SUPPLEMENTAL MATERIAL for "Observation of dually decoded regions of the human genome using ribosome profiling data."

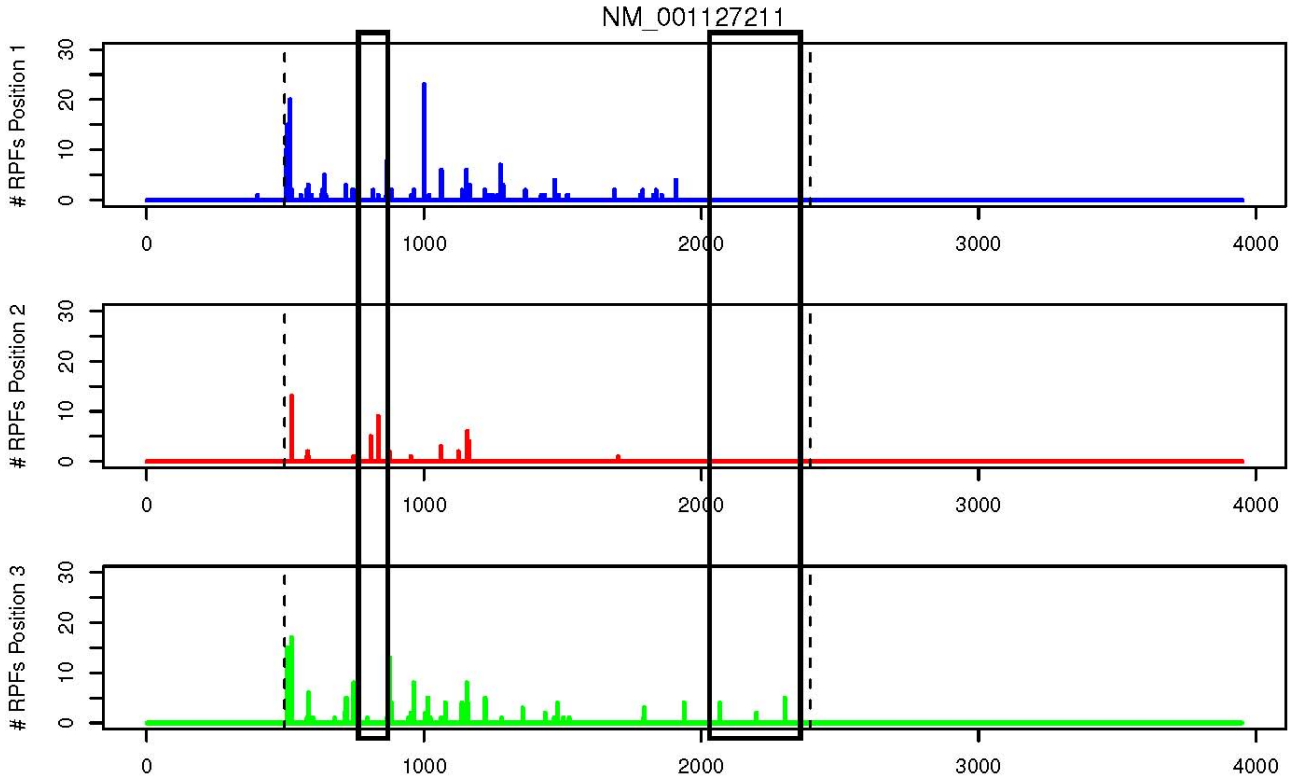
by

Audrey M. Michel[1], Kingshuk Roy Choudhury[2], Andrew E. Firth[3], Nicholas T. Ingolia[4], John F. Atkins[5,6], Pavel V. Baranov[1].*

List of items:

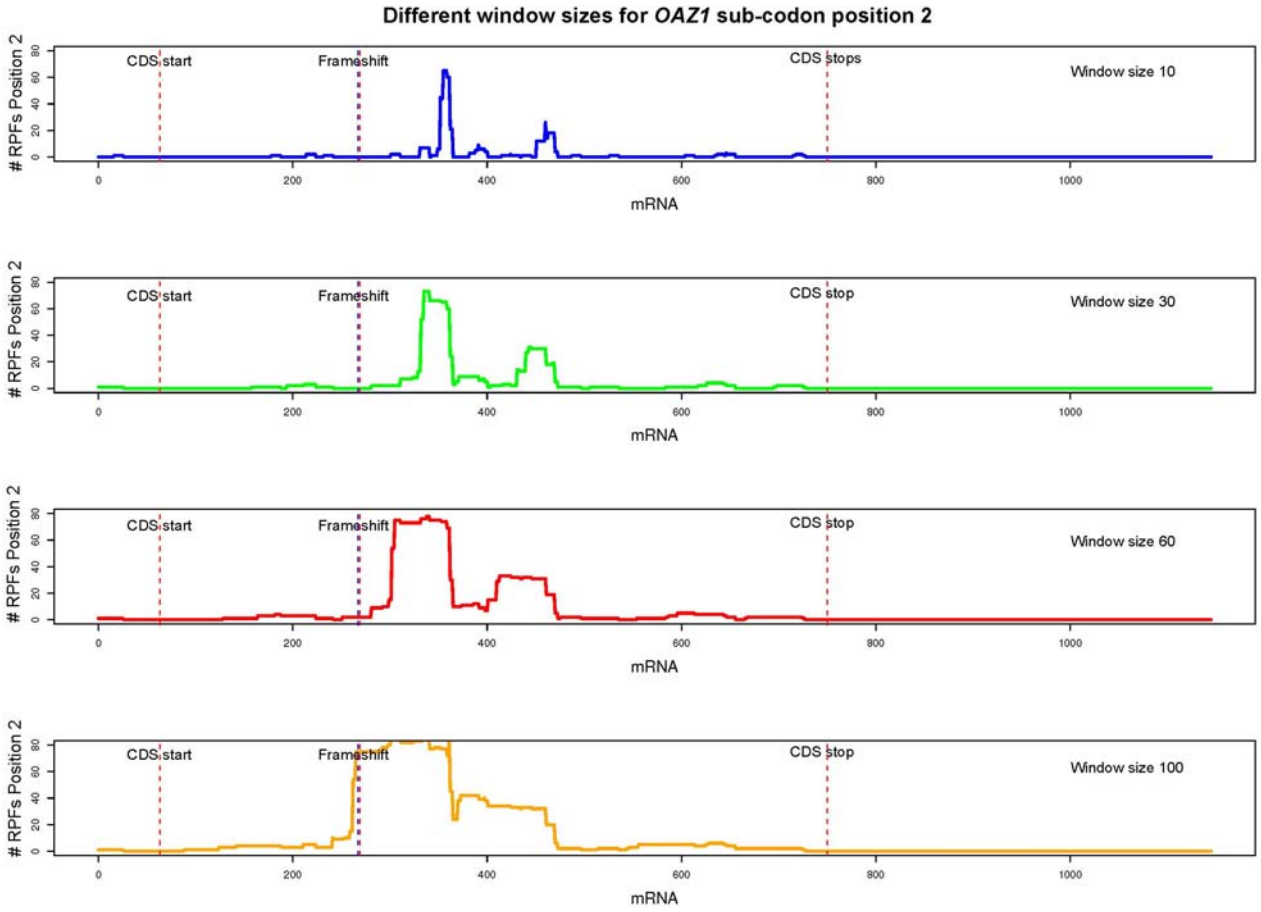
Supplemental Figure 118	MLOGD coding likelihood and synonymous site conservation for *THAP7*
Supplemental Figure 119	MLOGD coding likelihood and synonymous site conservation for *C11orf48*
Supplemental Figure 120	MLOGD coding likelihood for *PHPT1*
Supplemental Figure 121	Genomic multiple codon alignments for *NPAS2*
Supplemental Figure 122	Genomic multiple codon alignments for *THAP7*
Supplemental Figure 123	Genomic multiple codon alignments for *C11orf48*
Supplemental Figure 124	Genomic multiple codon alignments for *PHPT1*
Supplemental Discussion	False positives, unexplained profiles, staccato reads and potential ambiguity in the functional classification of dually coded genes.
R scripts	R scripts for computing the CSCPD and PTS

Supplemental Figure 1: Limitations of a sliding window approach



This example illustrates the unsuitability of a sliding window approach to monitor the transition of the lowest proportion of RPFs from one sub-codon position to another. Depending on the window size, two sub-codon positions may have no RPFs (large window) or sub-codon position 2 could locally have the highest number of RPFs (small window) while overall having the lowest proportion of RPFs. This non-uniformity of RPF distribution poses problems for detecting reading frame transitions using a sliding window approach.

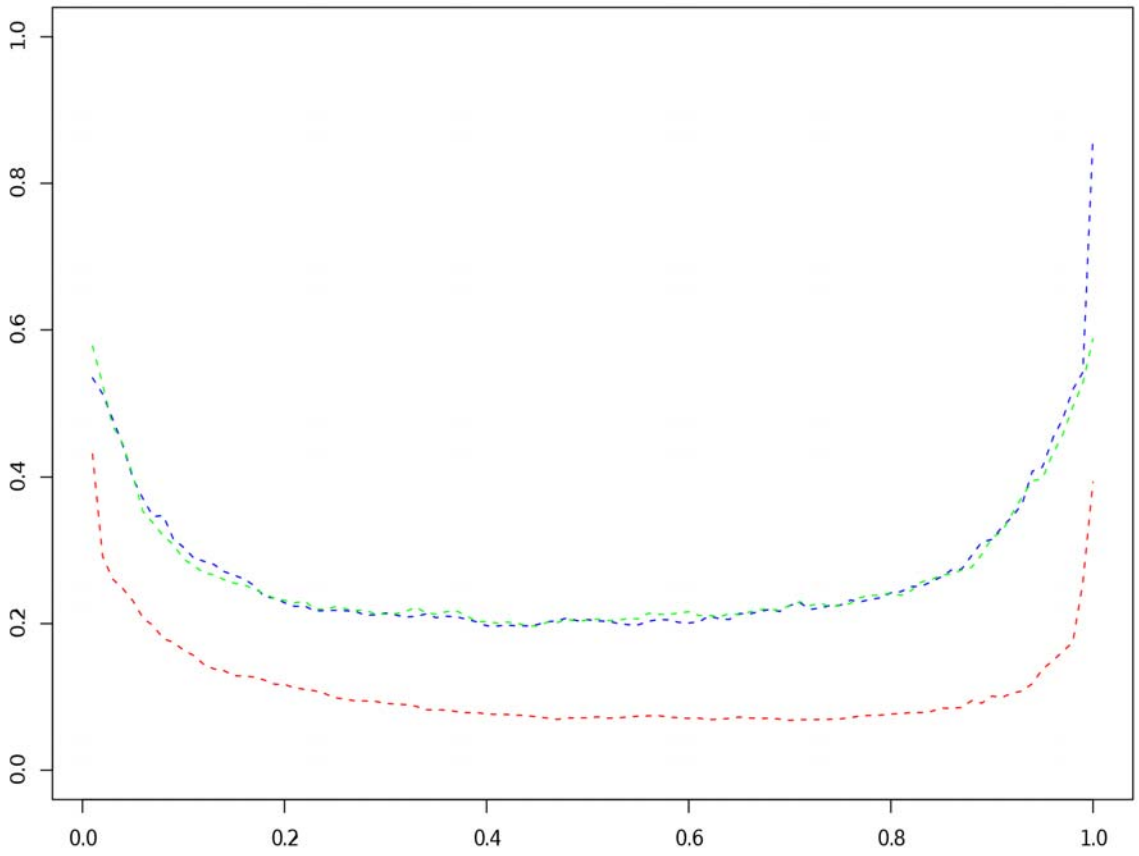
Supplemental Figure 2: Investigation of a sliding window approach on *OAZ1*



