience Area, +Pec Proteomics Research Group (+PPRG), Faculty of Medicine, Biomedical Research Institute of Lleida Dr. Pifarré Foundation (IRB Lleida), University of Lleida, Av. Rovira Roure, 80, Lleida 25198, Spain

Author for correspondence:
James P. Tam
Email: jptam@ntu.edu.sg

Received: *18 August 2022*
Accepted: *17 February 2023*

New Phytologist (2023)
doi: 10.1111/nph.18841

Key words: asparaginyl endopeptidase, butelase, ligase-activity determinant, peptide asparaginyl ligase, plant legumain.

Summary

- Peptide asparaginyl ligases (PALs) are useful tools for precision modifications of proteins and live-cell surfaces by ligating peptides after Asn/Asp (Asx). They share high sequence and structural similarity to plant legumains that are generally known as asparaginyl endopeptidases (AEPs), thus making it challenging to identify PALs from AEPs. In this study, we investigate 875 plant species from algae to seed plants with available sequence data in public databases to identify new PALs.
- We conducted evolutionary trace analysis on 1500 plant legumains, including eight known PALs, to identify key residues that could differentiate ligases and proteases, followed by recombinant expression and functional validation of 16 novel legumains.
- Previously, we showed that the substrate-binding sequences flanking the catalytic site can strongly influence the enzymatic direction of a legumain and which we named as ligase-activity determinants (LADs). Here, we show that two conserved substrate-binding Gly residues of LADs are critical, but negative determinants for ligase activity.
- Our results suggest that specific glycine residues are molecular determinants to identify PALs and AEPs as two different legumain subfamilies, accounting for *c.* 1% and 88%, respectively.

Introduction

Peptide ligases form peptide bonds, whereas proteases break them. Ligases are rare in nature, but can be found as bioprocessors of ribosomally synthesized and post-translationally modified peptides (RiPPs; Montalban-Lopez *et al.*, 2021) in bacteria (Mazmanian *et al.*, 1999), cyanobacteria (Lee *et al.*, 2009), fungi (Luo *et al.*, 2014), and plants (Barber *et al.*, 2013; Nguyen *et al.*, 2014). In particular, plants produce peptide asparaginyl ligases (PALs) for biosynthesis of N-to-C cyclic peptides, forming an Asn/Asp (Asx) peptide bond with N-terminal residues.

Together with asparaginyl endopeptidases (AEPs), PALs belong to Cys proteases of the C13 subfamily (MEROPS EC 3.4.22.34; Rawlings *et al.*, 2018). Asparaginyl endopeptidases are also known as vacuolar processing enzymes (VPEs) because they function in lytic vacuoles as degradative enzymes (Hara-Nishimura *et al.*, 1991), or more generally as legumains because they were discovered in legumes (Kembhavi *et al.*, 1993; Takeda *et al.*, 1994). The homologous proteases are also found in animals, designated as mammalian legumains (Chen *et al.*, 1997). In this work, we use 'legumain' to describe this family of enzymes because of simplicity. Legumains are expressed as proenzymes in which the core domain is protected by an inhibitory cap domain that can be removed by acid-induced auto-activation. In plants, legumains regulate diverse cellular processes including maturation of defense peptides and proteins, degradation of storage proteins during seed germination, and programmed cell death (Hara-Nishimura *et al.*, 1995; Hiraiwa *et al.*, 1997; Hatsugai *et al.*, 2004).

For over 40 yr, legumains have been reported to cut and join a protein precursor through a splicing mechanism in the maturation of lectins. However, this legumain-mediated splicing process requires a properly folded structure of the precursor protein (Carrington *et al.*, 1985, Min & Jones, 1994, Nonis *et al.*, 2021). Recent legumain-mediated bioprocessing examples include the biosynthesis of sunflower seed trypsin inhibitor SFTI-1 from storage proteins (Bernath-Levin *et al.*, 2015) and the cyclic trypsin inhibitor MCoTIs that involve a cleavage first at an Asn site and then a ligation at a downstream Asp to form a cyclic peptide (Du *et al.*, 2020; Liew *et al.*, 2021).

Although AEP was proposed to mediate the biosynthesis of cyclotide *in vivo* (Saska *et al.*, 2007; Craik & Malik, 2013), the Asx-specific ligase activity was not observed *in vitro* until the discovery of butelase-1 in 2014 (Nguyen *et al.*, 2014). Indeed, butelase-1 serves as the first PAL, displaying efficient *in vitro* ligase activity for both inter- and intramolecular ligation

*These authors contributed equally to this work.

reactions, under both acidic and neutral pH using an array of synthetic peptide and protein substrates with little to no detectable hydrolase activity. By contrast, most AEPs yield predominant hydrolytic products using a similar panel of substrates, even for the conformation-favored macrocyclization or intramolecular ligation reactions. Over the past 8 yr, different laboratories reported seven additional butelase-1-like ligases based on their *in vitro* ligation profiles using synthetic substrates (Harris *et al.*, 2015, 2019; Jackson *et al.*, 2018; Hemu *et al.*, 2019a).

Peptide asparaginyl ligases are versatile and precise tools for site-specific modification and can catalyze semi- and total syntheses of peptides and proteins. Peptide asparaginyl ligase-mediated modifications include protein cyclization (Nguyen *et al.*, 2015b, 2016; Hemu *et al.*, 2016), live-cell labeling (Bi *et al.*, 2017, 2020), conjugation (Nguyen *et al.*, 2015a; Bi *et al.*, 2018), and polymerization (Cao *et al.*, 2016; Hemu *et al.*, 2019b). Several chemical and enzymatic ligation methods for bio-orthogonal or chemoenzymatic ligation either in tandem or under one-pot conditions are also compatible with PALs (Cao *et al.*, 2015; Harmand *et al.*, 2018; Wang *et al.*, 2021).

Attempts to find a structural basis to distinguish PAL from AEPs have largely been unsuccessful. Comparison of 12 unique crystal structures of AEP and PAL core domains gave a RMSD of < 1 Å (Hemu *et al.*, 2020). Recently, we showed that certain substrate-binding sequences flanking the catalytic S1 cysteine pocket could serve as ligase-activity determinants (LADs) to distinguish a PAL from an AEP. We also showed that LAD motifs in PALs are generally hydrophobic and the ligation-promoting functions may attribute to their ability to exclude water from the catalytic center (Hemu *et al.*, 2019a, 2020). Here, we refine the substrate-binding residues central to the LAD hypothesis using bioinformatics study of 1500 plant legumains and then functional studies to validate the predicted PALs. Our results confirm that the LADs of AEP and PALs are distinctly different, suggesting that PALs constitute a distinct subfamily of legumains.

Materials and Methods

Retrieval of plant legumain sequences from existing protein and transcriptome databases

Core-domain sequences of butelase-1 (from G44 to N324; KF918345), butelase-2 (ALL55651.1), OaAEP1b (KR259377), AtVPE-α (NP_180165.1), AtVPE-β (NP_176458.1), and AtVPE-γ (NP_195020.1) were used as queries to mine transcriptomes deposited in NCBI and OneKP (Matasci *et al.*, 2014) databases using BLASTP (nr) and TBLASTX (nr and TSA) searches. The selection threshold was set to > 50% identity and > 95% sequence coverage as compared to the queries. Data from all searches were combined, and repeated homologs were removed to subsequently retrieve 1498 unique ORFs that encode plant legumains including 495 sequences from OneKP and 1003 sequences from NCBI. Two previously reported PALs, OaAEP4, and OaAEP5, were not found in the transcriptomic database and were added manually (Supporting Information Dataset S1). Except butelase, legumains mentioned in this study are named with a two-letter abbreviation in front denoting their host species: Bm, bitter melon (a common name of *Momordica charantia* L.); Br, *Brassica chinensis* L. (a synonym of *Brassica rapa var. chinensis*); Cr, *Catharanthus roseus* (L.) G. Don; Dc, *Dianthus caryophyllus* L.; Ha, *Helianthus annuus* L.; He, *Hybanthus enneaspermus* (L.) F. V. Muell. (a synonym of *Afrohybanthus enneaspermus* (L.) Flicker); Ls, *Lactuca sativa* L.; Oa, *Oldenlandia affinis* (Roem. & Schult.) *DC.* (a synonym of *Hedyotis affinis* Roem. & Schult.); Pe, *Petunia exserta* Stehmann; Pi, *Psychotria ipecacuanha* (a synonym of *Carapichea ipecacuanha (Brot.)* L. Andersson); Va, *Viola albida Palib.*; Vb, *Viola betonicifolia* J. E. Smith; Vo, *Viola orientalis (Maxim.)* W. Beck. V.; Vt, *Viola tricolor* L.; Vu, *Viola uliginosa* Besser; Vv, *Viola verecunda* A. Gray; Vy, *Viola yedoensis* Makino (a synonym of *Viola philippica Cav.*).

Bioinformatics analysis

Multiple sequence alignments were performed using CLUSTAL OMEGA with BLOSUM62 substitution matrix and default gap penalties (Sievers *et al.*, 2011). Aligned sequences were submitted to the Universal Evolutionary Trace (UET) server (http://lichtargelab.org/software/uet) for UET analysis to find functionally important residues. The real-value Evolutionary Trace (rvET) rank scores were mapped to the butelase-1 crystal structure (PDB code: 6DHI) and visualized using PYMOL. The gene phylogeny of 1500 legumains was analyzed by MEGA11 using a neighbor-joining method with 500 Bootstrap and visualized using iTOL. Sequence logos were extracted from JALVIEW v.2.10.

