Supplementary Materials for Manuscript

Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines

Xiongwen Cao[1,2,4], Alexandra Khitun[1,2,4], Zhenkun Na[1,2], Daniel G. Dumitrescu[1], Marcelina Kubica[2,5], Elizabeth Olatunji[2,5], Sarah A. Slavoff[1,2,3*]

[1]Department of Chemistry, Yale University, New Haven, Connecticut 06520, United States

[2]Chemical Biology Institute, Yale University, West Haven, Connecticut 06516, United

[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06529, United States

[4]These authors contributed equally

[5]These authors contributed equally

*Correspondence: sarah.slavoff@yale.edu

.

Table of Contents

Supplementary Figures

Supplementary Tables (all in separate Excel sheets)

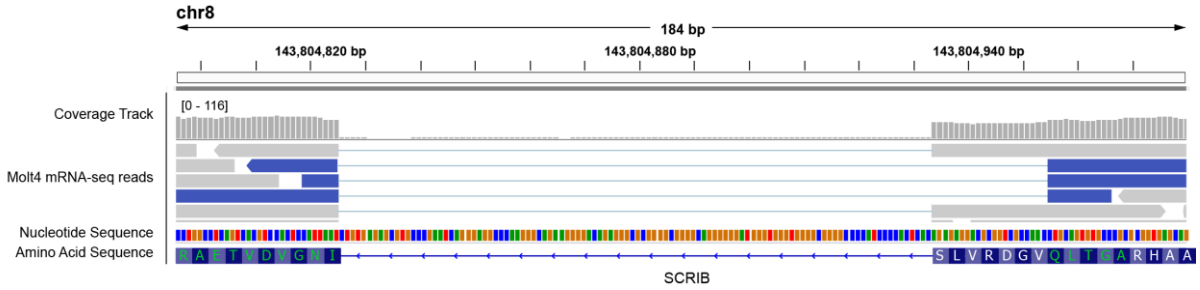


Figure S1. mRNA-Seq support for novel SCRIB transcript variant. Overall mRNA-seq coverage is shown in the top coverage track. mRNA-seq reads from MOLT4 cells are shown in the second track from the top. mRNA-seq reads supporting the novel splice junction between exons 20 and 21 of the SCRIB gene are highlighted in blue. Reads mapping to the annotated splice junction are shown in gray. Note that there is a small deletion in the novel variant, compared with the annotated transcripts. The genomic nucleotide sequence is displayed on the nucleotide sequence track with green, red, blue, and orange tiles representing A,T,C and G, respectively. The bottom track represents the annotated SCRIB protein sequence with amino acids highlighted in green corresponding to the peptide (sequence, AGTLQINGVDVTEAR) detected in our proteomics data which aligns exclusively to the alternatively spliced variant. Note that seven amino acids (VGDRVLS) are excluded in the protein translated by the novel variant, compared with the annotated SCRIB protein.