Sliding windows of different sizes (10 to 100nt) were applied to the RPFs in sub-codon position 2 for human ornithine decarboxylase antizyme 1 (*OAZ1*). Antizyme 1 is known to utilise +1 programmed ribosomal frameshifting in its expression and its sub-codon profile shows that there is an increase in RPFs aligning to sub-codon position 2 3' of the frameshift site (see Fig. 1C in the main text). However, the above illustrates that using a sliding window approach on sub-codon position 2 to detect framing transitions is impractical as there are numerous local peaks in the region of the frameshift site for the different window sizes. With small windows, there are several local maximums corresponding to the second proportion. Also while the first transition (elevation of RPF density in the 2nd sub-codon position) corresponds to frameshifting, the second transition (decrease of RPF density) cannot be clearly explained and is likely due to local stochastic fluctuations in the

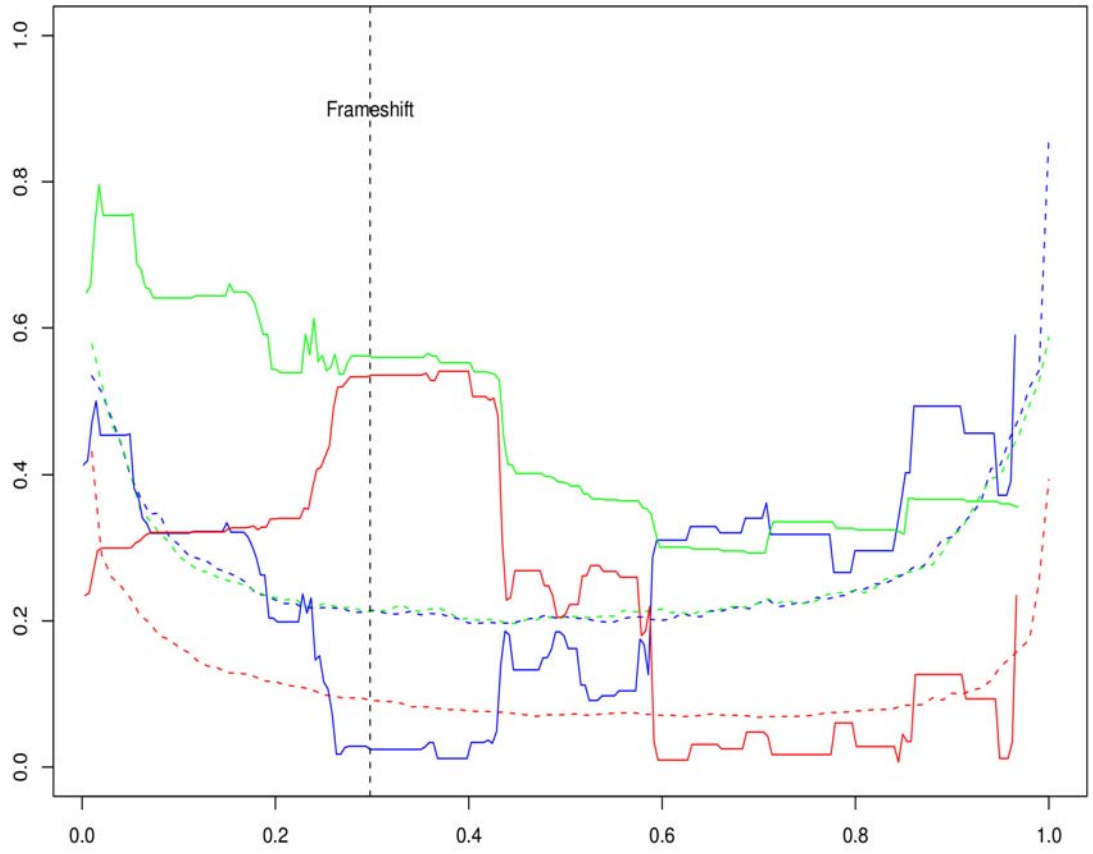
number of RPFs in the 3' end of CDS, where the overall density of RPFs is low, see Figure 1C of the main text.

Supplemental Figure 3: Interpolated 95th percentile CSCPD scores relative to CDS position



Interpolated 95th percentile CSCPD scores (vertical axis) relative to CDS position (horizontal axis) obtained from the analysis of the 1000 mRNA transcripts with the highest number of RPFs. For uniformity, relative positions of CDS are used (with total length of CDS as 1.0) Blue, red and green dotted lines correspond to sub-codon positions 1, 2 and 3, respectively.

Supplemental Figure 4: PTS plot for human Antizyme 1 (*OAZ1*) mRNA



PTS plot for human Antizyme 1 (*OAZ1*) mRNA whose translation is known to involve programmed ribosomal frameshifting. The frameshift site is indicated with a vertical interrupted line. Solid lines show the CSCPD scores for each sub-codon position in *OAZ1*. The dotted lines show the 95th percentiles (of the CSCPDs for the highest-coverage 1000 mRNAs) as in Supplemental Figure 3. The CSCPD scores for sub-codon positions 2 and 3 greatly exceed their corresponding 95th quantile interval. The area where *OAZ1* CSCPD scores exceed their 95th quantile intervals constitute the PTS score.

Testing PTS on simulated dual coding sequences.

To test PTS performance we decided to perform a number of simulations of dual coding regions.

We tested two types of dual coding, first those that occur due to frame transitions such as in the case of programmed ribosomal frameshifting and second, those that occur due to the translation of the same genomic region in two different reading frames (either from the same or from different transcript isoforms).

(1) Simulations for frame transitions that occur such as in programmed ribosomal frameshifting

Artificial 3000nt long sequences representing RPF data were generated (Supplemental Fig. 5), assuming the following:

i) The proportions of RPFs in the three sub-codon positions are a_1=0.45, a_2=0.1 and a_3=0.45.

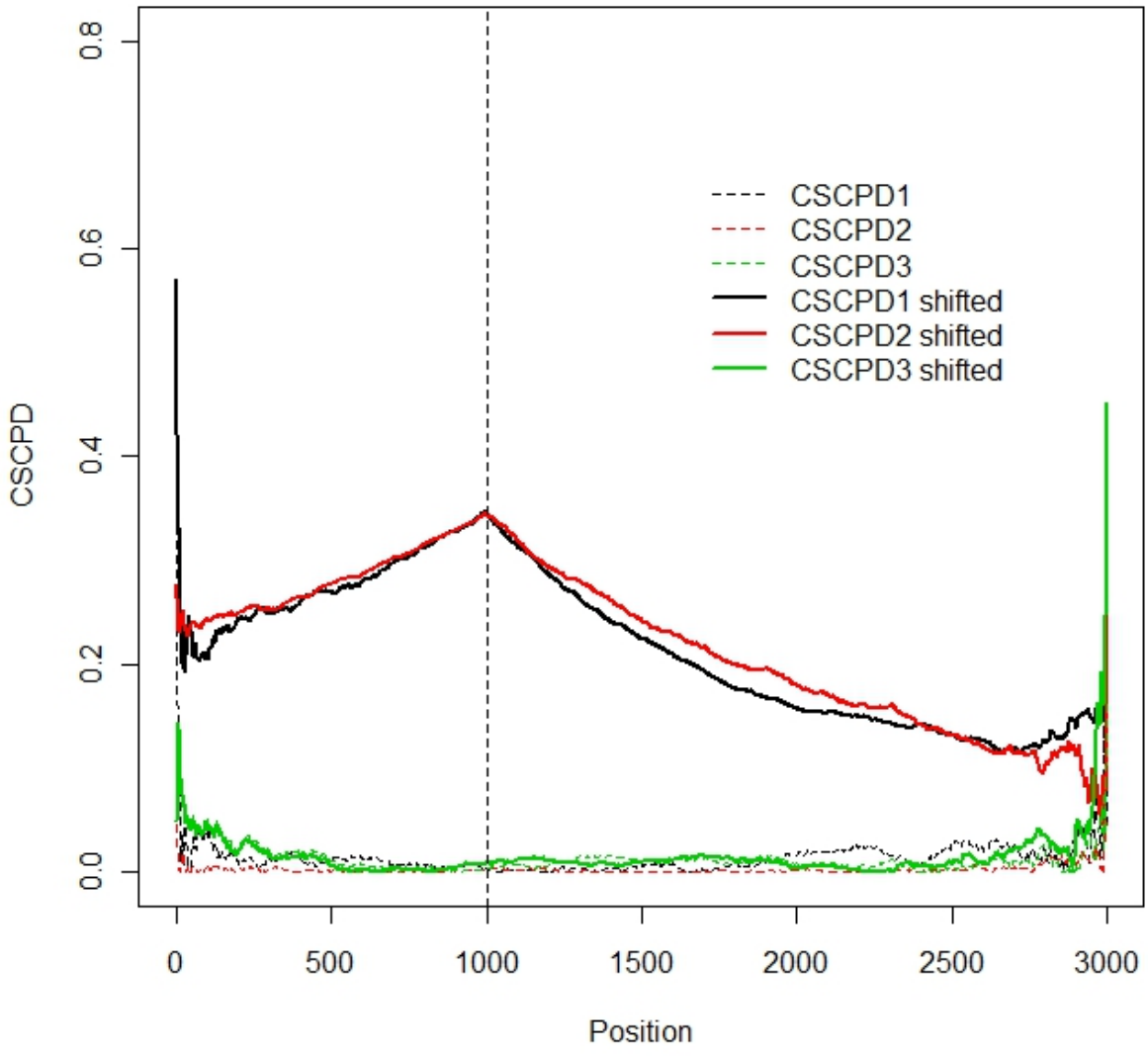
ii) RPFs at each position follow a Poisson distribution with a mean μa_i (i=1,2,3), μ=20.

