

Supplemental Material accompanying

Translation initiation downstream of annotated start codons in human mRNAs coevolves with the Kozak context.

by

Maria S. Benitez-Cantos, Martina M. Yordanova, Patrick B.F. O'Connor, Alexander V. Zhdanov, Sergey I. Kovalchuk, Dmitri B. Papkovsky, Dmitry E. Andreev & Pavel V. Baranov

Supplemental Figures

Supplemental Figure S1. Distribution of ORF lengths initiated at fdAUGs.

Supplemental Figure S2. Further experimental verification of downstream initiation at mRNAs containing sAUGs in evolutionary conserved weak context.

Supplemental Figure S3. GWIPS-viz screenshot of *AASDHPPT* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S4. GWIPS-viz screenshot of *ASPHD1* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S5. GWIPS-viz screenshot of *C1orf94* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S6. GWIPS-viz screenshot of *CMPK1* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S7. GWIPS-viz screenshot of *FRMD3* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S8. GWIPS-viz screenshot of *ISL2* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S9. GWIPS-viz screenshot of *LIMK1* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S10. GWIPS-viz screenshot of *PABPCL4L* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S11. GWIPS-viz screenshot of *RELB* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S12. GWIPS-viz screenshot of *ZBTB80S* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG.

Supplemental Figure S13. Peptide observation frequency in PeptideAtlas for *AASDHPPT* protein products.

Supplemental Figure S14. Peptide observation frequency in PeptideAtlas for *ASPHD1* protein products.

Supplemental Figure S15. Peptide observation frequency in PeptideAtlas for *FRMD3* protein products.

Supplemental Figure S16. Peptide observation frequency in PeptideAtlas for *ISL2* protein products.

Supplemental Figure S17. Peptide observation frequency in PeptideAtlas for *LIMK3* protein products.

Supplemental Figure S18. Peptide observation frequency in PeptideAtlas for *RELB* protein products.

Supplemental Figure S19. Peptide observation frequency in PeptideAtlas for *ZBTB8OS* protein products.

Supplemental Figure S20. Localization of *ISL2-SNAP* derived products in cells.

Supplemental Tables (Excel files)

Supplemental Table S1. “Strong” set.

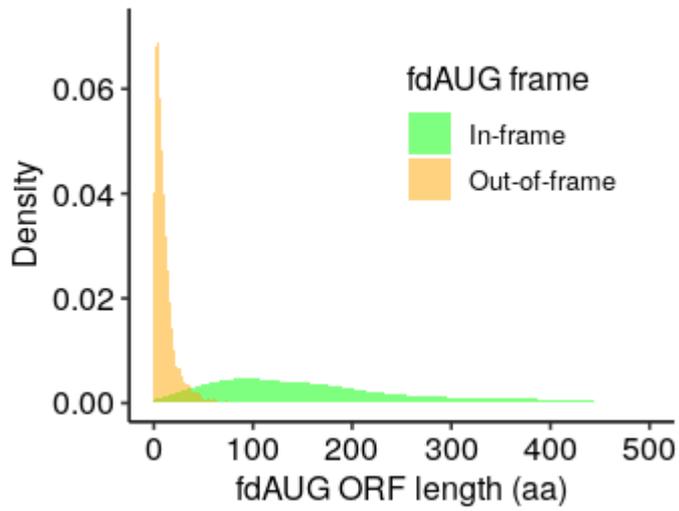
Supplemental Table S2. “Weak” set.

Supplemental Table S3. Conserved “strong” set.

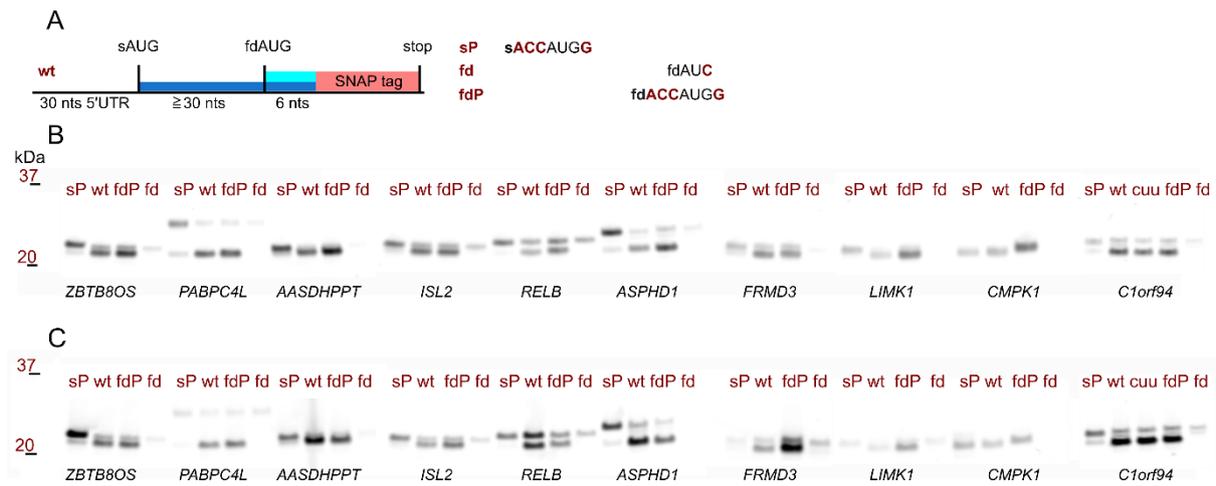
Supplemental Table S4. Conserved “weak” set.

Supplemental Table S5. Studies used in the ribosome profiling analysis.

Supplemental Table S6. Studies used for ribosome footprint density visualization.



Supplemental Figure S1 | Distribution of ORF lengths initiated at fdAUGs. ORFs initiated at out-of-frame fdAUGs are significantly shorter than those initiated at in-frame fdAUGs suggesting the lack of negative evolutionary selection acting on stop codon occurrence downstream of fdAUGs.



