Supporting Information

The Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue

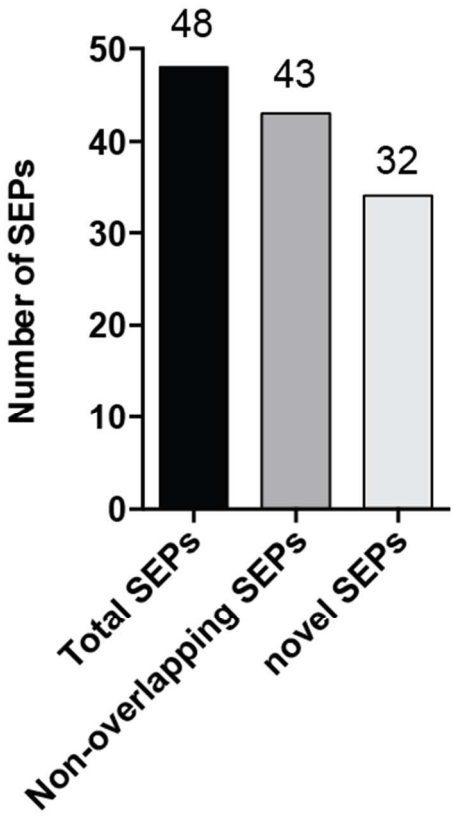
Jiao Ma[1], Carl Ward[1], Irwin Jungreis[3,4], Sarah Slavoff[1], Adam Schwaid[1], John Neveu[2,], Bogdan A. Budnik[2], Manolis Kellis[3,4] and Alan Saghatelian[1]

1 Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA

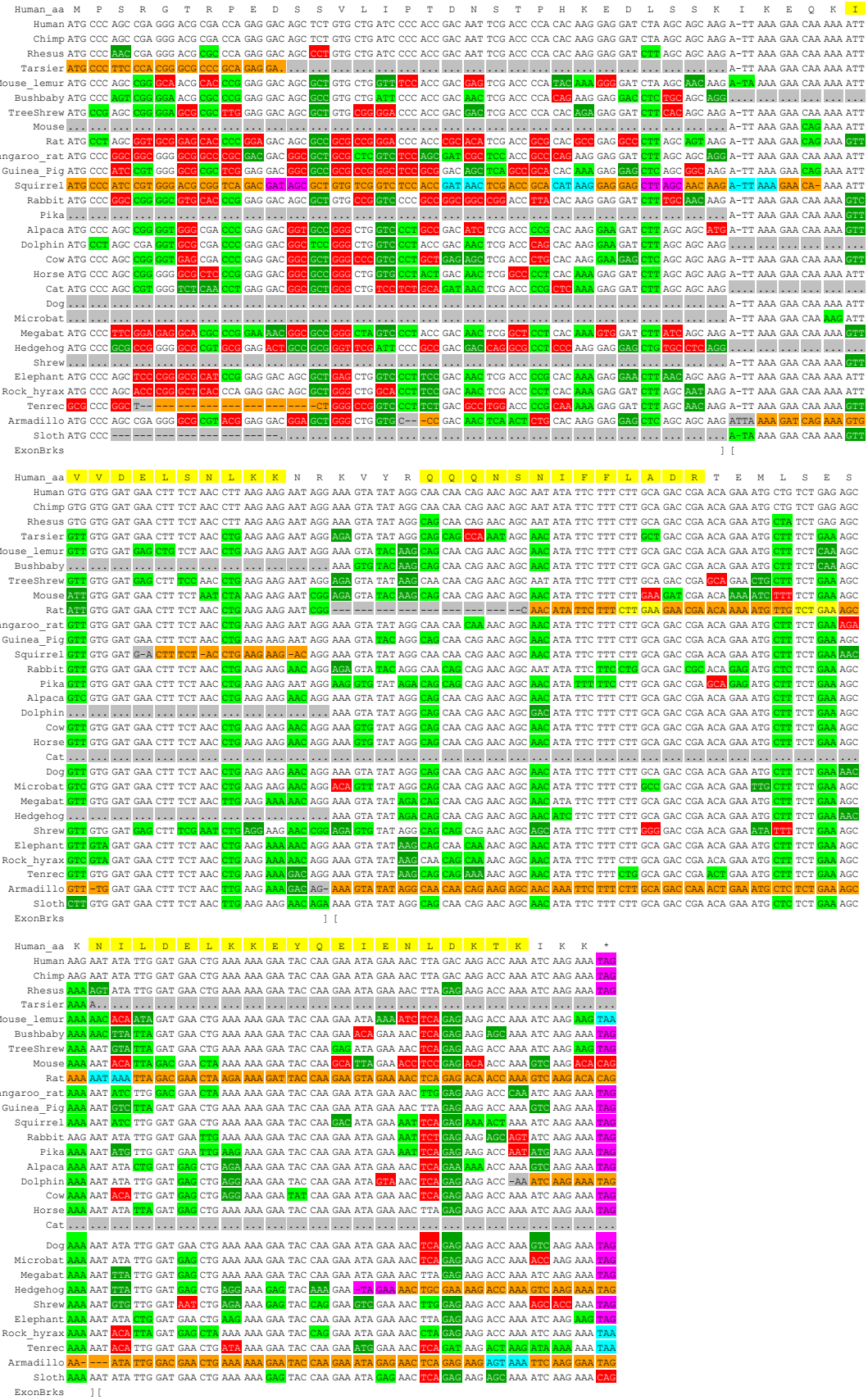
2 MSPRL, Center for Systems Biology, Harvard University, 52 Oxford Street, Cambridge, Massachusetts 02138, USA