A frameshift was created by eliminating coordinate 1001. The CSCPD scores at each sub-codon position were computed for the sequences with and without a frameshift. It can be clearly seen that with the exception of the 5' and 3' ends of CDS, the CSCPD score is close to 0 for all original sequences. For the sequences with a frameshift, the CSCPD score for position 3 is unchanged (because its proportion remains 0.45 throughout even after deletion), but there is a large increase in the CSCPD score for positions 1 and 2, whose proportions undergo a change downstream of the frameshift. In addition, the CSCPD scores for positions 1 and 2 are at their maximum at the location of the deletion.

To examine how the CSCPD score would behave in real mRNAs with and without simulated reading frame transitions, we generated CSCPD plots for mRNAs with artificially introduced frameshifts (either a deletion or an insertion of a nucleotide introduced in the middle of the corresponding CDS). Examples of CSCPD plots with and without an artificial frameshift for *SMARCA1* mRNA (RefSeq ID NM_003069) are shown in Supplemental Figures 6 and 7, respectively.

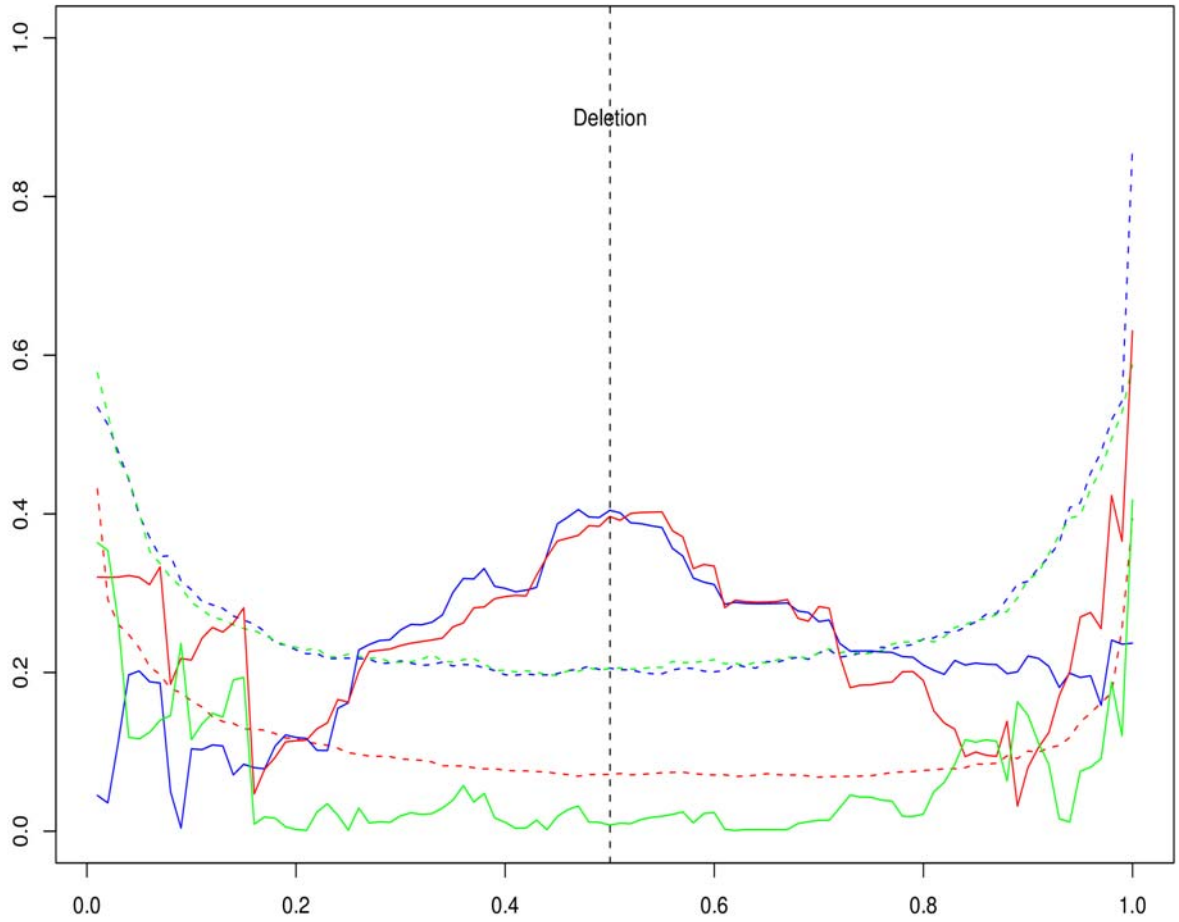
In Supplemental Figures 6 and 7, the deletion causing the artificial frameshift was introduced in the middle of the coding region of the real mRNA. To explore how the position of a frameshift affects PTS, we introduced artificial frameshifts at different locations of the coding region of real mRNAs (Fig 2D main text). As can be seen in Figure 2D, PTS performance correlates negatively with the distance of the frameshift from the middle of the CDS.

Supplemental Figure 5: CSCPD scores for artificial mRNAs with and without a simulated frameshift



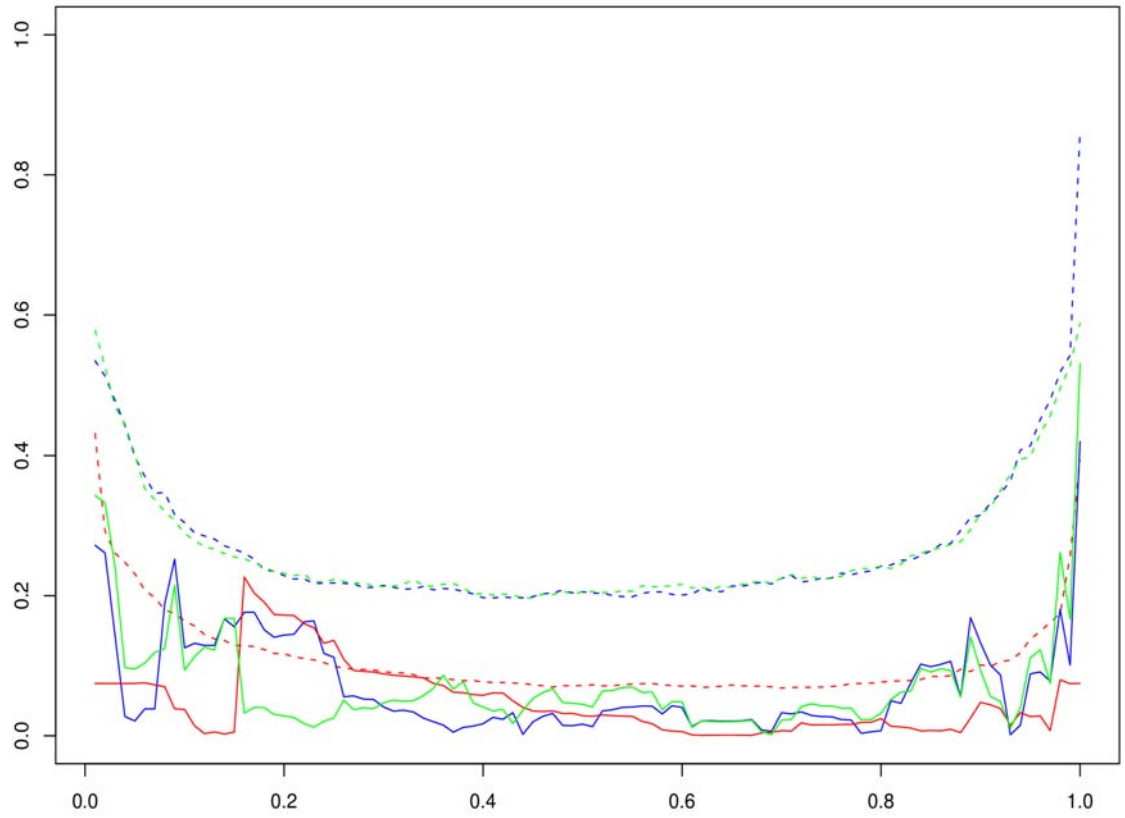
CSCPD scores for artificial 3000nt long mRNAs for each sub-codon position with and without a deletion. It can be seen that except at the beginning and the end, the CSCPD scores were close to 0 for all the sequences without a deletion (broken black, red and green lines). For the sequences with a deletion, the CSCPD score for position 3 is unchanged (solid green line), but there is a large increase in the CSCPD score for positions 1 and 2 (solid black and red lines), and their maximum CSCPD is at the location of the deletion.

Supplemental Figure 6: CSCPD scores for *SMARCA1* mRNA with an artificial deletion in the middle of its CDS



CSCPD scores for *SMARCA1* mRNA (RefSeq ID NM_003069) with an artificial deletion in the middle of its CDS. Solid lines show CSCPD scores, dotted lines show the 95th percentiles (of the CSCPDs for the highest-coverage 1000 mRNAs) as in Supplemental Figure 3.

Supplemental Figure 7: CSCPD scores for *SMARCA1* mRNA with no artificial deletion

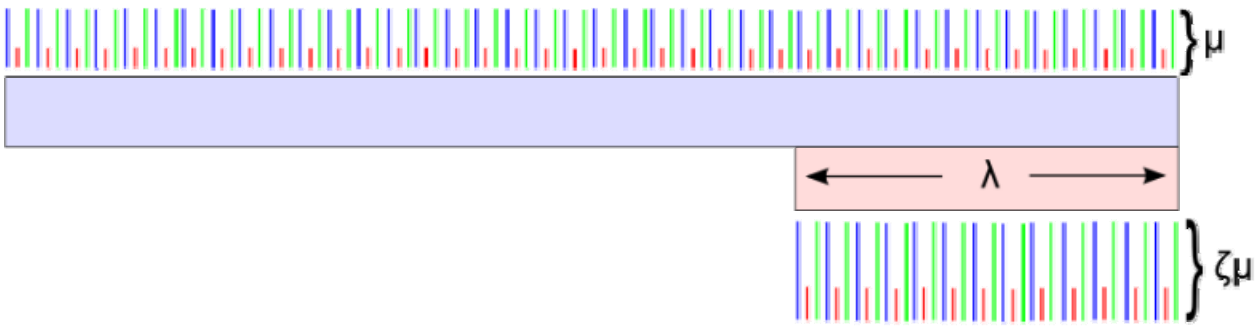


Same as Supplemental Figure 6, but without an artificial frameshift.

