Supplemental Figure 1. The construction and characterization of Lib 10AA for identifying protease substrate specificity. (a) Kunkel mutagenesis was used to construct Lib 10AA, that displays randomized 10AA peptides encoded by NNK codons. N encodes A, T, C, G while K encodes T and G, thus encoding 32 different tri-nucleotides encompassing all 20 amino acids. (b) The amino acid abundancy of Lib 10AA assayed by NGS closely matches the theoretical values (3.33%, 6.66% and 9.99%, depending on amino acid code degeneracy). (c) The amino acid composition established by NGS is calculated for each of the 10 positions in Lib 10AA and found to closely match the synthetic oligonucleotide input. (d, e) The nucleotide abundancy of Lib 10AA assayed by NGS closely matches the theoret Lib 10AA assayed by NGS closely matches the theoret Lib 10AA assayed by NGS closely matches the synthetic oligonucleotide input. (d, e) The nucleotide abundancy of Lib 10AA assayed by NGS closely matches the theoret Lib 10AA assayed by NGS closely matches the theoret Lib 10AA assayed by NGS closely matches the theoret Lib 10AA and found to closely match the synthetic oligonucleotide input. (d, e) The nucleotide abundancy of Lib 10AA assayed by NGS closely matches the theoretical values and NNK pattern.



Supplemental Figure 2. The construction and characterization of Lib hP for identifying protease substrate specificity. (a) Construction of the Lib hP, that displays tiled peptides covering the human proteome in 49AA blocks. (b) Assessment of the library quality before cloning and after library generation of Lib hP by NGS shows high recovery of sequences. (c, d) The amino acid composition calculated for each of the 49 positions in Lib hP is representative of the composition found in the human proteome. (e) Protein expression gene coverage of Lib hP, a single global proteomics experiment¹, and a single RNA-seq experiment^{2, 3}. f) Minimum numbers of cell lines that are required to achieve different gene coverage in RNA-seq. Data analysis was based on RNA-seq of 1556 cell lines (NIH RNA-seq of 934 human cancer cell lines from the Cancer Cell Line Encyclopedia² and Genentech RNA-seq of 675 commonly used human cancer cell lines³).



Supplemental Figure 3. Phagemid DNA and protein sequences used for phage library

construction. Pink: stll signaling peptide, yellow: avi-tag, black: linker, red: 10mer or 49mer, blue:

truncated pIII coat protein

Lib 10AA (for Caspases)

NT sequence:

TGAAAAAGAATATCGCATTTCTTCTTGCATCTATGTTCGTTTTTTCTATTGCTACAAATGCCTATGCA<mark>GGCCTGAACG</mark> ATATTTTTGAAGCGCAGAAAATTGAATGGCATGAANNKNNKNNKNNKNNKNNKNNKNNKNNKGAGAATCTGTACT TTCAGAGCGGCGGATCCGATTTTGATTATGAAAAGATGGCAAACGCTAATAAGGGGGGCTATGACCGAAAATGCCGAT GAAAACGCGCTACAGTCTGACGCTAAAGGCAAACTTGATTCTGTCGCTACTGATTACGGTGCTGCTATCGATGGTTTC ATTGGTGACGTTTCCGGCCTTGCTAATGGTAATGGTGCTACTGGTGATTTTGCTGGCTCTAATTCCCAAATGGCTCAA **GCGTTTCTTTATATGTTGCCACCTTTATGTATGTATTTTCTACGTTTGCTAACATACTGCGTAATAAGGAGTCTTAA**

AA sequence:

MKKNIAFLLASMFVFSIATNAYAGLNDIFEAQKIEWHE???????ENLYFQSGGSDFDYEKMANANKGAMTENADENAL QSDAKGKLDSVATDYGAAIDGFIGDVSGLANGNGATGDFAGSNSQMAQVGDGDNSPLMNNFRQYLPSLPQSVECRPFV FGAGKPYEFSIDCDKINLFRGVFAFLLYVATFMYVFSTFANILRNKES*

Lib 10AA (for ADAMs)

NT sequence:

ATATTTTTGAAGCGCAGAAAATTGAATGGCATGAA**TCTGGTGGTGGT**NNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK **GGAGGCGGATCC**GATTTTGATTATGAAAAGATGGCAAACGCTAATAAGGGGGCTATGACCGAAAATGCCGATGAAAA CGCGCTACAGTCTGACGCTAAAGGCAAACTTGATTCTGTCGCTACTGATTACGGTGCTGCTATCGATGGTTTCATTGG TGACGTTTCCGGCCTTGCTAATGGTAATGGTGCTACTGGTGATTTTGCTGGCTCTAATTCCCAAATGGCTCAAGTCGG TCTTTTATATGTTGCCACCTTTATGTATGTATTTTCTACGTTTGCTAACATACTGCGTAATAAGGAGTCTTAA

AA sequence:

MKKNIAFLLASMFVFSIATNAYAGLNDIFEAQKIEWHESGGG???????GGGSDFDYEKMANANKGAMTENADENALQS DAKGKLDSVATDYGAAIDGFIGDVSGLANGNGATGDFAGSNSQMAQVGDGDNSPLMNNFRQYLPSLPQSVECRPFVFG AGKPYEFSIDCDKINLFRGVFAFLLYVATFMYVFSTFANILRNKES*

Lib hP

NT sequence:

CGATGATAAGAAGCTTGATTTTGATTATGAAAAGATGGCAAACGCTAATAAGGGGGGCTATGACCGAAAATGCCGATGA AAACGCGCTACAGTCTGACGCTAAAGGCAAACTTGATTCTGTCGCTACTGATTACGGTGCTGCTATCGATGGTTTCAT TGGTGACGTTTCCGGCCTTGCTAATGGTAATGGTGCTACTGGTGATTTTGCTGGCTCTAATTCCCAAATGGCTCAAGT **GTTTCTTTTATATGTTGCCACCTTTATGTATGTATTTTCTACGTTTGCTAACATACTGCGTAATAAGGAGTCTTAA**

AA sequence:

