ACS chemical biology

Articles

pubs.acs.org/acschemicalbiology

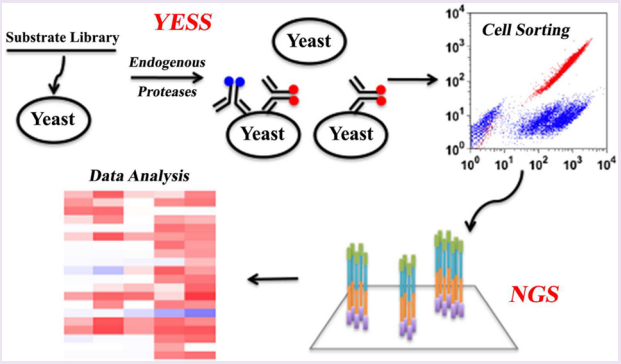
Profiling Protease Specificity: Combining Yeast ER Sequestration Screening (YESS) with Next Generation Sequencing

Qing Li,[†] Li Yi,[†,⊥] Kam Hon Hoi,[‡] Peter Marek,[†] George Georgiou,[*,§,∥] and Brent L. Iverson[*,†]

[†]Department of Chemistry, [‡]Department of Biomedical Engineering, [§]Department of Chemical Engineering, and [∥]Section of Molecular Genetics and Microbiology, University of Texas, Austin, Texas 78712, United States

Supporting Information

ABSTRACT: An enzyme engineering technology involving yeast endoplasmic reticulum (ER) sequestration screening (YESS) has been recently developed. Here, a new method is established, in which the YESS platform is combined with NextGen sequencing (NGS) to enable a comprehensive survey of protease specificity. In this approach, a combinatorial substrate library is targeted to the yeast ER and transported through the secretory pathway, interacting with any protease(s) residing in the ER. Multicolor FACS screening is used to isolate cells labeled with fluorophore-conjugated antibodies, followed by NGS to profile the cleaved substrates. The YESS-NGS method was successfully applied to profile the sequence specificity of the wild-type and an engineered variant of the tobacco etch mosaic virus protease. Proteolysis in the yeast secretory pathway was also mapped for the first time *in vivo* revealing a major cleavage pattern of Ali/Leu-X-Lys/Arg-Arg. Here Ali is any small aliphatic residue, but especially Leu. This pattern was verified to be due to the well-known endogenous protease Kex2 after comparison to a newly generated Kex2 knockout strain as well as cleavage of peptides with recombinant Kex2 *in vitro*. This information is particularly important for those using yeast display technology, as library members with Ali/Leu-X-Lys/Arg-Arg patterns are likely being removed from screens via Kex2 cleavage without the researcher's knowledge.



The analysis of enzyme substrate specificity is interesting, because there is no quantitative measure of absolute substrate specificity. Rather, specificity must be discussed in relative terms in which ratios of catalytic parameters with multiple substrates are presented to ascertain patterns of reactivity. It follows, then, that enzyme substrate specificity is defined better when more substrates are considered. Taken to the logical limit, the best possible characterization of enzyme substrate specificity would involve screening all possible substrates using a quantitative analysis followed by a comprehensive deconvolution of reactivity patterns.

In an effort to screen as many protease substrates as possible, we have combined the recently reported yeast endoplasmic reticulum (ER) sequestration screening (YESS) technology[1,2] with NextGen sequencing (NGS) (Figure 1a,b) and a comparative sequence analysis to profile protease specificity using a large number of possible sequences in a single experiment. In this approach, the YESS reporter substrate fusion construct consists of an Aga2 protein, the Flag antibody epitope sequence, a randomized putative substrate sequence, the HA epitope, and an ER retention signal peptide, in that order. The N-terminal Aga2 sequence ensures that following transit through the ER and secretion, the uncleaved substrate/cleaved products are covalently attached to the outer surface. Cells are probed simultaneously with anti-FLAG and anti-HA antibodies that are conjugated to phycoerythrin (PE) and fluorescein (FITC), respectively. Cleavage is detected via two-dimensional FACS analysis by monitoring the ratio of PE to FITC fluorescence. A high amount of both fluorescent signals indicates a lack of cleavage, while a high PE signal accompanied by a low FITC signal indicates cleavage at the substrate site. After FACS-based sorting and isolation, the cleaved sequences are identified by NGS followed by a comparative sequence analysis to deconvolute cleavage patterns.

Any recombinant protease of interest being analyzed in the YESS system will be hydrolyzing substrates above a background of endogenous yeast protease cleavage, in particular, the endogenous proteolysis involved with the yeast cellular secretion pathway. The cellular secretion machinery, including associated processing enzymes, is crucial for successful operation of the eukaryotic secretome.[3] Even minimal modification of a secretory pathway can drive global change in protein secretion and create wide-ranging cellular effects.[4−6] Studies of cellular secretory processes are essential to better understand the factors contributing to effective secretion, with application to recombinant protein production and engineering[7,8] as well as helping to uncover potential secretome alterations in diseases

