Running title: Legumain for specific proteome digestion

ExteNDing proteome coverage with legumain as highly specific digestion protease

Wai Tuck Soh[†,‡], Fatih Demir[§,‡] Elfriede Dall[†], Andreas Perrar[§], Sven O. Dahms[†], Maithreyan Kuppusamy[§], Hans Brandstetter[†], Pitter F. Huesgen[§,‖,⊥,*]

[†]Department of Biosciences, University of Salzburg, 5020 Salzburg, Austria.

[§]Central Institute for Engineering, Electronics and Analytics, ZEA-3, Forschungszentrum Jülich, 52428 Jülich, Germany.

[‖]Cologne Excellence Cluster on Cellular Stress Responses in Aging Associated Diseases, Medical Faculty and University Hospital, University of Cologne, 50931 Cologne, Germany.

[⊥]Institute for Biochemistry, Faculty of Mathematics and Natural Sciences, University of Cologne, 50674 Cologne, Germany.

‡ These authors contributed equally.

Corresponding author:

*p.huesgen@fz-juelich.de

ABSTRACT

Bottom-up mass spectrometry-based proteomics utilizes proteolytic enzymes with well characterized specificities to generate peptides amenable for identification by high throughput tandem mass spectrometry. Trypsin, which cuts specifically after the basic residues lysine and arginine, is the predominant enzyme used for proteome digestion, although proteases with alternative specificity are required to detect sequences that are not accessible after tryptic digest. Here, we show that the human cysteine protease legumain exhibits strict substrate specificity for cleavage after asparagine and aspartic acid residues during in-solution digestions of proteomes extracted from E.coli, mouse embryonic fibroblast cell cultures and *Arabidopsis thaliana* leaves. Generating peptides highly complementary in sequence, yet similar in their biophysical properties, legumain enabled complementary proteome and protein sequence coverage as compared to trypsin or GluC. Importantly, legumain further enabled the identification and enrichment of protein N-termini not accessible in GluC- or trypsin-digested samples. Legumain cannot cleave after glycosylated Asn residues, which enabled robust identification and orthogonal validation of N-glycosylation sites based on alternating sequential sample treatment with legumain and PNGaseF and vice versa. Taken together, we demonstrate that legumain is a practical, efficient protease for extending the proteome and sequence coverage achieved with trypsin, with unique possibilities for the characterization of post-translational modification sites.

Current “bottom-up” mass spectrometry-based proteomics, also termed shotgun proteomics, can achieve near-complete proteome coverage and allows extensive mapping of post-translational modification sites.[1] The basis of this approach is the selective protease-mediated digestion of isolated proteomes into peptides, which are then typically separated by reverse-phase liquid chromatography under acidic conditions and analyzed by tandem mass spectrometry (MS/MS). Peptides are subsequently identified by computational matching of the acquired spectra to proteome databases or spectral libraries, and the proteins present in the sample are inferred based on the identified peptides.[2] The serine protease trypsin has become the dominant workhorse for the proteome digestions due to its high cleavage efficiency, high specificity for cleavage after Arg or Lys and affordable price even for high quality preparations.[3] Proteomes digested with trypsin therefore consist of predictable peptides with a C-terminal basic residue favorable for ionization and generation of a dominant y-ions series, which facilitates database searches and peptide identification. However, about half of the peptides generated by trypsin are less than six residues long, and therefore too small for identification and/or unambiguous assignment to specific protein sequences.[4] Thus, many protein segments, including critical post-translational modification sites, and even whole proteins remain invisible in proteome analyses relying on trypsin alone.[3] This is especially true for proteolytic processing, a site-specific post-translational protein modification that can irreversibly alter protein function, interaction and localization[5, 6] and thereby exert important signaling functions[7]. Processed proteoforms are unambiguously identified by their new protease-generated neo- N- or C-termini.[8, 9] Identification of neo- N- and C-terminal peptides, which constitute a minor fraction

among all peptides in a proteome digest, is facilitated by a variety of methods that have been developed to allow their selective enrichment.[5] However, many neo-N- or C-terminal peptides are too short for mass spectrometry-based identification when only a single protease is used.[9]

Alternative proteases with high sequence specificity are therefore of great interest and increasingly applied in bottom-up proteomics, including termini profiling approaches.[3, 10] Established proteases include AspN for cleavage before Asp and Glu, chymotrypsin for cleavage after Phe, Tyr, Leu, Trp and Met; GluC (also known as *Staphyloccoccus aureus* protease V8) for cleavage after Asp and Glu; Lys-C for cleavage after Lys; Lys-N for cleavage before Lys[3, 11]; LysargiNase for cleavage before Arg and Lys[12]; and the prolyl endopeptidase neprosin that selectively cleaves after Pro and Ala.[13] Also proteases with broader sequence specificity such as elastase and thermolysin[14], proteinase K[15], subtilisin[16] and thermolysin WaLP and MaLP[17] are occasionally applied, but less favored due to the increased sample complexity with overlapping peptides and the less efficient spectrum-to-sequence matching due to the lack of a defined cleavage specificity as a restraint.[18] Notably, digest with a single additional protease increases the number of protein identifications by an average of 7-8%[11] and enables discovery of critical PTMs including phosphorylation site[16, 19] and N-terminal processing sites[10, 20] that are missed in tryptic digests. Hence, there is a persistent strong demand for new, highly specific proteolytic enzymes with improved, complementary or unexplored sequence specificity.[3]

Human legumain, also known as asparaginyl endopeptidase (AEP), is a well characterized caspase-like human cysteine protease known to cleave model substrates selectively after Asn and Asp residues.[21] Recently, legumain cleavage specificity was further characterized by in-gel digestion of denatured complex proteomes that revealed pH-dependent differences in sequence specificity, with optimal pH for cleavage after Asn and Asp at pH 6 and 4.5, respectively.[22] Based on this data, it was further suggested that legumain may be a suitable choice as precision digestion enzyme in proteomics applications.[22] Encouraged by these reports, we reasoned that legumain might also be an attractive enzyme for standard in-solution digestion proteomics workflows. We show that parallel digestion of proteomes isolated from *Arabidopsis thaliana* leaves, mouse embryonic fibroblasts (MEF) or *Escherichia coli* cell cultures with legumain, trypsin and GluC results in the identification of distinct peptides that together increase protein sequence and proteome coverage. Legumain retained its remarkable specificity even under unfavorable conditions. N-terminome profiling demonstrated strong complementarity to trypsin and superior performance compared to Glu-C. Asn is also the site of N-linked glycosylation, a common protein post-translational modification important in protein stability, folding, and protein-protein interaction.[23] By sequential processing with PNGase F and legumain, and vice versa, we demonstrate that N-glycosylation prevents legumain cleavage and propose that this tandem treatment strategy can provide orthogonal validation of N-glycosylation sites. Taken together, our data demonstrate that legumain is an attractive and reliable protease for specific digestion of proteomes after Asn and Asp, with particular advantages for PTM site identification including processed N-termini and N-glycosylation sites.