?DYKDDDDKKLDFDYEKMANANKGAMTENADENALQSDAKGKLDSVATDYGAAIDGFIGDVSGLANGNGATG DFAGSNSQMAQVGDGDNSPLMNNFRQYLPSLPQSVECRPFVFGAGKPYEFSIDCDKINLFRGVFAFLLYVATFMYVFSTF ANILRNKES*

Supplemental Figure 4. Biotinylation of the library before screening. (a) The experiment setup of phage ELISA for quantifying biotinylation level. (b) Phage ELISA confirmed that peptides of Lib 10AA and Lib hP were successfully displayed. The entire library (1 mL) was biotinylated by 5 μ L, 10 mg/mL BirA in buffer with 100 μ M D-biotin. (c) The library was first biotionylated with BirA in 100 μ M D-biotin before Round 1 selection, while for Round 2&3 phage selection, biotinylation was done during phage amplification using engineered XL1-Blue (BirA) strain. (d) In cell biotinylation was doable although not 100% complete.



Supplemental Figure 5. Phage titer after three rounds of selection show significant enrichment in Round 2 with 1µM or 100 nM Caspase-3 treatment. The number of released phage was determined by counting infectious units as a function of round of selection.



Round 2

Round 3

Supplemental Figure 6. Heatmap reveals the amino acid composition change compared to the input library after three rounds of selection.



Supplemental Figure 7. (a) Overall amino acid composition of the input library (Lib 10AA) and the outputs of Round 1, 2, 3. (b) Crystal structure of caspase-3 with DEVD substrate binding to the pocket (adapted from PDB: 3PD1). (c) The sequence logo generated based on MSA using DECIPHER package⁴. There's a clear DEVDGG motif showing up. We extracted this motif and defined P4-P2' based on existing knowledge. (d, e) The heatmap for caspase-3 substrates generated based on top 20,000 sequences identified from screening Lib 10AA (Round 3, d) and from MEROPS dataset (e).



Supplemental Figure 8. Generation of the PSSM. To illustrate the process, we use caspase-3 as an example. (a) Top 20,000 potential Caspase-3 substrates identified from Lib 10AA (Round 3) were aligned. (b) To more accurately reflect the characteristics at each position, a matrix that contains the number of observed amino acids at each position is calculated, known as PFM. (c) A PPM was created based on PFM. (d) The PPM was converted to a PWM using a formula that converts it to a log-scale. PWMs are also known as position specific scoring matrices (PSSM). (e) The calculation of a score of an example peptide. (f) Identification of potential cleavage site(s).

а

Aligning top ~20,000 Caspase-3 substrates identified from Lib 10AA

	P4	P3	P2	P1	P1'	P2'
substrate 1	D	G	V	D	G	А
substrate 2	D	W	Р	D	н	G
substrate 3	D	L	V	D	F	G
substrate 4	D	V	F	D	G	L
substrate 5	D	т	м	D	Y	S
substrate 19,999	D	М	V	D	E	G
substrate 20.000	D	D	м	D	G	L

Position frequency matrix (PFM)

	А	С	D	Е	F	G	н	1	к	L	М	Ν	Р	Q	R	s	т	V	W	Y
P4	388	1187	9050	854	67	149	306	101	8	261	288	297	120	147	50	2358	1646	543	26	83
P3	1202	1083	669	2271	881	714	656	266	170	1337	753	329	20	1151	729	1976	1010	1535	473	704
P2	770	902	6	12	360	256	320	876	54	1829	424	151	1938	151	1076	495	2478	5454	150	227
P1	9	7	17655	109	1	14	10	2	10	3	1	21	12	9	24	26	7	5	1	3
P1'	1425	735	180	116	399	7976	487	94	117	260	146	289	35	132	551	3216	676	256	346	493
P2'	1718	902	360	310	387	3239	419	381	345	938	394	456	1320	395	1374	2048	912	1285	332	414
										ł		РРМ	(N) =	$\frac{C_N}{\sum C}$						

С

d

е

b

Position Probability matrix (PPM %) A C D E F G H I K L M N P Q R S T V W Y P4 **2.16** 6.62 **50.48** 4.76 **0.37** 0.83 **1.71** 0.56 **0.04** 1.46 **1.61** 1.66 **0.67** 0.82 **0.28** 13.15 **9.18** 3.03 **0.15** 0.46 **P**3 6.70 6.04 3.73 12.67 4.91 3.98 3.66 1.48 0.95 7.46 4.20 1.84 0.11 6.42 4.07 11.02 5.63 8.56 2.64 3.93 P2 **4.29** 5.03 **0.03** 0.07 **2.01** 1.43 **1.78** 4.89 **0.30** 10.20 **2.36** 0.84 **10.81** 0.84 **6.00** 2.76 **13.82** 30.42 **0.84** 1.27 P1 0.05 0.04 98.47 0.61 0.01 0.08 0.06 0.01 0.06 0.02 0.01 0.12 0.07 0.05 0.13 0.15 0.04 0.03 0.01 0.02 P1' 7.95 4.10 1.00 0.65 2.23 44.49 2.72 0.52 0.65 1.45 0.81 1.61 0.20 0.74 3.07 17.94 3.77 1.43 1.93 2.75 P2' 9.58 5.03 2.01 1.73 2.16 18.07 2.34 2.13 1.92 5.23 2.20 2.54 7.36 2.20 7.66 11.42 5.09 7.17 1.85 2.31