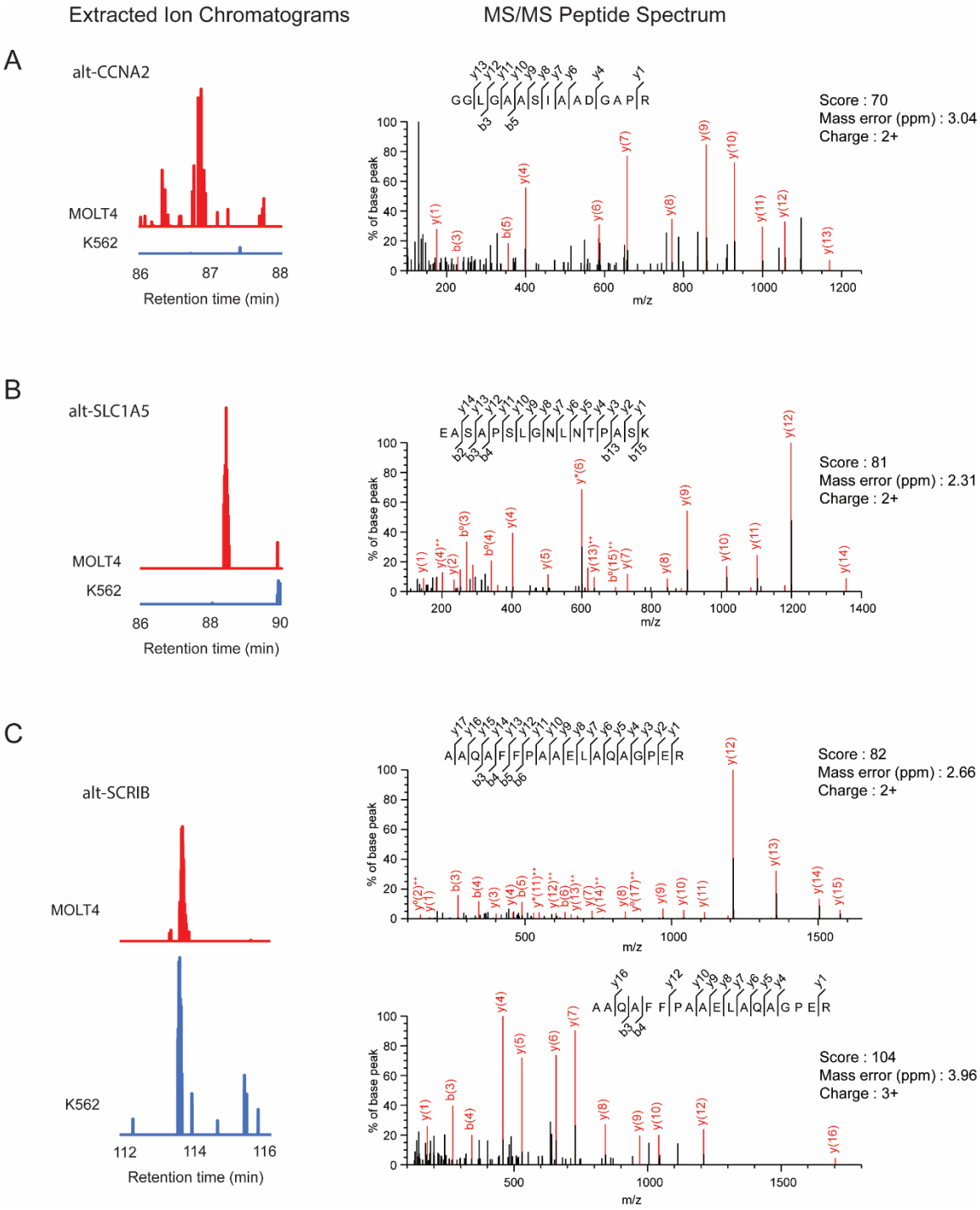


Figure S2. Confirmation of differential expression of three additional detected alt-proteins with extracted ion chromatograms (EICs). Shown are EICs (left) from MS_1 spectra corresponding to MS/MS spectra (right) of three tryptic peptides identified in MOLT4 cells only (A-B) or both cell lines (C). The same y-axis scale is used for each matched EIC pair. The EIC intensity at the same retention time was compared for the paired samples.

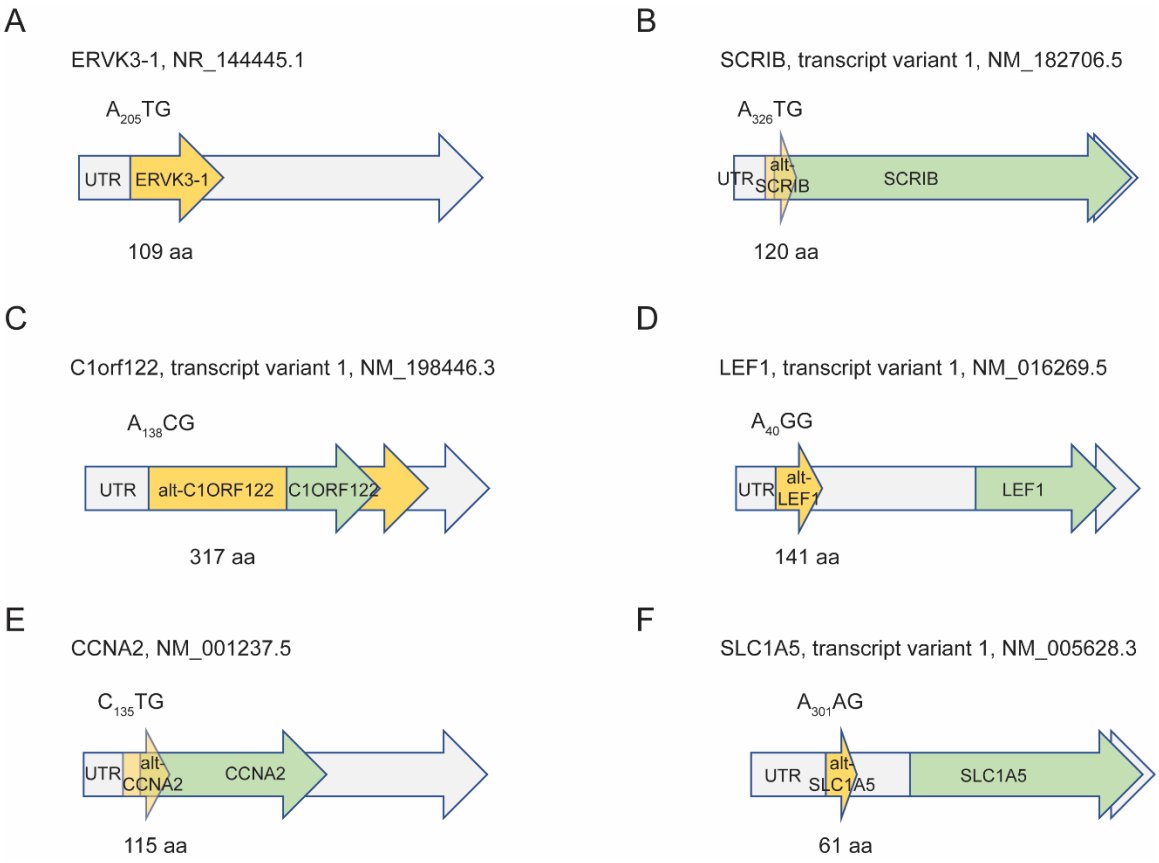


Figure S3. Predicted start codons used by the six selected alt-ORFs. Shown are schematic representations of the transcripts encoding the six selected alt-proteins. Top, transcript name and the accession number; light gray arrow, untranslated regions (UTR); yellow arrow, alt-ORF; green arrow, annotated coding sequence; the predicted start codon used by the alt-ORF is indicated above the transcript, numbered relative to the first nucleotide of the corresponding cDNA, and the predicted full length of the alt-protein is indicated at the bottom. aa, amino acid. (A-B) Two alt-proteins detected in both K562 and MOLT4 cells, (C) one detected in K562 only, and (D-F) three detected in MOLT4 only.