Received: June 23, 2016
Accepted: December 15, 2016
Published: December 15, 2016

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510−518

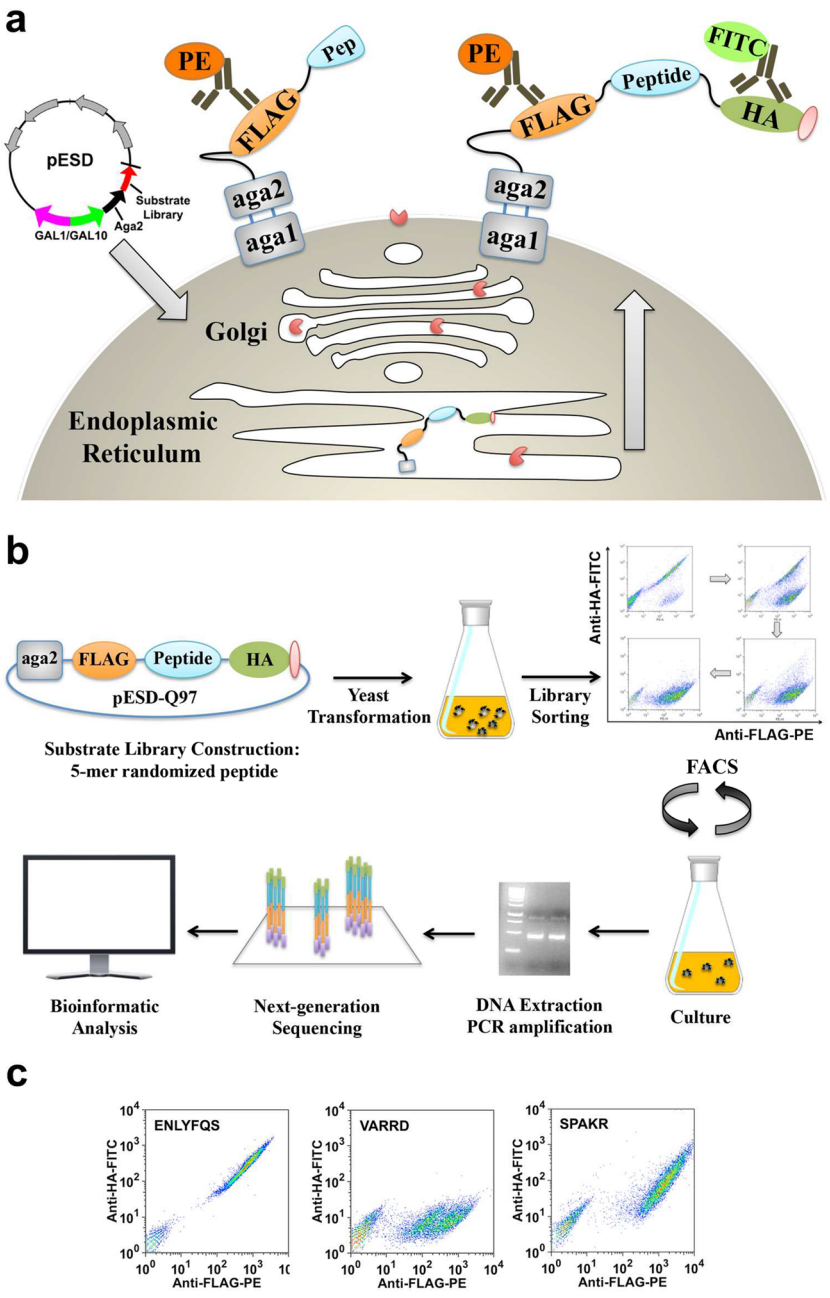


Figure 1. Yeast endoplasmic reticulum sequestration screening (YESS) system for mapping endopeptidase cleaveOme in the yeast secretory pathway. (a) Concept: The Aga2-substrate polypeptide library is expressed from the pESD shuttle vector, and translocated to the ER secretory pathway. The proteolytic cleavage of the substrate fusion polypeptide by the endogenous proteases gives rise to a product with cleaved signal that is displayed on the cell surface by virtue of the N-terminal Aga2. The presence of epitope tags in the processed substrate fusion is detected with fluorescently labeled antibodies to identify the cleaved or noncleaved signals. (b) Overview of the method: Substrate library is screened and enriched by selecting the library pool of clones showing the cleaved signals. Next generation sequencing is performed to sequence the substrate libraries. Bioinformatic processing is used to analyze the cleaveOme in the yeast secretory pathway. (c) In EBY100 cells, 2-color FACS analysis of cells with cleaved substrate and noncleaved substrate signals. From left to right panel: wild-type TEV substrate (ENLYFQS); VARRD (Arg-Arg pattern); SPAKR (Lys-Arg pattern).

such as cancer.[9] In eukaryotes, proteolytic processing in cellular secretory pathways plays an important role in protein maturation and protein sorting into secretory vesicles.[10,11] Most secreted proteins, including growth factors, receptors, enzymes, and neuropeptides, require proteolytic processing at specific sites.[12] Emphasizing their importance, null mutation in certain of these known convertase genes have lethal effects on embryos.[6]

Genetic and biochemical studies have led to the identification and characterization of endogenous convertases such as Kex2 (also known as kexin, peptidase 3.4.21.61) existing in the yeast secretory pathway.[10] Previous reports have indicated that the Kex2 convertase catalyzes cleavage after two basic residues, especially Lys-Arg, so dibasic sites were generally considered to be classical processing sites in precursors of secreted proteins.[13−15] However, due to a lack of any previous comprehensive analysis of the endogenous convertase cleaving patterns inside the secretory pathway of a living yeast cell, questions remain with respect to exactly which sequences are cleaved and how many different endogenous convertases are involved.

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510−518

Various chemical and biological based approaches, including microarray, phage display, and bacterial display, have been developed to characterize protease substrate specificity.[16−18] CLiPS is a particularly interesting approach using bacterial display of genetically encoded substrate libraries followed by FACS sorting to identify cleaved peptides.[19] More recent methods involve mass spectral analysis of peptide libraries,[20] endogenously cleaved protein substrates,[21] *Escherichia coli* based surface electrostatic capture,[22,23] phage displayed substrate library,[24] or TAILS technology.[25] Each of these previously reported methods has its advantages, but none combine a protease engineering platform with a comprehensive substrate specificity profiling technology, simplifying the development and analysis of engineered proteases. Additionally, no previous protease specificity analysis approach has taken advantage of the opportunity afforded by NGS to analyze extremely large numbers of sequences in a single experiment.

Herein we report the successful use of the YESS-NGS system to provide the first comprehensive analysis of the endogeneous cleavage specificity of convertases of the yeast secretory pathway, confirming Kex2 as the major endogenous protease in this pathway as well as identifying a more refined model for Kex2 specificity that was confirmed with recombinant Kex2 *in vitro*. These studies have provided the necessary foundation for a thorough profiling of protease substrate specificity, as exemplified by our comprehensive analysis of both the wild-type and an engineered tobacco etch mosaic virus protease (TEV-P). A Kex2 knockout strain produced in the course of these studies could find use when unwanted cleavage needs to be avoided such as during secreted protein production or protein engineering efforts using yeast display approaches.

RESULTS

System Validation. Negative and positive controls were run to validate the YESS-NGS approach (Figure 1). For a negative control, a YESS substrate fusion construct was created without an exogenous protease but with a substrate sequence not expected to have an endogenous yeast cleavage site (the TEV-P cleavage sequence ENLYFQS). Antibody labeling following incubation yielded cells with equally high PE and FITC signals as expected for a substrate that is not cleaved (Figure 1c). As a positive control for cleavage, a YESS substrate fusion construct was created incorporating a known Kex2 cleavage sequence, VARRD.[26] As expected, yeast cells containing the VARRD cleavage sequence displayed relatively high PE fluorescence and low FITC fluorescence in the FACS fluorescence scatter plots, consistent with proteolysis within the VARRD sequence (Figure 1c).

Understanding Background Cleavage in the Yeast Secretory Pathway. To characterize the endogeneous convertase(s) in the yeast secretory pathway, a substrate library was prepared by combinatorial NNS randomization of five sequential amino acid positions within the substrate region of the reporting construct (labeled as "peptide" in Figure 1 cartoon). A total of 3×10^8 cells were analyzed for the substrate library that has a theoretical diversity of 3.2×10^6 different members. Within the presorting library, the percentage of each amino acid appearing in the five positions was compared with the expected frequency of each amino acid in the NNS library (Table S1). All residues except Pro were found to be present at or near the expected frequency. Nevertheless, the sequencing results from each round of enrichment were normalized based on the presorting library when determining the specific substrate enrichment or de-enrichment, so the observed abundance deviation of Pro in the original library did not affect any of the final conclusions.