(2) Frame transitions that occur due to the translation of the same genomic region in two different reading frames

We examined how PTS performs on simulated dual coding sequences with varying parameters (Fig.2E-F of the main text and Supplemental Fig. 8). We generated 3000nt long artificial mRNA sequences as described in the previous section using a Poisson distribution to mimic RPF data, with a mean μa_i (i=1,2,3), where are a_1=0.45, a_2=0.1 and a_3=0.45. The simulations were carried out for three different RPF densities with μ=1, 10, 20 (corresponding to 0.3 RPFs/coordinate, 6 RPFs/coordinate. 9 RPFs/coordinate). To mimic dual coding, RPFs for a second (alternative) CDS were generated and superimposed onto the original mRNA RPF profile to yield a partial overlap of a length λ. The second CDS was generated using a Poisson distribution with a mean $\zeta\mu a_i$ (i=1,2,3). We explored the effect of the overlap length λ when $\zeta = 1$ (Fig. 2E). PTS reaches a maximum when the length of the dual coding region is similar to the length of the single coding region. We also explored the effect of an alternative CDS RPF density factor increase, ζ, on the PTS score (Fig. 2F). It can be seen that PTS positively correlates with the level of translation of the alternative CDS.

Supplemental Figure 8: Scheme of simulated dual coding mRNAS.



The main ORF (CDS) is shown as a blue box. The alternative frame is shown as a pink box. μ corresponds to the RPF density generated by ribosomes translating the main frame. ζμ is the RPF density at the alternative frame. The RPF densities for individual coordinates corresponding to the same sub-codon positions are shown at the same height for simplicity.

Estimating p-values for PTS

To estimate the p-value for a PTS of 10 and higher, for each sub-codon profile we permuted the RPF locations of the top 1000 RPF covered mRNA 1000 times generating 1,000,000 artificial mRNA profiles for which we calculated the PTS scores. This gave a p-value of 0.057. We then eliminated cases where the 1st and the 3rd sub-codon positions contributed to a high PTS but not the 2nd position (see explanation in the "Further refinements of PTS" section in the main manuscript). We then divided the remaining number of permuted transcripts that had a PTS of 10 and higher by the remaining total number of artificial mRNA profiles. This resulted in a re-estimated p-value of 0.0084.

Supplemental Data: summary of dual coding candidates.

108 candidates identified in our study are summarized in Supplemental Table 1 below and classified into six categories indexed in column 3 and coloured as follows.

1 (yellow) – known instances of programmed ribosomal frameshifting

2 (red) – non-upstream overlapping or nested nORFs

3 (pink) – regulatory uORFs overlapping the main pORF

4 (green) – dual coding due to a frame transition caused by alternative splicing or alternative transcription events

5 (gray) – instances where sub-codon profiles show dual coding behaviour but cannot be explained either by the mRNA ORF organization or by differences among alternatively spliced variants.

6 – (white) false positives where a high PTS was obtained for sub-codon profiles that do not exhibit dual coding behaviour.

Supplemental Table 1: 108 candidates with high PTS

RefSeq mRNA ID	Gene Name	Index	PTS	No. RPFs location 2	Corrected PTS
NM_024099	*C11orf48*	4	51.04	23	48.17
NM_000421	*KRT10*	4	49.12	14	44.62
NM_030573	*THAP7*	3	48.59	15	41.95
NM_004010	*DMD*	6	37.88	16	39.86
NM_004152	*OAZ1*	1	39.37	26	38.24
NM_001135861	*PHPT1*	4	35.48	17	36.95
NM_001002244	ANAPC11	4	62.00	13	36.49
NM_001172437	*PEG10*	1	36.27	70	34.80
NM_032237	*SGK196*	5	40.23	18	33.64
NM_006868	*RAB31*	3	33.52	19	28.25
NM_001011667	*CHCHD7*	4	37.79	12	26.60
NM_003721	*RFXANK*	5	22.87	13	25.87
NM_001025248	*DUT*	6	24.83	18	25.67
NM_016556	*PSMC3IP*	3	28.33	16	25.26
NM_152653	*UBE2E2*	3	26.39	21	24.90
NM_001032280	*TFAP2A*	5	23.55	12	24.55
NM_024947	*PHC3*	3	25.90	12	23.91
NM_005001	*NDUFA7*	6	29.32	12	23.56
NM_014761	*IST1*	4	22.26	44	22.53
NM_138558	*PPP1R8*	4	34.32	12	22.24
NM_181353	*ID1*	4	20.14	25	21.45
NM_001136262	*ATXN7L3B*	3	17.63	12	20.65
NM_016828	*OGG1*	3	25.60	12	20.04
NM_001042510	*ZNF706*	3	21.04	22	19.76
NM_032231	*FAM96A*	3	25.24	14	19.27
NM_022044	*SDF2L1*	4	16.56	17	18.13
NM_016468	*COX16*	3	29.33	15	17.98
NM_001003408	*ABLIM1*	6	58.80	13	16.97
NM_002346	*LY6E*	3	16.26	24	16.96
NM_002668	*PLP2*	3	26.49	17	16.62
NM_001001330	*REEP3*	6	16.25	14	16.23
NM_020548	*DBI*	6	14.98	23	16.04
NM_002518	*NPAS2*	2	28.94	15	15.50
NM_018838	*NDUFA12*	2	16.10	18	15.48
NM_004849	*ATG5*	5	38.01	17	15.14
NM_024066	*ERI3*	6	18.81	22	14.96
NM_014153	*ZC3H7A*	3	18.04	24	14.50
NM_001145678	*KIAA0825*	5	25.42	12	13.90
NM_203413	*C17orf81*	3	15.87	20	13.90
NM_006965	*ZNF24*	2	12.69	15	13.65
NM_053052	*SNAP47*	6	18.20	18	13.65
NM_005698	*SCAMP3*	3	12.33	13	13.53
NM_007145	*ZNF146*	5	15.47	18	13.39
NM_198400	*NEDD4*	6	15.11	14	13.17
NM_001065	*TNFRSF1A*	2	17.98	35	12.89
NM_001033112	*PAIP2*	3	22.57	15	12.74
NM_001037540	*SCML1*	5	17.37	19	12.66
NM_001080497	*MEGF9*	3	12.61	19	12.55

NM_006856	*ATF7*	3	19.33	12	12.22
NM_022455	*NSD1*	4	10.11	14	12.19
NM_004322	*BAD*	4	59.84	15	11.99
NM_024348	*DCTN3*	4	35.95	17	11.94
NM_006808	*SEC61B*	6	16.63	22	11.79
NM_001031806	*ALDH3A2*	5	16.03	23	11.70
NM_012235	*SCAP*	6	22.11	15	11.68
NM_018983	*GAR1*	6	11.02	18	11.57
NM_031892	*SH3KBP1*	6	12.14	13	11.33
NM_001006109	*C3orf37*	3	11.77	16	11.23
NM_015085	*RAP1GAP2*	5	32.25	12	10.57
NM_020182	*PMEPA1*	3	11.23	13	10.50
NM_006936	*SUMO3*	6	10.72	15	9.96
NM_004635	*MAPKAPK3*	3	14.40	16	9.96
NM_153273	*IP6K1*	5	12.93	14	9.90
NM_001163790	*PTBP3*	5	11.46	22	9.88
NM_002250	*KCNN4*	3	17.91	12	9.67
NM_001127211	*KIAA1598*	6	11.30	14	9.42
NM_018179	*ATF7IP*	6	11.74	13	9.39
NM_005837	*POP7*	3	27.14	19	9.30
NM_019026	*TMCO1*	3	11.14	19	9.21
NM_004846	*EIF4E2*	2	11.44	34	9.02
NM_000169	*GLA*	2	11.33	40	8.79
NM_000285	*PEPD*	6	10.85	14	7.96
NM_020382	*SETD8*	2	11.93	16	7.95
NM_006701	*TXNL4A*	6	10.82	15	7.79
NM_018268	*WDR41*	6	11.63	18	7.76
NM_007350	*PHLDA1*	6	11.64	35	7.71
NM_016558	*SCAND1*	4	16.03	14	7.49
NM_003860	*BANF1*	3	10.14	16	7.38
NM_002599	*PDE2A*	6	10.23	90	7.38
NM_018361	*AGPAT5*	6	10.10	14	7.35
NM_001134479	*LRRC8D*	2	18.04	12	7.25
NM_199229	*RPE*	6	15.17	15	6.98
NM_000883	*IMPDH1*	4	16.51	28	6.82
NM_052848	*CCDC97*	6	12.35	13	6.74
NM_031299	*CDCA3*	6	11.03	13	6.30
NM_018170	*RPRD1A*	2	11.86	18	6.18
NM_006764	*IFRD2*	2	11.95	24	5.59
NM_001098801	*FAM210A*	5	11.71	15	5.57
NM_172373	*ELF1*	6	10.50	17	5.45
NM_002695	*POLR2E*	2	12.23	24	5.29
NM_005094	*SLC27A4*	3	12.10	16	5.17
NM_018489	*ASH1L*	6	12.15	18	5.00
NM_004486	*GOLGA2*	6	12.52	25	4.77
NM_001012507	*CENPW*	3	15.08	15	4.17
NM_015957	*APIP*	3	10.40	18	4.03
NM_001145316	*DSN1*	5	16.22	18	3.72
NM_006923	*SDF2*	2	16.91	14	3.18
NM_014903	*NAV3*	6	34.18	12	2.69
NM_022170	*EIF4H*	6	10.47	49	2.34
NM_000404	*GLB1*	2	10.28	21	2.11
NM_014180	*MRPL22*	2	11.97	21	2.03
NM_002117	*HLA-C*	4	11.15	21	1.97

NM_032375	*AKT1S1*	2	27.02	15	1.55
NM_001114752	*CD55*	4	12.89	74	1.54
NM_016640	*MRPS30*	6	10.31	28	1.19
NM_013354	*CNOT7*	6	17.13	14	1.03
NM_005953	*MT2A*	6	17.41	27	0.48
NM_153713	*LIX1L*	3	35.37	14	0.14