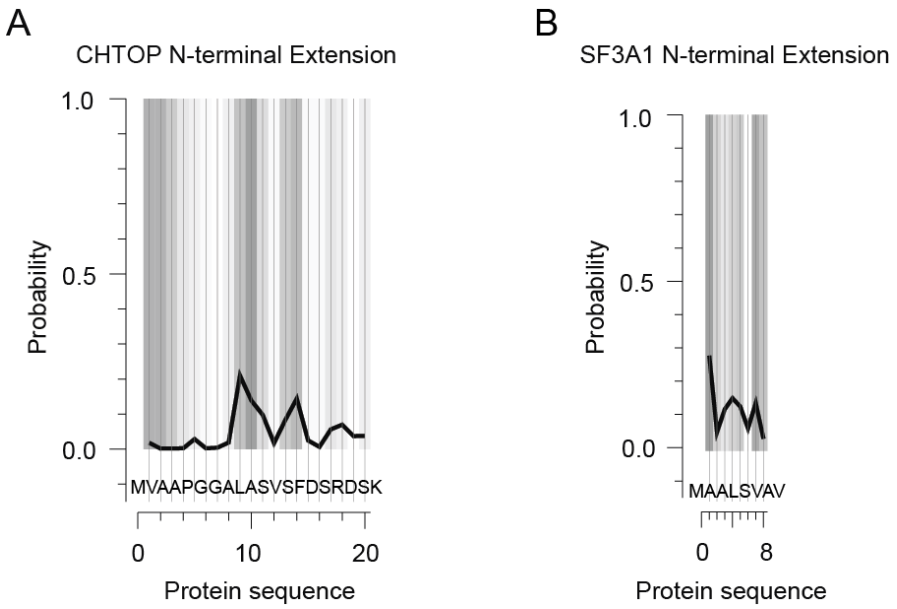


Figure S4. Results of TargetP bioinformatic analysis. (A) N-terminal extension of CHTOP protein is predicted to encode a mitochondrial transfer peptide (B) TargetP prediction result for SF3A1 N-terminal extension where the peak corresponds to a predicted signal peptide. The protein names are indicated at the top of each diagram. TargetP 2.0 probabilistically scores localization signals using a deep-learning model trained on proteins with known localization within three classes 1) mitochondrial transfer peptide 2) signal peptide and 3) other. Only sequences predicted to encode mitochondrial or signal peptides with >40% probability are represented. See Table S4 for full TargetP results.

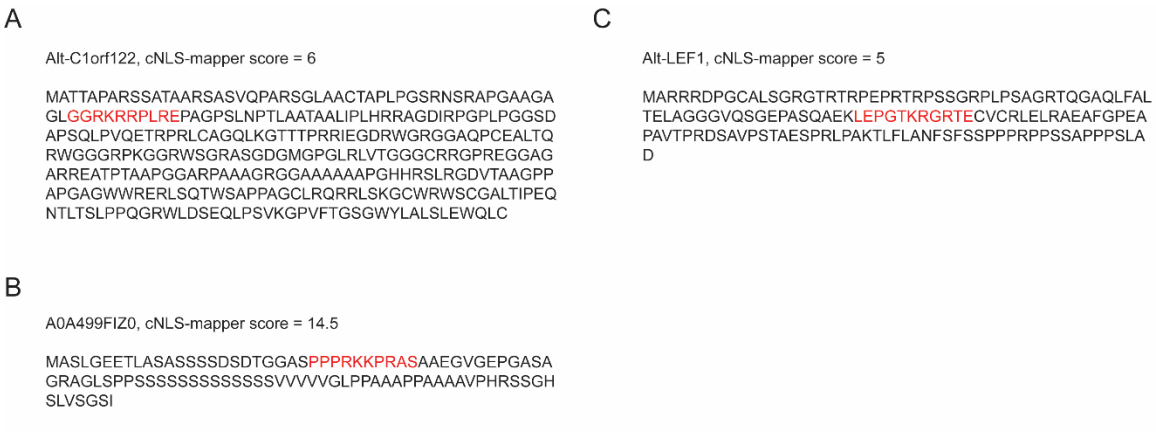


Figure S5. Three alt-proteins are predicted to encode nuclear localization signals (red) by cNLS-mapper. (A-C) The cNLS-mapper prediction results. The name of alt-proteins and the cNLS mapper score are indicated above each sequence. The default cutoff score of cNLS-mapper is 5.