				F	Posi	tior	n W	eigl	ht n	hatr	'ix (I	PW	<i>м</i> = М) с	= log or P	2 	PPM B	(C _N	$\frac{D}{2}$		
P4	-1.2	0.4	3.3	-0.1	-3.7	-2.6	-1.6	-3.1	-6.8	-1.8	-1.6	-1.6	-2.9	-2.6	-4.2	1.4	0.9	-0.7	-5.1	-3.4
P 3	0.4	0.3	-0.4	1.3	0	-0.3	-0.5	-1.8	-2.4	0.6	-0.3	-1.4	-5.5	0.4	-0.3	1.1	0.2	0.8	-0.9	-0.3
P2	-0.2	0	-7.2	-6.2	-1.3	-1.8	-1.5	0	-4.1	1	-1.1	-2.6	1.1	-2.6	0.3	-0.9	1.5	2.6	-2.6	-2
P1	-6.6	-7	4.3	-3	-9.8	-6	-6.5	-8.8	-6.5	-8.2	-9.8	-5.4	-6.2	-6.6	-5.2	-5.1	-7	-7.5	-9.8	-8.2
P1'	0.7	-0.3	-2.3	-3	-1.2	3.2	-0.9	-3.3	-2.9	-1.8	-2.6	-1.6	-4.7	-2.8	-0.7	1.8	-0.4	-1.8	-1.4	-0.9
P2'	0.9	0	-1.3	-1.5	-1.2	1.9	-1.1	-1.2	-1.4	0.1	-1.2	-1	0.6	-1.2	0.6	1.2	0	0.5	-1.4	-1.1
	Α	С	D	Е	F	G	Н	Ι	Κ	L	М	Ν	Ρ	Q	R	S	Т	V	W	Υ

C			the	e١	/al	ue	of	Α	fo	r th	ne	ре	pti	ide	D	θE	VD)G	G	i								
			А	С	D	E	FG	н	1	к	L	м	NF	- a	R	s	Т	v	w	Y						<u>.</u>		
		P4	0	0	1	0	0 0	0	0	0	0	0	0 0	0 0	0	0	0	0	0	0						3.3		
		Р3	0	0	0	1	0 0	0	0	0	0	0	0 0	0 0	0	0	0	0	0	0						+1.3		
		P2	0	0	0	0	0 0	0	0	0	0	0	0 0	0 0	0	0	0	1	0	0	X		PSSM			+2.6	=	16.6
		P1	0	0	1	0	0 0	0	0	0	0	0	0 0	0 0	0	0	0	0	0	0	· ^	•	1 000			+4.3		10.0
		P1'	0	0	0	0	0 1	0	0	0	0	0	0 0	0 0	0	0	0	0	0	0						+3.2		
		P2'	0	0	0	0	0 1	0	0	0	0	0	0 0	0 0	0	0	0	0	0	0						+1.9		
f	10AA sequen	ICe)				p	os	sik	ole	6	me	ər			s	cc	ore	es	0	f each 6	in	ner	m	naximur	n score		potential cleavage site(s)
	EFSDEVDGG	s					E	F F		E E	/ /D										0.8 -15.9				16.6	6		EFSDEV

EVDGGS

-13.1

16.6

-8.1

Supplemental Figure 9. (a) The Venn diagram of unique peptides identified from Lib hP that decrease due to increasing enrichment with each of three rounds of the selection. Sequences with low counts (i.e., <3 counts) being pre-filtered out increases the overlap of the sequences in three rounds. (b) Roughly, the 49mer with more cleavage sites on it got enriched faster. NGS count was affected by many other factors, like PCR, phage amplification by E coli and whether the cleavage site is a good substrate. (c) The distribution of three different structures throughout the whole input library. The GOR (Garnier-Osguthorpe-Robson) method was used for 2nd structure prediction. The secondary structure throughout the library is comparably even but the calculation result of the C- and N-terminal of 49mer is affected by the constant flanking regions.



Supplemental Figure 10. (a) The protein sequence of WARS. Two caspase-3 cleavage sites identified in substrate phage screening were highlighted in pink and blue. (b) Representative ESI mass spectra of the intact protein and the fragments after caspase-3 proteolysis.

а

b

MADMSNGEQGCGSPLELFHSIAAQGELVRDLKARNAAKDEIDSAVKMLLSLKTSYKAATGEDYKVDCPPGDPAPESGEGLDA TEADEDFVDPWTVQTSSAKGIDYDKLIVRFGSSKIDKELVNRIERATGQRPHRFLRRGIFFSHRDMHQILDAYENKKPFYLYTGR GPSSEAMHVGHLIPFIFTKWLQDVFNVPLVIQMTDDEKYLWKDLTLDQAYGYAVENAKDIIACGFDINKTFIFSDLDYMGMSP GFYKNVVKIQKHVTFNQVKGIFGFTDSDCIGKISFPAIQAAPSFSNSFPQIFRDRTDVQCLIPCAIDQDPYFRMTRDVAPRIGYPK PALLHSTFFPALQGAQTKMSASDPNSSIFLTDTAKQIKTKVNKHAFSGGRDTVEEHRQFGGNCDVDVSFMYLTFFLEDDDKLE QIRRDYTSGAMLTGELKKELIEVLQPLIAEHQARRKEVTDEIVKEFMTPRKLSYDFQLEHHHHHH



Supplemental Table 1. Examples of known caspase-3 substrates identified by SPD-NGS	