Three consecutive rounds of FACS sorting for high PE and low FITC signal intensity yielded 8.5×10^5 DNA sequences. Recall that in analogy to the VARRD positive control sequence, this high PE and low FTIC signal is consistent with cleavage within the substrate region of the reporting construct. A total of 1.0×10^7 sequences from the same library were also analyzed before sorting to provide an accurate basis for comparative sequence analysis. For both libraries, isolated DNA fragments containing the substrate sequences were amplified and analyzed with a Hiseq NextGen DNA sequencer (Illumina).

Increased prevalence, that is, enrichment, of particular residues at any of the five randomized positions of the sorted library relative to the unsorted library was taken as evidence of enhanced representation within the cleaved substrates, indicated by blue color in Figure 2. Conversely, several amino

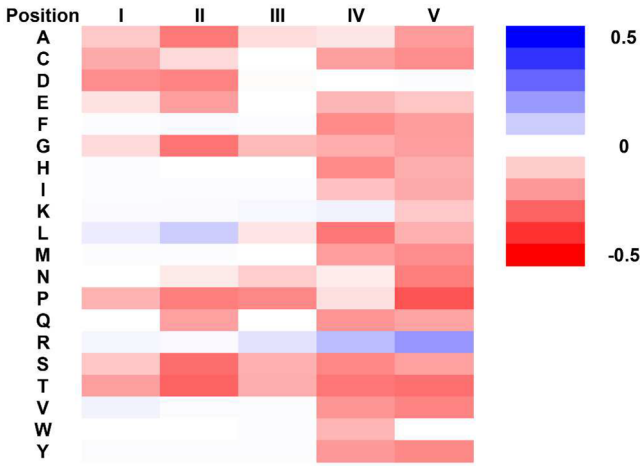


Figure 2. Yeast secretory pathway endopeptidase cleaveOme specificity profile. Heat map shows the specificity scores compiled from all sequences identified in the selection. Specificity scores were calculated by dividing the change in frequency of the amino acid at each position in the postselection pool compared to the preselection pool by the maximal possible change in frequency from preselection library to postselection library of the AA at each position. Blue and red boxes indicate enrichment for and against an AA at a given position, respectively, as indicated by the color scale.

acids were found to be substantially de-enriched following the FACS sorting, and these residues are shaded red in Figure 2. The detailed positive and negative specificity scores are summarized in Table S2. In addition, to evaluate the statistical significance of the enrichment and de-enrichment of each residue at different positions, a statistical test based on the underlying *t* test concept was implemented (Table S3). Based on our results, the greatest enrichments observed in the sorted library were for the basic residue Arg at positions III, IV, and V, with positive specificity scores of 0.06, 0.13, and 0.20, respectively. In addition, the hydrophobic residue Leu at positions I and II was also enriched, with positive specificity scores of 0.03 and 0.10, respectively. Note that the Roman numerals relate to the position of the substrate randomization, consisting of five consecutive positions, I−V, in the YESS substrate-reporting construct. Enrichment for the basic residue Lys at positions III and IV was also seen. However, with positive specificity scores less than 0.03, its enrichment is nowhere near the same as that seen