3 MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

4 The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA



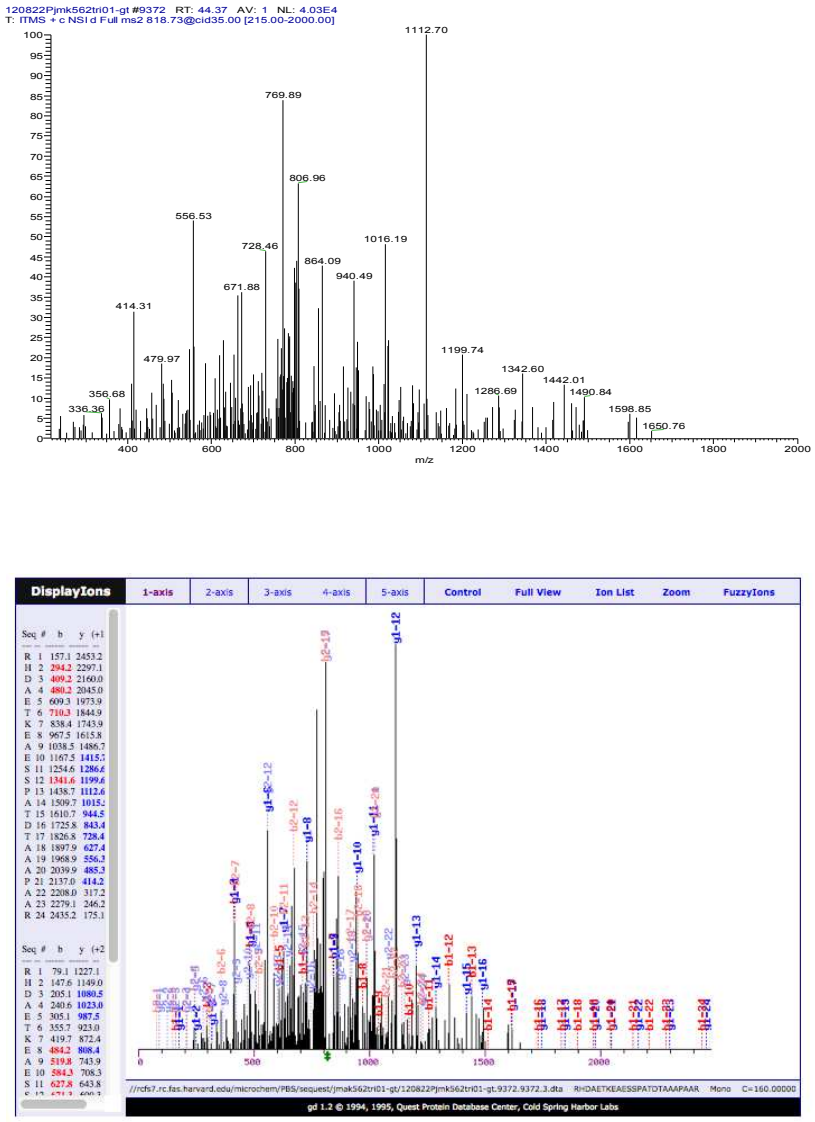
Supplementary Figure 1. Total number of SEPs detected in K562 cells using PAGE + LC-MS/MS workflow after performing an additional six technical replicates.