Substrate	Cleavage site	PSSM score	PDB	Solvent Accessibility (Acc)	Acc_ sum	seondary structure
presenilin 1	AQRDSH	4.5	2KR6	74 122 220 126 64 191	797	SSSSLS
TAR DNA-binding protein 43	DETDAS	12.3	5MRG	122 170 116 165 94 118	785	LLLLL
BH3-interacting domain death agonist	LQTDGN	6.5	2BID	144 171 115 136 58 133	757	SLLLLS
poly [ADP-ribose] polymerase 1	DEVDGV	15.3	2CS2	148 160 109 166 33 131	747	LLLLL
interleukin-18	DMTDSD	9.4	1J0S	147 63 164 140 155 30	699	STTTTL
RAC-alpha serine/threonine- protein kinase	EEMDFR	3.9	3QKK	132 121 70 115 49 175	662	TTLLLS
TNF receptor-associated factor 3	EEADSM	6.0	1FLK	173 88 60 96 171 74	662	ТТТТТТ
caspase-3 precursor	NSVDSK	6.9	1130	12 122 25 227 181 71	638	EELLLL
ataxin-3	DLPDCE	7.5	2KLZ	119 103 111 88 97 110	628	нннннн
receptor tyrosine-protein kinase erbB-2	DVFDGD	8.9	10VC	146 75 68 133 102 86	610	SLSSLL
catenin beta-1 (Beta-catenin)	YPVDGL	1.2	2G57	72 115 79 106 180 54	606	TSTTSL
polypyrimidine tract-binding protein 1	AAVDAG	8.6	2AD9	22 160 83 153 61 74	553	LLLLL
microtubule-associated protein tau	DMVDSP	12.4	5DMG	106 211 116 42 69 6	550	LLLLL
Bcl-2-like protein 1	SSLDAR	9.1	1LXL	129 84 73 99 61 103	549	SLLLLL
prothymosin alpha	DDVDTK	8.0	2L9I	89 100 92 150 60 41	532	GGGTLH
Transcription intermediary factor 1-beta	DGADST	9.0	2RO1	125 61 90 71 46 115	508	SSSSSS
T-cell surface glycoprotein CD3 zeta chain precursor	YLLDGI	4.4	2HAC	190 97 69 22 24 102	504	нннннн
Actin, cytoplasmic 2	LVIDNG	3.5	5JLH	58 168 118 117 21 7	489	LLLSLE
PC4 and SFRS1-interacting protein	DAQDGN	7.7	2M16	172 56 11 98 76 57	470	TTTTSL
Multifunctional protein ADE2	ELLDSP	8.2	2H31	41 20 94 144 52 117	468	LLTTSL
paxillin	SLLDEL	4.4	2VZD	157 95 36 131 44 3	466	LHHHHH
elongation factor 1-beta	DETDMA	8.8	1B64	107 128 3 94 47 70	449	TTSLHH
retinoblastoma-associated protein	DEADGS	13.1	2QDJ	75 101 111 24 34 103	448	LННННН
SLC4A1	EQGDGG	7.8	1HYN	154 51 33 105 14 86	443	LLSSLS
cellular tumor antigen p53	SDSDGL	7.6	1GZH	90 36 83 73 64 93	439	SLLLSS
flavin reductase	DEYDGH	9.1	1HDO	119 99 30 30 70 88	436	STTTTL
dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex, mitochondrial	IETDKA	2.0	1FYC	0 63 15 141 167 22	408	EELSSL

protein farnesyltransferase/geranylge ranyltransferase type-1 subunit alpha	VSLDSP	8.1	1JCQ	34 54 10 116 39 143	396	LLTTST
cell division control protein 42 homolog	DLRDDP	6.7	1A4R	32 39 49 125 47 83	375	GGGGLH
protein kinase N2	DEVDSL	13.5	4CRS	105 74 8 127 46 6	366	тіннн
tubulin beta chain	ELVDSV	9.8	5N5N	155 19 13 92 64 0	343	нннннн
26S proteasome non-ATPase regulatory subunit 12	MEVDYS	6.9	5GJQ	52 64 61 17 56 68	318	нннннн
coproporphyrinogen III oxidase, mitochondrial precursor	DDLDSP	10.6	2AEX	7 30 25 88 70 98	318	EEELLS
tubulin beta-2B chain	ELVDSV	9.8	6E7C	118 11 7 107 64 2	309	ттнннн
tubulin beta-3 chain	ELVDSV	9.8	5IJ0	153 19 3 91 40 0	306	нннннн
vimentin	CEVDAL	9.4	1GK4	80 89 51 61 3 12	296	LHHHHH
apoptotic chromatin condensation inducer in the nucleus	DELDYH	8.0	6G6S	105 27 0 27 91 39	289	нннннн
caspase-6 precursor	TEVDAA	10.7	3NR2	66 85 58 7 44 28	288	EEEEL
heterogeneous nuclear ribonucleoproteins A2/B1	AEVDAA	8.6	1X4B	62 53 4 104 45 2	270	ннннн
cyclin-dependent kinase inhibitor 1	DHVDLS	9.2	1AXC	26 0 45 105 67 20	263	LGGGTS
bifunctional methylenetetrahydrofolate dehydrogenase/cyclohydrola se, mitochondrial	DNVDGL	12.0	1ZN4	127 109 7 8 0 2	253	TTLSLE
Serine/arginine-rich splicing factor 1	DLKDHM	2.1	2M7S	36 0 88 81 34 4	243	ннннн
apoptotic protease-activating factor 1	SVTDSV	10.3	1Z6T	48 2 12 120 49 1	232	GGGTTL
DNA repair protein RAD51 homolog 1	DVLDNV	8.3	1N0W	92 2 22 57 50 0	223	ннннте
vav proto-oncogene	DQIDDT	5.7	3KY9	50 10 9 101 37 15	222	TGGGGS
serine-protein kinase ATM	DIVDGM	10.5	5NP0	36 48 18 50 16 22	190	HHHHHL
E3 ubiquitin-protein ligase NEDD4	GQVDVP	3.4	3B7Y	0 46 5 57 11 65	184	EEEEE
myosin light chain 3	DFVEGL	6.1	5TBY	64 1 49 37 7 23	181	нннннн
DNA-dependent protein kinase catalytic subunit	GDSDGP	4.1	5LUQ	15 12 35 58 28 3	151	нннннн
CCNE1	LDVDCL	4.5	1W98	1 45 7 71 0 6	130	HSGGGG

Supplemental Figure 11-14. Using SPD-NGS to identify protease specificity for caspases-2, -6, -7, and -8. (a) Heatmap reveals the amino acid composition change compared to the input library after three rounds of selection. (b) The Venn diagram of unique peptides identified from Lib 10AA as a function of rounds of selection. (c) The Sequence Logo (Probability) for caspase substrates generated by aligning to top ~20,000 sequences identified from screening Lib 10AA (Round 3). (d) A similar motif is obtained by selecting only proteolytic sites with a glutamate (E) at P1 position as seen previously by Nterminomics data of caspase-6. (e) The PSSM of caspase substrates is generated according to the alignment. (f) Violin plot of the maximum scores of 10AA peptides in input library, Round 1, 2, 3 outputs and the scores of peptides in MEROPS database. (g) The plot of maximum score vs NGS RPM for peptides in input library and Round 1, 2, 3 outputs. The peptides with higher scores get enriched faster. (h) The Venn diagram of unique peptides identified from Lib hP decreases with each of three rounds of the selection as enrichment increases as was seen for Lib 10AA. (i) The Sequence Logo (Probability) of caspase substrate consensus generated by all cleavage events identified from Lib hP. (j) Violin plot of the maximum scores of the Lib hP input library, and progressive enrichment of substrates as one progresses from outputs of Round 1, 2, and 3. (k) The plot of maximum score vs NGS RPM for peptides in input library and Round 1, 2, and 3 outputs. (I) Caspase has a known structural preference for cutting loop>helix>sheets and this matches structural bioinformatics for sites identified from Lib hP. GOR method was used for 2nd structure prediction. (m) Distribution of frequency observed as a function of number of cuts in the 49AA peptides show most have one cut but some have multiple that decays monotonically.