Supplemental Figure S2 | Further experimental verification of downstream initiation at mRNAs

containing sAUGs in evolutionary conserved weak context. **A.** Schematic of the test sequence

cassette fused to SNAP tag. The **wt** test sequence includes 30 nt upstream of sAUG and 6 nt

downstream of fdAUG. **sP** has the sAUG in perfect Kozak context, **fd** has the fdAUG changed to AUC

and **fdP** has the fdAUG in perfect Kozak context. **wt**, **sP**, **fd** and **fdP** test cassettes were designed for

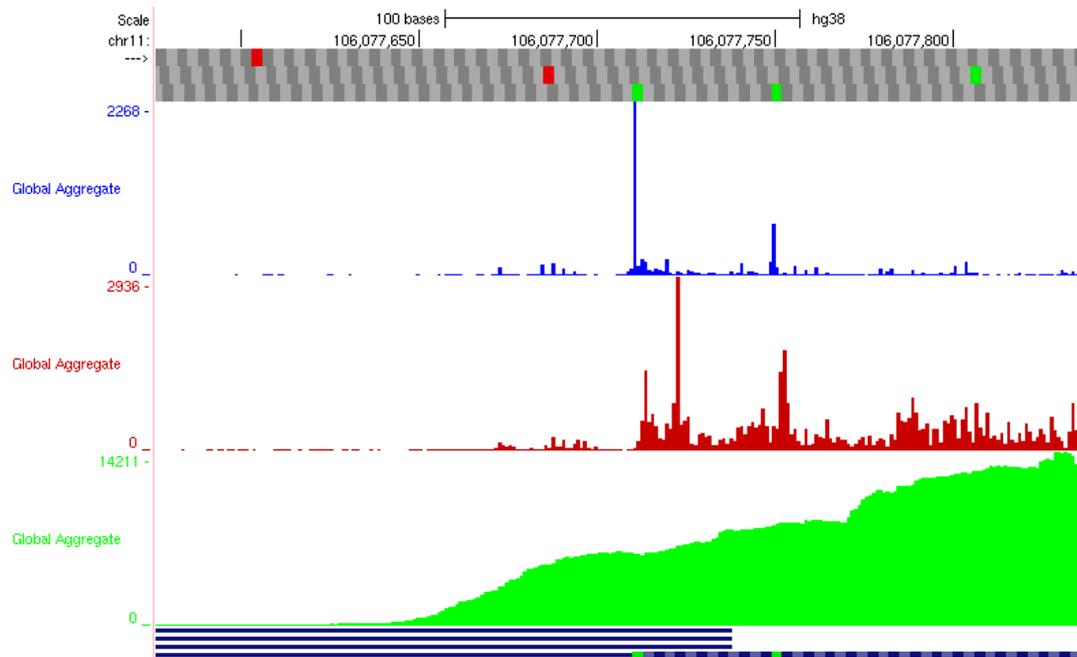
each of the 10 selected genes. **B.** Scans of protein gels used to separate SNAP-tagged protein

products of the test constructs expressed in HeLa cells. **C.** Scans of protein gels used to separate

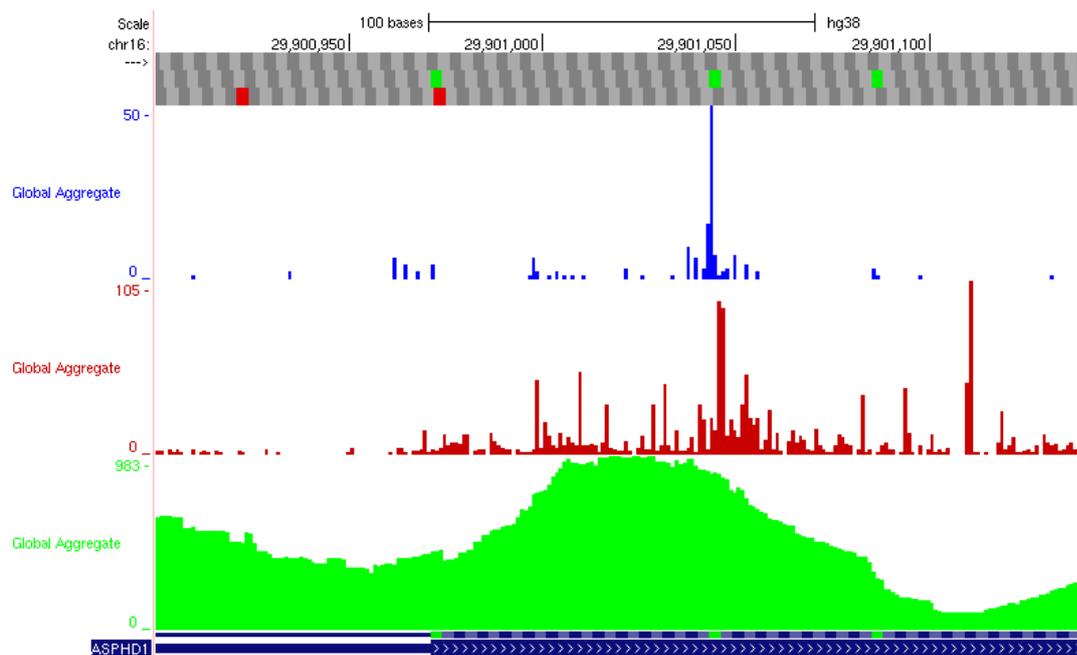
SNAP-tagged protein products of the test constructs expressed in HEK293A cells. Gene names are

shown below the lanes. For *C1orf94* loci in **B** and **C** there is an additional construct **cuu** in which a