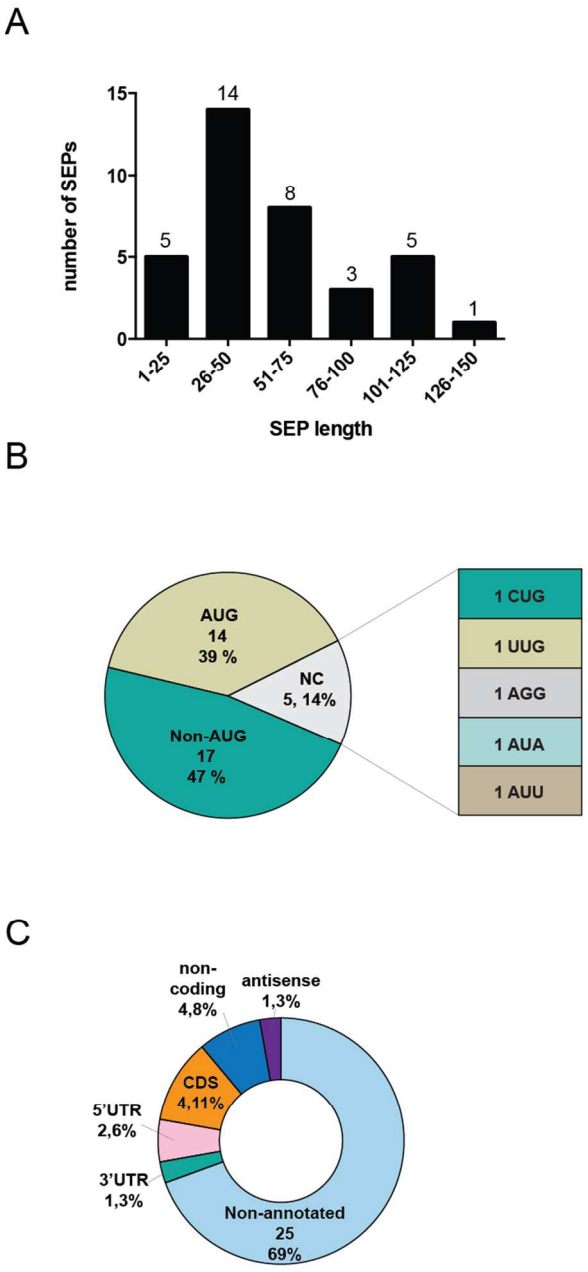
Supplementary Figure 2. Alignment for ASNSD1-SEP shows protein-coding signature.

Alignment for ASNSD1-SEP across 29 eutherian mammals, color-coded by CodAlignView ("CodAlignView: a tool for visualizing protein-coding constraint", I Jungreis, M Lin, M Kellis, in preparation). The amino acid sequences of the four tryptic peptides detected, NILDELKK, IVVDELSNLKK, QQQNSNIFFLADR, and EYQEIENLDKTK, are highlighted in yellow. The high concentrations of synonymous substitutions (light green) and conservative amino acid changes (dark green), and relatively low concentrations of radical amino acid changes (red) and frame-shifted regions (orange) is characteristic of protein-coding regions. The region's evolutionary coding potential as measured by per-codon PhyloCSF score, 4.315, is higher than 99.97% of non-coding regions, implying that it has been functional at the amino acid level in much of the eutherian mammal tree.