EXPERIMENTAL SECTION

Expression, purification and activation of human legumain. Human legumain was produced using the *Leishmania tarentolae* expression system (LEXSY) following a previously published protocol.[24] Briefly, legumain was recombinantly expressed as a secreted protein by LEXSY suspension culture at 26 ºC. The supernatant containing prolegumain protein was harvested by centrifugation and subjected to Ni^{2+}-NTA affinity purification followed by desalting using PD-10 columns (GE Healthcare). Purified legumain was activated at 20 ºC in a buffer containing 100 mM citric acid, pH 4.0, 100 mM NaCl, and 2 mM DTT. Progress of auto-activation was monitored by SDS-PAGE. Activated legumain was further purified using a PD-10 column (GE Healthcare) followed by size exclution chromatography to have the active protein in a final buffer composed of 20 mM citric acid pH 4.0, 50 mM NaCl and 2 mM DTT. Legumain activity was evaluated using the legumain specific fluorescent substrate Z-Ala-Ala-Asn-AMC (AAN-AMC; Bachem) at a concentration of 50 µM in assay buffer composed of 50 mM citric acid, pH 5.5, 100 mM NaCl, and 2 mM DTT at 37 °C. Fluorescence was detected using an Infinite M200 Plate Reader (Tecan) at 460 nm after excitation at 380 nm.

A. thaliana proteome preparation. *A. thaliana* Columbia (Col-8) leaves were harvested from 10 week old plants grown on soil under short day conditions (9 h/15 h photoperiod, 22 °C/18 °C, 120 µmol photons m-2 s-1) and snap frozen in liquid nitrogen. Leaves were ground in liquid nitrogen and resuspended in 10 ml/g fresh weight of extraction buffer (6 M Gua-HCl, 0.1 M HEPES pH 7.4, 5 mM EDTA, 1 mM DTT, HALT protease inhibitor cocktail; ThermoFisher, Dreieich, Germany). The suspension was homogenized using a Polytron PT-2500 (Kinematica, Luzern, Switzerland), filtered through Miracloth (Merck, Darmstadt, Germany), debris and nuclei removed by centrifugation at 500 x g, 4ºC for 10 min. Proteins in the supernatant were purified by chloroform-methanol precipitation[25], resuspended in extraction buffer and reduced with 5 mM DTT at 56 ºC, 30 min followed by alkylation with 15 mM iodoacetamide for 30 min at 25 ºC. The reaction was quenched by addition of 15 mM DTT for 15 min. The proteome extract was purified again with chloroform-methanol precipitation, resuspended in 0.2 mL of 0.1 M NaOH and diluted with water and 1 M Hepes pH 7.4 to a final concentration of 4 mg/ml in 0.1 M HEPES pH 7.4. The protein concentration was quantified using the BCA assay (ThermoFisher, Dreieich, Germany). For digestion, aliquots of the concentrated *A. thaliana* proteome extracts were diluted at least 4x to reach the required digestion buffer conditions and the pH confirmed with pH strips (Merck, Darmstadt, Germany).

Mouse embryonic fibroblast proteome preparation. Mouse embryonic fibroblast (MEF) cells were cultured in DMEM GlutaMax™ high glucose (Gibco 61965-026) supplemented with 10% FBS and 1x Penicilin/Streptomycin (Gibco 15140-122) at 37°C, 5% CO_2. Once the cells reached a confluency of up to 90 % the media was removed, washed with warm PBS and trypsinized (Gibco 25300-054). The trypsinized cells were pelleted, washed twice with warm PBS to remove excess media, and lysed with 1% SDS 100 mM HEPES pH 7.5 containing 1:50 (v/v) protease inhibitor cocktail (Sigma P8340). The sample was heated to 95 °C for 5 min, cooled, sonicated for 2 min and heated again to 95°C for 5 min to shear DNA. Protein concentration was measured and 100 µg

of protein was used for each proteome digestion. Proteins were reduced with 10 mM DTT for 30 min at 37°C and before alkylation by addition of 50 mM chloroacetamide (CAA) for 30 min at RT in the dark. The reaction was quenched by incubation with 50 mM DTT for 20 min at RT before purification with SP3 beads[26] and elution in the required digestion buffer.

E. coli proteome preparation. *E. coli* Dh5α cells were grown 200 ml LB media until an optical density of OD_{600nm} of 0.7. Cells were harvested by centrifugation at 400 x g for 15 minutes at 4 °C, washed by adding ice-cold PBS, and resuspended in 1 ml lysis buffer (4 % (v/v) SDS, 50 mM HEPES pH 7.4, 5 mM EDTA, 1x HALT protease inhibitor cocktail (ThermoScientific)) per 0.1 g fresh weight. The cells were lysed by heating two timed to 95°C for 5 min, with 10 min cooling of ice. Proteins were purified by chloroform-methanol precipitation, resuspended in 6 M Gua-HCl, 100 mM HEPES (pH 7.4), 5 mM EDTA and the concentration estimated using the BCA assay (ThermoFisher, Dreieich, Germany). 100 µg proteome was reduced by addition of 10 mM DTT for 30 min at 37°C, alkylated by addition of 50 mM CAA for 30 min at RT in the dark, and the reaction quenched by incubation with 50 mM DTT for 20 min at RT. The proteome was purified by chloroform-methanol precipitation and resolubilized in the appropriate digestion buffer.

Proteome digestions. Proteome aliquots of 100 µg were individually digested by legumain, GluC or trypsin. The digestion with legumain was carried out in a reaction containing 0.1 M MES pH 6.0, 0.1 M NaCl, 2 mM DTT at a protease to proteome ratio of 1:50 (m:m), unless otherwise stated. For GluC (SERVA Electrophoresis, Heidelberg, Germany) digestion, the same amount of proteome was digested in PBS pH 7.4 with a protease to proteome ratio of 1:50, whereas a 1:100 ratio was used for trypsin (SERVA Electrophoresis, Heidelberg, Germany) digestion in 0.1 M HEPES pH 7.4 supplemented with 5% acetonitrile and 5 mM $CaCl_2$. The pH was confirmed using pH strips (Merck, Darmstadt, Germany) and the digestions were carried out at 37 ºC overnight. For pH shift assays with legumain, an aliquot of the MEF proteome where digested at pH 6.0 for 5 h at 37°C, then the pH was lowered by stepwise addition of 1 M HCl until pH 4.0 was reached. Additional 2 µg Legumain and 1 mM DTT were added and incubated for another 5 h at 37°C.