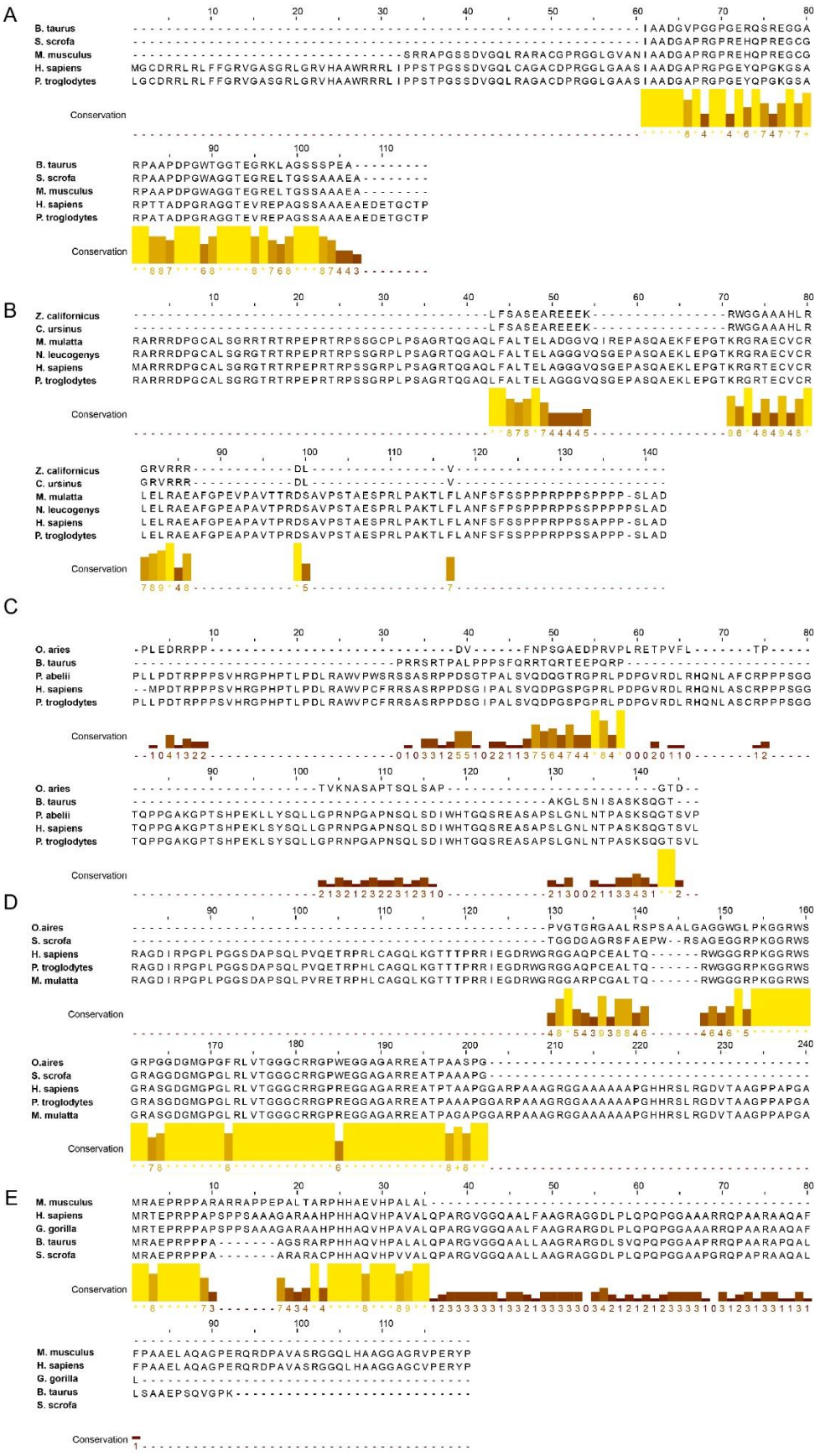

A
B. taurus
S. scrofa
M. musculus
H. sapiens
P. troglodytes
Conservation
B
Z. californicus
C. ursinus
M. mulatta
N. leucogenys
H. sapiens
P. troglodytes
Conservation
C
O. aries
B. taurus
P. abelii
H. sapiens
P. troglodytes
Conservation
D
O.aires
S. scrofa
H. sapiens
P. troglodytes
M. mulatta
Conservation
E
M. musculus
H. sapiens
G. gorilla
B. taurus
S. scrofa
Conservation

Figure S6. Amino acid-level conservation of alt-proteins in other species. (A) alt-CCNA2, (B) alt-LEF1, (C) alt-SLC1A5, (D) alt-C1orf122 (partial sequence) (E) alt-SCRIB. Alignments from Clustal Omega were quantified on the basis of physico-chemical properties using JalView. * indicates sequence identity.

Figure S7. Structure prediction of two nuclear alt-proteins. Predicted structural models for alt-C1orf122, C-score = -1.79 (A) or alt-LEF1, C-score = -3.03 (B) using I-TASSER. UCSF chimera was used to visualize and generate the figure. C-score is a confidence score for estimating the quality of predicted models by I-TASSER, which is typically in the range of [-5, 2].