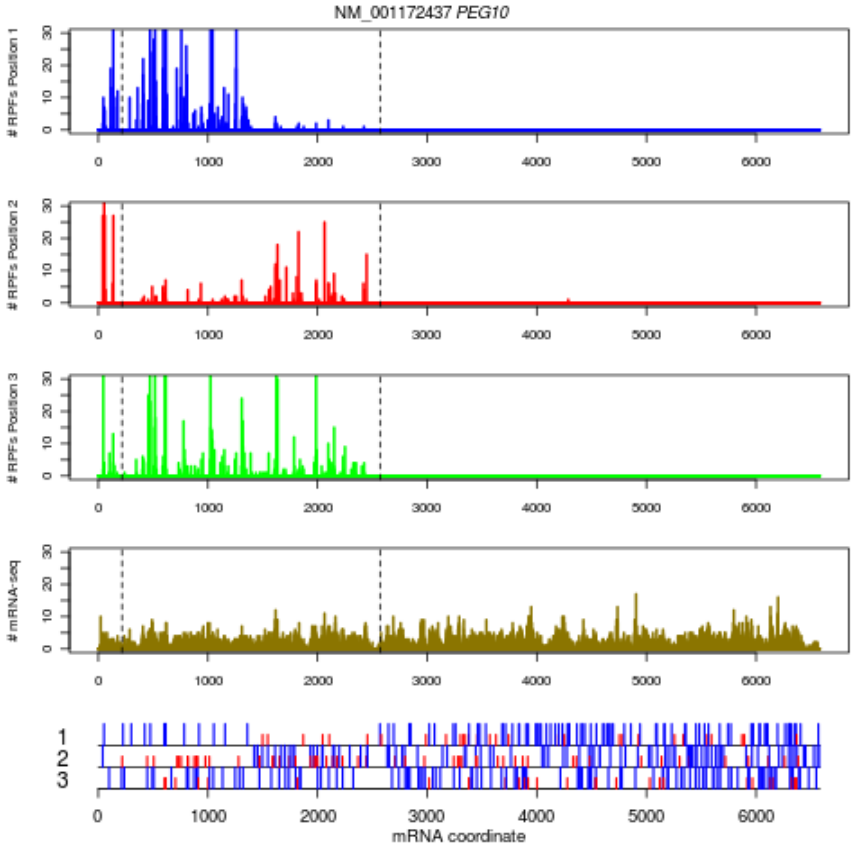
The RefSeq accession number, gene name, classification index, initial PTS, number of RPF locations in sub-codon position 2 and the corrected PTS after removal of the highest RPF peak in position 2 for 108 genes. The candidates are ranked according to the corrected PTS.

Supplemental Figures 9-116: Individual sub-codon and mRNA-seq profiles

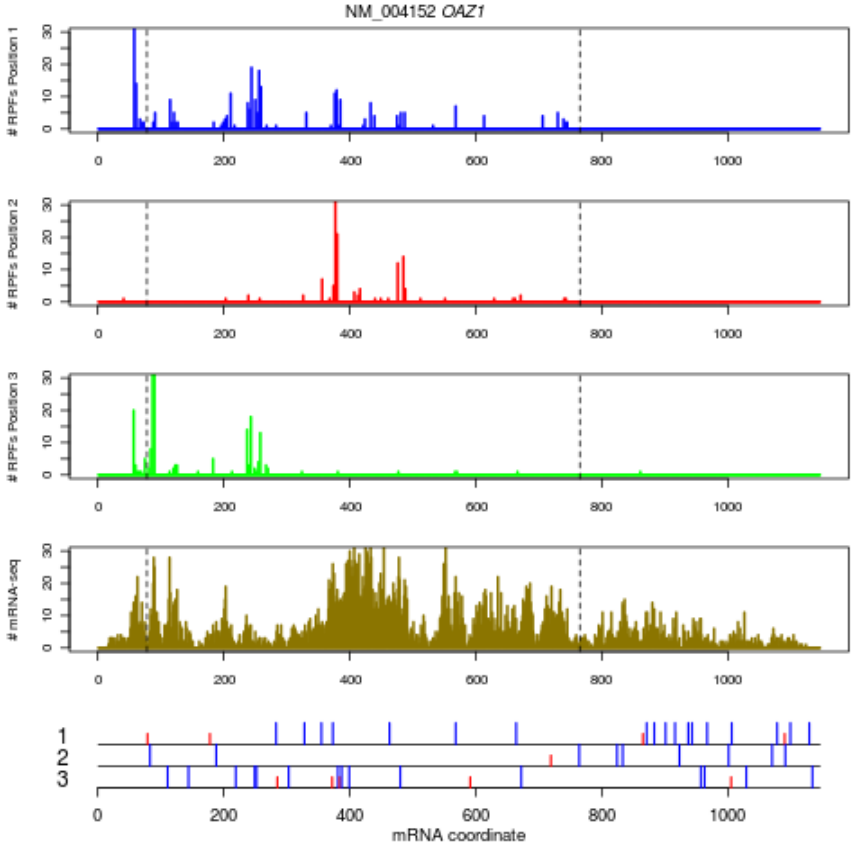
The following pages contain the sub-codon and mRNA-seq profiles and ORF organization of the 108 RefSeq candidates listed in Supplemental Table 1. The candidates are grouped according to the index classification outlined in Supplemental Table 1.