Cloning, expression, and purification of recombinant proteins

cDNA sequences of 16 selected enzymes (VyPAL4, MK085233.1, VyPAL5, MK085234.1; VaPAL1, GFWC01037197.1; VoPAL1, GFXR01024405.1; VuPAL1, GCAB01004088.1, VvPAL1, GFWF01025417.1; HaPAL1, XM_022113308.2; BmAEP1, XM_022292359.1; PiAEP1, OneKP:BQEQ_2002574; LsAEP1, JI582927.1; PeAEP1, GBRT01052954.1; PeAEP2, GBRT01019050.1; DcAEP1, OneKP:SHEZ_2097111; BrAEP1, GFUS01046649.1; CrAEP1, GACD01060809.1; VtAEP1, and OneKP:LPGY_2014619) were codon-optimized for *Escherichia coli* expression system. Sequences starting after the residue corresponding to butelase-1-L26 or after the signal peptide predicted using SIGNALP v.5.0 were cloned into the pET28a(+) vector at NdeI/XhoI restriction sites to generate a His6-fusion protein construct (Genscript, Nanjing, China). Point mutations were generated using a Q5 mutagenesis kit (New England Biolabs, Ipswich, MA, USA). Plasmids were amplified in DH5α *E. coli*, and protein was expressed from SHuffle T7 *E. coli* competent cells (S3026J; New England Biolabs). The transformed cells were grown to OD_{600} = 0.4 at 30°C in LB broth supplemented with 50 μg ml^{-1} kanamycin. The temperature was reduced to 16°C, and expression was induced by 0.1 mM isopropyl β-D-1-thiogalactopyranoside for 24–48 h. Cells were harvested by centrifugation at 6000 ***g*** for 15 min at 4°C, and the pellet was re-suspended in cold lysis buffer (20 mM Na HEPES, 100 mM

NaCl, 1 mM EDTA, 1 mM dithiothreitol (DTT), and 0.1% (v/v) Triton X-100, pH 7.0). Bacterial cells were lysed by sonication on ice, and cell debris was removed by centrifugation at 10 000 ***g*** for 30 min at 4°C. The filtered supernatant was loaded onto a column packed with cOmplete™ His-tag purification resin (Roche, Merck). His6-proteins were eluted with 4 × 2 bead volumes of elution buffer (20 mM Na HEPES, 1 mM EDTA, 1 mM DTT, 0.5 M imidazole, pH 7.5). Eluents were concentrated and buffer-exchanged with binding buffer (20 mM Na HEPES, 1 mM EDTA, 1 mM DTT, pH 7.5) using a centrifugal filter unit (Vivaspin-15; Satorius, Goettingen, Germany) to remove excessive imidazole. Additional purification was carried out by anion-exchange chromatography on a 1 ml HiTrap Q Sepharose column (GE Healthcare, Chicago, IL, USA). Yields of the recombinant-produced proenzymes ranged from 0.1 to 2 mg l^{-1} culture. Active butelase-2 used in this study was prepared as previously described using a Baculovirus system and protein expression in insect cells (Hemu *et al.*, 2020).

Acid-induced auto-activation and purification of active enzymes

Purified proenzymes in a concentration of 1–2 mg ml^{-1} were activated by lowering the pH to 4.0–4.5 and incubating at 37°C for 10 min to 2 h. For VaPAL-I342A and HaPAL1-V345A, activation was performed at pH 4.0–4.1 at 25°C overnight. Activation progress was analyzed by SDS-PAGE. Activated enzymes were purified on a HiLoad 16/600 Superdex 75 column with 20 mM sodium citrate buffer, 1 mM EDTA, 1 mM DTT, 100 mM NaCl and 5% (w/v) glycerol, pH 4.2. The concentration of purified enzymes was determined by measuring the absorbance at 280 nm on a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Shanghai, China). The eluents were neutralized to pH 5.0–6.0 and stored at 4°C or −80°C after the addition of 20% sucrose.

Determination of auto-activation sites

After separation by SDS-PAGE, the bands for activated enzymes were excised from the gels and subjected to in-gel tryptic digestion. Digested peptide fragments were extracted and sequenced by LC-MS/MS as described previously (Serra *et al.*, 2016).

Functional studies

Intramolecular cyclization of peptide substrates was performed at 37°C in reaction buffers having pH ranging from 4 to 8 (20 mM citrate or phosphate buffers with 1 mM EDTA and 5 mM β-mercaptoethanol). The peptide substrates GN12-GL (GLYRRGRLYRRN-GL) and SFTI1-GL (GRCTKSIPPICFPD-GL) were synthesized using Fmoc chemistry on automated synthesizer LibertyBlue (CEM; Matthews, NC, USA) and purified by preparative RP-HPLC. Activated recombinant legumains and substrates were mixed to the final concentrations of 40 nM and 20 μM, respectively. The reactions were monitored by MALDI-TOF mass spectrometry (5800; Applied Biosystems, Bedford, MA, USA) and quantitatively analyzed by RP-HPLC on a C18 analytical column (Aeris™ widepore; Phenomenex, Torrance, CA, USA) after being quenched with 1 : 1 v/v acetonitrile with 0.1% trifluoroacetic acid.

Results

LADs based on substrate-binding residues at P and P′ sites

Currently, eight PALs have been identified based on mutagenesis and *in vitro* functional studies on synthetic peptide or protein substrates. They include butelase-1 (Nguyen *et al.*, 2014), OaAEP1b (Harris *et al.*, 2015), HeAEP3 (Jackson *et al.*, 2018), OaAEP3/4/5 (Harris *et al.*, 2019), and VyPAL1/2 (Hemu *et al.*, 2019a). Sequence comparison and structural analysis of the reported PALs and AEPs suggested two LAD sites, LAD1 and LAD2, correspond to the substrate-binding residues (Fig. 1; Zauner *et al.*, 2018; Hemu *et al.*, 2020; Chen *et al.*, 2021).

LAD1 is the S2 substrate-binding pocket located at the non-prime, or amino-side, of the Asx–Xaa scissile peptide bond. It is a hydrophobic tripeptide forming the βIV strand that shapes the S2 substrate-binding pocket (Aaslanda *et al.*, 2002). Mutagenesis studies suggested that the middle residue of the LAD1 tripeptide, which in butelase-1 corresponds to Val237, and that is also known as the 'gate-keeper' (Yang *et al.*, 2017), exerts a stronger influence on enzymatic directionality than the first and the third residue. At this middle position, all eight known PALs possess a hydrophobic or bulky residue, including Val, Ile, Cys, and Pro (Hemu *et al.*, 2020). By contrast, known protease-legumains, or AEPs, such as AtVPEs, HaAEP1, and butelase-2, contain a Gly residue in the middle position of the LAD1 motif.

LAD2 covers the S1′ substrate-binding pocket located at the prime, or carboxyl side, of the scissile bond, is an aliphatic tripeptide forming the βI strand, which corresponds to butelase-1 Gly166–Gly167–Ala168. Because Gly166 in the S1′ pocket is conserved in both AEPs and PALs, we simplified LAD2 from a tripeptide motif to a dipeptide motif by excluding Gly166. Examples of PAL-like LAD2 motifs include Gly167–Ala168 (butelase-1), Ala–Pro (OaAEP3/4, HeAEP3, VyPAL1/2), and Ala–Ala (OaAEP1b, OaAEP5). By contrast, most known AEPs have Gly167–Pro168.

Data collection for 1500 plant legumains

To confirm the current LADs using a large data set, we collected plant legumain sequences from existing databases by BLAST search of the National Centre for Biotechnology Information (NCBI) and one-thousand-plants (OneKP) database using core-domain protein sequences (Gly44–Asn324, butelase-1 numbering) of known PALs and AEPs as queries (Matasci *et al.*, 2014). Sequences with < 50% identity were excluded because most belong to GPI-anchoring transamidases, another type of C13 Cys protease (Ohishi *et al.*, 2000). Together, we obtained 1500 nonredundant sequences including 1003 sequences from NCBI, 495 sequences from OneKP, and two additional PAL sequences, OaAEP4 and OaAEP5 (Harris *et al.*, 2019; Dataset S1).