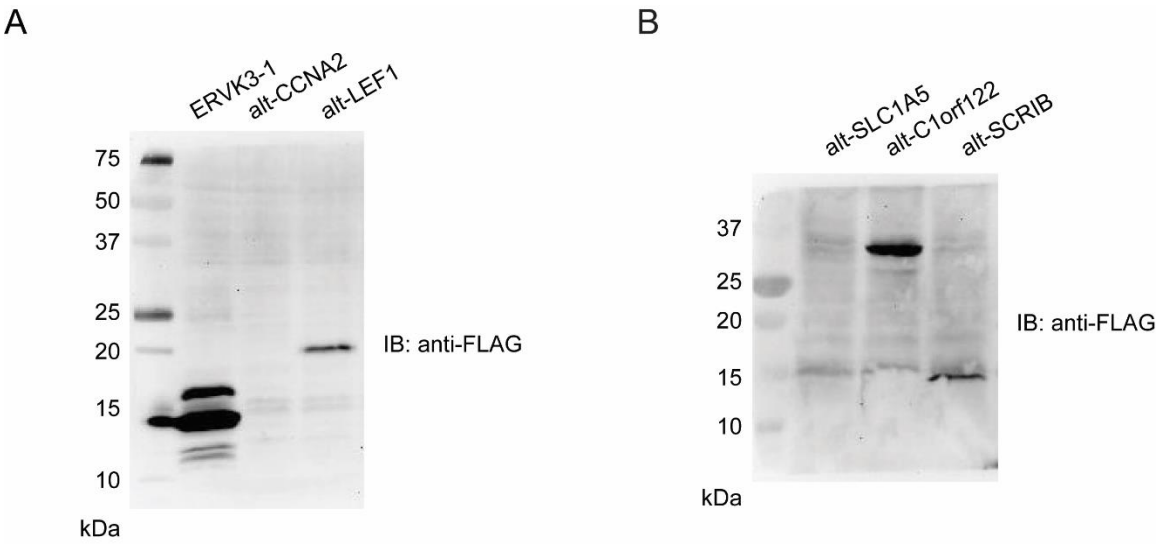
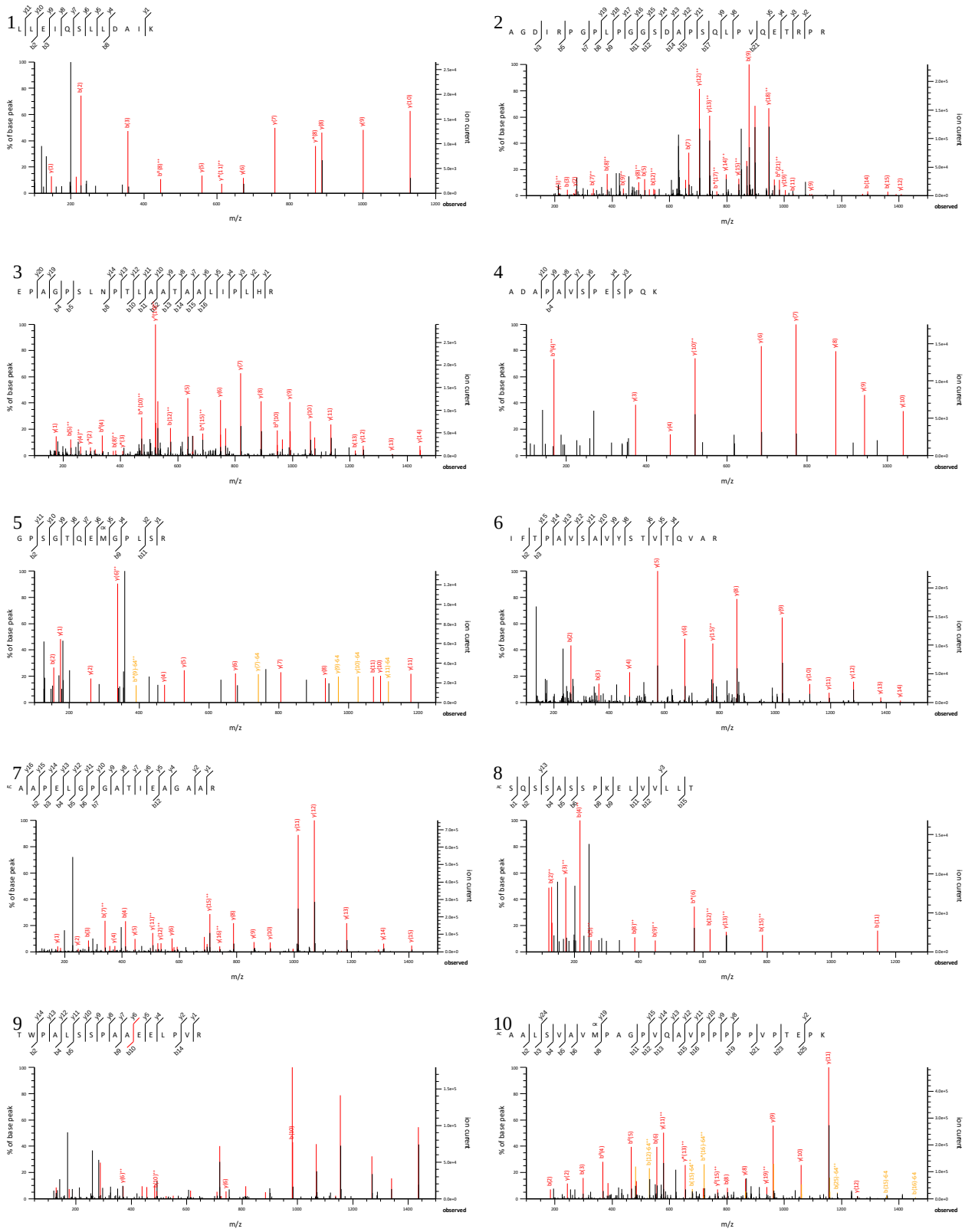


Figure S8. Uncropped Western blots for confirmation of alt-protein expression (related to Figure 5).