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510−518

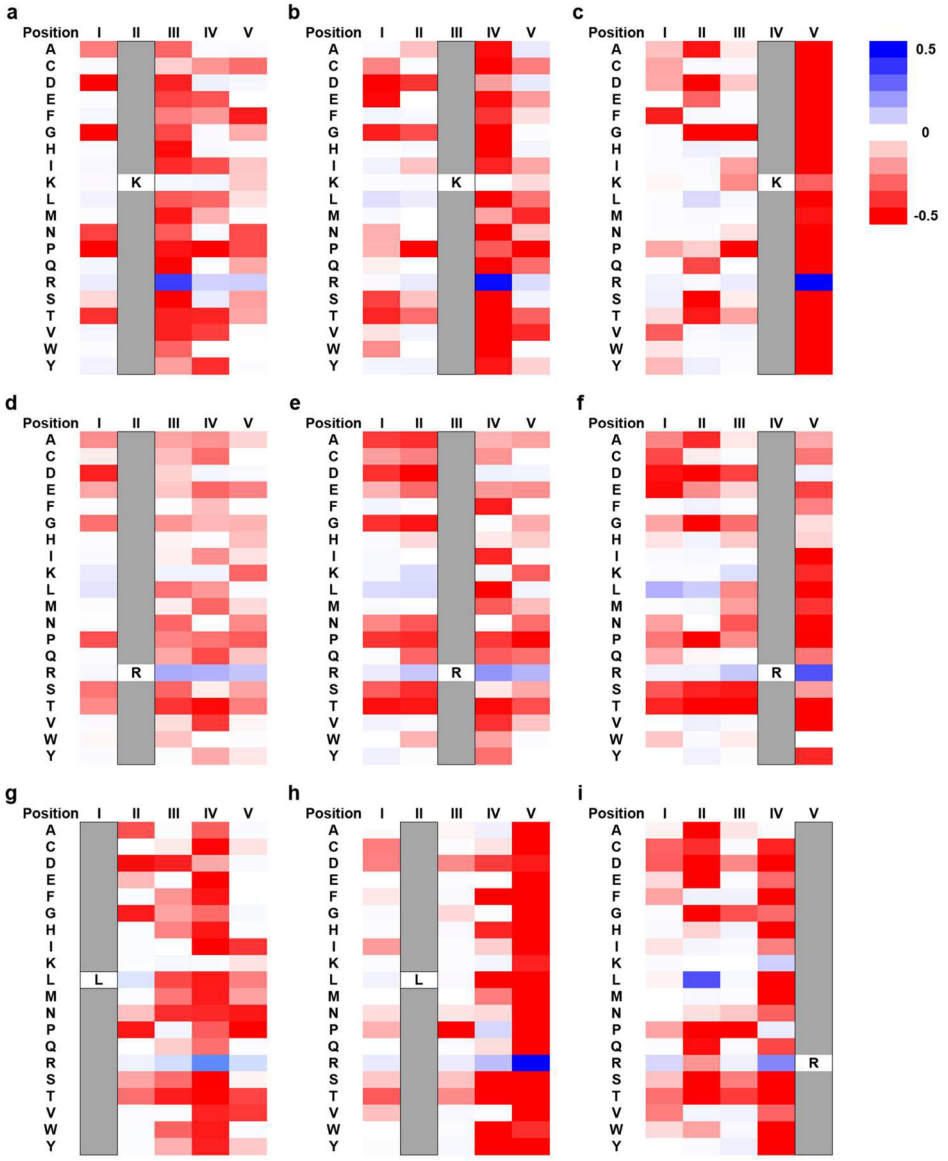


Figure 3. Analysis of cleavage sequence patterns in the yeast secretory cleaveOme when selected sequences are filtered for the presence of a particular residue (indicated by gray bar) at one of the randomized positions labeled as I–V. Blue color indicates a strongly enriched residue in the cleaved/selected pool relative to the unsorted pool, while red indicates a residue that is strongly de-enriched in the cleaved pool using the same scaling as in Figure 2. (a–c) Identification of enrichment of Arg (R) residues following Lys (K) at positions II–IV. (d–f) Identification of enrichment of one or more Arg (R) residues following Arg (R) at positions II–IV. (g,h) Identification of enrichment for Arg (R) in the third position following Leu (L) in positions I and II and (f,i) identification of corresponding enrichment of Leu (L) in the third position preceding Arg (R) in positions IV and V.

with Arg. The residue Val was also slightly enriched at positions I, II, and III, with positive specificity scores of 0.02, 0.01, and 0.01, respectively. No patterns were identified when the sorted libraries were analyzed after excluding all sequences containing the basic residues Lys and Arg (Figure S1). Note that there is a slight enrichment of Asp at positions IV and V, especially following Arg at positions III and IV, respectively. The most significant de-enrichment was seen at position II, especially for the small or hydrophilic residues Ala, Asp, Glu, Gly, Pro, Gln, Ser, and Thr.

To identify positive linkages between residues in cleaved sequences, all selected sequences with a specific residue in a given position (i.e., Leu at position I, etc.) were examined for the presence of residues appearing at a frequency above background in any other positions. The reported Lys/Arg-Arg specificity of Kex2[27] would predict that among the cleaved substrates, a strong enrichment for Lys/Arg-Arg would be found throughout the targeted area (Figure S2). As can be seen in the first three panels of Figure 3, when positions II, III, or IV were fixed as Lys, all amino acids immediately adjacent on the C-terminal side were strongly de-enriched with the exception of Arg, which was strongly enriched, and Lys, which was not de-enriched or enriched. Lys at position I showed modest enrichment of Arg at the II position (see Figure S3 for a presentation of the entire data set), but not at the level seen with Lys in positions II–IV. A related dibasic pattern exists for Arg at positions II–V, with Arg or Lys being enriched to the N-terminal side and usually Arg being enriched to the C-terminal side as well (Figure S2). Interestingly, Arg was observed to be enriched with positive specificity scores ranging from 0.01 to 0.20 at all

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510–518

positions in these substrates underscoring its importance in endogenous protease substrate recognition (Figure 3d–f). Note also that Arg is enriched to an overall significantly greater extent than Lys (Figure 2). Collectively, these data were interpreted to indicate the presence of a strong preference for a dibasic sequence in cleaved substrates, but only Lys-Arg or the much more common Arg-Arg, as only a small amount of Lys-Lys and no Arg-Lys enrichment was detected (Figure S3).

The highly enriched Leu residues in positions I and II (Figure 2) are strongly linked to at least one Arg residue later in the sequence. Filtering the data for all selected sequences containing Leu at position II shows significant enrichment for Arg at position V (Figure 3h). In corresponding fashion, looking at all selected sequences containing Arg at position V reveals enrichment of Leu at position II (Figure 3i). Similarly, the selected sequences with Leu at position I show a strong enrichment of Arg at position IV (Figure 3g) and the selected sequences with Arg at position IV show substantial enrichment of Leu at position I (Figure 3f).

A more in-depth analysis reveals that all sequences with Val at position II show a pronounced increase in Arg at both IV and V (Figure S3). The same is also true for Phe, Met, Ile, and maybe Tyr and Trp, Therefore, there appears to be justification for extending the overall cleavage pattern to be Ali/Leu-X-X-Arg in which Ali is any aliphatic residue, although Leu is the clearly dominant amino acid found in the cleaved sequences.

From analysis of the NGS data alone, it is not clear whether there are two or one actual predominant patterns. One might imagine that two different endogenous proteases are responsible for the dibasic Lys/Arg-Arg and Ali/Leu-X-X-Arg patterns. Alternatively, the same data could be interpreted as indicating a single endogenous protease responsible for a combined Ali/Leu-X-Lys/Arg-Arg pattern. Either way, Arg appears to be essential for proteolysis in the yeast secretory pathway.

Analysis of Recombinant Protease Specificity. As a proof-of-principle study, the combined YESS-NGS approach was next used to profile the sequence specificities of two recombinant proteases: wild-type (TEV-P) and an engineered variant (TEV-PE10: S120R, D148R, T173A, N177K, M218I)[1] of the tobacco etch mosaic virus (TEV) protease. Being previously engineered using the YESS system, purified TEV-PE10 exhibited a 5000-fold increase in reactivity (as k_{cat}/K_M) for a peptide substrate containing Glu at P1 instead of the wild-type preferred Gln.

To profile the substrate specificity of wild-type TEV-P as well as TEV-PE10, we introduced sequences encoding wild-type TEV-P or TEV-PE10 into the protease side of the YESS vector downstream of the GAL1 promoter. An abbreviated substrate library was generated by NNS randomization of four residues corresponding to the P1′, P1, P3, and P6 positions on the reporter construct side of the same YESS vector. Positions P2, P4, and P5 were fixed to be Phe, Leu and Asn, respectively, consistent with wild-type preferences at these positions.[28] After three rounds of FACS sorting for high PE and low FITC signal intensity, the enriched libraries were isolated, and the DNA fragments encoding the substrate sequences were amplified then analyzed by NGS. A large unsorted aliquot of the same library served as a reference. As before, sequences found to be enriched after sorting based on comparison to the unsorted reference were assumed to have undergone protease cleavage. In order to avoid the background signal from the endogeneous convertase(s) in the yeast secretory pathway, sequences were excluded that contained the basic amino acids Lys and Arg. The exclusion of Lys and Arg is expected to have only a negligible effect on our analysis of especially the P1 position, as previous work has shown absolutely no cleavage of substrates with Lys or Arg in the P1 position by wild-type TEV-P.[28]

The overall specificity profiles of the recombinant TEV proteases revealed that, as expected, wild-type TEV-P selectively recognizes Gln at P1 while the engineered TEV-PE10 variant prefers Glu at P1. Both recombinant TEV proteases exhibited strong preferences at P1′, P3, or P6 for Ser, Tyr, or Glu, respectively (Figure S4). To further deconvolute the positional correlations within the substrate profiles of recombinant TEV proteases, we analyzed the specificity profiles by looking at only selected sequences that contained a particular amino acid at one of the randomized positions (indicated by the gray boxes (Figure 4). Consistent with the overall specificity profiles, we