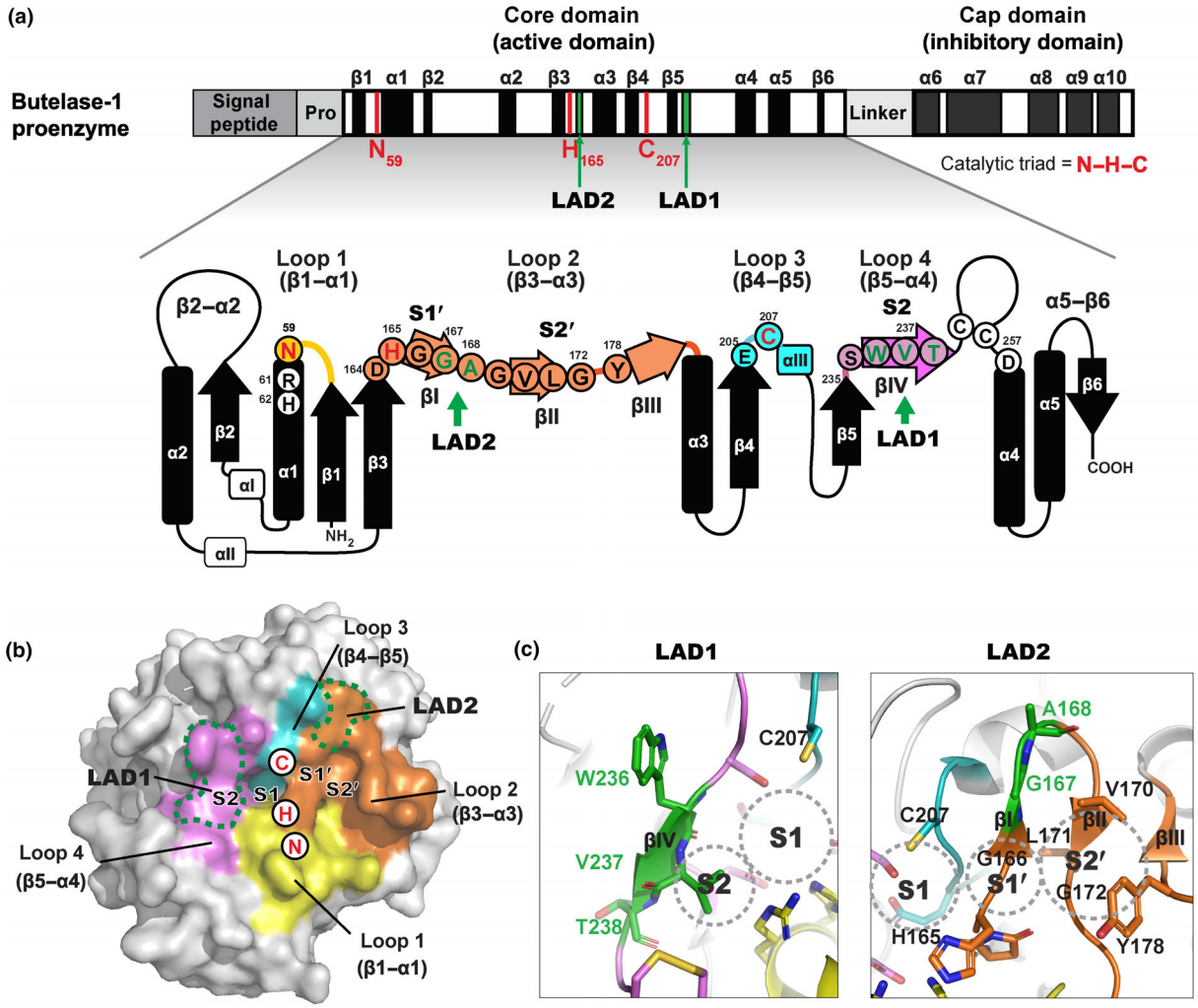


Fig. 1 1D, 2D, and 3D illustration of the ligase-activity determinants (LADs) LAD1 and LAD2 on butelase-1. (a) 1D schematic and 2D topology diagram of butelase-1. Location of LAD motifs in substrate-binding loop 2 and 4 on top of the core domain are labeled. Residues forming substrate-binding pockets S2–S1–S1′–S2′ are circled. Catalytic residues His165 and Cys207 are colored in red. LAD residues are colored in green. Other residues are in blank. (b) Four major substrate-binding loops mapped on the 3D structure of butelase-1 (PDB access code: 6DHI). Loop1-4 are colored in yellow, orange, cyan, and magenta, respectively. LAD motifs are marked with green dashed lines on the molecular surface near the S1 oxyanion hole. (c) Close-up view of LAD1 tripeptide Trp236–Val237–Thr238 that form the S2 pocket and LAD2 dipeptide Gly167–Ala168 next to S1′ and S2′ pockets. Location of the pockets are indicated by dashed grey circles.

The 1500 legumain sequences belong to 875 species from 249 plant families, ranging from primitive plants, such as green algae, moss, fern, and conifer, to flowering plants (Fig. S1). Crops of economic importance are highly represented in our dataset, including 13% (194/1500) from 75 species of the Poaceae family and 9% (132/1500) from 87 species of the Fabaceae family. There are 50 families containing between 5 and 80 sequences and the remaining 197 families contain <5 sequences. The six plant families that known for producing legumain-processed cyclic peptides, including Rubiaceae, Violaceae, Fabaceae, Solanaceae, Cucurbitaceae, and Asteraceae, together contain 370 sequences from 159 species.

Evolutionary trace analysis to reveal evolutionarily important residues for legumains

To identify molecular determinants for a PAL or an AEP using a different and global method to corroborate our proposed LADs, we performed UET (Lua *et al.*, 2016) analysis with the 1500 aligned sequences to determine the evolutionary importance of each residue. Universal Evolutionary Trace analysis gives each residue a rvET score (Mihalek *et al.*, 2004) that is calculated based on the level of diversity of a particular residue among distant evolutionary sequences (Dataset S1). An rvET score of 1.0 indicates a completely conserved residue that has a critical function. By contrast, a high rvET score of up to 200 indicates the presence of diverse residues among evolutionarily close analogs that have low functional importance. Fig. 2(a) shows the rvET score of each residue mapped onto the butelase-1 zymogen crystal structure (PDB ID: 6DHI; Yang *et al.*, 2017) using a color gradient ranging from red (most important) to white (least important) for visualization.

The substrate-binding sites S1 and S1′ comprise evolutionarily important residues that have low rvET scores <2 with surrounding residues having rvET scores of *c.* 30 (residues colored red and

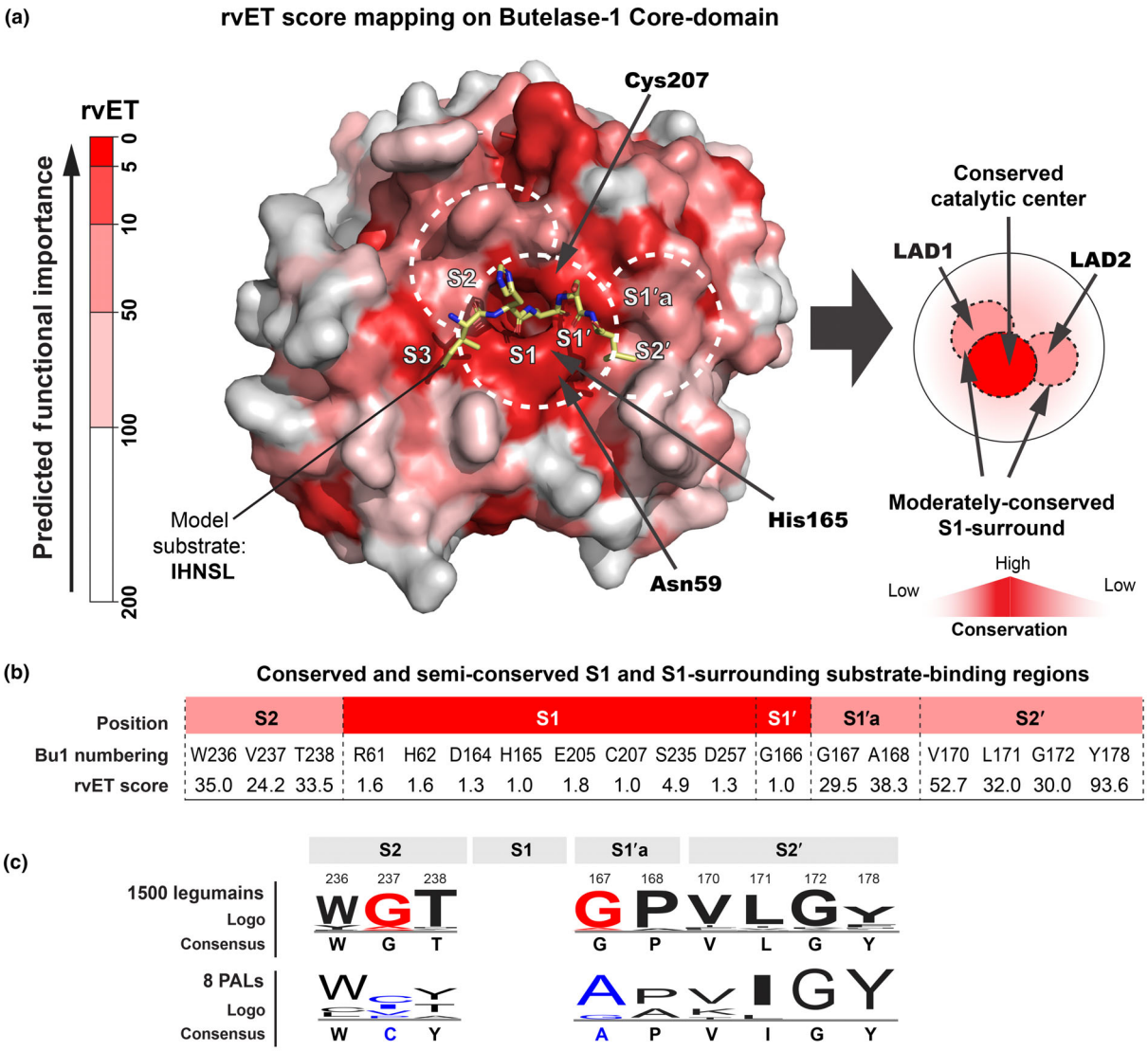


Fig. 2 Identification of evolutionarily important residues of plant legumains by Universal Evolutionary Trace analysis. (a) Ribbon structure of butelase-1 (PDB code: 6DHI) colored according to real-value Evolutionary Trace (rvET) scores. Red and white refer to the most and least conserved positions, respectively. A model peptide substrate IHNSL (yellow) was placed on top of the surface to show the location of substrate-binding pockets. An rvET-mapped illustration highlights the conserved central ring (red) and two semiconserved second rings covering S2 and S1′a–S2′ pockets (salmon). Ligase-activity determinant (LAD) motifs are located in the semiconserved rings. (b) rvET scores of the moderately conserved residues in substrate-binding pockets. Residues forming S1–S1′ pocket and catalytic dyad are highly conserved with rvET score < 5 (red). Away from the catalytic center, the conservation gradually decreased with an increasing of rvET score in the range of 24.2–93.6. (c) Comparison of sequence logos of moderately conserved S1-surrounding residues between 1500 legumains and eight confirmed peptide asparaginyl ligases (PALs). Highly conserved Gly237 and Gly167 (red) in plant legumains are occupied by non-Gly residues in PALs (blue).

pink, respectively, in Fig. 2b). This result is consistent with the fact that both AEPs and PALs share the same S1-pocket specificity toward Asx, but have substrate-binding pockets that vary in their enzymatic directionality and specificity (Dall & Brandstetter, 2013; Zauner *et al.*, 2018).