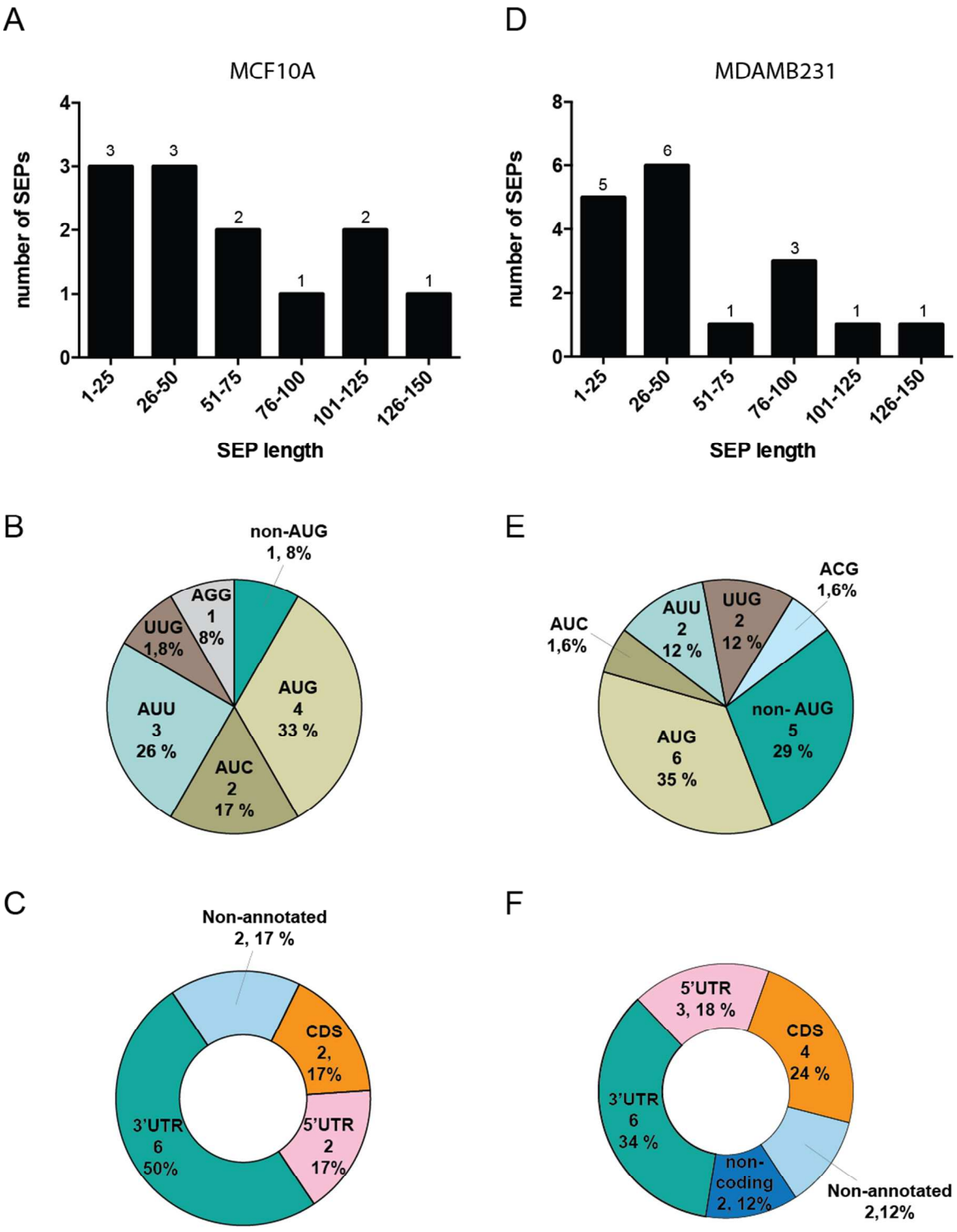
RHDAETKEAESSPATDTAAAPAAR, sf: .94, PRR3-SEP



Supplementary Figure 3. MS/MS spectra (raw spectrum: top, SEQUEST annotated spectrum: bottom) for the detected peptide: RHDAETKEAESSPATDTAAAPAAR for PRR3-SEP, with SF score of 0.94.



Supplementary Figure 4. **The characteristics of the 36 SEPs in K562 cells validated by Skyline-MRM.** (A) The length distribution of the SEPs, (B) the start codon usage of the SEPs, (C) the SEPs mRNA annotation by RefSeq.


A
MCF10A
number of SEPs
SEP length
1-25
26-50
51-75
76-100
101-125
126-150
3
3
2
1
2
1
D
MDAMB231
5
6
1
3
1
1
B
non-AUG
1, 8%
AGG
1
8%
UUG
1,8%
AUU
3
26 %
AUG
4
33 %
AUC
2
17 %
E
ACG
1,6%
AUC
1,6%
AUU
2
12 %
UUG
2
12 %
non- AUG
5
29 %
AUG
6
35 %
C
Non-annotated
2, 17 %
CDS
2,
17%
5'UTR
2
17%
3'UTR
6
50%
F
5'UTR
3, 18 %
CDS
4
24 %
3'UTR
6
34 %
non-
coding
2, 12%
Non-annotated
2,12%

Supplementary Figure 5. The characteristics of SEPs detected in MCF10A and MDAMB231 cell lines. (A) The length distribution of the SEPs, (B) the start codon usage of the SEPs, (C) the SEPs mRNA annotation by RefSeq in MCF10A cells. Similarly (D) The length distribution of the SEPs, (E) the start codon usage of the SEPs, (F) the SEPs mRNA annotation by RefSeq in MDAMB231 cells.