Supplemental Figure 12. SPD-NGS to identify protease specificity for Caspase-6.



Supplemental Figure 13. SPD-NGS to identify protease specificity for Caspase-7.



Supplemental Figure 14. SPD-NGS to identify protease specificity for Caspase-8.

Supplemental Figure 15. (a) Gel electrophoresis of active recombinant caspases-2, -3, -6, -7 and -8 purified from E. coli. b. (b) The activity of purified active recombinant caspases-2, -3, -6, -7 and -8 in caspase activity buffer. Ac-DEVD-R110 was used to monitor the caspase activity.



Supplemental Figure 16. Cluster the heat maps of each caspase.



Supplemental Figure 17. (a) Gel electrophoresis of active recombinant active ADAM 10 and ADAM 17 expressed by Expi293 cells. (b) The activity of purified active recombinant ADAM10 and ADAM 17 in activity buffer. Mca-KPLGL-Dpa-AR-NH2 was used to monitor the ADAMs activity.



Supplemental Figure 18. SPD-NGS to identify protease specificity for ADAM 17. (a) Heatmap reveals the amino acid composition change compared to the input library after three rounds of selection. (b) The Sequence Logo (Probability) for ADAM17 substrates generated by substrate sequences recorded in MEROPS database. (c) We cannot rule out the possibility that ADAMs family proteases may cleave some multipass transmembrane proteins.





Multipass membrane protein

Supplemental Figure 19. SPD-NGS to identify protease specificity for ADAM 10. (a) Heatmap reveals the amino acid composition change compared to the input library after three rounds of selection. (b) The Venn diagram of unique peptides identified from Lib 10AA as a function of rounds of selection. (c) The Sequence Logo (Bits) for ADAM10 substrates generated by aligning the top ~20,000 sequences identified from screening Lib 10AA (Round 3). (d) The Sequence Logo (Bits) for ADAM10 substrates recorded in MEROPS database. (e) The PSSM of ADAM10 substrates is generated according to the alignment. (f) Violin plot of the maximum scores of 10AA peptides in input library and Round 1, 2, 3 outputs. (g) The plot of maximum score vs NGS RPM for peptides in input library and Round 1, 2, 3 outputs. The peptides with higher scores get enriched faster. (h) The Venn diagram of unique peptides identified from Lib hP decreases with each of three rounds of the selection as enrichment increases as was seen for Lib 10AA. (i) The Sequence Logo (Bits) of ADAM10 substrate consensus generated by all cleavage events identified from Lib hP. (j) Violin plot of the maximum scores of the Lib hP input library, and progressive enrichment of substrates as one progresses from outputs of Round 1, 2, and 3. (k) The plot of maximum score vs NGS RPM for peptides in input library and Round 1, 2, and 3 outputs.



Supplemental Figure 20. The Sequence Logo (Probability) for caspase substrates generated with the alignment of top ~ 20,000 sequences enriched from SPD-NGS with Lib 10AA. Each of the sequence logo is generated by fixing the top abundant residues at different positions. The number above the sequence logo indicates the number of the sequences used for generating the logo.



Supplemental Figure 21. Primer design for NGS



Supplemental Table 2. Protein expression conditions

Protein	IPTG (mM)	Temperature (°C)	Time (hour)
∆CARD-caspase-2	0.2	30	4-6
caspase-3	0.2	30	4-6
caspase-6	0.05	30	16
caspase-7	0.2	30	16
∆DED-caspase-8	0.4	30	3-5
WARS	0.2	18	16

Protein expression and purification

ΔCARD-caspase-2, caspase-3, 6, 7, ΔDED-caspase-8 and WARS were cloned into His6-affinity tag containing vector pET23b. The active enzymes were expressed in *E. coli* BL21 (DE3) pLysS cells (Promega, # L1195) and WARS was expressed in *E. coli* strain BL21(DE3). Cells were grown in 2xYT media supplemented with 200 µg/mL ampicillin and 50 µg/mL chloramphenicol (for pLysS only) at 37 °C to an OD600nm at approximately ~0.6–0.8. Expression was then induced for the caspases and WARS with IPTG (details of concentration, temperature, and duration are available in **Supplemental Table 2**). Cells were harvested by centrifugation and lysed by sonication. The cell lysates were clarified by centrifugation, and the soluble protein fractions were purified by Ni-NTA resin (Qiagen, # 30230). The vector used to express ADAM10, 17 ECD-10His (extracellular domain) was generated by Gibson cloning and adapted from the commercially available pFUSE-hIgG1-Fc (InvivoGen) vector. Each ECD was subcloned between an N-terminal IL2 signal sequence and a C-terminal 10His-tag replacing the original Fc domain. Suspension Expi293 cells were grown to 2.5 million cell density and transiently transfected with ADAMs protease expression vectors using ExpiFectamine[™] 293 Transfection Kit (Thermo Fisher Scientific, # A14525). Medium was collected after 5-7 days and protein was purified by Ni-NTA chromatography.

Phagemid construction.

The pKM128 vector (phagemid) from Kurt Mou was modified to parental vectors, PhMd0007 and PhMd0003, for Lib10AA and Lib hP, respectively. An Avi tag was first inserted into the phagemid after a PhoA promoter, followed by a TAA stop codon, an EcoRI restriction site, a TEV cleavage site and the truncated gIII protein for PhMd0007 (Lib 10AA). For Lib hP, PhMd0003 has an Avi tag, followed by an EcoRI restriction site, a BamHI restriction site, a TAA stop codon, a HindIII restriction site, followed by the truncated gIII. All oligonucleotides for cloning were from Integrated DNA Technologies (IDT).