Further analysis of consensus sequences in the substrate-binding sites of the 1500 legumains shows that the moderately conserved residues are Gly167/Pro168 next to S1′ pocket, Leu171/Gly172 in S2′ pockets, and Trp236/Gly237/Thr238 in S2 pocket. Importantly, two moderately conserved Gly residues, the S1'-substrate-binding Gly167 and the S2-substrate-binding Gly237, are found to be non-Gly in the consensus sequences of PALs, supporting their importance as the key residues in LAD1 and LAD2 that influence ligase and protease activity of legumains (Fig. 2c). Overall, the global sequence analysis by UET agrees with the previously proposed substrate-binding amino acids in LAD motifs.

Occurrence and distribution of PAL-like LADs in legumains

We performed detailed analyses of amino acid occurrences at S2 pocket residue 237 of LAD1 and S1′ pocket residue 167 of

LAD2, which are occupied by Gly in 90% and 88% of legumains, respectively. Besides Gly, 8% of legumains possess small residues like Ala and Ser at S2 pocket position 237. The remaining 2% (29/1500) possess a PAL-like amino acid (Val, Ile, Pro, or Cys) S2 pocket position 237, which are referred as LAD1+ motifs (Fig. 3a). On the prime side, Gly167 followed by a Pro at S1′ pocket position 168 was the most common combination and was found in 84% of legumain sequences. Meanwhile, 11% of legumains (166/1500) contained PAL-like LAD2 dipeptides (Ala167–Xaa168 or Gly167–Ala168), including 92 sequences with an Ala167–Xaa168 (Xaa = Pro, Ala, Thr, Val, and Ser) dipeptide and 74 sequences with a butelase-1-like Gly167–Ala168 dipeptide, which are grouped as LAD2+ motifs. The remaining 5% of sequences had other residues at S1′ pocket position 167 including Ser, Tyr, Thr, Asp, Asn, and Val (Table S1).

Based on the amino acid composition at S2 pocket position 237 and S1′ pocket position 167–168, we grouped the 1500 legumains into three types: AEP-like (hydrolytic-), PAL-like (ligating-), and hybrid legumains (Fig. 3b). Asparaginyl endopeptidase-like legumains, that lack LAD+ motifs, are most common, and present in 88% of the 1500 legumains. By contrast, only 18 sequences correspond to PAL-like legumains, accounting for only 1.1% of the sequences. These include eight functionally characterized PALs, and 10 predicted PALs (Fig. 3c). S2 pocket residue 237 of LAD1 in these PAL-like sequences is occupied by bulky and hydrophobic amino acids, whereas the LAD2 motif (S1′ pocket residues 167–168) has either an Ala–Pro or Ala–Ala dipeptide sequence. The remaining 10% of legumains are hybrids, with 14 sequences bearing a LAD1+ motif and 148 sequences bearing a LAD2+ motif.

The distribution of 180 legumains with one or both LAD+ motifs is scattered in the phylogenetic tree (Fig. S1). Butelase-1 and Violaceae PALs display a closer evolutionary relationship, while HaPAL1 and OaAEPs locate in separate clades. The presence of LAD1+ mutations is more common in eudicots and can be found in angiosperms, while LAD2+ mutations are present in all phyla, ranging from green algae to dicots, indicating that the variations of LAD motifs evolved through random mutations.

Functional validation of LAD-based predictions

To validate the LAD-based classification, we selected 15 predictions for recombinant expression and activity tests. This group included seven putative PALs, three putative AEPs, and five

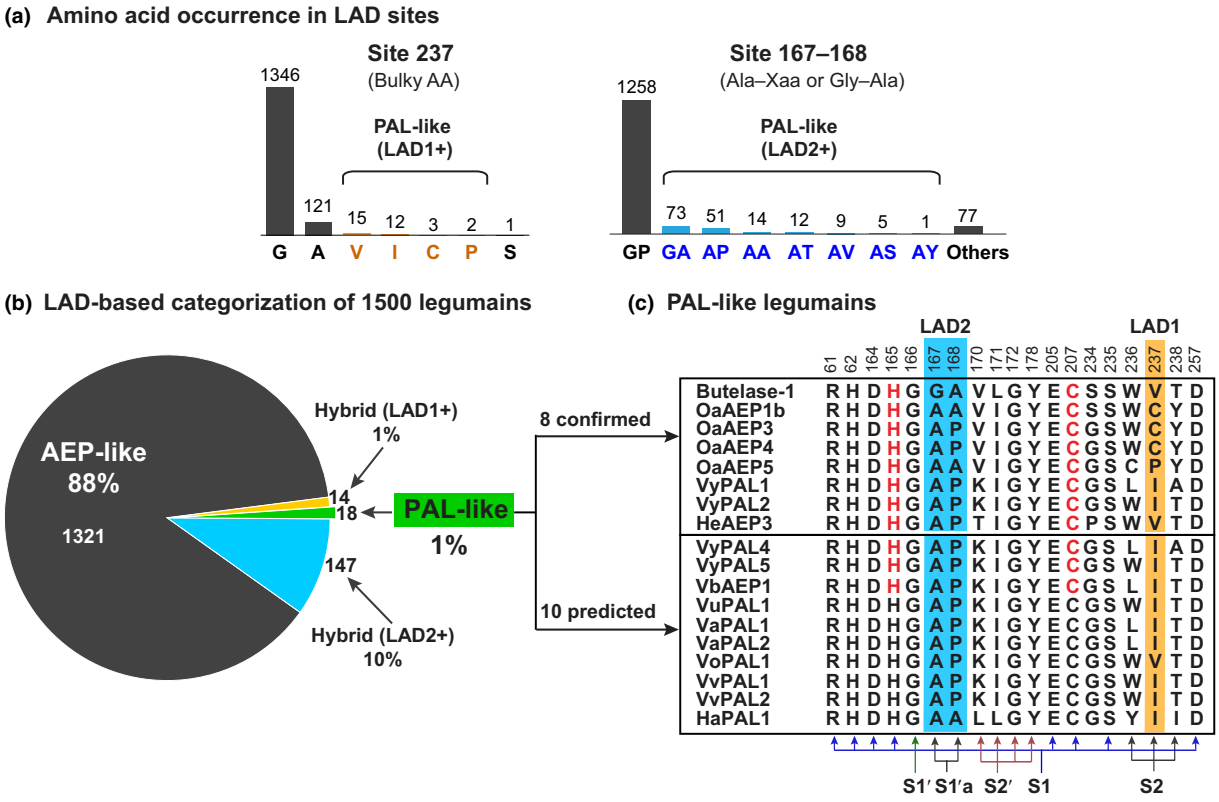


Fig. 3 Screening of new peptide asparaginyl ligases (PALs) based on amino acid composition of ligase-activity determinants (LADs). (a) Amino acid occurrences at position 237 (LAD1) and 167–168 (LAD2) of plant legumains. At position 237, Val, Ile, Cys, and Pro are predicted to be PAL-like LAD1 motif (LAD1+, highlighted in orange). At position 167–168, Gly–Ala, and Ala–Xaa are predicted to be PAL-like LAD2 motif (LAD2+, highlighted in blue). The rest are predicted to be asparaginyl endopeptidase (AEP)-like (LAD−). (b) Categorization of 1500 plant legumains into AEP-like, PAL-like, and two hybrid subtypes based on LAD compositions. Eighteen sequences are classified as PAL-like legumains with both PAL-like LAD1 and LAD2 motifs. They include all eight confirmed PALs and 10 predicted PALs as listed in the right panel with the most differentiating residues in LAD1 and LAD2 motifs highlighted in orange and blue, respectively. Catalytic residues are highlighted in red. Pocket-forming residues, and LAD sites are aligned and labeled with butelase-1 numbering. Abbreviations used for nomenclature of PAL-like legumains: Ha, *Helianthus annuus*; He, *Hybanthus enneaspermus*; Oa, *Oldenlandia affinis*; Va, *Viola albida*; Vb, *Viola betonicifolia*; Vo, *Viola orientalis*; Vu, *Viola uliginosa*; Vv, *Viola verecunda*; Vy, *Viola yedoensis*.

putative hybrid legumains. For comparison, we also included the previously characterized butelase-2 (Nguyen *et al.*, 2014; Serra *et al.*, 2016) as an AEP control.

Proenzymes of each legumain were expressed in *E. coli* using a protocol we previously described (Hemu *et al.*, 2021). Acid-induced auto-activation at pH 4.0–4.5 was successful for all enzymes except VaPAL1 and HaPAL1, which may require trans-activation by other vacuolar proteases in their *in vivo* activation process (Fig. S1). For these two legumains, we found that substitution of a conserved hydrophobic residue in the linker region, VaPAL1–Ile342 or HaPAL1–V345, with alanine could facilitate auto-activation (Fig. S1). This mutation site is located after the major autolytic cleavage sites as confirmed by MS/MS sequencing (Fig. S1) and thus will not remain in the activated enzyme.