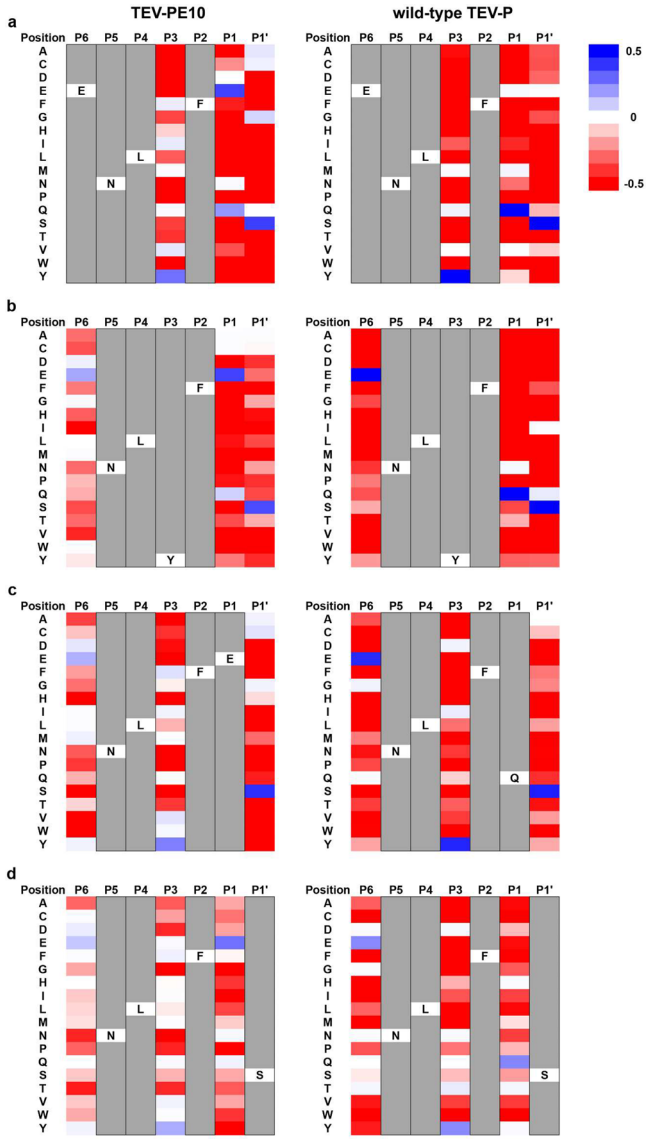


Figure 4. Specificity profiles of engineered TEV-PE10 (left panels) and wild-type TEV-P (right panels) based on the substrate library being randomized at P1′, P1, P3, and P6 within the substrate region, and mutation at position P6 with amino acid E (a) or position P3 with amino acid Y (b) or position P1 with amino acid E/Q (c) or position P1′ with amino acid S (d). After filtration of Arg and Lys containing sequences, the total number of peptide sequences from the postselection library for the wild-type TEV-P and TEV-PE10 are 473011 and 1786504, respectively.

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510–518

observed significant enrichments of the ENLYFQS sequence for the wide-type TEV-P and ENLYFES sequence for the engineered TEV-PE10 (Figure 4). Note that while TEV-PE10 did appear to cleave substrates with Gln at P1 to some extent, no detectable enrichment was seen for Glu at P1 with TEV-P.

The bottom line is that by far the most important change in specificity observed among the P1′, P1, P3, and P6 residues is that TEV-P prefers Gln, while TEV-PE10 prefers Glu at P1. The other three positions examined appeared similar or identical in specificity between the two, indicating that the engineered TEV-PE10 maintained a specificity profile that is only altered at the P1 position relative to wild-type.[28] No other residues were enriched to a significant extent at any of the randomized positions when cleaved by either protease. Further, the TEV-PE10 specificity for Glu at P1 is not the result of relaxed specificity at that position but represents a genuine alteration of specificity in favor of Glu while retaining some activity for Gln, as no other amino acids are enriched at P1. It is worth noting that the TEV-PE10 variant was isolated using YESS according to a strategy that incorporates counter-selection as well as selection for a new substrate, an approach that is intentionally designed to avoid isolating engineered variants with relaxed substrate specificity profiles.

Identification of Kex2 as the Major Endogenous Protease in the Yeast Secretory Pathway. The profiling of the sorted library without an added exogenous protease clearly indicated significant enrichment of Ali/Leu-X-X-Arg as well as dibasic Lys/Arg-Arg sequences (Figure 2). As mentioned previously, these two main cleavage patterns could be the result of two or more endogenous proteases in the yeast secretory pathway, or they could represent two related patterns recognized by the same protease. Of the known yeast proteases, Kex2 was the most likely to be involved based on its documented activity in yeast secretory processing,[10] further implicated by its reported dibasic sequence preference.[29] However, to the best of our knowledge, a strong preference for Ali/Leu-X-X-Arg by Kex2 had not been noted previously. As our first step toward understanding the origin of the observed secretory pathway proteolysis, a Kex2 knockout EBY100 strain (EBY100^{Kex2-}) was created.

Both the EBY100 and Kex2 knockout EBY100^{Kex2-} strains were analyzed by FACS with the 20 most common individual substrates identified in the NGS sequence analysis of endogenous yeast secretory pathway proteolysis described above (Table 1). In each case, the identified sequence was cloned into the substrate construct and the FACS signal monitored in both the EBY100 and the EBY100^{Kex2-} strain. As before, cleavage activity was indicated by a high PE and low FITC signal, while a lack of cleavage was revealed if both the PE and FITC signals were high (Figure S5). Table 1 lists the 20 most common substrates and whether the FACS analysis revealed cleavage with either yeast strain.

Of the top 20 most common substrates isolated, 15 contained the Leu-X-X-Arg pattern, and *all* of these were cleaved efficiently in EBY100 as expected. Of these 15 substrates, *none* exhibited cleavage in EBY100^{Kex2-}. Notice that ten of the cleaved substrates in the EBY100 strain fit the Leu-X-X-Arg motif (LRPRA, RLRPR, RLLPR, RLTPR, PLLPR, PLRPR, RLAPR, ALLPR, PLLAR, PLYPR) and five others contained an Arg-Arg sequence within a Leu-X-Arg-Arg motif (ALARR, ALSRR, RLSRR, PLLRR, and SLRRR). It is interesting that there were five of the 20 most common substrates, ARKPA, GSFRP, NAFSH, YPVCV, and SPAWR, which presented no cleavage in either EBY100 or EBY100^{Kex2-}. Note that these five substrates did not have either an Ali/Leu-X-X-Arg or Lys/Arg-Arg motif.