1
L L E I Q S L L D A I K
2
A G D I R P G P L P G G S D A P S Q L P V Q E T R P R
3
E P A G P S L N P T L A A T A A L I P L H R
4
A D A P A V S P E S P Q K
5
G P S G T Q E M G P L S R
6
I F T P A V S A V Y S T V T Q V A R
7
A A P E L G P G A T I E A G A A R
8
S Q S S A S S P K E L V V L L T
9
T W P A L S S P A A E E L P V R
10
A A L S V A V M P A G P V Q A V P P P P V P T E P K
% of base peak
m/z
Ion current
observed

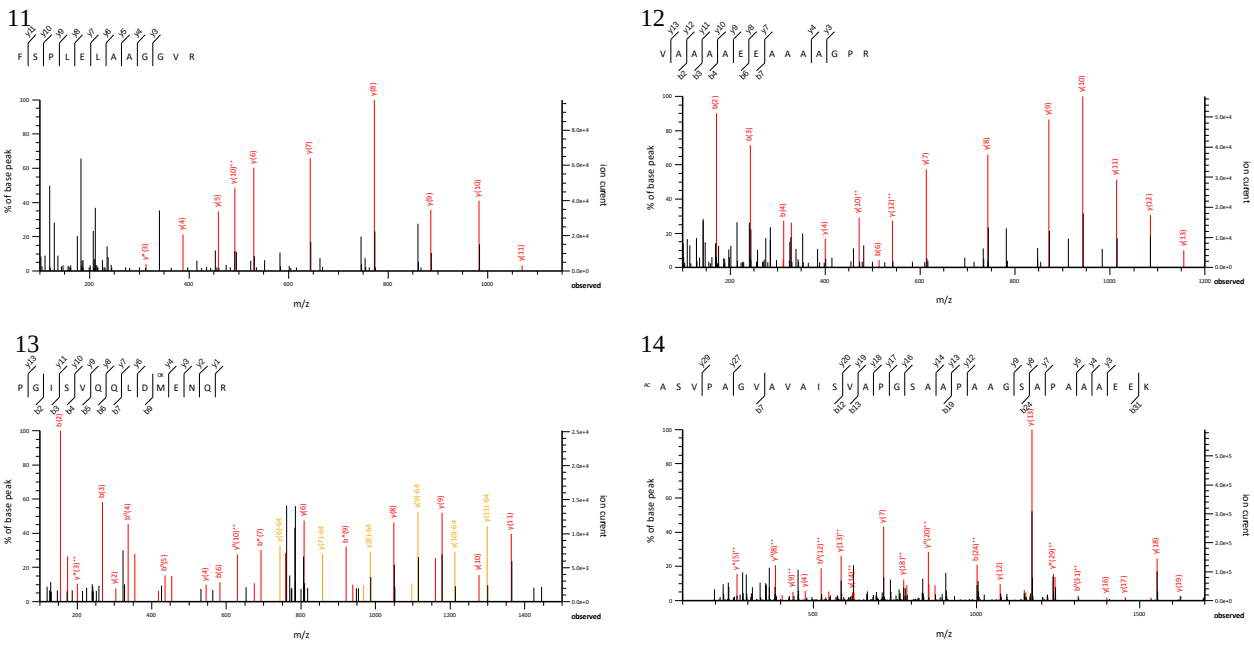
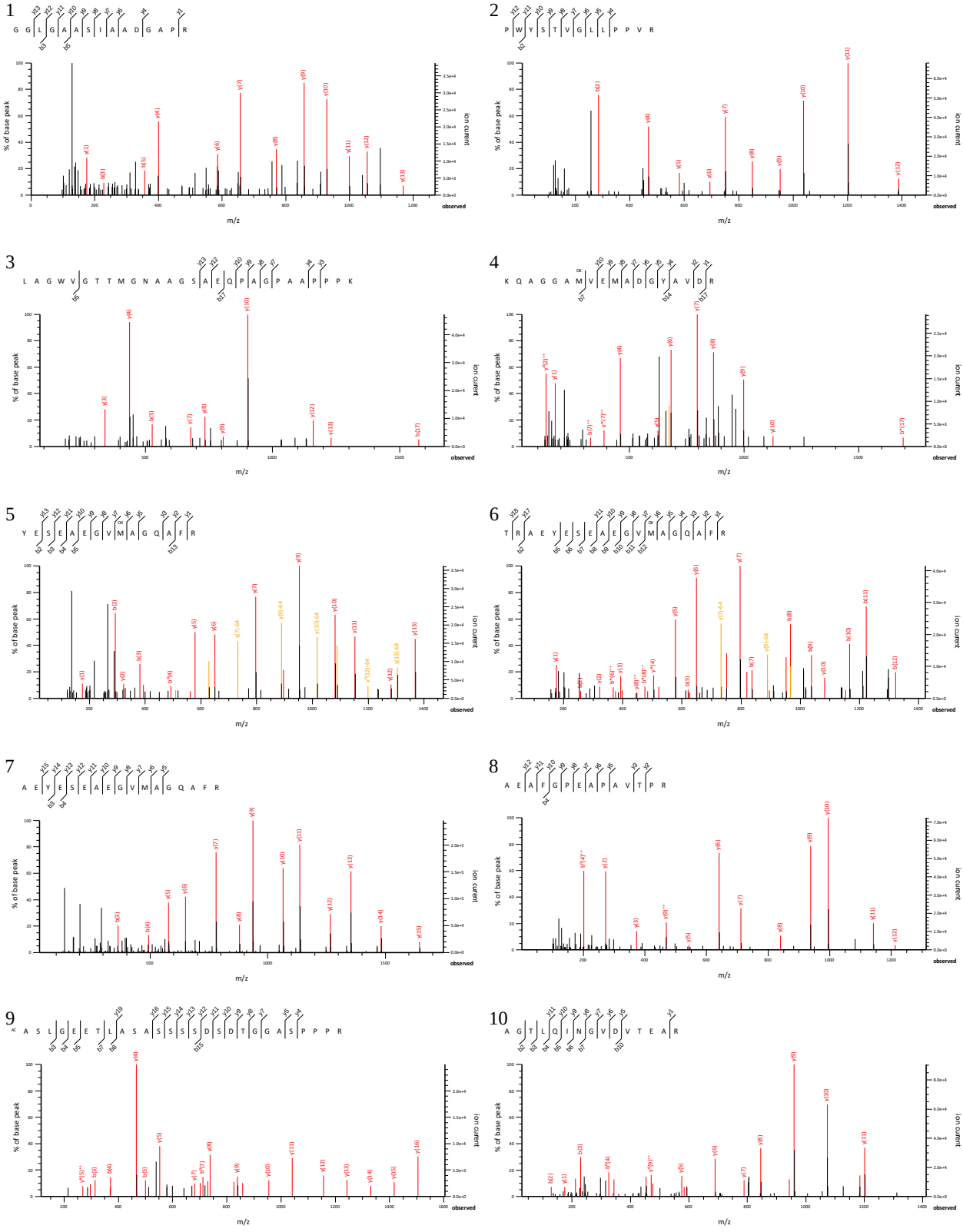