Mass spectrometry. All samples were desalted using self-packed C18 STop And Go Extraction tips as described.[27] Analysis was performed on a two-column nano-HPLC setup (Ultimate 3000 nano-RSLC system with Acclaim PepMap 100 C18, ID 75 µm, particle size 3 µm: trap column of 2 cm and analytical column of 50 cm length; ThermoFisher) with a binary gradient from 5-32.5% B for 80 min (A: H_2O + 0.1% FA, B: ACN + 0.1% FA) and a total runtime of 2 h per sample coupled to a high resolution Q-TOF mass spectrometer (Impact II, Bruker) as described.[28] Data was acquired with the Bruker HyStar Software (v3.2, Bruker Daltonics) in line-mode in a mass range from 200-1500 m/z at an acquisition rate of 4 Hz. The Top17 most intense ions were selected for fragmentation with dynamic exclusion of previously selected precursors for the next 30 sec unless an intensity increase of factor 3 compared to the previous precursor spectrum was observed. Intensity-dependent fragmentation spectra were acquired between 5 Hz for low intensity precursor

ions (> 500 cts) and 20 Hz for high intensity (> 25k cts) spectra. Fragment spectra were acquired with stepped parameters, each with 50% of the acquisition time dedicated for each precursor: 61 µs transfer time, 7 eV collision energy and a collision RF of 1500 Vpp followed by 100 µs transfer time, 9 eV collision energy and a collision RF of 1800 Vpp.

Mass spectrometry data analysis. Database searches were performed with MaxQuant [29] v.1.6.0.16 using standard Bruker Q-TOF settings that included peptide mass tolerances of 0.07 Da in first search and 0.006 Da in the main search. *A. thaliana, M. musculus* and *E. coli* protein database were downloaded from UniProt (*A. thaliana* release 2018_01, 41350 sequences) with appended common contaminants as embedded in MaxQuant. The "revert" option was enabled for decoy database generation. For shotgun proteome samples, specificity was set to "unspecific" for the characterization of the cleavage specificity, otherwise according to the enzyme used (cleavage at K/R|X for trypsin, D/E|X for GluC, or D/N|X for legumain). Oxidation (M), acetylation (protein N-term) were set as variable modifications and the "match between runs" option was disabled. Analysis of the label-free shotgun data was performed with Perseus[30] v.1.6.1.1: validation of protein identification required at least 2 unique peptides for each protein and label-free quantification (LFQ) in at least 2 replicates. Searches for the N-termini were performed as described above, except that the enzyme specificity was set as Arg-C/GluC (DE)/legumain semi-specific with free N-terminus and duplex dimethyl labeling with light $^{12}CH_2O$ formaldehyde or heavy $^{13}CD_2O$ formaldehyde (peptide N-term and K). Oxidation (M), acetyl (N-term), Gln->pyro-Glu, and Glu->pyro-Glu were set as dynamic modifications and the re-quantify option turned off; the unspecific search window was set to 8-40 amino acids. Data evaluation and positional annotation for N-termini analyses was performed using an in-house Perl script (MANTI.pl; available at http://MANTI.sourceforge.io) that combines information provided by MaxQuant and UniProt to annotate and classify identified N-terminal peptides. In short, MaxQuant peptide identifications are consolidated by removing non-valid identifications (peptides identified with N-terminal pyro-Glu peptides that do not contain Glu or Gln as N-terminal residue, peptides with dimethylation at N-terminal Pro), contaminant, reverse database peptides, and non-quantifiable acetylated peptides in multi-channel experiments (no K in peptide sequence to determine labeled channel). For N-terminal peptides mapping to multiple entries in the UniProt protein database, a "preferred" entry was determined in a binary decision tree. Protein entries where the identified peptide matched position 1 or 2 were preferred over alternative positions, and then manually reviewed UniProt protein entries were favoured over alternative models. If multiple entries persisted, the alphabetically first entry was used to retrieve positional annotation information. For visualization of protein sequence coverage, protein structures were modeled with the Phyre2 server.[31]

Enrichment of N-terminal peptides. Protein N-terminal peptides were enriched using the High-efficiency Undecanal based N Termini EnRichment (HUNTER) method essentially as described.[32] Briefly, equal amounts of *A. thaliana* proteome were dimethyl labeled with 20 mM heavy ($^{13}CD_2O$) or light (CH_2O) formaldehyde and 20 mM sodium cyanoborohydride at 37 ºC for 16 hours to block all primary amines. To ensure complete reaction, the same concentration of reagents was added

again and incubated for another 2 hours. Proteins were purified by chloroform-methanol precipitation to remove excess reagents, dissolved in 0.1 M HEPES pH 7.4 and protein concentration was estimated using the BCA assay according to manufacturer instructions (ThermoFisher, Dreieich, Germany). The samples (400 µg/sample) were digested with legumain, GluC, and trypsin at 37 ºC for 16 hours in the respective digestion buffers and protease to proteome ratios as described above. The protease-generated peptides were hydrophobically tagged with undecanal using an undecanal to proteome ratio of 50:1 and supplemented with 20 mM sodium cyanoborohydride in 40% ethanol at 50 ºC for 45 min. The reaction was extended by addition of 20 mM sodium cyanoborohydride for another 45 min. The reaction was then acidified with final 1% TFA and centrifuged at 21000 x g for 5 min to precipitate free undecanal. Supernatant was injected to a pre-activated HR-X (M) cartridge (Macherey-Nagel, Düren, Germany). The flow-through containing N-terminal peptides was collected. Remaining N-terminal peptides on the HR-X (M) cartridge were eluted with 40% ethanol containing 0.1% TFA, pooled with the first eluate and subsequently evaporated in the SpeedVac to a small volume suitable for C18 StageTips purification.

Identification of glycosylation sites. Apoplastic fluid proteome enrichment was carried out as described[33] with some modifications. The whole *A. thaliana* rosettes were infiltrated with cold sterile water in a SpeedVac for 3 min at pressure between 600-2500 Pa. The infiltrated rosettes were then centrifuged at 4 °C, 3000 x g for 10 min into a collection tube containing Halt protease inhibitor cocktail (ThermoFisher, Dreieich, Germany). Extracted apoplastic fluid proteins were purified by chloroform-methanol precipitation and resuspended in 50 mM HEPES pH 7.4. The protein concentration was quantified by using the BCA assay. The sample was then reduced with 5 mM DTT at 56 ºC for 30 min, alkylated with 15 mM iodoacetamide at 25 ºC for 30 min in the dark and the reaction quenched with 15 mM DTT at 25 ºC for 15 min. The protein extract was then separated into two aliquots. One aliquot of 100 µg apoplast proteome was treated with PNGase F (SERVA Electrophoresis, Heidelberg, Germany) for 2 hours at 37 ºC before legumain digestion with protease to extract ratio of 1:50 at 37 ºC, pH 6 (pH adjusted with final concentration of 0.1 M MES pH 6.0). In parallel, another 100 µg of protein extract was pre-digested with legumain and then treated with PNGase F using the same conditions. The samples were subsequently dimethyl labeled with 20 mM heavy ($^{13}CD_2O$) and light (CH_2O) formaldehyde and 20 mM sodium cyanoborohydride at 37 ºC for 2 hours. The reactions were quenched with 0.1 M Tris pH 7.4 at 37 °C for 1 hour, pooled in a 1:1 ratio and peptides were purified by C18 StageTips.