1. Known cases of programmed ribosomal frameshifting.

Supplemental Figure 9: Sub-codon profile, mRNA-Seq and ORF organization for

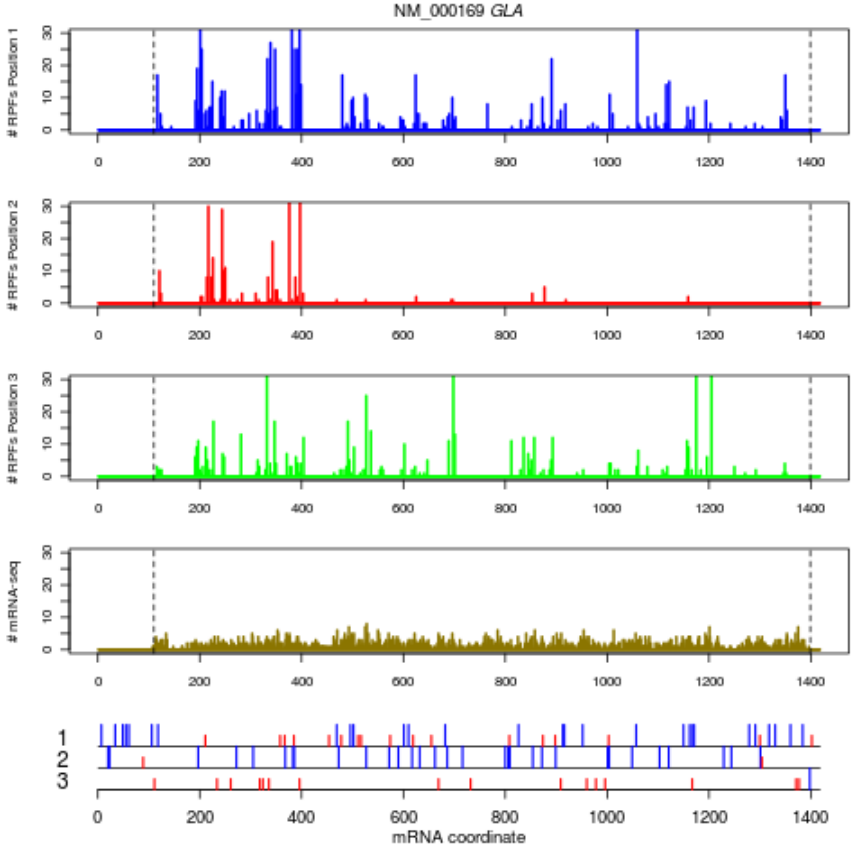


Supplemental Figure 10: Sub-codon profile, mRNA-Seq and ORF organization for

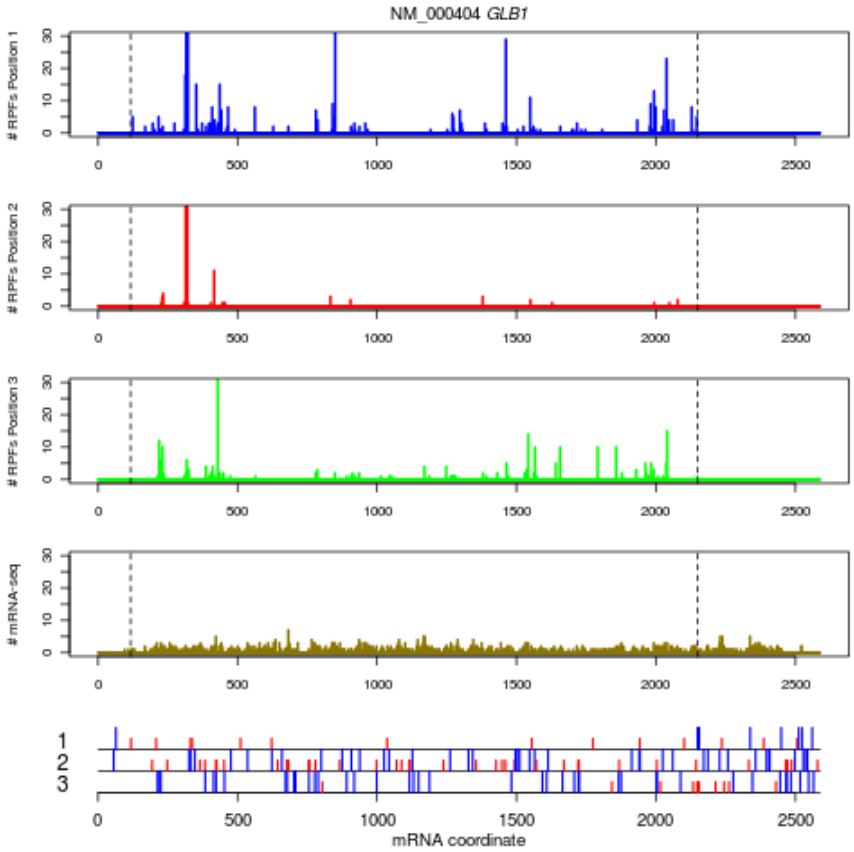


2. Non-upstream ORFs (nORFs).

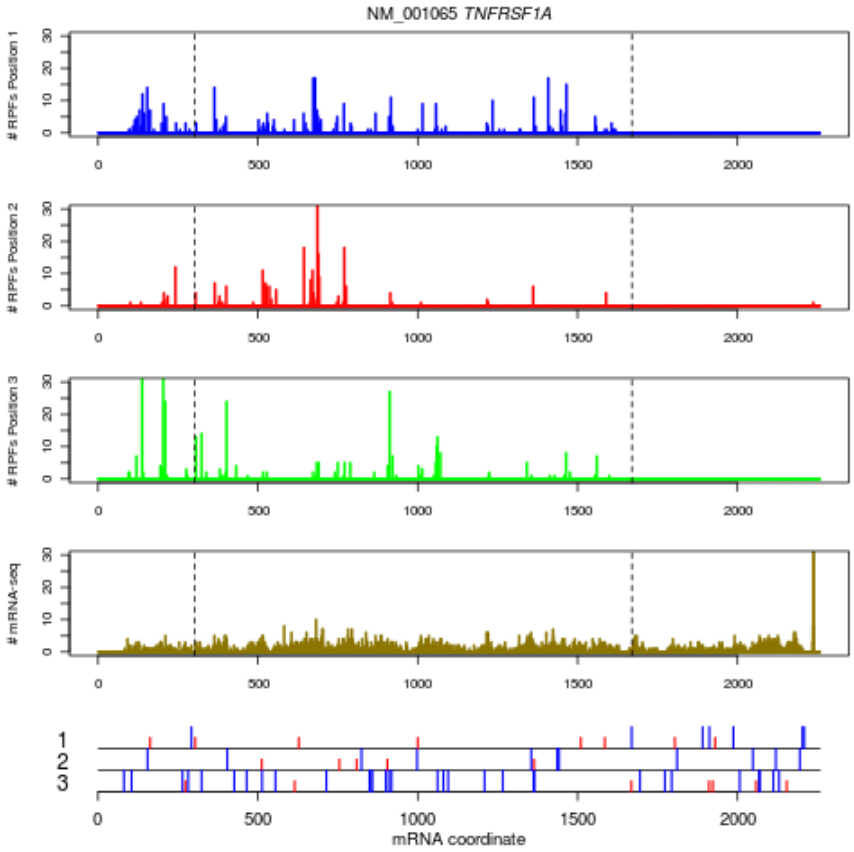
Supplemental Figure 11: Sub-codon profile, mRNA-Seq and ORF organization for



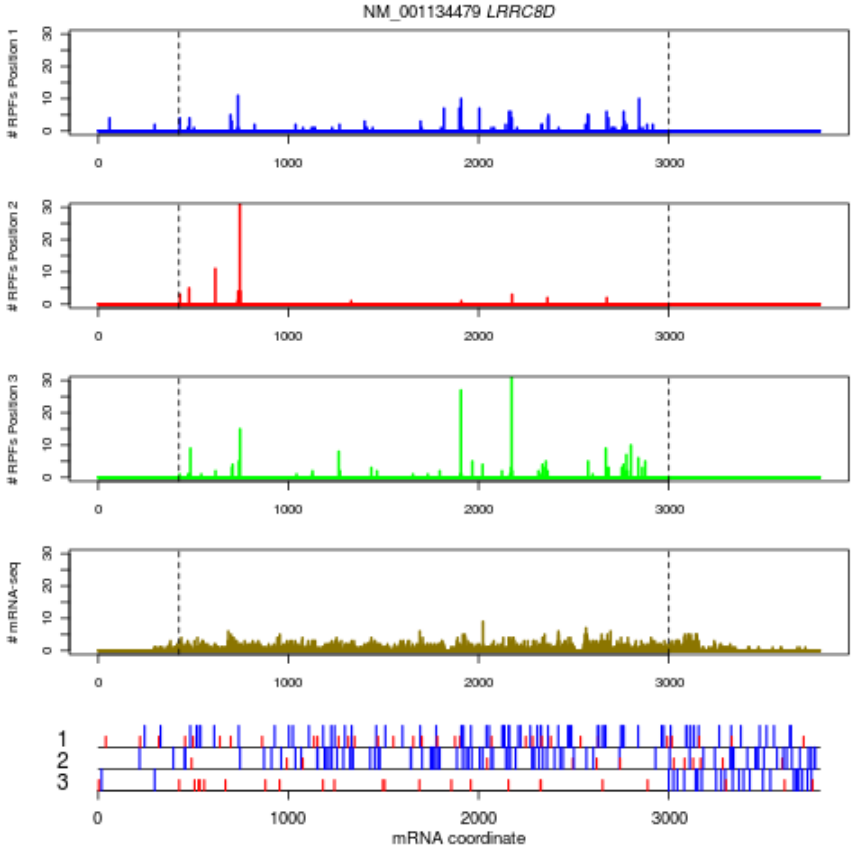
Supplemental Figure 12: Sub-codon profile, mRNA-Seq and ORF organization for



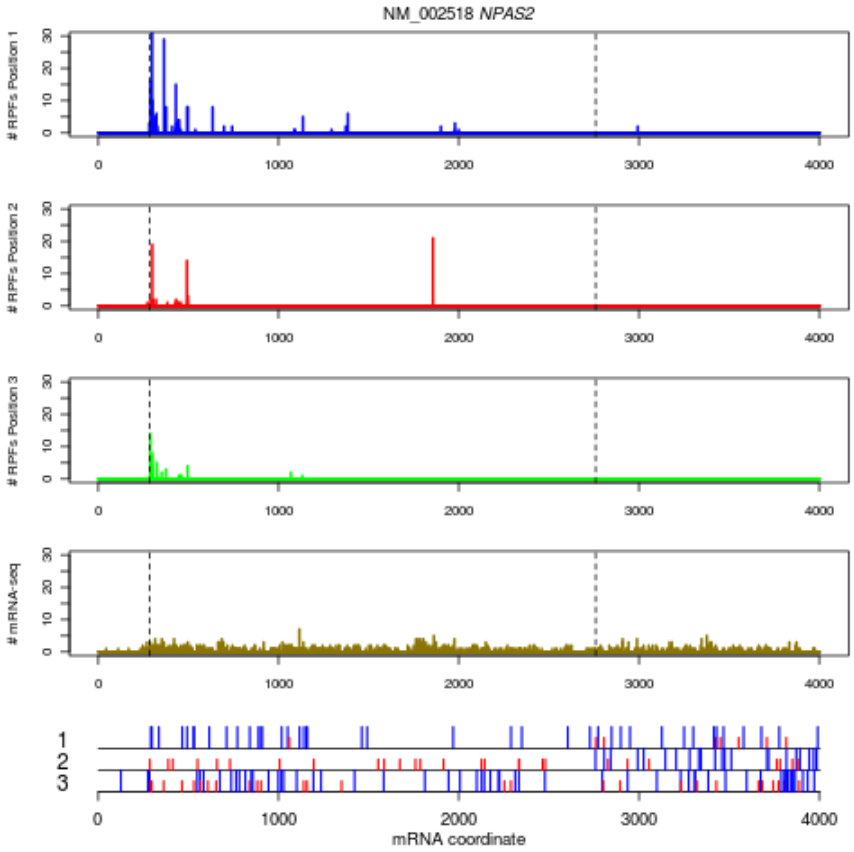
Supplemental Figure 13: Sub-codon profile, mRNA-Seq and ORF organization for



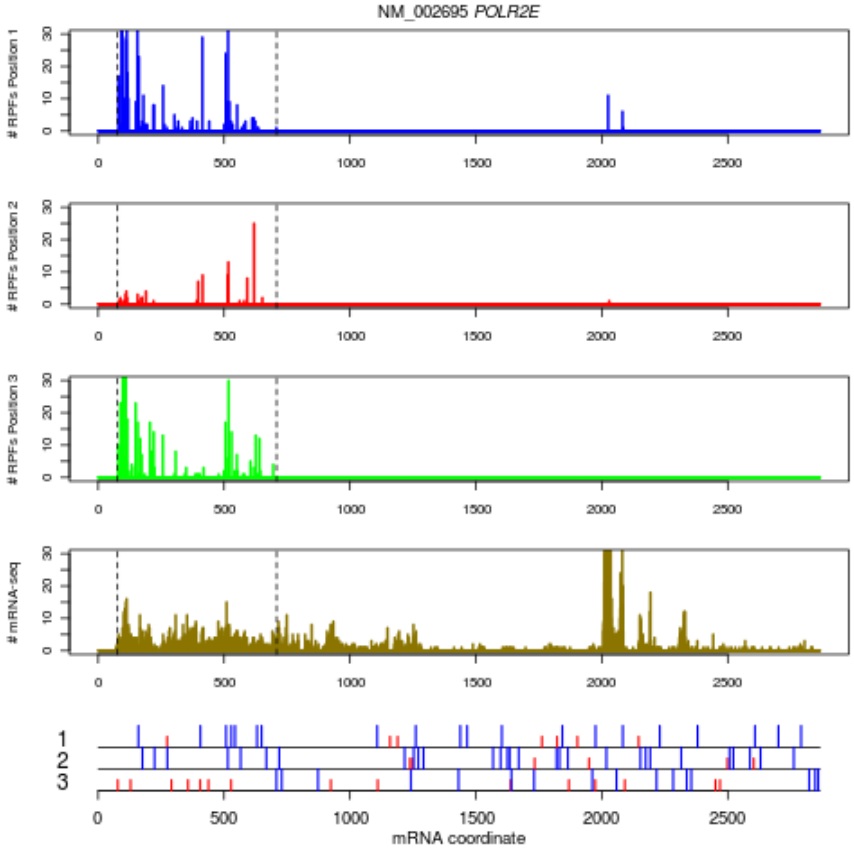
Supplemental Figure 14: Sub-codon profile, mRNA-Seq and ORF organization for