XL-1Blue BirA strain construction

To facilitate the screening, we have engineered the XL-1 Blue cells to express the biotin ligase BirA with pBirAcm, an engineered pACYC184 plasmid with an IPTG inducible birA gene. Briefly, competent XL-1Blue cells were transformed with pACYC184 plasmid harboring chloramphenicol resistance. On the day prior to round 1 selection, a single colony of XL-1 Blue BirA was inoculated into 100 µL 2xYT. The 100 µL

28

culture was then split into 3 2-mL 2xYT cultures containing tetracycline (5 μ L/mL)/chloramphenicol, carbenicillin (50 μ L/mL), or kanomycin (25 μ L/mL) in a 12mL falcon E coli culture tubes. These tubes were incubated in a 37°C, 250 RMP shaking incubator overnight to allow tetracycline/chloramphenicol resistant clones to grow out.

Construction of randomized 10AA synthetic substrate phage library (Lib 10AA)

The 10AA randomized sequences in Lib 10AA were constructed using synthetic DNA containing 10 NNK degenerate codons (where N=A/G/C/T and K=T/G) encoding all 20 amino acids at each of the 10 positions (**Supplemental Figure 1**). The synthetic DNA was incorporated into the phagemid template by Kunkel mutagenesis (**Supplemental Figure 1a**).^{5, 6} The theoretical amino acid sequence diversity of Lib 10AA is ~10¹³ but the true diversity was titered after transfection at ~10⁹. NGS sequencing of the unselected library confirmed the broad representation of sequences in the library and revealed an amino acid or nucleotide distribution very close to the theoretical values based on the input synthetic DNA (**Supplemental Figure 1b, 1d**). The NGS also showed the amino acid and nucleotide diversity was uniform across all 10 positions as programmed by the synthetic DNA (**Supplemental Figure 1c, 1e**).

Isolating ssDNA of phagemid

The parental phagemid was first constructed in which an avi-tag was linked directly to a sequence including an EcoRI restriction site, a TAA stop codon, followed by a truncated M13 gene III protein. After transformation, a single colony of *E. coli* CJ236 (dut/ung⁻) was picked and grown in 1 mL 2XYT supplemented with appropriate antibiotics. After the culture reached O.D. 0.6, M13K07 helper phage (10¹⁰ PFU ml⁻¹) was added and the culture was incubated at 37 °C, 250 RPM. After 1 hour, the culture was transferred into 500 mL 2XYT supplemented with appropriate antibiotics, uridine and grown overnight. The next day, phage particles were precipitated with 1/5 volume of PEG/NaCI solution and collected by centrifugation. The phage pellet was resuspended in PBS buffer and ssDNA was purified using a MiniPrep column. Briefly, phage solution was applied to MiniPrep column to bound the phage to the column matrix and 0.7 mL of buffer MLB to the column twice to lyse the phage particles. The ssDNA on the column was washed by 0.7 mL PE buffer twice and eluted by ddH₂O.

Kunkel mutagenesis to generate heteroduplex ccc-dsDNA

29

The first step in the synthesis of heteroduplex CCC-dsDNA is the phosphorylation of the mutagenic oligonucleotide. The mutagenic oligonucleotide was 5'-phosphorylated to enable ligation by T4 DNA ligase at 37 °C for 1 hour.

Component	Volume (µL)	Final Concentration
Oligonucleotide (100 uM)	3	9 ug
10 X PNK Reaction Buffer	5	1X
ATP, 10 mM	5	1 mM
T4 PNK (10 U/uL)	5	50 U
ddH2	32	50 uL

Supplemental Table 3. Oligo phosphorylation setup.

After phosphorylation, the oligos were annealed to the ssDNA as following:

Supplemental Table 4. Setup for oligo annealing.

Component	Volume (µL)	Final Concentration
dU-ssDNA template	904	100 pmol
10X TM buffer	200	1X
Phosphorylated mutagenic oligonucleotide	150	300 pmol
ddH2O	846	2000 uL

The above annealing reaction mixture was prepared in a 1.5-ml microcentrifuge tube and aliquoted into 50 PCR tubes. The annealing reaction was conducted at 90 °C for 3 min, 65 °C for 5 min, 63 °C for 5 min, 61 °C for 5 min, 59 °C for 5 min, 57 °C for 5 min, 20 °C for 5 min. To complete the enzymatic synthesis of CCC-dsDNA, 6.5 μ L of the following solution was added to each of the 50 reactions above. The reaction was completed overnight at 20 °C.

Supplemental Table 5. Setup for enzymatic synthesis of CCC-dsDNA

Component	Amount (µL)	Final
ATP, 10 mM	100	346 mM
dNTP mix (25 mM of each)	100	865 mM of each nucleotide
DTT, 1M	15	5.19 mM
T4 DNA ligase	30	2,000 ligation units or 30 Weiss units

The ccc-dsDNA was purified by PCR purification column. To improve electroporation efficiency, drop dialysis was conducted to remove excess salt in the DNA. The DNA can be frozen at -20 °C for later use.

30

E. coli electroporation and phage propagation

A 1-mm gap electroporation cuvette used in electroporation was pre-chilled on ice. A 50 μ L aliquot of electrocompetent *E. coli* SS320 was thawed and mixed with 5 μ L DNA before being transferred into the cuvette. Electroporation was performed following the manufacturer's instruction with a Bio-rad Gene Pulser with the following settings: 1.0 mm cuvette, 10 μ F, 600 Ohms, 1800 Volts. The electroporated cells were immediately rescued by adding 1 mL prewarmed SOC and transferred to 10 mL SOC media in a baffled flask. The cuvette was rinsed twice with 1 mL SOC media. A serial dilution was made to determine the library diversity (typically 1-3 x 10⁹). After 30-min recovery, all the SOC culture was transferred into 1 L 2XYT/carb/kan/M13K07 helper phage media and shaken at 250 rpm at 37 °C for 20 hours.