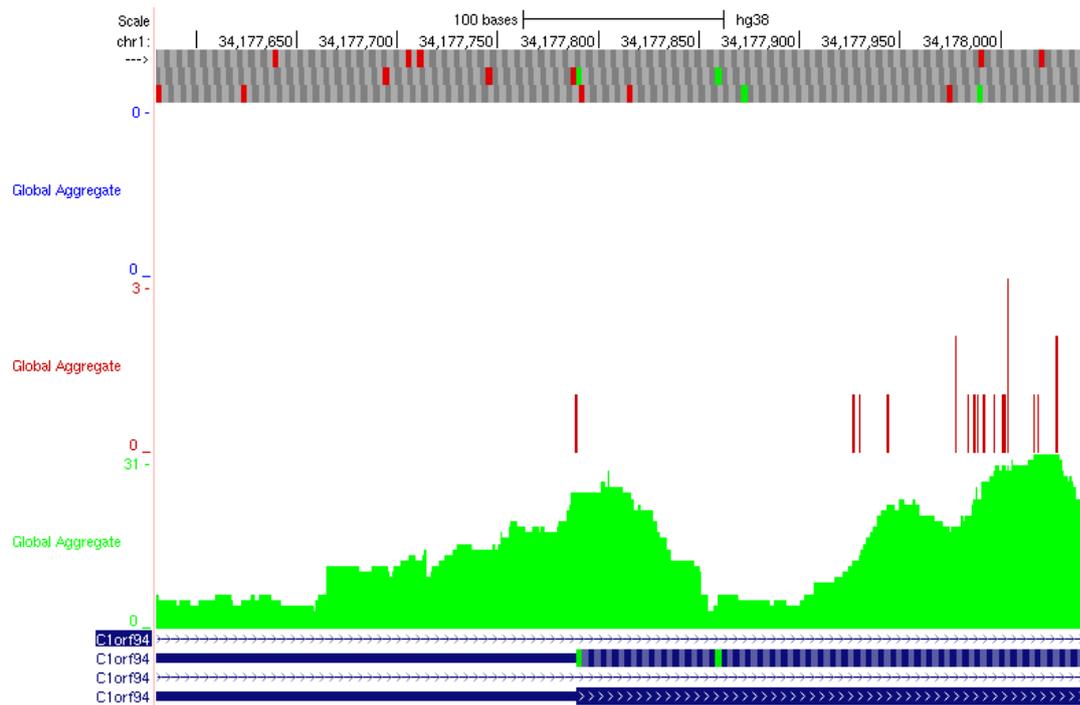
CUG codon between sAUG and fdAUG was changed to a CUU codon.



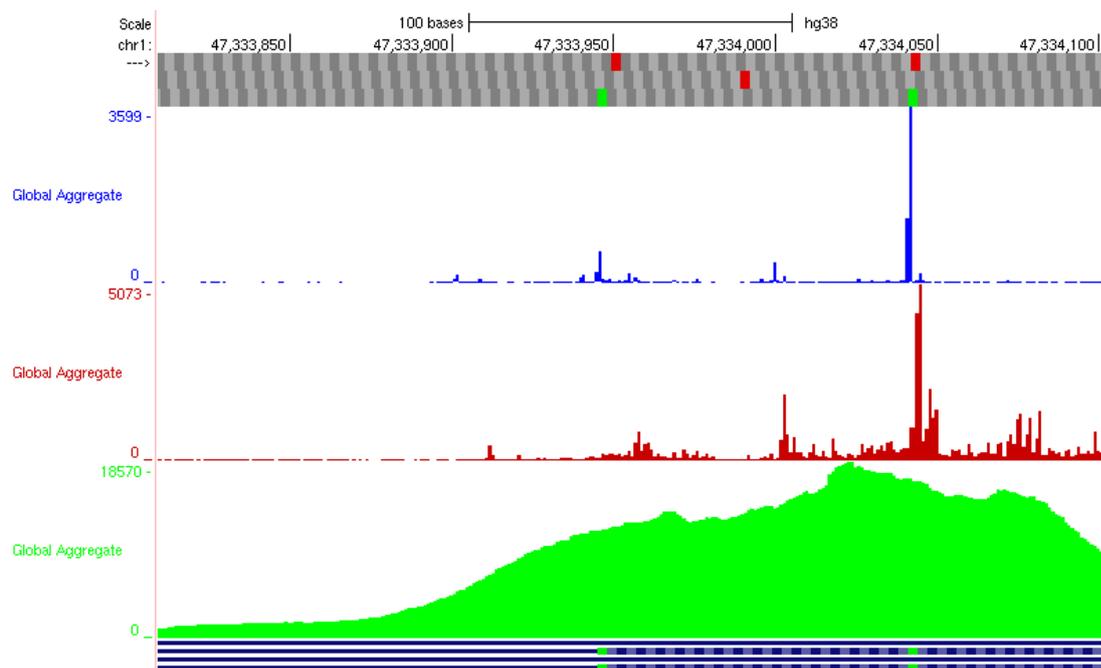
Supplemental Figure S3 | GWIPS-viz screenshot of *AASDHPPT* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



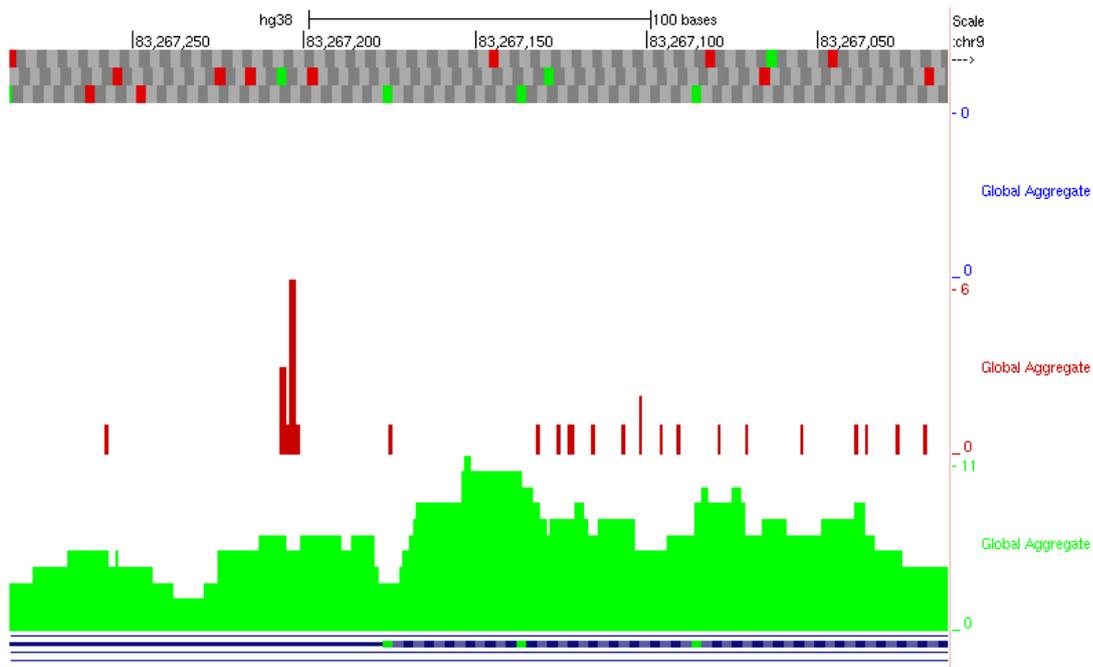
Supplemental Figure S4 | GWIPS-viz screenshot of *ASPHD1* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