All functional studies of recombinant legumains were performed with a synthetic peptide substrate GN12-GL (GLYRRGRLYRRN-GL, M.W. 1749 Da), which yields the hydrolyzed linear peptide GN12 (M.W. 1579 Da) or the N-to-C cyclized peptide cGN12 (M.W. 1561 Da) through Asn-specific hydrolysis or cyclization, respectively (Fig. 4a). The common recognition signal of Asn–Gly–Leu was chosen based on the known substrate specificity of characterized legumains (Nguyen *et al.*, 2014; Hemu *et al.*, 2019a; Dall *et al.*, 2020). To show that legumain-mediated catalytic reactions are pH-dependent, we performed functional studies at nine discrete pHs, from pH 4.0 to 8.0, with an interval of 0.5 pH unit. This pH range covers most of the reactive pHs of legumains reported in literature. A fixed enzyme-to-substrate ratio of 1 : 500 (mol : mol) was applied and reactions were allowed to proceed at 37°C for 5 or 10 min. Progress of the reactions was monitored by MALDI-TOF MS and yields were quantified by RP-HPLC (Fig. S1).

The seven selected PAL-like legumains (LAD1+ and LAD2+) are VaPAL1, VoPAL1, VuPAL1, VvPAL1, HaPAL1, VyPAL4, and VyPAL5. Two other predicted PALs, VaPAL2, and VvPAL2, were not selected because they are isoforms that share 96.1% and 98.9% sequence identity with VaPAL1 and VvPAL1, respectively. VbAEP1 was recently reported as the only PAL present in transcriptome analyses of its host species (Rajendran *et al.*, 2021). Functional studies showed that all seven predictions indeed acted as PALs, affording high ligation selectivity with a cyclization/hydrolysis (C : H) ratio > 20, a ratio suggesting the products contain < 4% GN12 from hydrolysis and > 96% cGN12 from ligation. This selectivity was maintained from pH 4.0 to 8.0 with no observable hydrolysis at pH ≥ 6.0 (Fig. 4b; Table S1). Similar to butelase-1, six predicted *Viola* PALs displayed pH optima between pH 6 and 6.5. HaPAL1 was the exception with a pH optimum of 7.0–7.5. At pH 6.0, HaPAL1 could accept both Asp and Asn when tested with a substrate based on the native linear precursor of the sunflower trypsin inhibitor SFTI-GL (GRCTKSIPPICFPD-GL, M.W. 1703 Da) with P1-Asp (Fig. S1) to give > 90% cyclic SFTI-1, suggesting that it could be a cyclase for SFTI-1 bioprocessing. Overall, LAD+ motifs work well in predicting PALs. Among the seven new PALs, VyPAL5, and VuPAL1 are the most efficient, affording > 90% cyclization yield within 10 min, whereas the rest gave 60–80% yield.

Among four predicted AEPs (LAD1− and LAD2−), butelase-2 and BrAEP1 contain the most common LAD combinations of S2-Gly237/S1′–Gly167–Pro168. This combination is found in 81% (1213/1500) of legumains in our data set. CrAEP1 and BmAEP1 have an Ala-substitution at position 237, which is the second most common amino acid at this position (8%). BmAEP1 has an additional Ser-substitution at position 167, which is the third most commonly found residue with 3.4% occurrence, after Gly (90%) and Ala (6%). For four AEPs, protease activity predominates with C : H ratios < 0.25 at their optimal pH, which ranged from pH 4.0 to 5.5 (Fig. 4c). Similar to butelase-2, all three putative AEPs behaved as predicted.

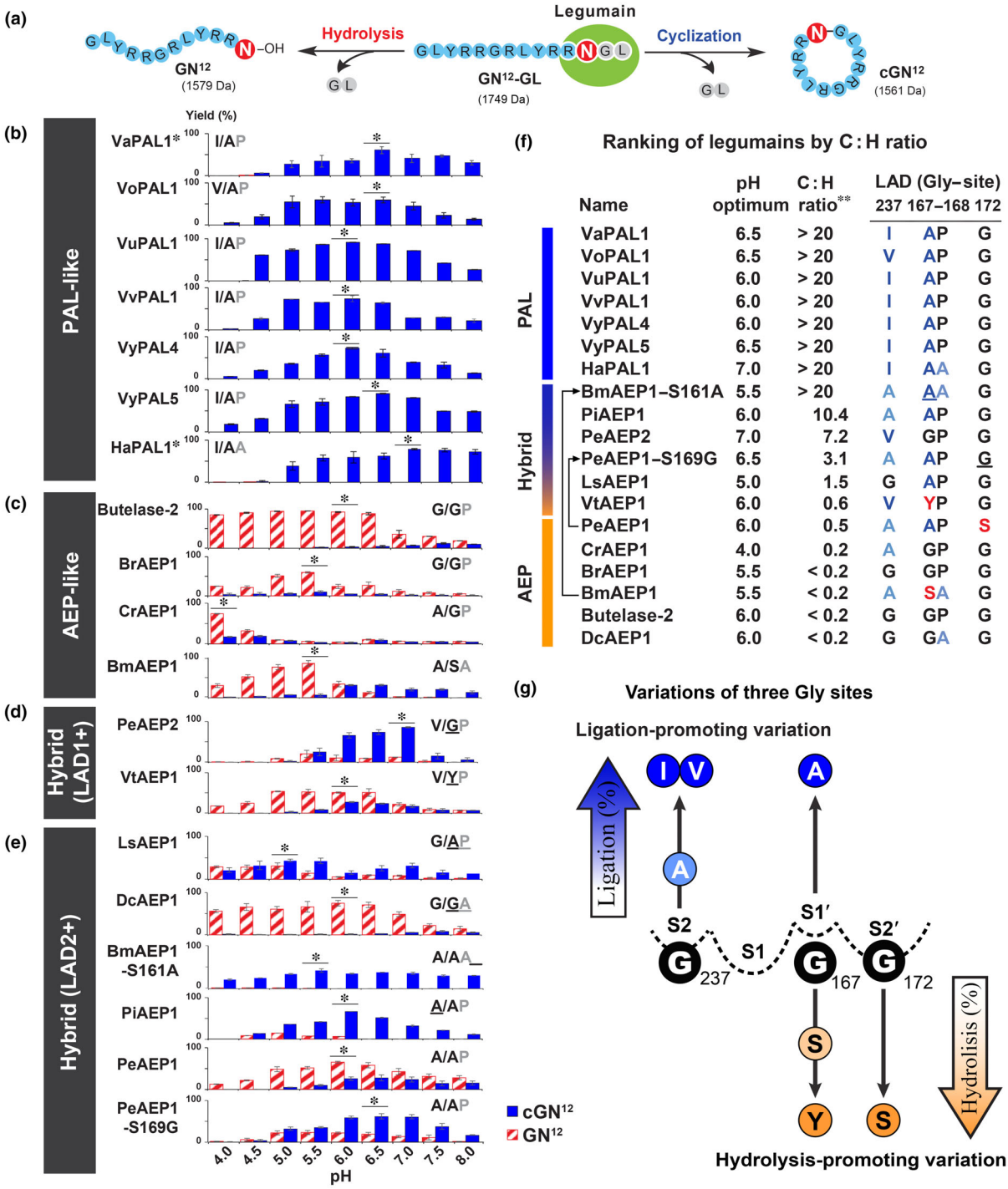
Since functional studies of hybrid legumains (LAD1+/− and LAD2+/−) could shed light on the relative importance of each LAD, we selected two legumains with LAD1+ and four with LAD2+. LAD1+ hybrids include PeAEP2 and VtAEP1. They have same LAD1-Val237 but different LAD2 motifs as Gly167–Pro168 and Tyr167–Pro168 in the S1′ pocket, respectively. PeAEP2 is bifunctional, acting predominantly as a protease at pH < 5.0, but as a ligase at pH > 5.5, with a C : H ratio 7.2 *c.* pH 7.0 (Fig. 4d). By contrast, VtAEP1, which is a predicted hybrid of LAD1+, acted predominantly like an AEP, with C : H ratios that remained < 1 throughout the tested pH range. These results suggest that LAD1+ is not the major determining factor, at least not under acidic conditions of pH < 5.5.

Of the four selected LAD2+ hybrids (LsAEP1, DcAEP1, PiAEP1, and PeAEP1), LsAEP1 and DcAEP1 share the same LAD1− motif of Trp236–Gly237–Thr238, but a different LAD2+ dipeptide motif of Ala167–Pro168 and Gly167–Ala168, respectively. We thus examined which LAD2+ motif confers more PAL-like activity. Functional studies showed that LsAEP1 and DcAEP1 displayed contrasting profiles. LsAEP1 (Ala167–Pro168) had a bifunctional profile with a C : H ratio 1.5 at its optimal pH 5.0, whereas DcAEP1 (Gly167–Ala168) had a butelase-2-like profile with C : H ratios < 0.25 from pH 4.0 to 8.0 (Fig. 4e).

To determine the relative importance of Ala167 or Ala168 of LAD2 motif in the S1′ pocket, we performed a LAD2+ Ala-substitution at BmAEP1–Ser161 (corresponding to Ala167 in butelase-1 numbering) to give BmAEP1–S161A (Fig. S1). This single mutation efficiently suppressed protease activity of wild-type (WT) BmAEP1 and conferred a PAL-like activity with C : H ratios > 20 from pH 4.0 to 8.0. These results suggested the importance of LAD2+ and that Ala at S1′ pocket position 167 imposed a more profound ligation-promoting effect than Ala at position 168, a result that agrees with the UET analysis.