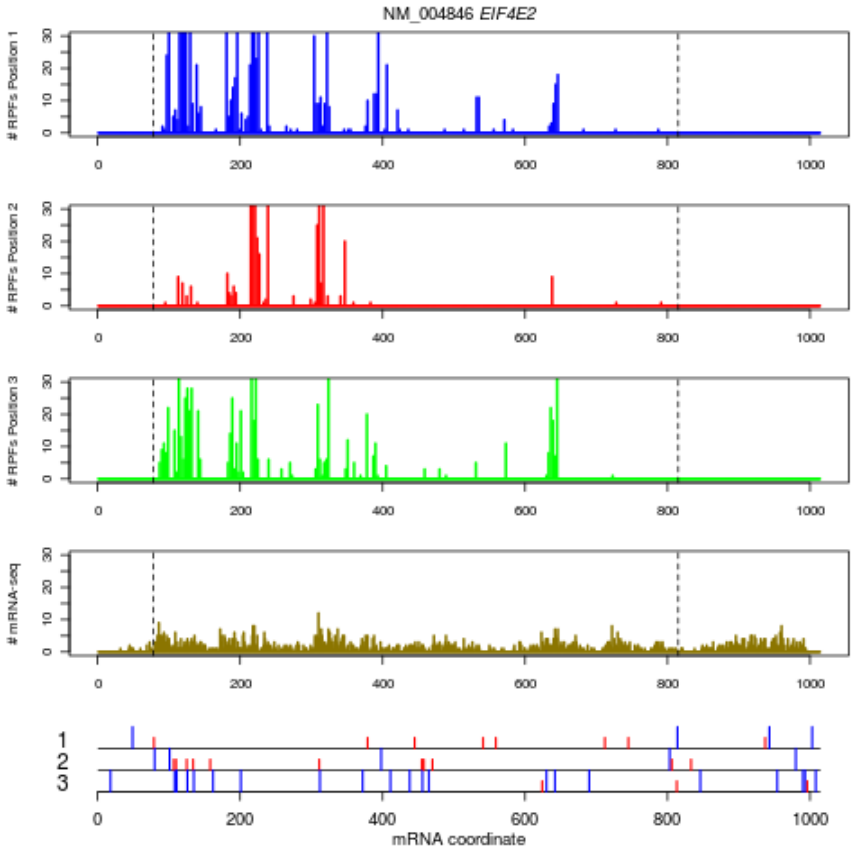
Supplemental Figure 15: Sub-codon profile, mRNA-Seq and ORF organization for



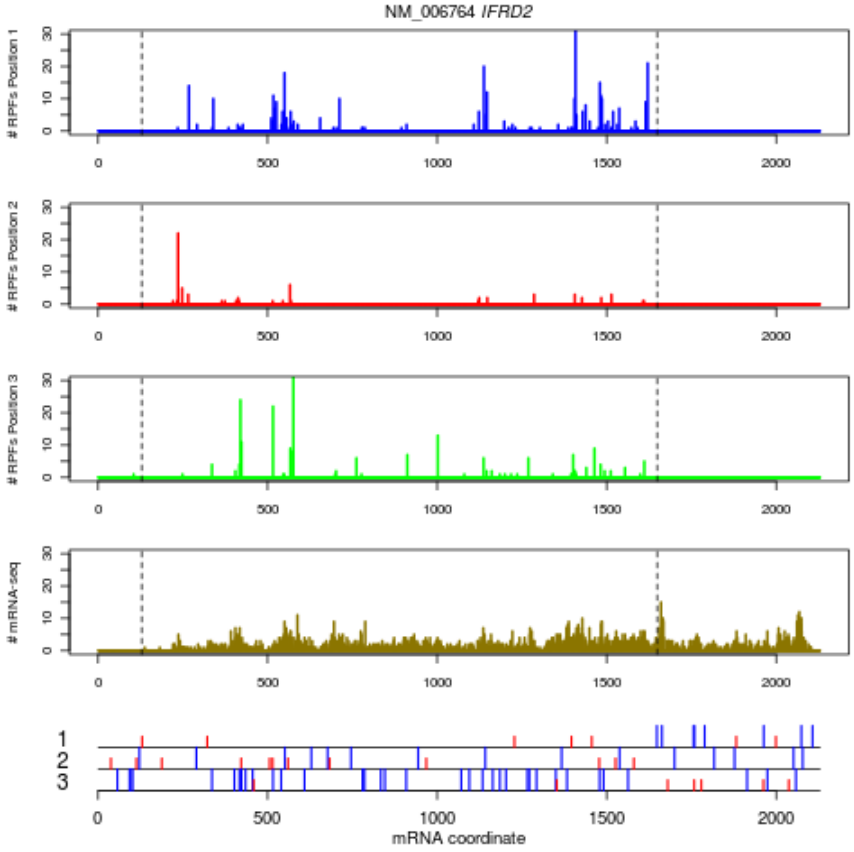
Supplemental Figure 16: Sub-codon profile, mRNA-Seq and ORF organization for



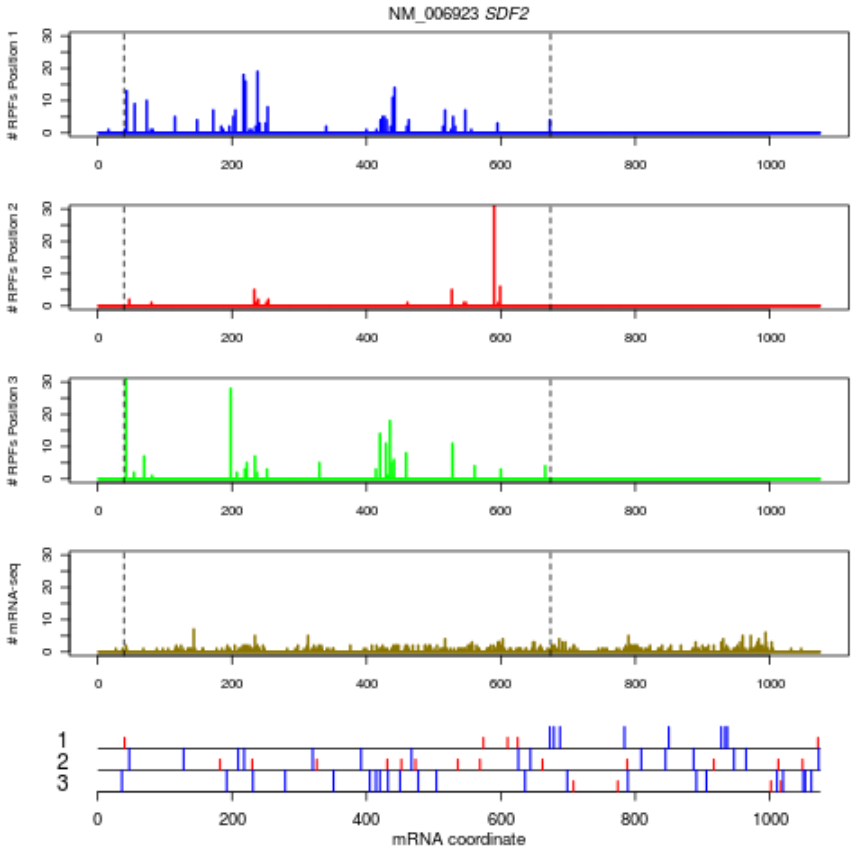
Supplemental Figure 17: Sub-codon profile, mRNA-Seq and ORF organization for



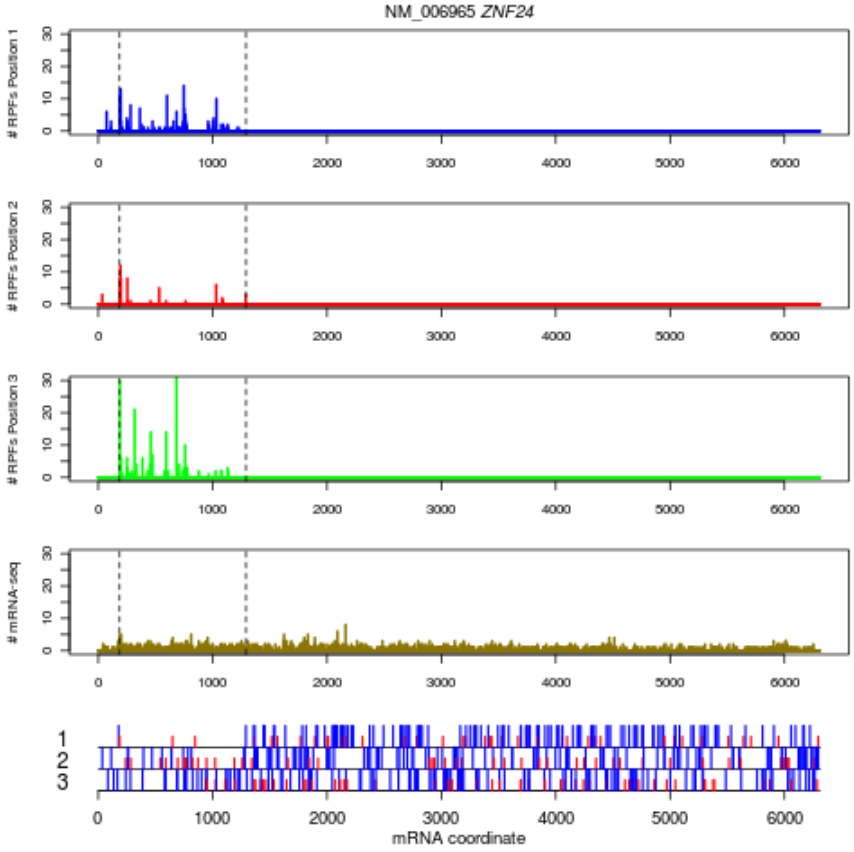
Supplemental Figure 18: Sub-codon profile, mRNA-Seq and ORF organization for



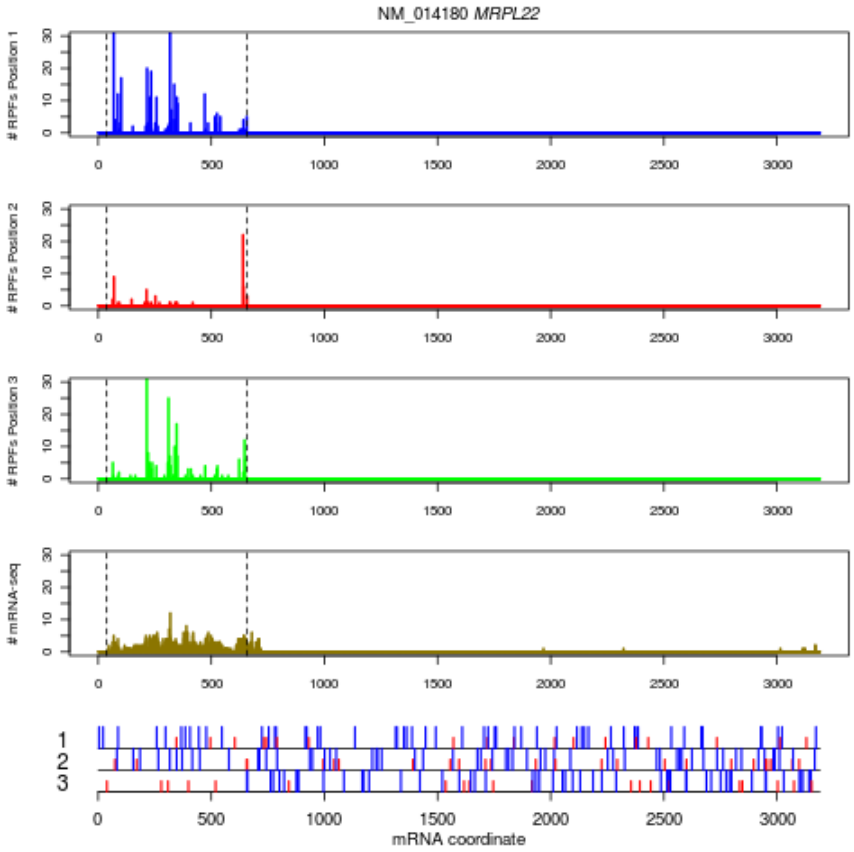
Supplemental Figure 19: Sub-codon profile, mRNA-Seq and ORF organization for