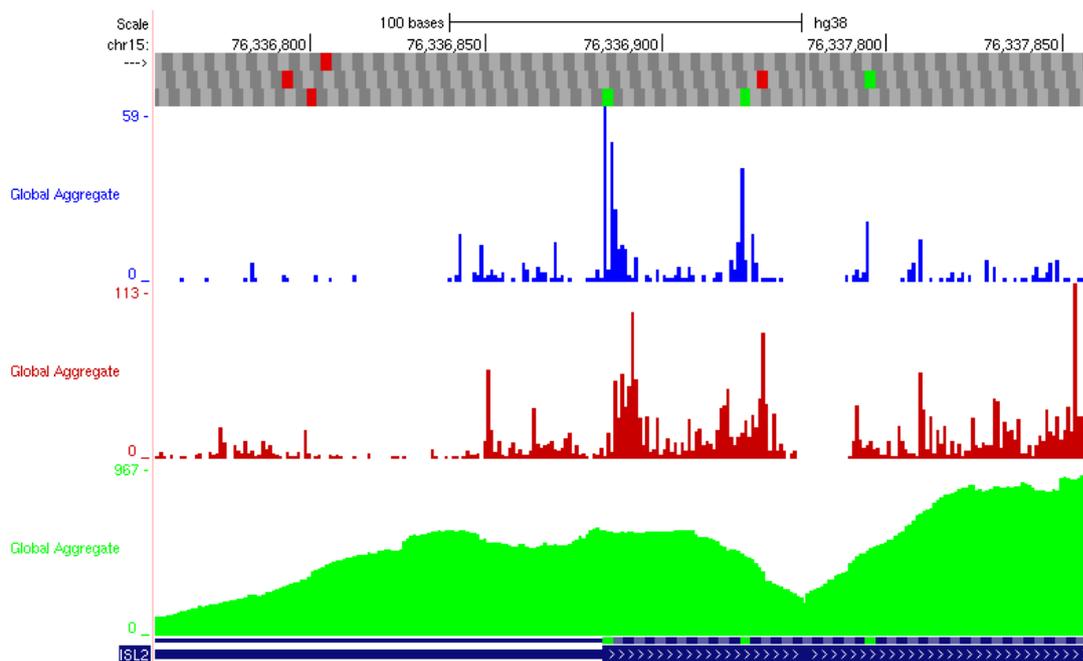
Supplemental Figure S5 | GWIPS-viz screenshot of *C1orf94* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



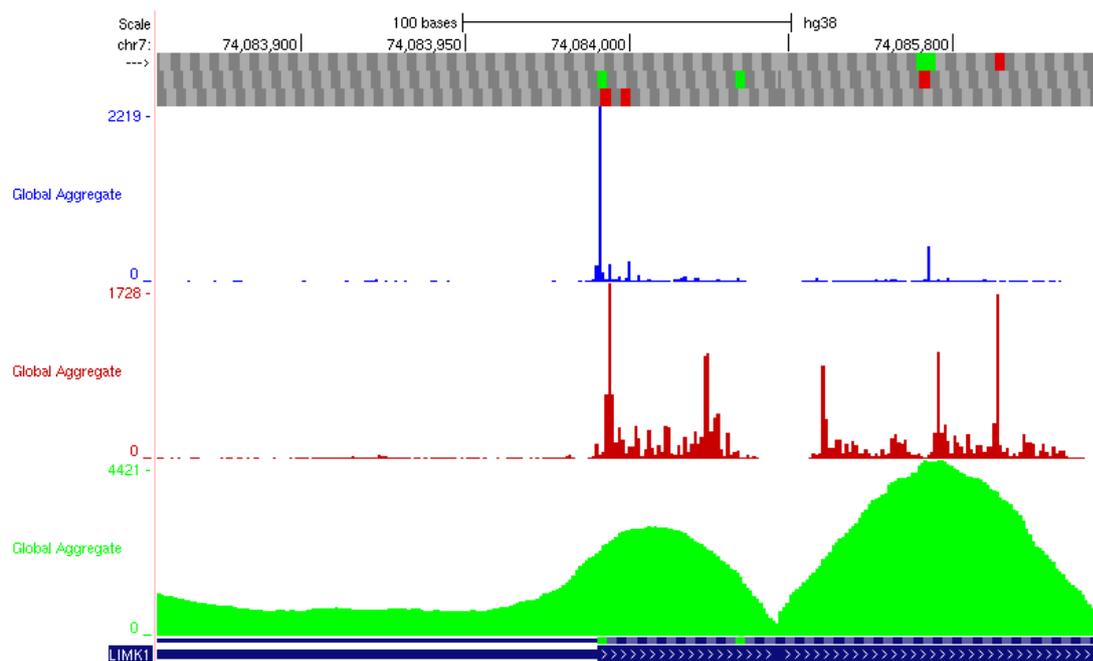
Supplemental Figure S6 | GWIPS-viz screenshot of *CMPK1* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