Supplementary Table 1. **A list of 36 SEPs detected in K562 cells that were validated by Skyline-MRM.** The peptides detected by shotgun proteomics were shown in blue, the additional peptides validated in Skyline-MRM were shown in red, the overlapping peptides detected in both shotgun method and Skyline-MRM were highlighted in purple.

Detected peptide by shotgun	Detected peptide by Skyline-MRM	Predicted SEP length	SEP sequence
IVYGDIRK KIVYGDIR	IVYGDIR	41	LYQRREELKLQWRSFNLDKIV YGDIRKEGILNALETMGSSF
ADAPAVSPESPQK	KPPFQPR ADAPAVSPESPQK	101	MSRGQFGQGQEPLDMFFWV NEISGEITYPPQKADAPAVSPE SPQKKPPFQPRSVQEAPCSPQ GPPAQRPALAPPSKPSLKDSG SRNPCPSAPTWARPKPEE
KNFFISK	ALAYIGAR	20	MPKNFFISKIKALAYIGARC
EAESSPATDTAAAPAAR RHDAETKEAESSPATDTAAA PAAR	APSSLPSK EAESSPATDTAAAPAAR	93	MPGRDGGRQLAPECGRRRCS PQPSGSGMAGASALLPTDPP EAETESPRAPSSLPSKSRCSHC RRHDAETKEAESSPATDTAAA PAARAGRDWR
FSPLELAAGGVR	FSPLELAAGGVR LLLPCPCCFG	127	MAGGDPAAGRAVGAAKAKR AAGICGDTAAPALSGHPPDCL PESPGRGGLTCSPSQSRFSPLE LAAGGVRGCGAQLVTVGCSE PSWSPGEGAGPGEQQSLDVE AAEALQAGPASPRCRLLLPCP CCFG
YLDLHER	QIFHSNK	33	IRKYLDLHERQIFHSNKHLNNII NCKAQTTRVK
QGSLVQQVALR	EVLEQLMK	66	MMKNQWMEIGHHQLHQPS RFLHLKHPEILPTVQGSLVQQ VALRLEVLLRKEVLEQLMKMIL
RNLVVVINL	DPMNAK TLEVVIS	71	MNVRNVRRNLVVVINLFYITG FMSLRDPMNAKSVGRTFVVA INLLYIKDFILVRNPMNVQNV GRTLEVVIS

IYNRLYFLEK	VNFCNK	24	VNFCNKASWDLKKIYNRLYFLEKF
EGGWRQVEGTGTPK	DSAKPIR QVEGTGTPK	75	MTTGERDSAKPIRATATRQEDRSPEGGWRQVEGTGTPKSKQGSRVLAAQEETQHPEAVPQRADPKGASASPLRRQ
ELSQYLK	MDELSQYLK	42	MDELSQYLKVILPSTVVLDVILLQLSFLLYIANFLFSLGSLP
SRQVDQEVRSSR	QVDQEVR	24	SQHFGRSRQVDQEVRSSRTAWPRW
ENIPDITK	DFVFNSLSK	32	DFVFNSLSKILVENRPAFVNENIPDITKPKHF
VAEIIIER	TLYTCLLR LVSHGINLALIFSIWK	116	RVAEIIIERLVSHGINLALIFSIWKCLKENHFHCRKSFFKYLLYPREISLYLPPQAVISCFREWNPPCPSIFWFLGLNSSLVKSPWLGILSWEQILSCSLMCLHSPKTLYTCLLRA
AIVVARVVTIPK	MVAIVVAR	43	MVAIVVARVVTIPKIMHQPVLSFLNFHVPLYTFMSVYVDLSLV
GDFLNLR AGDFLNLR	AGDFLNLR IGISYQFCK	41	AGDFLNLRIGISYQFCKFSPINYFFFLFSPCLLYGILLDIS
GFLAGYVVAK TLRDYLQLLR NQLESLQR RVEDEVNSGVGQDGSLLSSPFLK	NQLESLQR NQLESLQRR GFLAGYVVAK	103	MADDKDSLPKLKDLAFLKNQLESLQRRVEDEVNSGVGQDGSLLSSPFLKGFLAGYVVAKLRASAVLGFAVGTCTGIYAAQAYAVPNVEKTLRDYLQLLRKGPD
RLLFAGK IRLLFAGK	GTTFSWVIR	22	GTTFSWVIRLLFAGKLNYSSMS