The next day, the culture was centrifuged and the supernatant was collected. Phage were precipitated by adding 1/5 volume 5X PEG/NaCl precipitation buffer (20% PEG-8000, 2.5 M NaCl) and centrifuged at 9,000 RPM for 20 min. Pellets were resuspended in ¼ starting volume storage buffer (1 X PBS, 0.05% tween-20, 0.2% BSA) supplemented with protease inhibitor. The library was stored at -80 °C freezer supplemented 10% glycerol.

Construction of human proteome substrate phage library (Lib hP)

The Lib hP was generated by subcloning from a human tiled T7 phage library in 49AA blocks with 25AA overlaps as previously described⁷ (**Supplemental Figure 2**). Briefly, the synthetic DNA for the T7 library was generated from all human sequences (variants and isoforms) in the NCBI protein database (Nov 2015) and redundancy collapsed based on 95% sequence identity. These were split into 49-residue blocks and tiled in 25-residue overlaps to reduce the risk of losing library diversity and to provide duplicate coverage. This yielded a total of 731,000 sequences that were codon optimized for expression in *E. coli*. We PCR amplified the oligos pools for T7 library provided by the DeRisi Lab with proper flanking region

and cloned into our M13-phagemid to generate Lib hP using ligation method. The quality of Lib hP was evaluated by NGS (150nt single reads on Ilumina HiSeq4000 or NextSeq). Considering the 25-residue overlap and NGS errors, we would expect Lib hP to have higher sequence coverage. As expected, the amino acid composition was very uniform across the 49 positions except for the enrichment of methionine at the first position that was programed as a start codon for protein translation from the T7 library (**Supplemental Figure 2c, 2d**). The tiled peptide library (Lib hP) covers the entire human proteome, thus should give us more comprehensive profile than with proteomics because no cell line expresses all the proteins at one time. A single bulk cell proteomics experiment or RNAseq from any given cell line would only identify 8,000-10,000 proteins or 10,000-15,000 transcripts, respectively (**Supplemental Figure 2e**). Using RNA-seq, the minimum numbers of cell lines that are required to achieve different gene coverage is shown in **Supplemental Fig. 2f.** One can estimate for 95% gene coverage would require more than 40 cell lines.

Vector digestion

The parental phagemid was first constructed in which an avi-tag was linked directly to a sequence including an EcoRI restriction site, a TAA stop codon, a BamHI restriction site, and a HindIII restriction site, followed by a truncated M13 gene III protein (**Figure 1a**). The EcoRI and HindIII restriction sites was for cloning and the middle BamHI site was designed for removing parental phagemid in the cloning process. The parental phagemid was then subjected to restriction digestion under the following reaction conditions:

Supplemental Table 6. Setup for vector digestion.

Reagent	Volume (µL)	Concentration
Purified PCR product	2	2000 ng/µL (by Midi Prep)
NEB CutSmart buffer	5	10X
EcoRI-HF (NEB)	1	20,000 U/mL
HindIII-HF (NEB)	1	20,000 U/mL
*BamHI-HF (NEB)	1	20,000 U/mL
ddH ₂ O	41	/

20 reactions were combined and incubated at 37 °C for 3 hours

*BamHI-HF was added two hours later than the other two restriction enzymes

After digestion, the digested vector was size selected by agarose gel electrophoresis and concentrated by GeneVac before ligation/cloning.

Insert preparation

The single-stranded DNA oligonucleotide library was generously provided by DeRisi Lab at UCSF. Oligos were amplified by PCR with primers adding relevant EcoRI and HindIII restriction sites. 1 µL of 5 nM oligo is used for 96 25-µL PCR reactions:

Supplemental Table 7. PCR setup for insert amplification.

Reagent	Volume (µL)	Concentration
Input DNA library	1	0.05 nM
Phusion® High-Fidelity PCR Master Mix with HF Buffer	12.5	2X
Forward Primer	1	10 µM
Reverse Primer	1	10 µM
ddH ₂ O	9.5	/

Primer design:

Forward Primer: ATGGCATGAAGAATTCTGGAGCCATCCGCAGTTCG (underscored: EcoRI restriction site)

Reverse Primer: AATCAAAATCAAGCTTCTTATCATCGTCGTCCTTGTAGTC (underscored: HindIII restriction site)

Supplemental Table 8. Thermo cycle setup for PCR

Temperature (°C)	Times (mm:ss)	cycles
98	03:00	1
98	00:30	25
72	00:30	
72	10:00	1
4	∞	/

All the PCR products were combined and purified by AMPure PCR purification magnetic beads. DNA was then subject to restriction digestion under the following reaction conditions:

Supplemental Table 9. Setup for insert digestion

Volume (µL)	Concentration
10	300 ng/µL
5	10X
1	20,000 U/mL
1	20,000 U/mL
33	/
	Volume (μL) 10 5 1 1 33

Reaction was incubated at 37 °C for 2 hours

The digestion mixture was purified by Qiagen PCR purification kit. The sizes of the fragments were

confirmed by agarose gel electrophoresis before ligation.

Cloning/Ligation

The ligation reaction was set up in the PCR tubes according to the table below:

Supplemental Table 10. Ligation setup

Reagent	Volume (µL)	Concentration
Digested vector	1	400 ng/µL
Insert	1	228 ng/µL
T4 DNA ligase Buffer	1	10X
T4 DNA ligase	1	400,000 U/mL
ddH ₂ O	6	1

Reaction was incubated at 16 °C overnight

The ligation mixture was purified by Qiagen PCR purification kit and then concentrated by GeneVac. To improve electroporation efficiency, drop dialysis was conducted to remove excess salt in the DNA. The DNA can be frozen at -20 °C for later use.