Table 1. Analysis of Top 20 Peptide Substrates of the Sorted Library in the EBY100 and EBY100^{Kex2-} Strains

substrate	EBY100	EBY100^{Kex2-}	substrate	EBY100	EBY100^{Kex2-}
ARKPA	X	X	RLTPR	√	X
GSFRP	X	X	PLLPR	√	X
NAFSH	X	X	PLLRR	√	X
ALARR	√	X	PLRPR	√	X
LRPRA	√	X	SPAWR	X	X
ALSRR	√	X	RLAPR	√	X
RLRPR	√	X	ALLPR	√	X
RLLPR	√	X	PLLAR	√	X
YPVCV	X	X	PLVPR	√	X
RLSRR	√	X	SLRRR	√	X

Although not in the top 20 most common isolated substrates, two additional substrates, VARRD and SPAKR, were chosen to investigate dibasic recognition *in vivo* by Kex2 (Figure 1c). These were chosen because no Leu residue is present although there is a dibasic site in each case. The VARRD sequence, previously used as a positive control due to its known activity as a Kex2 substrate,[26] exhibited the expected extensive cleavage in EBY100 and not in EBY100^{Kex2-}. The SPAKR sequence was cleaved to a modest but measurable extent in EBY100 but again not in EBY100^{Kex2-}.

Analysis of Kex2 Cleavage of Enriched Peptides *in Vitro*. In order to further investigate whether the secretory pathway cleavage pattern observed in yeast *in vivo* can be attributed to Kex2, two of the most highly enriched substrates (ALARR and LRPRA) seen with YESS, along with three control peptides (VARRD, AAARR, and ARPRA), were synthesized with an EDANS/DABCYL fluorophore/quencher pair at the N- and C-termini, respectively. Cleavage by recombinant Kex2 was monitored by fluorescence *in vitro* (Table 2, Figure S6).

Table 2. Michaelis–Menten Kinetics of the Recombinant Kex2 with Peptide Substrates

	k_{cat} (s^{-1})	K_m (μM)	k_{cat}/K_m (s^{-1} M^{-1})
ALARR	0.72 ± 0.04	2.44 ± 0.07	(2.95 ± 0.11) × 10^5
VARRD	0.13 ± 0.03	4.06 ± 0.12	(3.20 ± 0.09) × 10^4
LRPRA	0.098 ± 0.01	3.68 ± 0.24	(2.66 ± 0.24) × 10^4
AAARR	0.025 ± 0.009	4.75 ± 0.21	(5.26 ± 0.05) × 10^3

The most highly enriched of the peptide substrates, containing the ALARR sequence, was cleaved the fastest (k_{cat}/K_m = (3.0 ± 0.11) × 10^5 s^{-1} M^{-1}), followed by VARRD, the known Kex2 substrate (k_{cat}/K_m = (3.20 ± 0.09) × 10^4 s^{-1} M^{-1}), and LRPRA, another highly enriched sequence in our YESS results (k_{cat}/K_m = (2.7 ± 0.24) × 10^4 s^{-1} M^{-1}). An AAARR control peptide, lacking the Leu at P4 but containing Arg-Arg at P2–P1, reacted the slowest (k_{cat}/K_m = (5.26 ± 0.05) × 10^3 s^{-1} M^{-1}) compared to the other three. The final control peptide ARPRA, lacking a Leu at P4 as well as a Lys or Arg at P2, was not cleaved. Taken together, these results indicate a qualitative link between our library enrichment results derived using YESS *in vivo*, which identified a preponderance of Ali/Leu-X-Lys/Arg-Arg cleavage, and the substrate specificity profile of the Kex2 enzyme *in vitro*, which preferred the sequence containing Leu-X-Arg-Arg.

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510–518

In line with the knockout strain experiments described above, these *in vitro* peptide cleavage results add strong support to the hypothesis that Kex2 is the major endogenous protease in the yeast secretory pathway.

DISCUSSION

The potential of proteases to serve as targeted, catalytic biotherapeutics will critically depend on a highly restricted substrate preference for an intended protein or peptide sequence.[30] It is therefore essential that efficient and comprehensive substrate specificity profiling methods are developed to complement protease engineering platforms in order to characterize more fully any possible therapeutic candidates. The high throughput YESS protease engineering platform technology was combined with NGS and comparative sequence analysis. To the best of our knowledge, the YESS-NGS approach reported here is the first to combine the power of a high throughput substrate-sorting platform with NGS and comparative sequence analysis. Merging a substrate profiling technology with protease engineering in one yeast platform (YESS) not only is efficient but allows for protease cleavage analysis under physiological conditions within a yeast cell.

As a prelude to using the YESS-NGS technology to analyze an exogenous protease, endogenous proteolysis within the yeast secretory pathway was profiled. By comparing NGS results from the same five-residue randomized library before and after three rounds of sorting in the YESS platform, strong preferences for Ali/Leu-X-X-Arg as well as a related or perhaps different Lys/Arg-Arg were identified. The 20 most highly represented substrates in the sorted population, as well as two other substrates, VARRD and SPAKR (chosen to investigate dibasic cleavage without a Leu residue), were selected and examined in both EBY100 as well as EBY100^{Kex2-} strains, in which the Kex2 knockout strain was prepared specifically for this purpose. A perfect 15 out of 15 sequences with a Leu-X-X-Arg or Leu-X-Arg-Arg motif were cleaved in the EBY100 but not the EBY100^{Kex2-} strain. In addition, both of the known Kex2 substrates VARRD and to a lesser extent SPAKR were cleaved in the EBY100 but not the Kex2 knockout EBY100^{Kex2-} strain. Because the only known difference between the EBY100 and EBY100^{Kex2-} strains is the presence or absence of the Kex2 protease, respectively, the most straightforward explanation for these results is that Kex2 recognizes the Leu-X-X-Arg substrates as well as the Leu-X-Lys/Arg-Arg sequences. It is therefore proposed that Ali/Leu-X-Lys/Arg-Arg describes the specificity for Kex2 *in vivo* and that Kex2 is the major protease operating in the yeast secretory pathway.

We are not certain why five of the most common sequences isolated in the YESS sorting were highly represented in the NGS results yet were not cleaved in even the EBY100 strain. These five might somehow be favored during the library preparation, amplification, or sorting steps, or perhaps during NGS sample preparation or sequencing steps. No pattern in these sequences has been identified. It is also interesting to note the preponderance of Pro residues in the 20 most highly represented sequences of the selected pool (Table 1). One explanation for the observed Pro preference might be an unusually high abundance of Pro residues in the original substrate library (Table S1). However, a preponderence of Pro residues was not apparent in the overall NGS enrichment analysis (Figure 2 and Figure 3) presumably because the enrichment/de-enrichment analysis normalized against the unsorted library.