GLIENPALIR	LMQEGK QGLIENPALIR	100	QRVQAERLAIRARLKREYLLQY NDPNRQGLIENPALIRWAYAR TTNVYPNFRPTPKNSLMGALC GFGPLIFIYYIIKTERDRKEKLM QEGKLDRTVHLSY
QNIKGLENILQK	GLENILQK VVTLTLQSSENQR	81	IWSRVVTLTLQSSENQRQNIK GLENILQKEAATCVDNGLFMP LLLSVDLVQETCSGDGCEGG MRIDIDTPVSQTCLFITLL
SWLTPVAGK	MAPLGLK	26	MAPLGLKDPLSSWLTPVAGKL VMAVS
HALPLLK	QEFHALPLLK NSTNFFLLIK	52	NSTNFFLLIKQRSFGGFIPIADK RGKDGKCSRFLSFHKQEFHAL PLLKQRKE
GAGILLLR	TGAGILLLR	44	TGAGILLLRWLTHWLLAGSLR SSPGVPLHVLLHGLMMWHEP HSV
KQNSLIANMEK	SGYINR SCLHSIK GEEAAEEK MQIEATR GQTLFSSTK QNSLIANMEK	114	LIKKQNSLIANMEKVLVVWM EDQTSHNIPLSQSLIQSKGQTL FSSTKNEKGEEAAEEKFEASRV WLMRFKERSCLHSIKMQIEAT RADEEGTASDPEDPAKLIDKS GYINRFTM
MKNFLAVTITGK	TSVQGITTVILK	110	IDWRRKKRKKIEKRKSFRRAEV NTKNISPLLPHLPPPPLPLLRL QKAVVRVRVTIKKKKYKGRKE RKTSVQGITTVILKRRTSLRRES

			FMKNFLAVTITGKKPRKSPGS
KDLHLSWEPK	FEFFPK NGLPSVLLVK	43	KQQPPLFSLYKFEFFPKLKDLH LSWEPKEKNGLPSVLLVKEIL
KNEFLLK	DHVLFFK	27	IIFKNEFLLKDHVLFFKSIFSSYF CYC
AEIIILK	VFDLQDF MAEIIILK	17	MAEIIILKAKVFDLQDF
TPLLAYIQ	TPLLAYIQPDTSAF	49	MNLEMEKKAGLFQRVDLSEL DSTIELCCIFCGSSKTPLLAYIQP DTSAF
HAFLNLR	HALFLNLR NLQTPGAVGEDK	54	HAFLNLRAIPSPQSNNLERPQ VQLLHSPDPLLSTPRNLQTPG AVGEDKKKSGVA
EVEGAVSR	QSEVMSQK IFNNHTLIIK	42	TQEVEGAVSRDCITALQPGKQ SEVMSQKQTTKIFNNHTLIIK
RKPLYTIGWNL	DFTSHQLER	64	SSGKGKSNSQRDFTSHQLERL SSKRQNIKRVGKNAEKRKPLY TIGWNLNWYSHYKKQHGGSS KN
KINALLK	GNILLSNK	50	SQPPLKCLCLIKINALLKGNILLS NKCGCVFYHTSILRKCWTSEY HKTGN
FQPPHHVQSSPDVK	GLSFQPPHHVQSSPDVK	32	ESCPEPTEQKGLSFQPPHHVQ SSPDVKSQFWF
ARDQYGHLIPTK	KPSFPSPR DQYGHLIPTK GSCHFLSQVGGWGI	61	MCAEIEEGAEGVTARDQYGH LIPTKVASGPQGLSGARKPSFP SPRLRGSCHFLSQVGGWGI
LAFIFLPDR	NDLAFIFLPDR	65	AKIVPLHSSLGDRVRPCLKTKQ TKEFRNDLAFIFLPDRQCIHQD GTLTGNQVLAPLLAGKEHEVF

Supplementary Table 2. A list of all novel SEPs detected in this study.