Two other hybrids, PiAEP1 and PeAEP1, have identical LAD combinations of LAD1− and LAD2+ of Ala237/Ala167–Pro168. However, they displayed very different functional profiles: PiAEP1 was PAL-like (C : H ratio 10.4 at pH 6.0) and PeAEP1 was AEP-like (C : H ratio 0.49 at pH 6.0; Fig. 4f). To explain this 20-fold difference in ligase activity, we re-examined other moderately conserved substrate-binding residues, particularly those in the S2′ pocket. PiAEP1 and PeAEP1 differ at position 172, which is a conserved Gly in 93% of legumains including PiAEP1. At this position, PiAEP1 has Gly, but

New Phytologist



Name	pH optimum	C : H ratio**	LAD (Gly–site) 237	167–168	172
VaPAL1	6.5	> 20	I	AP	G
VoPAL1	6.5	> 20	V	AP	G
VuPAL1	6.0	> 20	I	AP	G
VvPAL1	6.0	> 20	I	AP	G
VyPAL4	6.0	> 20	I	AP	G
VyPAL5	6.5	> 20	I	AP	G
HaPAL1	7.0	> 20	I	AA	G
BmAEP1–S161A	5.5	> 20	A	AA	G
PiAEP1	6.0	10.4	A	AP	G
PeAEP2	7.0	7.2	V	GP	G
PeAEP1–S169G	6.5	3.1	A	AP	G
LsAEP1	5.0	1.5	G	AP	G
VtAEP1	6.0	0.6	V	YP	G
PeAEP1	6.0	0.5	A	AP	S
CrAEP1	4.0	0.2	A	GP	G
BrAEP1	5.5	< 0.2	G	GP	G
BmAEP1	5.5	< 0.2	A	SA	G
Butelase-2	6.0	< 0.2	G	GP	G
DcAEP1	6.0	< 0.2	G	GA	G

Fig. 4 Activity study of selected plant legumains. (a) Legumain-mediated reactions with the model substrate GN12-GL could yield both hydrolytic product GN12 and cyclized product cGN12 with 18 Da mass difference. Legumain is illustrated by a green oval. Residues forming the peptide substrate and the cyclic peptide product are circled in blue. The Asn at the enzyme-recognition site is circled in red. (b) The product distribution profiles of peptide asparaginyl ligase (PAL)-like legumains, (c) asparaginyl endopeptidase (AEP)-like legumains, (d) hybrid legumains with LAD1+ motifs and (e) hybrid legumains with LAD2+ motifs. Reactions were performed with a fixed enzyme-to-substrate molar ratio of 1 : 500 at 37°C under nine different pHs (pH 4.0–8.0) for 5 or 10 min. The product yields (%) were quantified by analytical RP-HPLC. Mean and SD values were calculated based on three to five independent reactions. Ligase-activity determinant (LAD) composition of each enzyme is marked in the top corner of the plot. Mutation sites are underlined. Optimal pH for each enzyme is marked with *. (f) Ranking of C : H ratios of 19 recombinant legumains showed over 100-fold difference between PALs and AEPs. **, C : H ratio > 20 was an estimated value indicating that no hydrolytic products were detected by HPLC, and so the hydrolysis yield was set as < 4% based on the detection limit of HPLC. Ligation-promoting variations are highlighted in blue. Hydrolysis-promoting variations are highlighted in red. Bm, bitter melon (*Momordica charantia*); Br, *Brassica chinensis*; Cr, *Catharanthus roseus*; Ha, *Helianthus annuus*; He, *Hybanthus enneaspermus*; Ls, *Lactuca sativa*; Oa, *Oldenlandia affinis*; Pe, *Petunia exserta*; Pi, *Psychotria ipecacuanha*; Va, *Viola albida*; Vo, *Viola orientalis*; Vt, *Viola tricolor*; Vu, *Viola uliginosa*; Vv, *Viola verecunda*; Vy, *Viola yedoensis*. Arrow, indication of the wild-type and mutant pair. (g) Directionality of legumains is primarily determined by amino acid variations at three conserved and substrate-binding Gly sites.

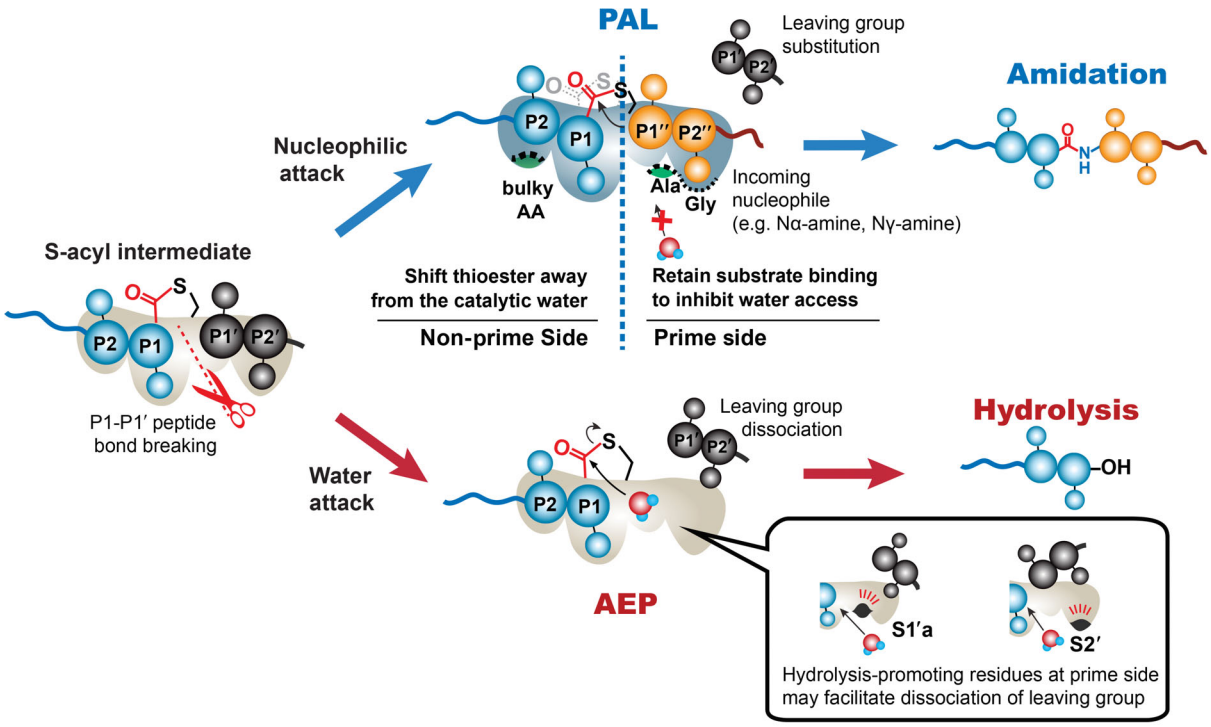


Fig. 5 Proposed roles of ligase-activity determinant (LAD) motifs based on functional and mutagenesis study. Upon formation of an S-acyl intermediate between the substrate (blue balls) and the enzyme (pockets in wheat), peptide asparaginyl ligase (PAL) mediates amidation (blue arrows) between substrate and the incoming peptide (orange balls), while and asparaginyl endopeptidase (AEP) mediates hydrolysis (red arrows). PAL-like LAD1 (bulky AA in S2) may stabilize the Asx-thioester intermediate and facilitates the nucleophilic attack by displacing the thioester intermediate away from catalytic water. PAL-like LAD2 (Ala in S1′ pocket) may favor peptide binding rather than water molecules. Similarly, conserved S2′-Gly172 also stabilizes enzyme-substrate interaction. Asparaginyl endopeptidase (AEP)-like LAD2/S2′ mutations favor hydrolysis by facilitating a rapid substrate dissociation.

PeAEP1 has Ser (corresponding to Ser169; Fig. S1). Substitution of Ser169 to Gly rescued part of the ligase activity to give a C : H ratio to 3.1 at pH 6.5, which represents a sixfold increase relative to WT (Fig. S1). These results suggested that variations of the conserved Gly in or near the S2′ pocket could shift legumains toward hydrolysis. The replacement of Gly at this position could disturb the stability of enzyme-substrate interaction on the prime side. A recent mutagenesis study by Dall *et al.* (2021) on human legumains also showed that V155G (homologous to butelase-1 Gly172) mutation could increase ligase activity.

Overall, our functional study using the C : H ratios on synthetic peptide substrates validated LADs of three legumain types. At their optimal pHs, PALs have a C : H ratio of > 20 and AEPs' C : H ratios are < 0.25. The hybrid legumain with C : H ratios < 1 to 7.2 showed that the LAD2 at the prime side could also serve as a 'gatekeeper' of ligase activity through Ala167 and residues in the S2′ substrate-binding pocket.

Discussion

Legumains display both Asx-specific hydrolase and ligase activity. This atypical behavior that enables a bidirectional enzymatic reaction is not found in other endopeptidases such as trypsin, chymotrypsin, and pepsin, all of which act entirely as a hydrolase. This report provides a structural basis to distinguish a hydrolase from a ligase in the legumain family, and an explanation for the atypical behavior of legumains, in part because of a very active Asx-thioester intermediate at the catalytic site during the catalytic transition state.

An accepted mechanism for legumain catalysis is the formation of an Asx-thioester intermediate (Fig. 5). However, this *S*-acyl intermediate is chemically unstable and prone to intramolecular cyclization with its side-chain amide (Asn) or carboxylic acid (Asp) to form rapidly, a five-member heterocyclic as a succinimide or an anhydride, respectively. The ability of an Asx side chain to participate a side reaction through a five-member intermediate is not found in other families of endopeptidases. In this regard, legumains have evolved a highly sophisticated substrate-binding pocket pattern, S2–S1–S1′–S2′, which not only stabilizes the thioester acyl intermediate to prevent intramolecular side-chain cyclization, but also facilitates intermolecular reactions to accept either an attacking water or amine nucleophile during a hydrolase or ligase reaction. This rationale forms a basis of our proposed molecular determinants in exploiting substrate-binding pockets for the legumain catalytic actions.