In vitro biotinylation of the phage library with BirA ligase

10 times more BirA ligase was used for the biotinylation reaction than a typical condition because BirA activity was partially inhibited by 10% glycerol. Generally, 120 μ L or BioMix A (10X concentration: 0.5M bicine buffer, pH 8.3), 120 μ L of BioMix B (10X concentration: 100mM ATP, 100mM MgOAc, 500 μ M dbiotin)), and 10 μ L (10 mg/mL) BirA ligases were added into 1 mL substrate phage library with an Avi-tag on the N-terminal of the displayed peptides. The reaction mixture was incubated at 37 °C for 1 hour. The phage were then precipitated in 30 mL PBS buffer supplemented with 0.05% Tween20 and 0.2% BSA (PBSTB) and 1/5 volume 5X PEG/NaCl. The phages were pelleted with centrifuge at 9000 RPM, 20 min and the supernatant was completely removed. To move extra free d-biotin, the phage were resuspended in PBSTB 1mL, precipitated with 1/5 volume 5X PEG/NaCl, pelleted by centrifuge twice.

NGS sample preparation

Propagating the phage mixture to saturation in culture followed by PCR was more accurate than direct PCR without propagation which required more PCR cycles. We thus used overnight cultures⁴² for NGS sample preparation after boiling the phage to extract ssDNA as templates for PCR reactions. For samples from both Lib 10AA and Lib hP, a barcoding PCR process was performed using primers listed in **Supplemental Figure 21.** The thermal PCR profile for Lib 10AA was: 98 °C (20 s), 63 °C (15 s), 72 °C (30 s), 15 cycle. The thermal profile for Lib hP was: 98 °C (30 s), 72 °C (30 s), 15 cycle to prevent heterodimerization due to the 25-residue overlap. The number of cycles was determined empirically to prevent product laddering, assessed by agarose gel electrophoresis. PCR products were gel purified on 1.2 % agarose. Illumina library quality was assessed by Agilent DNA 1000 Bioanalyzer kit (5067-1504, Agilent), according to manufacturer's instructions. Libraries were sequenced on a NextSeq or HighSeq 4000 (Illumina) using single-read 50 or 150 base pair reads. Custom sequencing primers were of Tm =67 °C, GC% = 50-52.

Sequencing analysis pipeline.

Sequence filtering and peptide analysis were performed using an in-house informatics pipeline written mainly in R. Sample scripts are available for download at:

https://github.com/crystaljie/NGS_data_process_sample_script_for_substrate_phage_paper_JZH OU.git). The initial data processing for Lib 10AA and Lib hP are different. Raw NGS data from Lib 10AA, "*.fastq.gz" sequencing files were converted into a table with DNA sequences, amino acid sequences, counts/frequency, RPM as four columns, which were saved as .csv files for further analysis. For Lib hP, reads were aligned to a reference dataset using bowtie2 to generate Sam files. Sam files were then converted into Bam files using Samtools, which were then parsed using a suite of in-house analysis tools (R) to make a table with gene, ref.name, individual phage counts/frequency, peptide sequence, etc. as columns. These tables are saved as .csv files for further analysis. Some quality filters were applied: i) all the sequences with a stop codon are removed (Lib 10AA); ii) only the sequences that were in-frame are

kept (Lib 10AA); iii) the bowtie2 alignment mapping quality should be greater than 10. Although we applied this stringent cut-off, it does not necessarily mean sequences that showed up only once in NGS may not be a possible substrate.

Generation of PSSM

The first step towards building Scoring Matrix for predicting proteolytic sites involves an unbiased sequence alignment and we illustrate it with the caspase-3 data (**Supplemental Figure 8**). For practical reasons of computing time we aligned the top 20,000 most enriched sequences in Round 3 from Lib 10AA (**Supplemental Figure 8a**) to generate a position frequency matrix (PFM), the most basic representation of a motif and simply for each position the total counts of each amino acid (C_N) (**Supplementary Fig. 8b**). 20,000 sequences guarantee the variability of the collection of proteolytic sites, which enhances the accuracy for predicting additional sites. Based on the alignment, the position frequency matrix (PFM) was generated, converted to a position probability matrix (PPM, **Supplemental Figure 8c**) and then position weight matrix (PWM) using formula shown below, where B_N , a background probability for amino acid is 0.05 (**Supplemental Figure 8d**).

$$PPM(N) = \frac{C_N}{\sum C}$$

$$PWM = log2\left(\frac{PPM(C_N)}{B_N}\right)$$

PWMs are also known as position specific scoring matrices (PSSMs). Since starting from such a large pool of sequences, all the amino acids are scored in the matrix and there no need to use a pseudocount. Using this equation, the log of fractions where the probability of a certain amino acid in a sequence is higher than that of the background probability of that amino acid result in positive scores, and vice versa for negative scores.

Reference

- 1. Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins, *Molecular & Cellular Proteomics* **11** (2012).
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M. W., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jane-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. Y. K., Yu, J. J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N. X., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., and Garraway, L. A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, *Nature* 483, 603-607 (2012).
- Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., Fullgrabe, A., Fuentes, A. M. P., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., Jaiswal, P., Huber, W., Choudhary, J., Parkinson, H. E., and Brazma, A. Expression Atlas update-an integrated database of gene and protein expression in humans, animals and plants, *Nucleic Acids Research* 44, D746-D752 (2016).
- 4. Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment, *Bmc Bioinformatics* **16** (2015).
- 5. Tonikian, R., Zhang, Y. N., Boone, C., and Sidhu, S. S. Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries, *Nat. Protoc.* **2**, 1368-1386 (2007).
- 6. Kunkel, T. A., Roberts, J. D., and Zakour, R. A. Rapid and efficient site-specific mutagenesis without phenotypic selection, *Methods Enzymol.* **154**, 367-382 (1987).
- O'Donovan, B., Mandel-Brehm, C., Vazquez, S. E., Liu, J., Parent, A. V., Anderson, M. S., Kassimatis, T., Zekeridou, A., Hauser, S. L., Pittock, S. J., Chow, E., Wilson, M. R., and DeRisi, J. L. Exploration of Anti-Yo and Anti-Hu paraneoplastic neurological disorders by PhIP-Seq reveals a highly restricted pattern of antibody epitopes, *bioRxiv*, 502187 (2018).