Data deposition. MS data have been deposited to the ProteomeXchange Consortium[34] (http://www.proteomexchange.org) via PRIDE [35] (https://www.ebi.ac.uk/pride/archive/) partner repository: PXD014696 for data relating to comparative proteome digestion with legumain, GluC and trypsin, PXD014699 for *A. thaliana* proteome digested by Legumain in the presence of various denaturants and PXD014698 for various pH, PXD014697 for HUNTER N-termini profiling of *A. thaliana* leaves, PXD014680 for N-glycosylation site mapping.

RESULTS

Legumain cleaves denatured proteomes exclusively after Asn and Asp. Previous data obtained by in-gel protein digestion-based specificity profiling[22] and by biochemical characterization with test peptides[21] suggested that legumain cleaves substrates C-terminally to Asn and Asp residues in a pH-dependent manner, with optimal activity and high selectivity for Asn-containing substrates near pH 6. To test whether this exquisite specificity holds true under in-solution proteome digest conditions, we digested three aliquots of a denatured *A. thaliana* proteome with legumain at pH 6.0 for 18 h. In parallel, we digested three aliquots of the same proteome with trypsin and GluC at pH 7.4. To determine protease cleavage site specificity, peptides were analyzed by nano-LC-MS/MS and the acquired spectra were matched to the UniProt *A. thaliana* proteome database using non-specific search settings, i.e. without defining an enzyme cleavage specificity. This unbiased search identified 4452, 4078, and 7985 peptide sequences in legumain, GluC and tryptic digests, respectively, from which we compiled 6300, 5673 and 12107 unique non-redundant cleavage sites based on the sequence surrounding both ends of the identified peptides. For legumain, 93.3% of the observed cleavage sites were Asn and Asp (51.0% after Asn, 42.3% after Asp). A small percentage of unspecific cleavage is expected because of endogenous background proteolysis. The percentage of specific cleavage in a whole proteome is comparable to 96.7% cleavages after Lys and Arg observed for trypsin (58.0% after Lys, 38.7% after Arg) and more stringent than the 85.4% cleavages after Glu and Asp (72.7% after Glu, 12.7% after Asp) observed for GluC. Visualization of the relative amino acid abundance surrounding the cleavage sites with IceLogos reflected the strict specificity at the P1 position preceding the hydrolyzed peptide bond in all three enzymes (Fig. 1). While GluC (Fig. 1b) and trypsin (Fig. 1c) do not allow cleavage before proline (P1' position), this is not the case for legumain (Fig. 1c). We further analyzed a single replicate of a mouse embryonic fibroblast proteome and identified 1893, 1722 and 4377 peptides using nonspecific database searches after digestion with legumain, trypsin and GluC. Similar specificity profiles were obtained based on the 3244, 2999 and 7965 non-redundant cleavage sites derived from the peptides in legumain (Fig. 1d), GluC (Fig. 1e) and trypsin (Fig. 1f) digests, again showing that legumain tolerates Pro at P1´(Fig. 1d). 94.5% of the cleavages observed in legumain digest matched the expected specificity (63.6% after Asn, 30.9% after Asp), 97.6 % in the tryptic digest (51.9% after Arg, 45.7% after Lys) and 85% in the Glu-C digest (76.6% after Glu, 8.4% after Asp). These observations were further confirmed by analyses of an *E.coli* proteome (Supporting Information, SI Fig S1), where 2681 peptides identified after legumain digestion yielded 4187 cleavage sites with 86.2% cleavage after Asn and Asp (53.1% after Asn, 33.1% after Asp), while 85.3% of the 8597 unique cleavages observed in 5374 peptides identified after tryptic digest matched the expected specificity (44.1% after Arg, 41.2 % after Lys).

Complementary protein sequence coverage by legumain digestion compared to GluC and trypsin. With the strict cleavage specificity of legumain under proteome digest conditions confirmed by the unbiased database search, we repeated spectra-to-sequence matching using standard enzyme-specific settings with up to three missed cleavages, using cleavage after Asn and Asp as specificity rule for legumain. As expected, the smaller search space significantly increased the number of peptide identifications in the *A. thaliana* dataset by 64%, 8% and 66% to 7284, 4394, 12806 unique peptide sequences for legumain, GluC, and trypsin, respectively (Fig. 2a). Specific searches of the MEF proteome dataset increased peptide identifications by 129%, 73% and 61% to 4296, 2983, 8489 unique peptides for legumain, GluC, and trypsin, respectively, compared to non-specific searches. In *E.coli*, peptide identifications improved by 33% and 7% to 3568 and 5767 unique peptides for legumain and trypsin.

While trypsin showed the expected superior performance legumain digests resulted in the identification of more peptides than GluC, for example 66% more in the *A. thaliana* dataset. Interestingly, the legumain and GluC datasets showed only a minimal overlap of 66 identical peptides delimited by cleavages after Asp on both sides, which may occur with both enzymes but are not favored by GluC under the applied reaction conditions (Fig. 2a). Analysis of the length (Fig. 2b), hydrophobicity (Fig. 2c), and isoelectric point (Fig. 2d) of the identified *A. thaliana* peptides revealed very similar properties for all three enzymes. In contrast, the biophysical properties of all theoretical peptides in *in silico* digested *A. thaliana* and *M. musculus* proteomes predicted a higher number of peptides with pI>9 in GluC- and legumain-digested proteomes compared to trypsin (SI Fig. S2a, b). However, comparison to our data (Fig. 2b-d) suggests that such peptides are rarely identified with the standard experimental setup with reverse-phase chromatography under acidic conditions and ionization and mass spectrometric analysis in positive ion mode. Despite these physical similarities, peptides identified after digestion with the three proteases covered distinct amino acids in the identified *A. thaliana* proteins (Fig. 2e).