Four conserved glycine residues lie within or near the S2–S1–S1′–S2′ substrate-binding sites: S1′-Gly166 is invariant and S2′-Gly172 is also highly conserved in both AEPs and PALs. Our mutagenesis study suggested that S2′-Gly172 is a determinant for PALs to maintain a high C : H ratio. The remaining two Gly residues, S1′-Gly167 and S2-Gly237, are critical, but negative, determinants of ligase activity.

Previously, we suggested that the S2 pocket position 237 in the middle of LAD1 motif functions as a 'gate-keeper'. In this study, we show that the S2-Gly at position 237 occurs in 90% of legumains and is a negative ligase determinant. Substitution of S2-Gly237 with an aliphatic amino acid enhances ligase activity, and this increase correlates with the bulkiness of the substituted amino acid side chain. Thus, the ligase-promoting role of LAD1 at S2 pocket position 237 is likely associated with stabilizing the Asx-thioester intermediate.

Unlike S2-Gly237 of LAD1, the S1′-Gly167 of LAD2 next to the S1′ pocket is less well defined because it depends on residue 168 and, to a certain extent, residue 172 in S2′ pocket. Substitution of S1′-Gly167 with Ala facilitates ligase activity, but substitution into other residues, such Ser and Tyr, promotes protease activity. Our mutagenesis study also suggested that prime-side LAD residues have greater influence on the directionality than those on the nonprime side. Thus, we propose the S1′ position 167 could serve as a true 'gate-keeper' for the incoming nucleophiles (Fig. 5). Zauner *et al.*, (2018) proposed that retention of a leaving group or incoming nucleophile in prime-side pockets could exclude access by water and enhance ligation. Our results agree with this hypothesis by revealing the synergic effect of prime-side residues at S1′ position 167 and S2′ position 172. Our results also explain the low occurrence of PALs in nature because it requires co-existence of multiple ligase-determining residues at the substrate-binding pockets. It should be pointed out the difference between a PAL and an AEP often exceed a 100-fold variation in C : H ratio of a synthetic peptide substrate.

Peptide asparaginyl ligases are useful ligating enzymes. With the validated LAD hypothesis, PALs can be fished out for variety of applications that AEPs could not afford due to the restrictions in specific substrates and reaction conditions. Expanding the repertoire of PALs will broaden their utility in biochemical and biotechnological applications. For example, the optimal and operational pH for PALs such as butelase-1, VyPAL2, and OAaEP1b is near neutral. This work and recent progress on PALs have extended their operational pH from pH 5.5 (e.g. BmAEP1-S161A) to pH 9.0 (e.g. McPAL1), and substrate P1 site recognition from Asn/Asp to unnatural amino acids such as Asn(OH) and Asn(Me) in addition to P2′ recognition signals (Xia *et al.*, 2021). In turn, these advances allow various site-specific modifications as well as semi- and total synthesis of proteins in tandem, orthogonally, or under one-pot conditions.

In conclusion, plant legumains can catalyze hydrolysis and/or ligation of Asx–peptide bonds. In this study of 1500 plant legumains, we exploited, analyzed, and validated the usefulness of multiple sequence variants in the S2–S1–S1′–S2′ substrate-binding pockets as the molecular basis to distinguish hydrolase and ligase activity and for classification as PALs or AEPs. We observed that the enzymatic directionality of legumains could be 'edited' by engineering key residues in substrate-binding sites. Similar principles could be applied to the engineering of other 123 subfamilies of thioester-forming Cys proteases, such as structurally similar caspases from the C14 subfamily, to evoke their ligase potentials.

Acknowledgements

This research was supported by the Academic Research Grant Tier 3 (MOE2016-T3-1-003) from the Singapore Ministry of Education and Nanyang Technological University. We thank Dr Ka Ho Wong for his pioneer contribution to the bioinformatics analysis.

Competing interests

None declared.

Author contributions

JPT, XH and N-YC designed the research. N-YC and XH performed bioinformatic analysis, XH, N-YC, HTL and SH expressed and performed activity study of recombinant enzymes, XZ provided synthetic substrates, AS performed proteomic study. JPT, XH and N-YC wrote the manuscript. C-FL and JL provided revision suggestions to the manuscript writing. XH and N-YC contributed equally to this work. All authors read and approved the final version of the manuscript.

ORCID

Ning-Yu Chan https://orcid.org/0000-0001-9688-4397
Xinya Hemu https://orcid.org/0000-0003-1979-5854
Side Hu https://orcid.org/0000-0001-9320-5360
Julien Lescar https://orcid.org/0000-0002-9623-8130
Heng Tai Liew https://orcid.org/0000-0002-7711-9160
Chuan-Fa Liu https://orcid.org/0000-0001-7433-2081
Aida Serra https://orcid.org/0000-0002-4017-1096
James P. Tam https://orcid.org/0000-0003-4433-198X
Xiaohong Zhang https://orcid.org/0000-0002-0213-9140

Data availability

Data are available in article Supporting Information.

References

Aaslanda R, Abramsb C, Ampec C, Balld LJ, Bedforde MT, Cesarenif G, Gimonag M, Hurleyh JH, Jarchaui T, Lehtoj V-P *et al.* 2002. Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Letters* **513**: 141–144.

Barber CJS, Pujara PT, Reed DW, Chiwocha S, Zhang H, Covello PS. 2013. The two-step biosynthesis of cyclic peptides from linear precursors in a member of the plant family Caryophyllaceae involves cyclization by a serine protease-like enzyme. *Journal of Biological Chemistry* **288**: 12500–12510.

Bernath-Levin K, Nelson C, Elliott AG, Jayasena AS, Millar AH, Craik DJ, Mylne JS. 2015. Peptide macrocyclization by a bifunctional endoprotease. *Chemistry and Biology* **22**: 571–582.

Bi X, Yin J, Hemu X, Rao C, Tam JP, Liu CF. 2018. Immobilization and intracellular delivery of circular proteins by modifying a genetically incorporated unnatural amino acid. *Bioconjugate Chemistry* **29**: 2170–2175.

Bi X, Yin J, Nguyen GKT, Rao C, Halim NBA, Hemu X, Tam JP, Liu CF. 2017. Enzymatic engineering of live bacterial cell surfaces using butelase 1. *Angewandte Chemie (International Edition in English)* **56**: 7822–7825.

Bi X, Yin J, Zhang D, Zhang X, Balamkundu S, Lescar J, Dedon PC, Tam JP, Liu CF. 2020. Tagging transferrin receptor with a disulfide FRET probe to gauge the redox state in endosomal compartments. *Analytical Chemistry* **92**: 12460–12466.

Cao Y, Nguyen GK, Chuah S, Tam JP, Liu CF. 2016. Butelase-mediated ligation as an efficient bioconjugation method for the synthesis of peptide dendrimers. *Bioconjugate Chemistry* **27**: 2592–2596.

Cao Y, Nguyen GK, Tam JP, Liu CF. 2015. Butelase-mediated synthesis of protein thioesters and its application for tandem chemoenzymatic ligation. *Chemical Communications* **51**: 17289–17292.

Carrington DM, Auffret A, Hanke DE. 1985. Polypeptide ligation occurs during post-translational modification of concanavalin A. *Nature Australia* **313**: 64–67.

Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ. 1997. Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase. *The Journal of Biological Chemistry* **272**: 8090–8098.

Chen Y, Zhang D, Zhang X, Wang Z, Liu CF, Tam JP. 2021. Site-specific protein modifications by an engineered asparaginyl endopeptidase from *Viola canadensis*. *Frontiers in Chemistry* **9**: 768854.

Craik DJ, Malik U. 2013. Cyclotide biosynthesis. *Current Opinion in Chemical Biology* **17**: 546–554.

Dall E, Brandstetter H. 2013. Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways and regulation. *Proceedings of the National Academy of Sciences, USA* **110**: 10940–10945.

Dall E, Stanojlovic V, Demir F, Briza P, Dahms SO, Huesgen PF, Cabrele C, Brandstetter H. 2021. The peptide ligase activity of human legumain depends on fold stabilization and balanced substrate affinities. *ACS Catalysis* **11**: 11885–11896.

Dall E, Zauner FB, Soh WT, Demir F, Dahms SO, Cabrele C, Huesgen PF, Brandstetter H. 2020. Structural and functional studies of *Arabidopsis thaliana* legumain beta reveal isoform specific mechanisms of activation and substrate recognition. *Journal of Biological Chemistry* **295**: 13047–13064.

Du J, Yap K, Chan LY, Rehm FBH, Looi FY, Poth AG, Gilding EK, Kaas Q, Durek T, Craik DJ. 2020. A bifunctional asparaginyl endopeptidase efficiently catalyzes both cleavage and cyclization of cyclic trypsin inhibitors. *Nature Communications* **11**: 1575.

Hara-Nishimura I, Inoue K, Nishimura M. 1991. A unique vacuolar processing enzyme responsible for conversion of several proprotein precursors into the mature forms. *FEBS Letters* **294**: 89–93.

Hara-Nishimura I, Shimada T, Hiraiwa N, Nishimura M. 1995. Vacuolar processing enzyme responsible for maturation of seed proteins. *Journal of Plant Physiology* **145**: 632–640.

Harmand TJ, Bousbaine D, Chan A, Zhang X, Liu DR, Tam JP, Ploegh HL. 2018. One-pot dual labeling of IgG 1 and preparation of C-to-C fusion proteins through a combination of sortase A and butelase 1. *Bioconjugate Chemistry* **29**: 3245–3249.