Peptide cleavage results *in vitro* also support the hypothesis that Kex2 is the major protease operating in the yeast secretory pathway. In particular, the preference for Ali/Leu-X-Lys/Arg-Arg seen in the YESS results was consistent at the qualitative level with Kex2 cleavage of five related peptide substrates *in vitro*. Kex2 is well-known to prefer substrates with Lys-Arg or Arg-Arg sequences at P2−P1.[13−15,29] Although previous analysis had revealed a basic or aliphatic residue specificity at P4,[13,29] we are unaware that the strong specific preference for Leu at P4 seen in our YESS results has been reported for Kex2. It is therefore gratifying that *in vitro*, ALARR, with both Leu at P4 and Arg-Arg at P2−P1, was cleaved the fastest by recombinant Kex2 out of the peptide substrates examined, and LRPRA, the other highly enriched peptide from our YESS study, was also cleaved well, confirming that Leu at P4 and Arg at P1 alone can confer strong activity (Figure 6S). Consistent with the importance of Leu at P4, the absence of a P4 Leu in the control AAARR peptide resulted in diminished activity (compare to ALARR), while removing the P4 Leu in the control ARPRA peptide (compare to LRPRA) resulted in no Kex2 cleavage activity.

Yeast cells have been widely used for recombinant protein production and engineering; however, proteolytic degradation of the recombinant protein of interest has been a perpetual problem.[31] The Ali/Leu-X-Lys/Arg-Arg preference identified here could be applied to develop computational models to predict the potential cleavage sites in the proteins being transported through the yeast secretory pathway. This information is particularly important for those using yeast display technology, as library members with Ali/Leu-X-Lys/Arg-Arg patterns are likely being removed from screens without the researcher's knowledge.

As an initial proof-of-principle for protease substrate specificity analysis, the YESS-NGS approach was used to evaluate the sequence specificity of the wild-type TEV-P and an engineered variant TEV-PE10 of the tobacco etch mosaic virus protease in EBY100. This method should be extendable to other recombinant or engineered proteases. Beyond just confirming the different specificities at P1 that were previously identified using individual peptide substrates, the data reported here verify that P1 preference represents the only observed significant difference in specificity between TEV-PE10 and TEV-P when the analysis was expanded to include three other key substrate residues.[28,32] This latter conclusion could only be reached with as much certainty following a thorough substrate specificity analysis such as that reported here.

A word of caution: one should not assume that the extent of enrichment or de-enrichment observed in these analyses has a strictly linear correlation to protease substrate preference. In theory, other factors beyond protease catalytic rates such as relative representation in the original library or concentration differences of different substrate sequences in the ER might influence the absolute amount of enrichment observed, making quantitative comparisons within our data unreliable. Nevertheless, the qualitative correspondence between the enrichment results and the peptide preferences observed with recombinant Kex2 cleavage analyzed *in vitro* is reassuring. In addition, a previous quantitative analysis with individual peptide substrates different only at P1 indicated TEV-P displayed a roughly 380-fold preference for Gln relative to Glu, while TEV-PE10 exhibited a roughly 13-fold preference for Glu relative to Gln. These measured values track in a relative way with the data in Figure 4 in which some enrichment of P1 Gln substrates is

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510−518

noted with TEV-E10 but no enrichment of P1 Glu substrates is seen with TEV-P. It should also be pointed out that unlike technologies such as TAILS, the YESS-NGS method could not provide direct cleavage site information in a specific substrate. Following the YESS-NGS analysis, substrate cleavage sites must be confirmed using alternate methods, although they can usually be inferred by evaluating the consensus.

Defining substrate specificity with greater precision will be increasingly necessary as engineered proteases are developed for more sophisticated applications including therapies.[30] Having a comprehensive substrate profiling capability within the YESS protease engineering platform will facilitate the rapid identification and full characterization of engineered proteases with desirable cleavage activities. The present study is meant to serve as a proof-of-principle, demonstrating the potential of the YESS-NGS approach. Future work, currently underway, will combine the Kex2 knockout EBY100^{Kex2-} strain with larger substrate libraries for a more comprehensive substrate specificity analysis of exogenous proteases.

METHODS

Protocols for vector construction, library construction, yeast cell sorting, and protease characterization are described in SI Methods.

ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acschembio.6b00547.

Experimental methods, amino acid distribution in the presorting library, detailed specificity scores of the endopeptidase cleaveOme, 95% confidence intervals for the specificity scores, primers used, yeast secretory pathway endopeptidase cleaveOme specificity profile sequencing results, Recombinant TEV protease specificity profile sequencing results, results for cleaveOme of yeast secretory pathway for the substrate library not containing the basic residues lysine and arginine, Evidence for the monoR and dibasic patterns recognized by the endogenous proteases, heat maps showing the cleaveOme specificity profile upon mutation, specificity profiles of engineered TEV-PE10 and wild-type TEV-P based on the substrate library being randomized at P1′, P1, P3, and P6, FACS analysis of cells with single substrates to validate the patterns observed in cleaveOme, and cleavage of peptide substrates by Kex2 under *in vitro* conditions (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: gg@che.utexas.edu.

*E-mail: iversonb@austin.utexas.edu.

ORCID

Li Yi: 0000-0002-7835-6985

Present Address

⊥Hubei Collaborative Innovation Center for Green Transformation of Bioresources, Hubei Key Laboratory of Industrial Biotechnology, Hubei University, Wuhan, China

Author Contributions

Q.L. and L.Y. contributed equally. Q.L., L.Y., G.G., and B.L.I. designed research; Q.L., L.Y., K.H.H., and P.M. performed research; Q.L., L.Y., K.H.H., G.G., and B.L.I. analyzed data; Q.L., L.Y., G.G., and B.L.I. wrote the paper.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Clayton Foundation (to B.L.I.) and National Institutes of Health Grant R01CA189623 (to G.G.)

REFERENCES

(1) Yi, L., Gebhard, M. C., Li, Q., Taft, J. M., Georgiou, G., and Iverson, B. L. (2013) Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A. 110*, 7229−7234.

(2) Yi, L., Taft, J. M., Li, Q., Gebhard, M. C., Georgiou, G., and Iverson, B. L. (2015) Yeast Endoplasmic Reticulum Sequestration Screening for the Engineering of Proteases from Libraries Expressed in Yeast. *Methods Mol. Biol. 1319*, 81−93.

(3) Girard, V., Dieryckx, C., Job, C., and Job, D. (2013) Secretomes: the fungal strike force. *Proteomics 13*, 597−608.