In total, the parallel application of legumain, GluC, and trypsin in technical triplicates identified 1524, 1090 and 2380 protein groups in the *A. thaliana* proteome, respectively, combining to a total of 2785 protein groups, with legumain contributing 8.8% exclusive identifications (Fig. 2g, h, SI Table S1). As expected from the number of peptide identifications, a large majority of 2057 proteins (74.3%) had the highest sequence coverage in the tryptic digest, followed by 507 (18.3%) in legumain digests and 206 (7.4%) in GluC digests (SI Table S1). For example, sequence coverage of SUPEROXIDE DISMUTASE (At1g08830) (Fig. 2f, SI Fig. S3a) was a remarkable 93% in legumain digests compared to 43% and 49% in the GluC and trypsin datasets, and sequence coverage of the GERMIN-LIKE PROTEIN 1 (At1g72610) was 63% with legumain compared to only 23% and 8% with GluC and trypsin (SI Fig. S3b). Notably, for each of the three proteases >80% of the proteins were identified in all three replicates, indicating high degree of reproducibility in the digests (SI Fig. S4a). On the single replicate level, the combination of any tryptic digest with any legumain or GluC digest resulted in a slightly higher number of protein identifications than any two tryptic replicate combined (SI Fig. S4b). We further compared reproducibility by label

free proteome quantification (LFQ) with MaxQuant after filtering for protein groups quantified by 2 or more peptides (SI Table S2). This demonstrated excellent correlation of the LFQ values between the technical digestion replicates and also a high correlation between the LFQ values obtained from digests of the three different proteases (Fig. 2h).

In the MEF proteome, 1469, 1140 and 2242 protein groups were identified in legumain, GluC and tryptic digests, combining to 2587 protein groups in total with 7.7% exclusively identified in the legumain digests (SI Fig. S4c). A larger overlap was observed between the *E.coli* proteome digests, where 842 and 1180 protein groups were identified after legumain and tryptic digestion, respectively, but only 37 (3%) of these were exclusive for legumain (SI Fig. S4d).

Legumain cleaves after Asn more efficiently than Asp. The digestion efficiency of a protease can be reflected by the number of missed potential cleavage sites within the identified peptides. In the *A. thaliana* dataset, legumain generated on average 53% of the peptides without missed cleavage sites, 34% with one missed potential cleavage site and 13% with more than one missed cleavage site (Fig. 3a). GluC performed worse, with only 30% of the peptides with no missed cleavage, but almost 12% of the identified peptides containing three missed cleavage sites. Trypsin was the best performing enzyme, with only 18% of the peptides containing one or more missed cleavage sites (Fig. 3a). When we further considered the identity of the amino acid residue, we noted that legumain reliably cleaved after Asn residues with only 5% of the peptides containing Asn uncleaved, but missed one or more Asp in 40% of the peptides (Fig. 3b). Most missed cleavage sites in GluC-digested proteomes were at Asp and even trypsin showed a higher fidelity at Arg than at Lys (Fig. 3b). Remarkably, legumain cleaved after Asn residues as efficiently as trypsin at the favored Arg-containing cleavage sites. Similar trends were observed in digests of MEF and *E.coli* proteomes, where legumain digests consistently showed a high cleavage efficiency at Asn sites with more missed cleavages at Asp (SI Fig.S5).

Assessing legumain efficiency in different reaction conditions. Previous publications have shown that legumain is more active at lower pH[24] and that cleavage after Asn is favored at higher pH.[22] To test if this is also the case with the digest conditions applied here, we digested whole-leaf *A. thaliana* proteome at varying pH between 5.0 and 6.5 for shorter (2h) and longer (24h) incubation times (SI Fig. S6a). We observed the highest number of peptide identifications at pH 6, which may have been caused by the higher propensity for proteome precipitation at lower pH that we observed in concentrated samples. As expected, legumain showed an increasing preference for Asn with increasing pH, and this kinetic preference was also reflected in different digestion times. Short proteome digestions (2 h) and/or lower pH (pH 5.0) resulted in a higher proportion of Asn cleavages (SI Fig. S6b), whereas longer incubation (24 hours) and/or lower pH yielded more complete cleavage after Asp (SI Fig. S6b). Based on this observation, we tested whether acidification of the

MEF proteome digest after an initial incubation at pH 6 would result in more complete cleavage at Asp residues. Indeed, this two-step incubation maintained efficient cleavage at Asn residues while decreasing the number of peptides containing missed Asp cleavage sites (SI Fig. S5).

Denaturants are commonly used for proteome preparations, but are problematic during digestion. We tested the tolerance of legumain to urea and guanidinium hydrochloride, but observed dramatically decreased digestion efficiency (SI Fig. S7a), reflected in decreased peptide identifications (SI Fig. S7b) with increased frequency of missed cleavage (SI Fig. S7c). In contrast, legumain tolerated the organic solvent acetonitrile quite well with little decrease in efficiency up to 10% acetonitrile concentration (SI Fig. S7).

We also assessed the amount of legumain necessary to achieve optimal digest by varying the protease to proteome ratio. Digestion appeared equally efficient in several dilutions down to a legumain to proteome ratio of 1:100 as judged by number of identified peptides from equal starting material (SI Fig. S8a). Another important enzyme property for routine use is the shelf life time, where our recombinant legumain preparations withstood ten freeze/thaw cycles without loss of peptidase activity (SI Fig. S8b).

Legumain is highly complementary for protein N-termini profiling. The complementarity of different digestion enzymes is particularly helpful for identification of specific post translational modification sites such as phosphorylations[14, 16] and protein termini[10, 20], as these may reside in sequences that are not accessible by trypsin. To demonstrate the value of legumain for this purpose, we profiled N-termini in the *A. thaliana* leaf proteome with our recently established HUNTER protocol (Fig. 4a).[36] In three replicates per enzyme, two aliquots of *A. thaliana* leaf proteome were differentially dimethyl labeled to block all unmodified primary amines. Thus, all protein N-termini are modified, either by endogenous modifications such as acetylation or in vitro by dimethylation. Differentially labeled duplicates are unified and digested in parallel with legumain, GluC or trypsin. This digestion generates new N-terminal primary amines in all internal and C-terminal peptides, which are then undecanal-labeled while the blocked N-terminal peptides remain inert. Undecanal-tagging increases the hydrophobicity of the digest-generated peptides, which enables their selective retention on a C18 cartridge while the dimethyl-labeled (or otherwise modified) protein N-terminal peptides are highly enriched in the flow-through for selective analysis (Fig. 4a). With this negative selection, we identified a total of 4773 N-terminal peptides (SI Table S3), with 1167, 1209 and 2342 N-terminal peptides identified in legumain, GluC and tryptic digests, respectively. The differential labeling demonstrated equivalent accuracy in quantification for all three enzymes (SI Fig. S9). For comparison of the overlap of identified protein N-termini, we extracted the first seven residues of each N-terminal peptide (Fig 4b). Only a minority of 100 protein N termini were identified by all three proteases, and an additional 632 were identified by 2 proteases, with a majority of 2101 N-termini identified only in digest of a single enzyme (Fig. 2b). For example the acetylated, native N-terminus of the GLUCOSINOLATE TRANSPORTER-1 NPF2.10 was only

identified in legumain digests, whereas multiple Glu in the N-terminal peptide excluded identification in GluC digests while the tryptic digest would deliver a very long peptide with unfavorably high content in acidic amino acids (Fig. 4c). Similarly, legumain digest uniquely identified an endoproteolytic processing site in CLPR3 (Fig. 4d).