Harris KS, Durek T, Kaas Q, Poth AG, Gilding EK, Conlan BF, Saska I, Daly NL, Van Der Weerden NL, Craik DJ *et al.* 2015. Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. *Nature Communications* **6**: 10199.

Harris KS, Guarino RF, Dissanayake RS, Quimbar P, Mccorkelle OC, Poon S, Kaas Q, Durek T, Gilding EK, Jackson MA *et al.* 2019. A suite of kinetically superior AEP ligases can cyclise an intrinsically disordered protein. *Scientific Reports* **9**: 10820.

Hatsugai N, Kuroyanagi M, Yamada K, Meshi T, Tsuda S, Kondo M, Nishimura M, Hara-Nishimura I. 2004. A plant vacuolar protease, VPE, mediates virus-induced hypersensitive cell death. *Science* **305**: 855–858.

Hemu X, El Sahili A, Hu S, Wong K, Chen Y, Wong YH, Zhang X, Serra A, Goh BC, Darwis DA *et al.* 2019a. Structural determinants for peptide-bond formation by asparaginyl ligases. *Proceedings of the National Academy of Sciences, USA* **116**: 11737–11746.

Hemu X, El Sahili A, Hu S, Zhang X, Serra A, Goh BC, Darwis DA, Chen MW, Sze SK, Liu C-F *et al.* 2020. Turning an asparaginyl endopeptidase into a peptide ligase. *ACS Catalysis* **10**: 8825–8834.

Hemu X, Qiu Y, Nguyen GK, Tam JP. 2016. Total synthesis of circular bacteriocins by butelase 1. *Journal of the American Chemical Society* **138**: 6968–6971.

Hemu X, Zhang X, Nguyen GKT, To J, Serra A, Loo S, Sze SK, Liu C-F, Tam JP. 2021. Characterization and application of natural and recombinant butelase-1 to improve industrial enzymes by end-to-end circularization. *RSC Advances* **11**: 23105–23112.

Hemu X, Zhang X, Tam JP. 2019b. Ligase-controlled cyclo-oligomerization of peptides. *Organic Letters* **21**: 2029–2032.

Hiraiwa N, Kondo M, Nishimura M, Hara-Nishimura I. 1997. An aspartic endopeptidase is involved in the breakdown of propeptides of storage proteins in protein storage vacuoles of plants. *European Journal of Biochemistry* **246**: 133–141.

Jackson MA, Gilding EK, Shafee T, Harris KS, Kaas Q, Poon S, Yap K, Jia H, Guarino R, Chan LY *et al.* 2018. Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases. *Nature Communications* **9**: 2411.

Kembhavi AA, Buttle DJ, Knight CG, Barrett AJ. 1993. The two cysteine endopeptidases of legume seeds: purification and characterization by use of specific fluorometric assays. *Archives of Biochemistry and Biophysics* **303**: 208–213.

Lee J, Mcintosh J, Hathaway BJ, Schmidt EW. 2009. Using marine natural products to discover a protease that catalyzes peptide macrocyclization of diverse substrates. *Journal of the American Chemical Society* **131**: 2122–2124.

Liew HT, To J, Zhang X, Hemu X, Chan NY, Serra A, Sze SK, Liu CF, Tam JP. 2021. The legumain McPAL1 from *Momordica cochinchinensis* is a highly stable Asx-specific splicing enzyme. *The Journal of Biological Chemistry* **297**: 101325.

Lua RC, Wilson SJ, Konecki DM, Wilkins AD, Venner E, Morgan DH, Lichtarge O. 2016. UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures. *Nucleic Acids Research* **44**: D308–D312.

Luo H, Hong SY, Sgambelluri RM, Angelos E, Li X, Walton JD. 2014. Peptide macrocyclization catalyzed by a prolyl oligopeptidase involved in alpha-amanitin biosynthesis. *Chemistry and Biology* **21**: 1610–1617.

Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al.* 2014. Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**: 17.

Mazmanian SK, Liu G, Ton-That H, Schneewind O. 1999. *Staphylococcus aureus* sortase, an enzyme that anchors surface proteins to the cell wall. *Science* **285**: 760–763.

Mihalek I, Reš I, Lichtarge O. 2004. A family of evolution–entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology* **336**: 1265–1282.

Min W, Jones DH. 1994. *In vitro* splicing of concanavalin A is catalyzed by asparaginyl endopeptidase. *Nature Structural & Molecular Biology* **1**: 502–504.

Montalban-Lopez M, Scott TA, Ramesh S, Rahman IR, Van Heel AJ, Viel JH, Bandarian V, Dittmann E, Genilloud O, Goto Y *et al.* 2021. New developments in RiPP discovery, enzymology and engineering. *Natural Product Reports* **38**: 130–239.

Nguyen GKT, Cao Y, Wang W, Liu CF, Tam JP. 2015a. Site-specific N-terminal labeling of peptides and proteins using butelase 1 and thiodepsipeptide. *Angewandte Chemie (International Edition in English)* **54**: 15694–15698.

Nguyen GKT, Hemu X, Quek JP, Tam JP. 2016. Butelase-mediated macrocyclization of D-amino-acid-containing peptides. *Angewandte Chemie International Edition* **55**: 12802–12806.

Nguyen GKT, Kam A, Loo S, Jansson AE, Pan LX, Tam JP. 2015b. Butelase 1: a versatile ligase for peptide and protein macrocyclization. *Journal of the American Chemical Society* **137**: 15398–15401.

Nguyen GKT, Wang S, Qiu Y, Hemu X, Lian Y, Tam JP. 2014. Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nature Chemical Biology* **10**: 732–738.

Nonis SG, Haywood J, Schmidberger JW, Mackie ERR, Costa TPSD, Bond CS, Mylne JS. 2021. Structural and biochemical analyses of concanavalin A circular permutation by jack bean asparaginyl endopeptidase. *Plant Cell* **33**: 2794–2811.

Ohishi K, Inoue N, Maeda Y, Takeda J, Riezman H, Kinoshita T. 2000. Gaa1p and Gpi8p are components of a glycosylphosphatidylinositol (GPI) transamidase that mediates attachment of GPI to proteins. *Molecular Biology of the Cell* **11**: 1523–1533.

Rajendran S, Slazak B, Mohotti S, Stromstedt AA, Goransson U, Hettiarachchi CM, Gunasekera S. 2021. Tropical vibes from Sri Lanka – cyclotides from *Viola betonicifolia* by transcriptome and mass spectrometry analysis. *Phytochemistry* **187**: 112749.

Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018. The *MEROPS* database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research* **46**: D624–D632.

Saska I, Gillon AD, Hatsugai N, Dietzgen RG, Hara-Nishimura I, Anderson MA, Craik DJ. 2007. An asparaginyl endopeptidase mediates *in vivo* protein backbone cyclization. *Journal of Biological Chemistry* **282**: 29721–29728.

Serra A, Hemu X, Nguyen GK, Nguyen NT, Sze SK, Tam JP. 2016. A high-throughput peptidomic strategy to decipher the molecular diversity of cyclic cysteine-rich peptides. *Scientific Reports* **6**: 23005.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Soding J *et al.* 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**: 539.

Takeda O, Miura Y, Mitta M, Matsushita H, Kato I, Abe Y, Yokosawa H, Ishii S. 1994. Isolation and analysis of cDNA encoding a precursor of *Canavalia ensiformis* asparaginyl endopeptidase (legumain). *Journal of Biochemistry* **116**: 541–546.

Wang Z, Zhang D, Hemu X, Hu S, To J, Zhang X, Lescar J, Tam JP, Liu C-F. 2021. Engineering protein theranostics using bio-orthogonal asparaginyl peptide ligases. *Theranostics* **11**: 5863–5875.

Xia Y, To J, Chan NY, Hu S, Liew HT, Balamkundu S, Zhang X, Lescar J, Bhattacharjya S, Tam JP *et al.* 2021. N^{γ}-hydroxyasparagine: a multifunctional unnatural amino acid that is a good P1 substrate of asparaginyl peptide ligases. *Angewandte Chemie (International Edition in English)* **60**: 22207–22211.

Yang R, Wong YH, Nguyen GKT, Tam JP, Lescar J, Wu B. 2017. Engineering a catalytically efficient recombinant protein ligase. *Journal of the American Chemical Society* **139**: 5351–5358.

Zauner FB, Elsässer B, Dall E, Cabrele C, Brandstetter H. 2018. Structural analyses of *Arabidopsis thaliana* legumain γ reveal differential recognition and processing of proteolysis and ligation substrates. *The Journal of Biological Chemistry* **293**: 8934–8946.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Accession number, species, ligase-activity determinant residues and amino acid sequences of 1500 legumains.

Dataset S2 Real-value evolutionary trace scores of 1500 legumains.

Fig. S1 Family distribution of 1500 plant legumains in 249 families.

Fig. S2 Phylogeny of 1500 plant legumains and the distribution of peptide asparaginyl ligases.

Fig. S3 SDS-PAGE data for the recombinant expression and activation of 16 legumains.

Fig. S4 Activation of VaPAL1 and HaPAL1 facilitated by Ala-substitution in the linker region.

Fig. S5 LC-MS/MS sequencing of VuPAL1 to identify the C-terminal cleavage site.

Fig. S6 MS and HPLC data for the functional study of 16 recombinant legumains.

Fig. S7 MS data for HaPAL1-mediated SFTI-1 cyclization.

Fig. S8 Activation and activity study of BmAEP1-S161A.

Fig. S9 Sequence alignment of 18 legumains.

Fig. S10 SDS-PAGE, MS, and HPLC data of PeAEP1 and PeAEP1-S169G.

Table S1 Summary of amino acid composition of ligase-activity determinants.

Table S2 Summary of enzymatic activity of 17 recombinant legumains and two mutants.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.