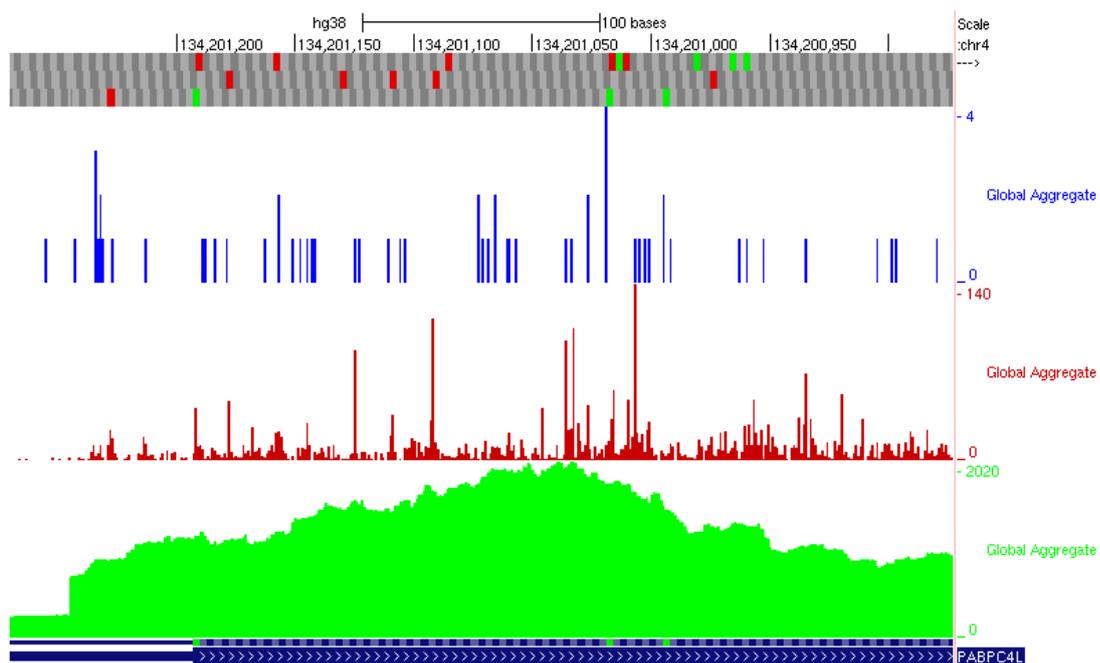
Supplemental Figure S7 | GWIPS-viz screenshot of *FRMD3* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



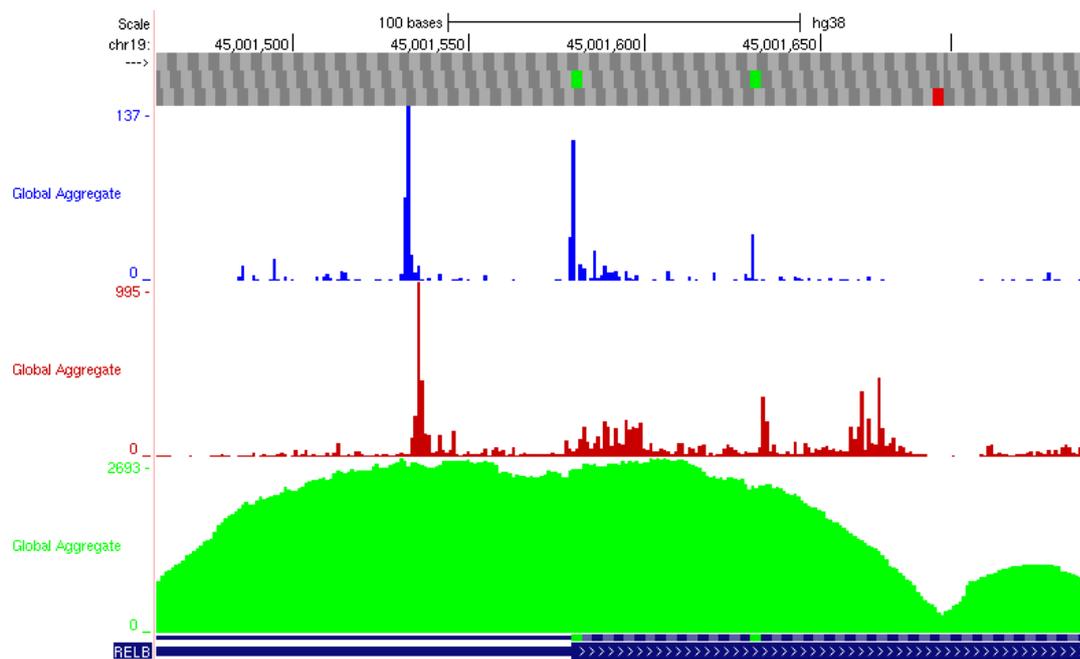
Supplemental Figure S8 | GWIPS-viz screenshot of *ISL2* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