Legumain as a tool for N-glycosylation site mapping. N-glycosylation is an important and frequent modification of secreted proteins[23, 37] Removal of the glycan by PNGase F results in deamidation of the Asn to Asp and facilitates mass spectrometry-based identification of occupied N-glycosylation sites.[38].We speculated that N-glycosylation would prevent legumain from hydrolyzing adjacent peptide bonds, based on the crystal structure of human legumain that revealed that the zwitterionic character of its S1 subsite provides an ideal binding site for Asn, but no space to accommodate a glycosylated Asn residue.[39] In contrast, Asp residues resulting from deglycosylation by PNGase F treatment would be cleaved. Thus, sequential treatment with legumain and PNGase F should result in longer peptides containing a missed deamidated Asn (Fig. 5a, workflow 1), whereas PNGase F treatment before legumain digest should result in shorter peptides ending with a deamidated Asn (Fig. 5a, workflow 2). In proof of concept, we isolated *A. thaliana* apoplastic fluid proteome enriched in secreted N-glycosylated proteins, and sequentially treated two aliquots with legumain and PNGase F and vice versa in two parallel reactions (Fig. 5a). Treated peptides were differentially dimethyl labeled with heavy and light formaldehyde and combined before nano-LC-MS/MS analysis. Indeed, we found several peptides that fulfilled the expectations (Fig. 5b, SI Table S4). Peptides from 45 proteins contained a deamidated Asn as missed cleavage in workflow 1, whereas peptides from 49 proteins ended with deamidated Asn. For 6 proteins including myrosinase 1 (TGG1), an important glycoprotein involved in plant defense,[40] we observed peptides matching to the same N-glycosylation sites in both workflows, providing intrinsic orthogonal validation (Fig. 5c). Notably, this glycosylation site has also been reported previously[41].

DISCUSSION

It is well established that the use of complementary proteases with different specificity in bottom up proteomic workflows can improve proteome coverage and provide access to sequences that are missed in tryptic digests.[3, 11] This not only allows identification of “missing proteins” that have not been identified by mass spectrometry before,[42] one of the central goals of the Human Proteome Project,[43] but is also important for comprehensive mapping of post-translational modification sites including phosphorylations[14, 16] and global identification of protein termini[10, 20]. Here we characterize human legumain as a new digestion protease in the proteomic toolbox. Legumain exhibited strict sequence specificity for cleavage after Asn and Asp and high cleavage efficiency that makes it a highly suitable alternative proteolytic enzyme for proteomics.

We have established conditions for reliable in-solution proteome digestion with legumain and show that the alternative cleavage site at Asn yields an entirely different set of peptides compared to trypsin, with only minimal overlap in the number of identified peptides delimited by Asp on both sides in comparison to GluC digests. In agreement with the kinetic cleavage preferences determined with peptide substrates[22, 24], Vidmar et al. reported only minimal cleavage at Asp residues at pH 6 during in-gel digestion of a denatured proteome[22]. In contrast, we have observed a much higher cleavage efficiency at Asp residues at pH6 in our dataset, which likely arises from the different digest conditions (in-gel digestion for 2h with citrate buffer[22] compared to in solution digestion for 16 h in a MES buffer). We noted that the dataset of Vidmar et al. contains a higher proportion of missed cleavages at Asn and Asp residues than our dataset (SI Fig. S10), suggesting that the prolonged reaction under more favorable in-solution conditions enables legumain more complete cleavage at Asp residues even at pH6.0. Notably, a similar effect was observed for Ulilysin/LysargiNase, which has a strong preference for Arg when tested with peptide substrates[44] but results in near-complete digestion at Lys residues under proteome digest conditions.[12]

Digestion with legumain consistently identified more peptides than digestion with GluC, but trypsin was far superior. This has been reported for various other digestion proteases, particularly those that do not select for cleavage at basic residues.[4, 11] One explanation is that digestion with enzymes such as legumain and GluC generates peptides with internal basic residues. This can give rise to internal fragment ions during collision-induced dissociation (CID) and result in unassignable, complex spectra.[4] In contrast, fragmentation by electron transfer dissociation (ETD) is not affected by the position of the basic residues and has been reported to improve peptide identifications after digestion with proteases that generate long peptides or peptides with internal basic residues.[4]

Parallel digests with all three enzymes increased proteome and protein sequence coverage and were particularly beneficial for protein N-termini identification, where a single digest often generated N-terminal peptides that are too short, too long or otherwise unfavorable for identification.[9] By extension, similar benefits may be expected for other post-translational modifications. Furthermore, using a sequential incubation with legumain and PNGase F we have demonstrated that legumain cannot cleave after glycosylated Asn residues, in contrast to deamidated deglycosylated Asn after PNGase F treatment. On a larger scale, evidence for N-glycosylation can be obtained by PNGase F treatment in ^{18}O-water, which results in deamidation of Asn to partially ^{18}O-labeled Asp.[38] However, this partial labeling makes relative quantification across samples challenging, while omission of the ^{18}O-labeling decreases confidence of the site identification as deamidation can also occur spontaneously. Based on our proof-of-concept experiment with *A. thaliana* apoplast proteome, we propose tandem sequential PNGase F/legumain treatment as an alternative strategy for experimental validation of N-glycosylation sites.

There are many further potential applications for legumain in peptide-centric proteome workflows. We have previously used legumain to generate high-quality *E. coli* proteome-derived peptide libraries, which enabled detailed cleavage specificity profiling of the vitamin K-dependent coagulation protease sirtilin that would not be possible in trypsin-generated libraries.[45] Legumain maintains

activity at low pH, down to pH 4.0 and is active in non-reducing conditions,[21] therefore it is also suitable for protein disulfide bond determination at the low pH environment required to prevent disulfide reshuffling.[46] Currently pepsin is used for these experiments due to its high activity under acidic conditions. However, pepsin generates a large number of overlapping peptides due to its broad specificity with non-exclusive preference for cleavage after Tyr, Phe, Trp and Leu that complicate the spectra assignment, whereas legumain´s high cleavage specificity would alleviate this problem. Taken together, we propose that recombinant human legumain is an attractive protease to complement trypsin in bottom-up mass spectrometry-based proteomics.

ASSOCIATED CONTENT

Supporting Information Available;

Supporting figures (PDF)

Supporting tables (.xlsx)

AUTHOR INFORMATION

Corresponding Author

*E-mail: p.huesgen@fz-juelich.de

Author Contributions

F.D. and P.F.H. conceived the project. W.T.S., F.D., E.D., S.O.D., H.B. and P.F.H. designed the experiments and analyzed the data. E.D. and S.O.D. provided the recombinant human legumain and W.T.S., F.D., M.K. and A.P. performed the experiments. The manuscript was written by W.T.S., F.D. and P.F.H., and edited by all authors. All authors have approved the final version of the manuscript.