Figure S9. K562 MS/MS spectra corresponding to peptides listed in Table S3.


1
G G L G A A S I A A D G A P R
2
P W Y S T V G L L P P V R
3
L A G W V G T T M G N A A G S A E Q P A G P A A P P P K
4
K Q A G G A M V E M A D G Y A V D R
5
Y E S E A E G V M A G Q A F R
6
T R A E Y E S E A E G V M A G Q A F R
7
A E Y E S E A E G V M A G Q A F R
8
A E A F G P E A P A V T P R
9
A S L G E E T L A S A S S S S D S D T G G A S P P P R
10
A G T L Q I N G V D V T E A R
% of base peak
Ion current
m/z
observed

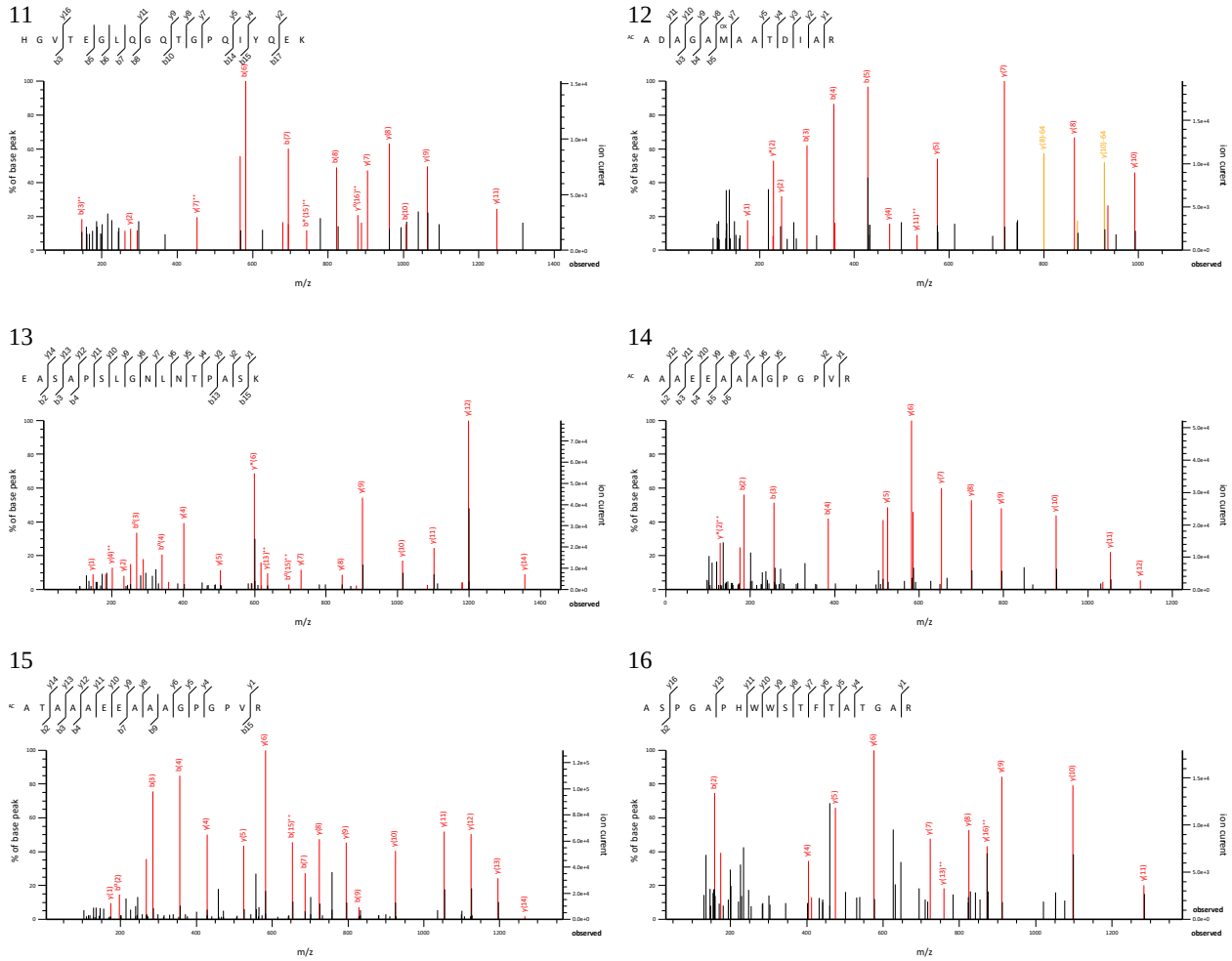


Figure S10. MOLT4 MS/MS spectra corresponding to peptides listed in Table S3.

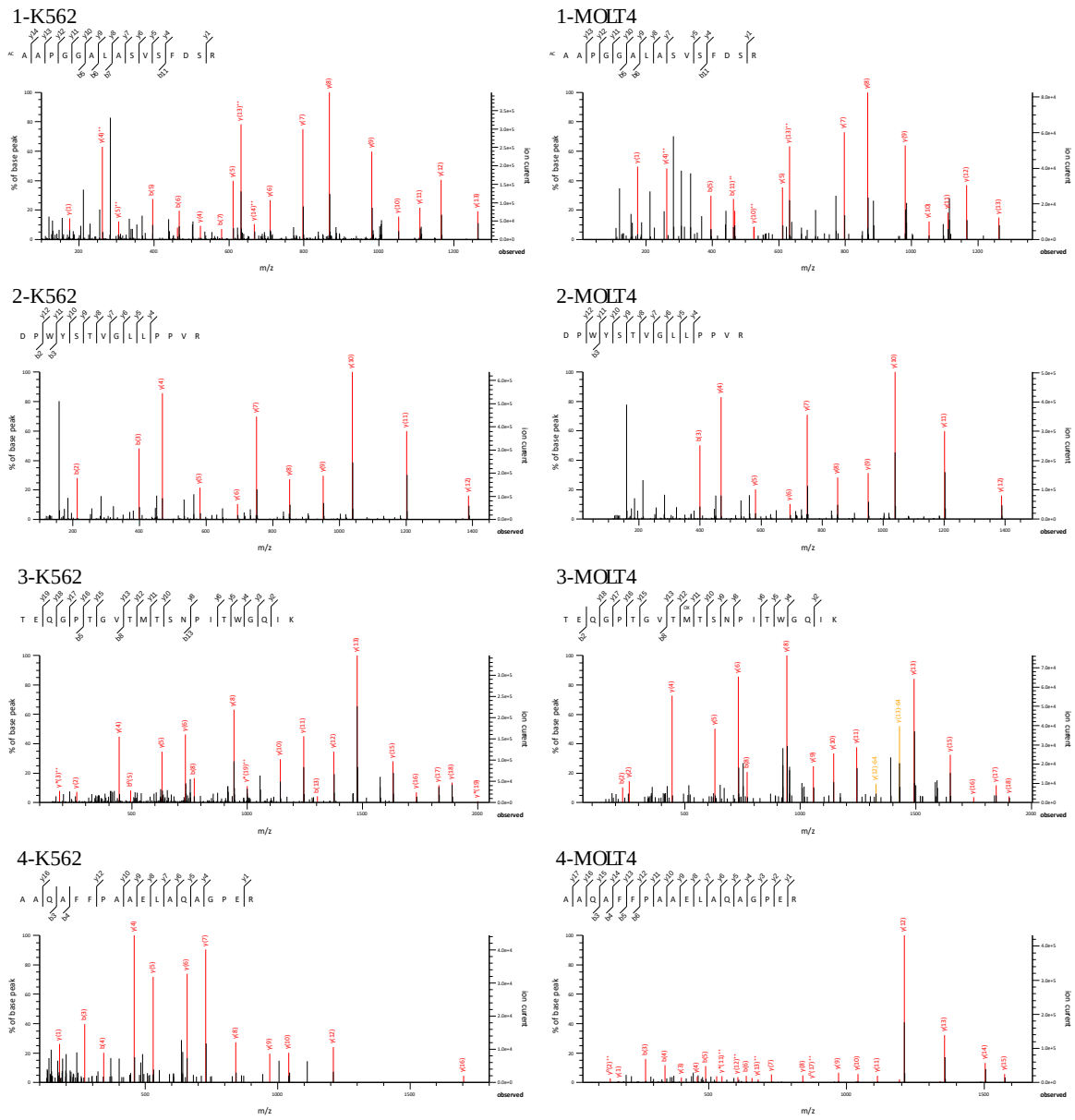


Figure S11. MS/MS spectra corresponding to peptides listed in Table S3 which are found in both K562 and MOLT4 cells.