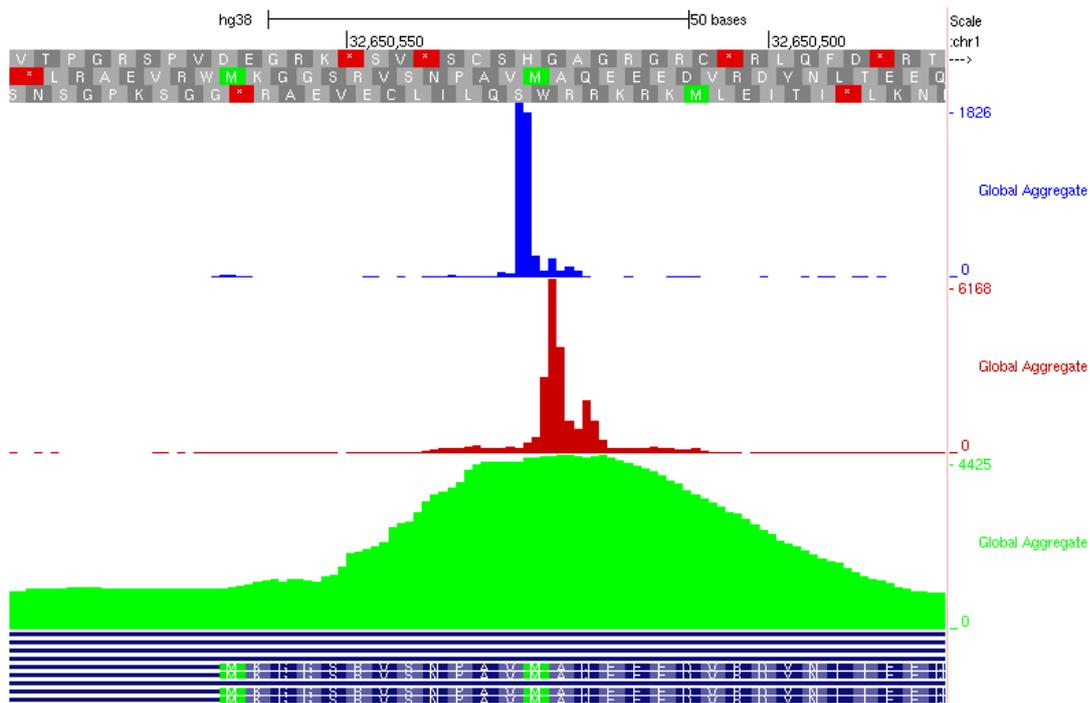
Supplemental Figure S9 | GWIPS-viz screenshot of *LIMK1* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



Supplemental Figure S10 | GWIPS-viz screenshot of *PABPC4L* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.

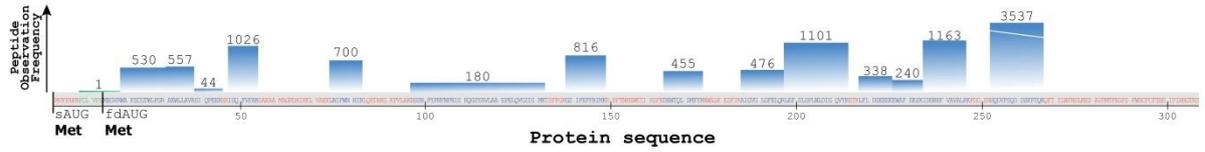


Supplemental Figure S11 I GWIPS-viz screenshot of *RELB* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.



Supplemental Figure S12 | GWIPS-viz screenshot of *ZBTB80S* locus of the reference human genome (hg38 assembly) in the vicinity of sAUG and fdAUG. Tracks of aggregated ribosome profiling data are shown as the following. Blue are the locations of ribosome P-sites from ribosome footprints enriched at sites of initiation. Dark red are the locations of A-sites inferred from footprints of arrested elongating ribosomes. Refseq annotations tracks are shown at the bottom and ORF plots are at the top.

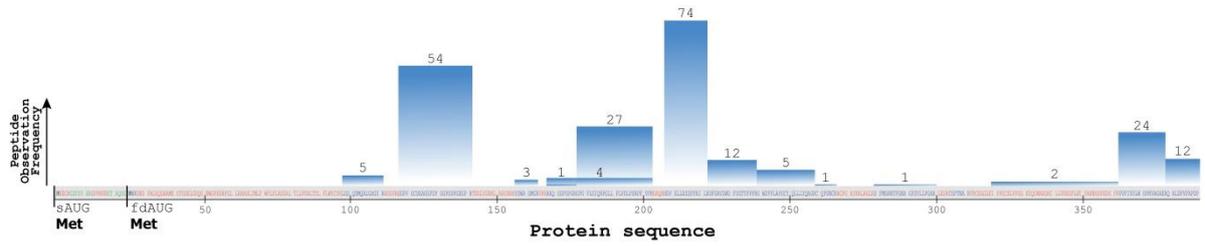
AASDHPPT M V F P A K R F C L V P S M E G V R W A F S C G T W L P S R A E W L L A V R S I Q P E E K E R I G Q F V F A R D A K A A M A G R L M I R K L V A E K L N I P W N H I R L Q R T A K G K P V L A K D S S N 100
 P Y P N F N F N I S H Q G D Y A V L A A E P E L Q V G I D I M K T S F P G R G S I P E F F H I M K R K F T N K E W E T I R S F K D E W T Q L D M F Y R N W A L K E S F I K A I G V G L G F E L Q R L E F 200
 D L S P L N L D I G Q V Y K E T R L F L D G E E E K E W A F E E S K I D E H H F V A V A L R K P D G S R H Q D V P S Q D D S K P T Q R Q F T I L N F N D L M S S A V P M T P E D P S F W D C F C P T E E 300
 I P I R N G T K S



Supplemental Figure S13 I Peptide observation frequency in PeptideAtlas for AASDHPPT protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides.

ASPHD1

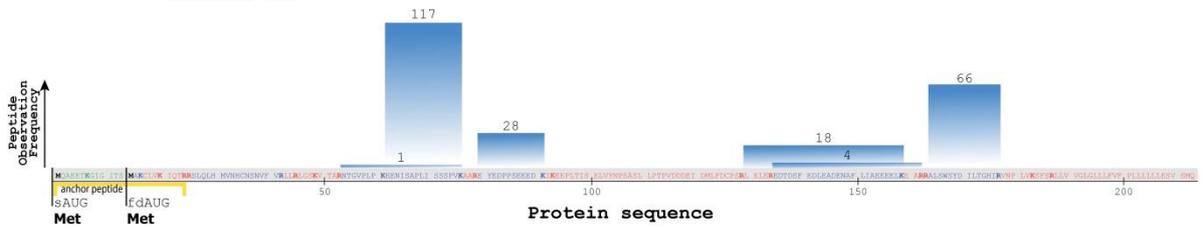
MKEGRSSFSV ERGPRKERET AQSGMVKGNS PAGESGAAME GTGGELGGQG NWPEDAPGL LARASLIMLP WPLPLASSAL TLLFGALTSI FLWICYRLGS 100
 QDMQALGAGS RAGGVRGGPV GCSEAGGPPS GPGDPGEGP RTEGLVSRRL RAYARRYSWA GMGRVRAAQ GPGPGRGPG VLGIQRPGLL FLPDLPSAFF 200
 VPRDAQRHV ELLESSFPAI LRDFGAVSWD FSGTTPPRG WSPPLAPGCY QLLYQAGRC QPSNCRCPG AYRALRGLRS FMSANTFGNA GFSVLLPGAR 300
 LEGRCGPTNA RVRCHLGLKI PPGCELVVGG EPQCWAEGHC LLVDDSLHT VAHNGSPEDG PRVVFIVDLW HPNVAGAERQ ALDFVFADP



Supplemental Figure S14 I Peptide observation frequency in PeptideAtlas for ASPHD1 protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides.