(4) Aridor, M., and Hannan, L. A. (2000) Traffic jam: a compendium of human diseases that affect intracellular transport processes. *Traffic 1*, 836−851.

(5) Aridor, M., and Hannan, L. A. (2002) Traffic jams II: an update of diseases of intracellular transport. *Traffic 3*, 781−790.

(6) Roebroek, A. J., Umans, L., Pauli, I. G., Robertson, E. J., van Leuven, F., Van de Ven, W. J., and Constam, D. B. (1998) Failure of ventral closure and axial rotation in embryos lacking the proprotein convertase Furin. *Development 125*, 4863−4876.

(7) Porro, D., and Mattanovich, D. (2004) Recombinant protein production in yeasts. *Methods Mol. Biol. 267*, 241−258.

(8) Sudbery, P. E. (1996) The expression of recombinant proteins in yeasts. *Curr. Opin. Biotechnol. 7*, 517−524.

(9) Paltridge, J. L., Belle, L., and Khew-Goodall, Y. (2013) The secretome in cancer progression. *Biochim. Biophys. Acta, Proteins Proteomics 1834*, 2233−2241.

(10) Seidah, N. G., and Prat, A. (2002) Precursor convertases in the secretory pathway, cytosol and extracellular milieu. *Essays Biochem. 38*, 79−94.

(11) Zhou, A., Webb, G., Zhu, X., and Steiner, D. F. (1999) Proteolytic processing in the secretory pathway. *J. Biol. Chem. 274*, 20745−20748.

(12) Beinfeld, M. C. (1998) Prohormone and proneuropeptide processing. Recent progress and future challenges. *Endocr. J. 8*, 1−5.

(13) Rockwell, N. C., and Fuller, R. S. (1998) Interplay between S1 and S4 subsites in Kex2 protease: Kex2 exhibits dual specificity for the P4 side chain. *Biochemistry 37*, 3386−3391.

(14) Rockwell, N. C., Wang, G. T., Krafft, G. A., and Fuller, R. S. (1997) Internally consistent libraries of fluorogenic substrates demonstrate that Kex2 protease specificity is generated by multiple mechanisms. *Biochemistry 36*, 1912−1917.

(15) Rozan, L., Krysan, D. J., Rockwell, N. C., and Fuller, R. S. (2004) Plasticity of extended subsites facilitates divergent substrate recognition by Kex2 and furin. *J. Biol. Chem. 279*, 35656−35663.

(16) Diamond, S. L. (2007) Methods for mapping protease specificity. *Curr. Opin. Chem. Biol. 11*, 46−51.

(17) Matthews, D. J., Goodman, L. J., Gorman, C. M., and Wells, J. A. (1994) A survey of furin substrate specificity using substrate phage display. *Protein Sci. 3*, 1197−1205.

(18) Scholle, M. D., Kriplani, U., Pabon, A., Sishtla, K., Glucksman, M. J., and Kay, B. K. (2006) Mapping protease substrates by using a biotinylated phage substrate library. *ChemBioChem 7*, 834−838.

(19) Boulware, K. T., and Daugherty, P. S. (2006) Protease specificity determination by using cellular libraries of peptide substrates (CLiPS). *Proc. Natl. Acad. Sci. U. S. A. 103*, 7583−7588.

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510−518

(20) O'Donoghue, A. J., Eroy-Reveles, A. A., Knudsen, G. M., Ingram, J., Zhou, M., Statnekov, J. B., Greninger, A. L., Hostetter, D. R., Qu, G., Maltby, D. A., Anderson, M. O., Derisi, J. L., McKerrow, J. H., Burlingame, A. L., and Craik, C. S. (2012) Global identification of peptidase specificity by multiplex substrate profiling. *Nat. Methods 9*, 1095−1100.
(21) Dix, M. M., Simon, G. M., and Cravatt, B. F. (2008) Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell 134*, 679−691.
(22) Varadarajan, N., Pogson, M., Georgiou, G., and Iverson, B. L. (2009) Proteases that can distinguish among different post−translational forms of tyrosine engineered using multicolor flow cytometry. *J. Am. Chem. Soc. 131*, 18186−18190.
(23) Varadarajan, N., Gam, J., Olsen, M. J., Georgiou, G., and Iverson, B. L. (2005) Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity. *Proc. Natl. Acad. Sci. U. S. A. 102*, 6855−6860.
(24) Li, H. X., Hwang, B. Y., Laxmikanthan, G., Blaber, S. I., Blaber, M., Golubkov, P. A., Ren, P., Iverson, B. L., and Georgiou, G. (2008) Substrate specificity of human kallikreins 1 and 6 determined by phage display. *Protein Sci. 17*, 664−672.
(25) Kleifeld, O., Doucet, A., auf dem Keller, U., Prudova, A., Schilling, O., Kainthan, R. K., Starr, A. E., Foster, L. J., Kizhakkedathu, J. N., and Overall, C. M. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N−termini and protease cleavage products. *Nat. Biotechnol. 28*, 281−288.
(26) Bostian, K. A., Elliott, Q., Bussey, H., Bum, V., Smith, A., and Tipper, D. J. (1984) Sequence of the preprotoxin dsRNA gene of type I killer yeast: multiple processing events produce a two-component toxin. *Cell 36*, 741−751.
(27) *MEROPS database*, http://merops.sanger.ac.uk/.
(28) Dougherty, W. G., Carrington, J. C., Cary, S. M., and Parks, T. D. (1988) Biochemical and mutational analysis of a plant virus polyprotein cleavage site. *EMBO J. 7*, 1281−1287.
(29) Bevan, A., Brenner, C., and Fuller, R. S. (1998) Quantitative assessment of enzyme specificity in vivo: P2 recognition by Kex2 protease defined in a genetic system. *Proc. Natl. Acad. Sci. U. S. A. 95*, 10384−10389.
(30) Li, Q., Yi, L., Marek, P., and Iverson, B. L. (2013) Commercial proteases: present and future. *FEBS Lett. 587*, 1155−1163.
(31) Sinha, J., Plantz, B. A., Inan, M., and Meagher, M. M. (2005) Causes of proteolytic degradation of secreted recombinant proteins produced in methylotrophic yeast Pichia pastoris: case study with recombinant ovine interferon−tau. *Biotechnol. Bioeng. 89*, 102−112.
(32) Phan, J., Zdanov, A., Evdokimov, A. G., Tropea, J. E., Peters, H. K., III, Kapust, R. B., Li, M., Wlodawer, A., and Waugh, D. S. (2002) Structural basis for the substrate specificity of tobacco etch virus protease. *J. Biol. Chem. 277*, 50564−50572.

DOI: 10.1021/acschembio.6b00547
ACS Chem. Biol. 2017, 12, 510−518