NOTES

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We thank Dr. Ulrich Eckhard for critically reading this manuscript. W.T.S. is a PhD student in the Immunity in Cancer and Allergy PhD program funded by the Austrian Science Fund FWF (project W_01213). This project was in part supported by a starting grant of the European Research Council with funding from the European Union´s Horizon 2020 program (grant 639905, to P.F.H.) and the German Research Foundation DFG (FOR2743, grant HU1756/3-1 to P.F.H.).

REFERENCES

1. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016,** *537* (7620), 347-55.
2. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd, Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews* **2013,** *113* (4), 2343-94.
3. Tsiatsiani, L.; Heck, A. J., Proteomics beyond trypsin. *FEBS J* **2015,** *282* (14), 2612-26.
4. Swaney, D. L.; Wenger, C. D.; Coon, J. J., Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res* **2010,** *9* (3), 1323-9.
5. Perrar, A.; Dissmeyer, N.; Huesgen, P. F., New beginnings and new ends - Methods for large-scale characterization of protein termini and their use in plant biology. *Journal of experimental botany* **2019,** *70* (7), 2021-38.
6. Lange, P. F.; Overall, C. M., Protein TAILS: when termini tell tales of proteolysis and function. *Curr Opin Chem Biol* **2013,** *17* (1), 73-82.
7. Turk, B.; Turk, D.; Turk, V., Protease signalling: the cutting edge. *The EMBO journal* **2012,** *31* (7), 1630-43.
8. Klein, T.; Eckhard, U.; Dufour, A.; Solis, N.; Overall, C. M., Proteolytic Cleavage-Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification. *Chemical reviews* **2018,** *118* (3), 1137-1168.
9. Niedermaier, S.; Huesgen, P. F., Positional proteomics for identification of secreted proteoforms released by site-specific processing of membrane proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2019,** *1867* (12), 140138.
10. Vogtle, F. N.; Wortelkamp, S.; Zahedi, R. P.; Becker, D.; Leidhold, C.; Gevaert, K.; Kellermann, J.; Voos, W.; Sickmann, A.; Pfanner, N.; Meisinger, C., Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* **2009,** *139* (2), 428-39.
11. Giansanti, P.; Tsiatsiani, L.; Low, T. Y.; Heck, A. J., Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nature protocols* **2016,** *11* (5), 993-1006.
12. Huesgen, P. F.; Lange, P. F.; Rogers, L. D.; Solis, N.; Eckhard, U.; Kleifeld, O.; Goulas, T.; Gomis-Ruth, F. X.; Overall, C. M., LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nature methods* **2015,** *12* (1), 55-58.
13. Schräder, C. U.; Lee, L.; Rey, M.; Sarpe, V.; Man, P.; Sharma, S.; Zabrouskov, V.; Larsen, B.; Schriemer, D. C., Neprosin, a Selective Prolyl Endoprotease for Bottom-up Proteomics and Histone Mapping. *Molecular & cellular proteomics : MCP* **2017,** *16*, 1162-1171.

14. Schlosser, A.; Vanselow, J. T.; Kramer, A., Mapping of Phosphorylation Sites by a Multi-Protease Approach with Specific Phosphopeptide Enrichment and NanoLC-MS/MS Analysis. *Analytical chemistry* **2005,** *77*, 5243-5250.
15. Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., 3rd, A method for the comprehensive proteomic analysis of membrane proteins. *Nature biotechnology* **2003,** *21* (5), 532-8.
16. Gonczarowska-Jorge, H.; Loroch, S.; Dell'Aica, M.; Sickmann, A.; Roos, A.; Zahedi, R. P., Quantifying Missing (Phospho)Proteome Regions with the Broad-Specificity Protease Subtilisin. *Analytical chemistry* **2017,** *89* (24), 13137-13145.
17. Meyer, J. G.; Kim, S.; Maltby, D. A.; Ghassemian, M.; Bandeira, N.; Komives, E. A., Expanding proteome coverage with orthogonal-specificity α-lytic proteases. *Molecular & cellular proteomics : MCP* **2014,** *13* (3), 823-35.
18. Wu, C. C.; Yates, J. R., The application of mass spectrometry to membrane proteomics. *Nature biotechnology* **2003,** *21*, 262-267.
19. Giansanti, P.; Aye, T. T.; van den Toorn, H.; Peng, M.; van Breukelen, B.; Heck, A. J., An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas. *Cell reports* **2015,** *11* (11), 1834-43.
20. Lange, P. F.; Huesgen, P. F.; Nguyen, K.; Overall, C. M., Annotating N termini for the human proteome project: N termini and nalpha-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *Journal of proteome research* **2014,** *13* (4), 2028-44.
21. Dall, E.; Brandstetter, H., Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. *Proc Natl Acad Sci U S A* **2013,** *110* (27), 10940-5.
22. Vidmar, R.; Vizovisek, M.; Turk, D.; Turk, B.; Fonovic, M., Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *EMBO J* **2017,** *36* (16), 2455-2465.
23. Hebert, D. N.; Lamriben, L.; Powers, E. T.; Kelly, J. W., The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis. *Nat Chem Biol* **2014,** *10* (11), 902-10.
24. Dall, E.; Brandstetter, H., Activation of legumain involves proteolytic and conformational events, resulting in a context- and substrate-dependent activity profile. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **2012,** *68* (Pt 1), 24-31.
25. Wessel, D.; Flugge, U. I., A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem* **1984,** *138* (1), 141-3.
26. Hughes, C. S.; Moggridge, S.; Muller, T.; Sorensen, P. H.; Morin, G. B.; Krijgsveld, J., Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature protocols* **2019,** *14* (1), 68-85.
27. Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature protocols* **2007,** *2* (8), 1896-906.
28. Rinschen, M. M.; Hoppe, A. K.; Grahammer, F.; Kann, M.; Volker, L. A.; Schurek, E. M.; Binz, J.; Hohne, M.; Demir, F.; Malisic, M.; Huber, T. B.; Kurschat, C.; Kizhakkedathu, J. N.; Schermer, B.; Huesgen, P. F.; Benzing, T., N-Degradomic Analysis Reveals a Proteolytic Network Processing the Podocyte Cytoskeleton. *J Am Soc Nephrol* **2017,** *28* (10), 2867-2878.