FRMD3

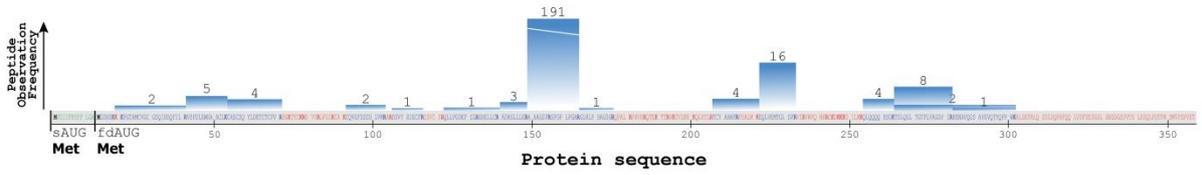
M~~Q~~AEE~~T~~KGIG ITS~~M~~AK~~C~~L~~V~~K IQ~~T~~RRSLQLH MVNHGNSNVF ~~V~~RL~~L~~RLGSKV TARNTGVPLP ~~K~~ENISAPLI SSSPV~~K~~AARE YEDPPSEED ~~K~~I~~K~~EEPTIS 100
 ELVYNPSASL LPTPVDDDEI DMLFDCPS~~R~~L ELER~~E~~EDTDSF EDLEADENAF LIAEEEE~~L~~KE ~~A~~RRALSWSYD ILTGHIRVNP LV~~K~~SFS~~R~~LLV VGLGLLLFVF 200
 PLLLLLLESV SMQ



Supplemental Figure S15 I Peptide observation frequency in PeptideAtlas for *FRMD3* protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides. Anchor peptides predicted by Signal P is marked in orange.

ISL2

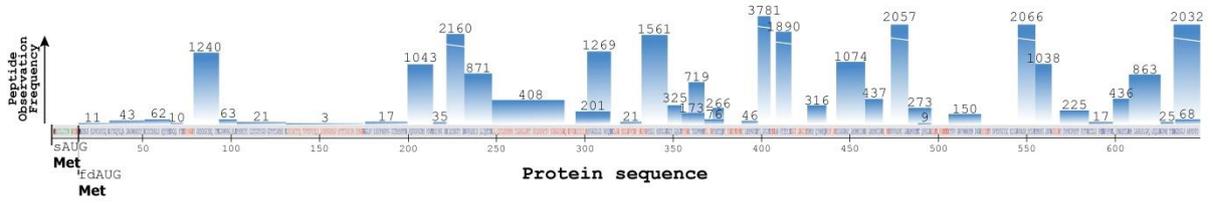
MVDIIFHYPF LGAMGDHSK KFGTAMCVGC GSQIHDQFIL RVSPDLEWHA ACLKCAEC3Q YLDETCTCFV RDGKTYCKRD YVRLFGIKCA KCQVGFSSSD 100
 LVMRARDSVY HIECFRC3VC SRQLLPDDEF SLREHELLCR ADHGLLEERA AAGSPRSPGP LPGARGLHLP DAGSGRQPAL RPHVHKQTEK TTRVRTVLNE 200
 KQLHTLRICY AANFRPDALM KEQLIVMTGL SPRVIRVWFQ NKRCCKDKK3S ILMKQLQQQ HSDKTSLQGL TGTPLVAGSP IRHENAVQGS AVEVOTYQPP 300
 WKALSEFALQ SLDLQPAFQQ LVSFSESGSL GNSSGSDVTS LSSQLPDTN SMVSPVET



Supplemental Figure S16 I Peptide observation frequency in PeptideAtlas for *ISL2* protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides.

LIMK1

MRLLTLLCCTW **REERM**SEEGS ELPVCASCGQ RIYDQYQLQA LNADWHADCF RCDCSASLS HQYYEKDGL FCKKDYWARY GESCHGCSEQ ITRGLVMVAG 100
 ELKYHPECFI CLTCGTFIGD GDTYTLVEHS KLYCGHCYIQ TVVTPVIEQI LPDSPGSHLP HTVTLVSIPA SSHGKRLSV SIDPPHGGPG CGTEHSHTVR 200
 VQGVDPGCMSPDVKNSIHVG DRILEINGTP IRNVPLDEID LLIQETSRL QLTLEHDPHD TLGHGLGPEP SPLSSPAYTP SGEAGSSARQ **KPVLR**SCSID 300
RSPGAGSLGS PASQ**RKDLGR** **SESLRVVCRP** **HRIFR**PSDLI HGEVLCKGCF GQAIKVTHRE TGEVMVMKEL IRFDEETQRT **FLKEVKVMRC** LEHPNVLKFI 400
 GVLV**KDKRLN** FITEYIKGGT **LRGIK**SMSD QYPWQRVVF AKDIASGMAY LHMNIHRD LNSHNCLVRE NKNVVVADFG LARLMVDEKT QPEGLRSL**KK** 500
PDRKKRYTVV GNPYMAPEM INGR**SYDEK**V DVFSFGIVLC EIIGRVNADP DYLPRTMDFG LNVR**GF**LDRY CPPNCPPSPF PITVRCCDLD **PEKRP**SEVKL 600
 EHWLETLRMH LAGHLPLGPQ LEQLDRGFWE TYR**R**GESGLP AHEVVPD



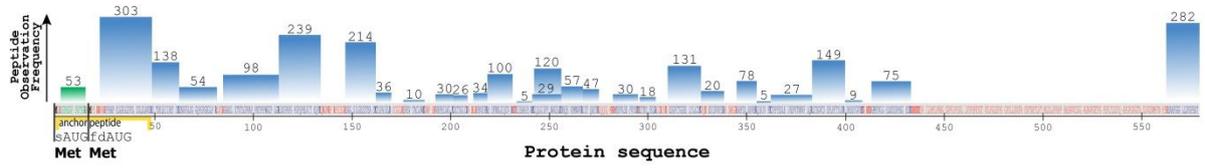
Supplemental Figure S17 | Peptide observation frequency in PeptideAtlas for *LIMK3* protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides.