29. Tyanova, S.; Temu, T.; Cox, J., The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols* **2016,** *11* (12), 2301-2319.
30. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature methods* **2016,** *13* (9), 731-40.
31. Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J., The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **2015,** *10* (6), 845-58.
32. Weng, S. S. H.; Demir, F.; Ergin, E. K.; Dirnberger, S.; Uzozie, A.; Tuscher, D.; Nierves, L.; Tsui, J.; Huesgen, P. F.; Lange, P. F., Sensitive determination of proteolytic proteoforms in limited microscale proteome samples. *Molecular & cellular proteomics : MCP* **2019,** *TIR119.001560*.
33. Joosten, M. H., Isolation of apoplastic fluid from leaf tissue by the vacuum infiltration-centrifugation technique. *Methods Mol Biol* **2012,** *835*, 603-10.
34. Deutsch, E. W.; Orchard, S.; Binz, P. A.; Bittremieux, W.; Eisenacher, M.; Hermjakob, H.; Kawano, S.; Lam, H.; Mayer, G.; Menschaert, G.; Perez-Riverol, Y.; Salek, R. M.; Tabb, D. L.; Tenzer, S.; Vizcaino, J. A.; Walzer, M.; Jones, A. R., Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J Proteome Res* **2017,** *16* (12), 4288-4298.
35. Vizcaino, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q. W.; Wang, R.; Hermjakob, H., 2016 update of the PRIDE database and its related tools. *Nucleic acids research* **2016,** *44* (D1), D447-56.
36. Weng, S. S. H.; Demir, F.; Ergin, E. K.; Dirnberger, S.; Uzozie, A.; Tuscher, D.; Nierves, L.; Tsui, J.; Huesgen, P. F.; Lange, P. F., Sensitive determination of proteolytic proteoforms in limited microscale proteome samples. *bioRxiv* **2019**, 566109.
37. Clerc, F.; Reiding, K. R.; Jansen, B. C.; Kammeijer, G. S. M.; Bondt, A.; Wuhrer, M., Human plasma protein N-glycosylation. *Glycoconj J* **2016,** *33*, 309-343.
38. Kuster, B.; Mann, M., 18O-Labeling of N-Glycosylation Sites To Improve the Identification of Gel-Separated Glycoproteins Using Peptide Mass Mapping and Database Searching. *Analytical chemistry* **1999,** *71*, 1431-1440.
39. Dall, E.; Brandstetter, H., Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. *Proceedings of the National Academy of Sciences of the United States of America* **2013**.
40. Barth, C.; Jander, G., Arabidopsis myrosinases TGG1 and TGG2 have redundant function in glucosinolate breakdown and insect defense. *Plant J* **2006,** *46* (4), 549-62.
41. Liebminger, E.; Grass, J.; Jez, J.; Neumann, L.; Altmann, F.; Strasser, R., Myrosinases TGG1 and TGG2 from Arabidopsis thaliana contain exclusively oligomannosidic N-glycans. *Phytochemistry* **2012,** *84*, 24-30.
42. Wang, Y.; Chen, Y.; Zhang, Y.; Wei, W.; Li, Y.; Zhang, T.; He, F.; Gao, Y.; Xu, P., Multi-Protease Strategy Identifies Three PE2 Missing Proteins in Human Testis Tissue. *Journal of proteome research* **2017,** *16* (12), 4352-4363.
43. Paik, Y. K.; Overall, C. M.; Deutsch, E. W.; Van Eyk, J. E.; Omenn, G. S., Progress and Future Direction of Chromosome-Centric Human Proteome Project. *Journal of proteome research* **2017,** *16* (12), 4253-4258.

44. Tallant, C.; Garcia-Castellanos, R.; Marrero, A.; Canals, F.; Yang, Y.; Reymond, J. L.; Sola, M.; Baumann, U.; Gomis-Ruth, F. X., Activity of ulilysin, an archaeal PAPP-A-related gelatinase and IGFBP protease. *Biological chemistry* **2007,** *388* (11), 1243-53.
45. Dahms, S. O.; Demir, F.; Huesgen, P. F.; Thorn, K.; Brandstetter, H., Sirtilins - the new old members of the vitamin K-dependent coagulation factor family. *J Thromb Haemost* **2019,** *17* (3), 470-481.
46. Gorman, J. J.; Wallis, T. P.; Pitt, J. J., Protein disulfide bond determination by mass spectrometry. *Mass Spectrom Rev* **2002,** *21* (3), 183-216.

FIGURE LEGENDS

Figure 1. Substrate cleavage specificity of legumain, GluC and trypsin. iceLogos visualize the amino acid frequencies based surrounding the cleavage sites inferred from peptides identified by nonspecific database searches after digestion of an *A. thaliana* leaf proteome (a-c) or mouse embryonic fibroblast cell lysate proteome (b-f) with (a,d) legumain, (b,e) GluC or (c,f) trypsin. The numbers of non-redundant cleavage sites for each logo are indicated.

Figure 2. Analysis of an *A. thaliana* leaf proteome digested with legumain, GluC and trypsin, each performed in three technical repeats. (a) Overlap of unique peptide sequences identified using enzyme-specific database queries. Analysis of the (b) length, (c) hydrophobicity, and (d) isoelectric point of the identified peptides. (e) Overlap in unique amino acid identified by digestion with the three proteases. (f) Protein sequence coverage observed for superoxide dismutase (At1g08830) in legumain (red, 93%), GluC (green, 43%) and trypsin (blue, 49%) proteome digests. (g) Upset plot showing the overlap in protein groups identified in individual technical digestion replicates. (h) Venn diagram showing the total overlap of protein groups identified by the three enzymes. (i) Reproducibility of proteome quantification (MaxQuant LFQ). Only proteins quantified with 2 or more peptides were considered. Value indicates the Pearson correlation between the LFQ values obtained for technical replicates.

Figure 3. Potential cleavage sites missed by legumain, GluC, and trypsin in *A. thaliana* leaf proteome digests. (a) Percentage of peptides containing up to three missed cleavage sites. (b) Missed cleavage sites sorted by missed amino acid residues.

Figure 4. Complementary N-terminome coverage by parallel digestion with legumain, GluC, and trypsin. (a) Experimental workflow for the enrichment of N-terminal peptides using HUNTER. For detailed description, see main text. Blue and red circles indicate differential stable isotope labeling by reductive dimethylation, magenta triangles indicate undecanal modification. (b) Overlap in N-termini identification based on the first seven amino acids of each N-terminal peptide identified in the experiments with the three proteases. Peptide MS/MS fragmentation spectra of (c) the acetylated mature N-terminus of GLUCOSINOLATE TRANSPORTER-1 and (d) a proteolysis-derived dimethylated N-terminus in the CLPR3 subunit of the ATP-dependent Clp protease. Both termini

were identified in legumain digests, with sequence context surrounding the identified peptide indicated in grey. UniProt accession code and gene accession numbers are indicated.

Figure 5. Identification of N-glycosylation sites by sequential processing with legumain and PNGase F. (a) Scheme of the experimental workflow. For details, see main text. Blue and red circles indicate differential stable isotope labeling by dimethylation. Asterisks indicate deamidated asparagine residue arising from PNGase F treatment. (b) Overlap of N-glycosylation identified with internal deamidated Asn in workflow 1 and with C-terminal deamidated Asn in workflow 2. (c) MS/MS fragmentation spectra of an N-glycosylation site in MYROSINASE 1 identified in both workflows. UniProt and *A. thaliana* gene accession codes are indicated.