RELB

```

M L R S G P A S G P S V P T G R M P S R R V A R P P A A P E L G A L G S P D L S S L S L A V S R S T D E L E I I D E Y I K E N G F G L D G G Q P G P G E G L P R L V S R G A A S L S T V T L G P V A P 100
P A T P P P W G C P L G R L V S P A P G P Q P H L V I T E Q F K Q R G M R F R Y E C E G R S A G S I L G E S S T E A S K T L P A I E L R D C G G L R E V E V T A C L V N K D W P H R V H P H S L V G 200
K Q C I D G I C R V R L R P H V S P R H S F N N L G I Q C V R K K E I E A A I E R K I Q L G I D P Y N A G S L K N H Q E V D M N V V R I C F Q A S Y R D Q Q G Q M R R M D P V L S E P V Y D K K S T N T 300
S E L R I C R I N K E S G P C T G G E E L Y L L C D K V Q K E D I S V V F S R A S W E G R A D F S Q A D V H R Q I A I V F K T P P Y E D L E I V E P V T V N V F L Q R L T D G V C S E P L P F T Y L P R 400
D H D S Y G V D K K R K R G M F D V L G E L N S S D P H G I E S K R R K K K P A I L D H F L P N H G S G P F L P P S A L L P D P D F F S G T V S L P G L E P P G G P D L L D D G F A Y D P T A P T L E T 500
M L D L L P P A P P H A S A V V C S G G A G A V V G E T P G P E P L T L D S Y Q A P G P G D G G T A S L V G S N M F P N H Y R E A A F G G G L L S P G P E A T

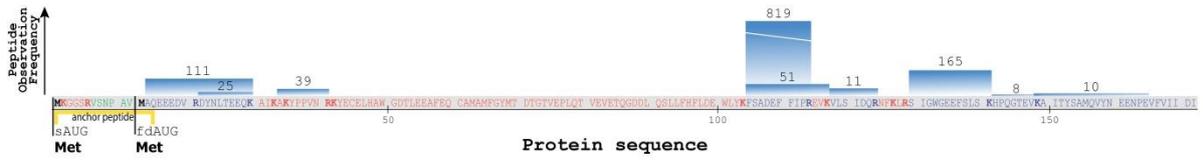
```



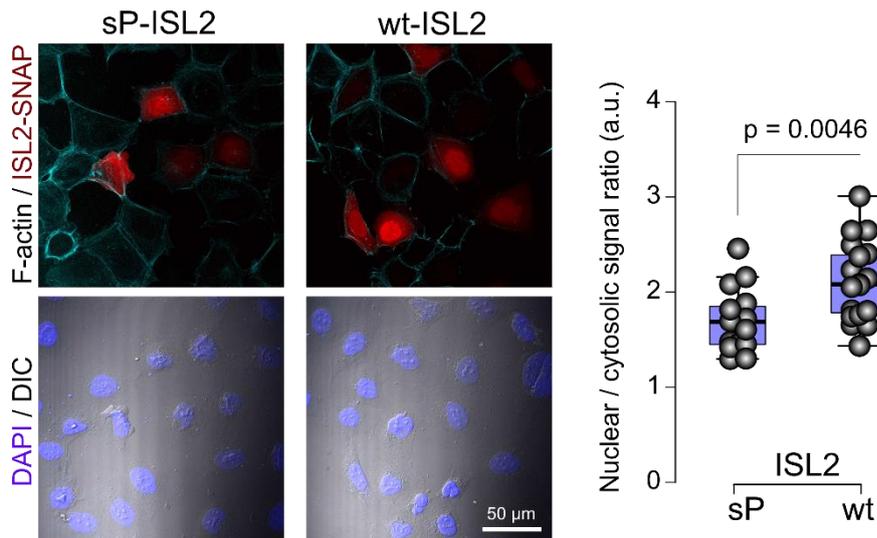
Supplemental Figure S18 | Peptide observation frequency in PeptideAtlas for RELB protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides.

ZBTB80S

MKGGSRVSNP AVMAQEEEDV RDYNLTTEEQK AIKAKYPPVN RKYECELHAW GDTLEEAFEQ CAMAMFGYMT DTGTVEPLQT VEVETQGGDL QSLLFHFLDE 100
WLYKFSADEF FIPREVKVLS IDQRNFKLRS IGWGEEFSL S KHPQGTEVKA ITYSAMQVYN EENPEVFVII DI



Supplemental Figure S19 | Peptide observation frequency in PeptideAtlas for ZBTB80S protein products. In red are the tryptic peptides not expected to be detectable. In green are detectable peptides unique for the long proteoform and in blue are shared peptides. Anchor peptide predicted by Signal P is marked in orange.



Supplemental Figure S20 I Localization of *ISL2-SNAP* derived products in cells. *ISL2-SNAP* (red, stained with SNAP-Cell® 647-SiR) is counterstained with Alexa Fluor 546 phalloidin (cyan) and DAPI (blue). Images represent a single plane (DIC) and stacks of five focal planes taken with a 0.5 μm step (fluorescence). In the right panel, relative proportion of *ISL2* P-SNAP (N = 16) and *ISL2* WT-SNAP (N = 19) products localized to the nuclei was cross-compared using *t*-test.