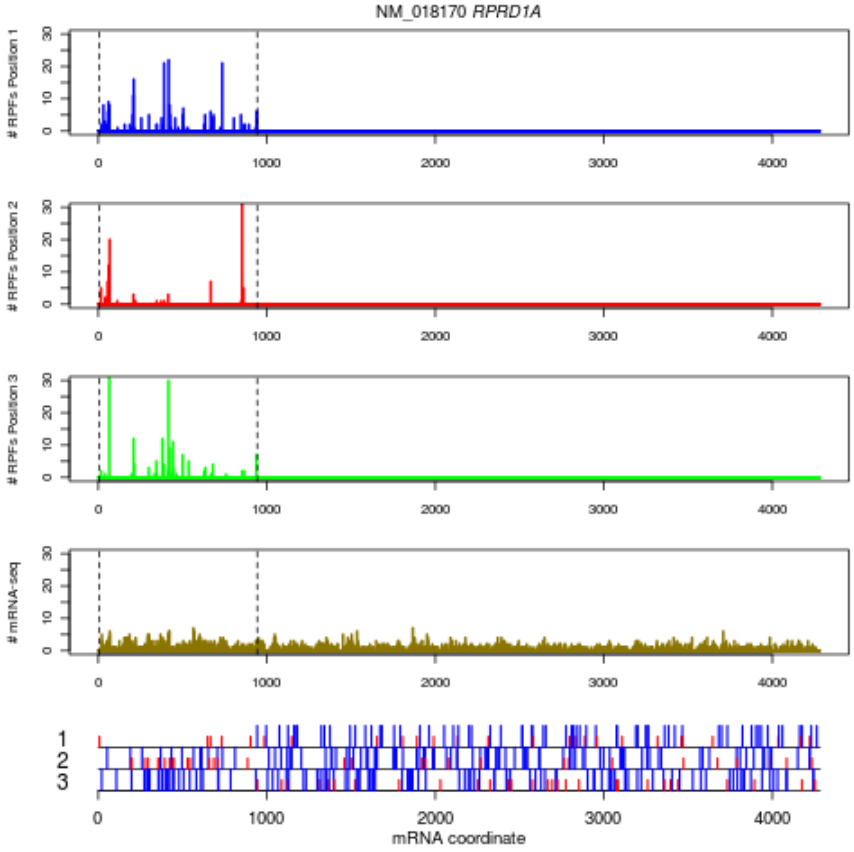
Supplemental Figure 20: Sub-codon profile, mRNA-Seq and ORF organization for



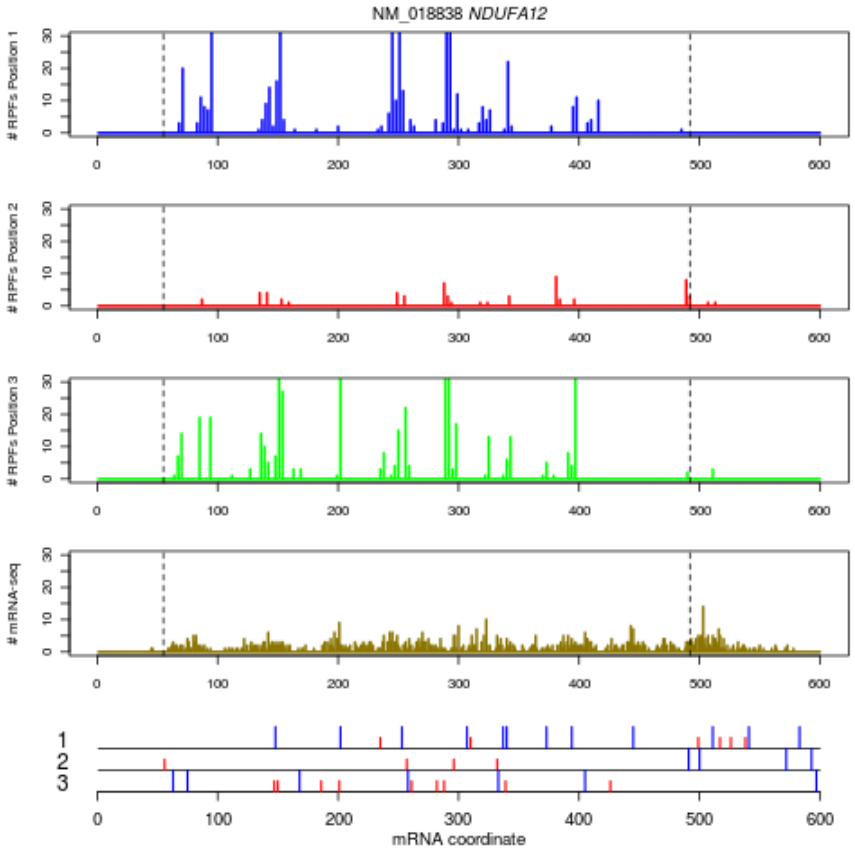
Supplemental Figure 21: Sub-codon profile, mRNA-Seq and ORF organization for



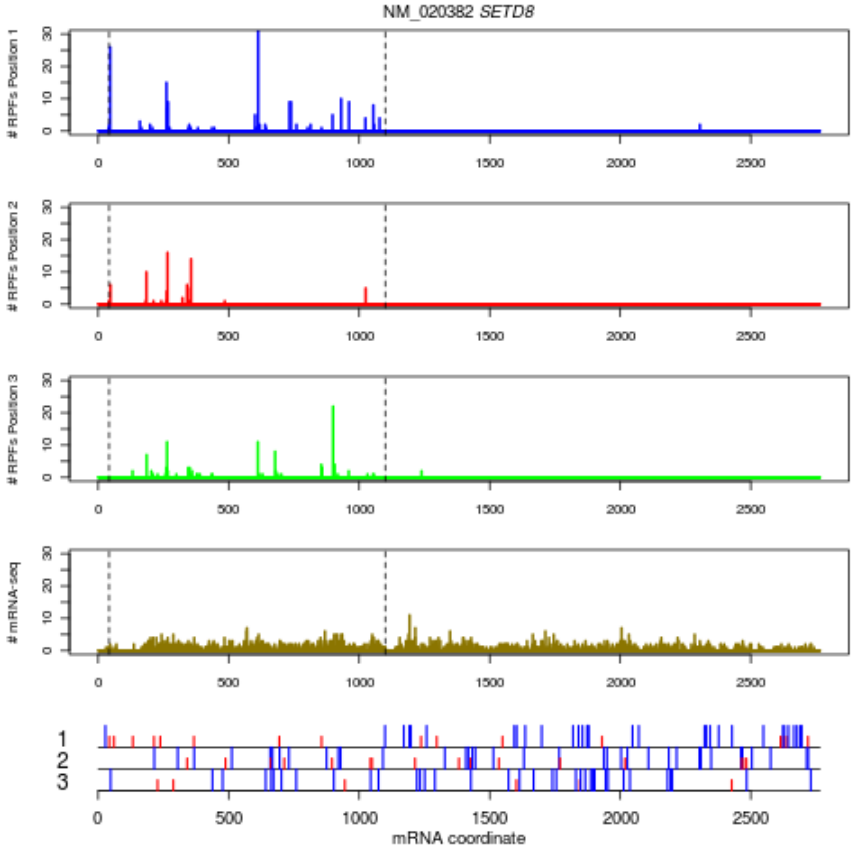
Supplemental Figure 22: Sub-codon profile, mRNA-Seq and ORF organization for



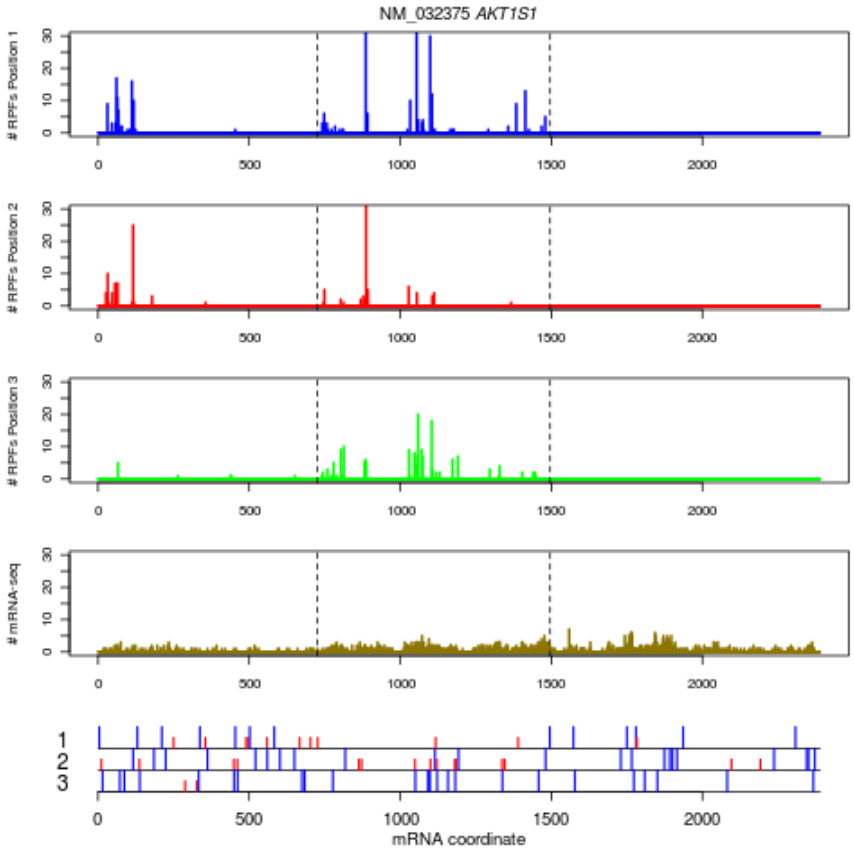
Supplemental Figure 23: Sub-codon profile, mRNA-Seq and ORF organization for



Supplemental Figure 24: Sub-codon profile, mRNA-Seq and ORF organization for

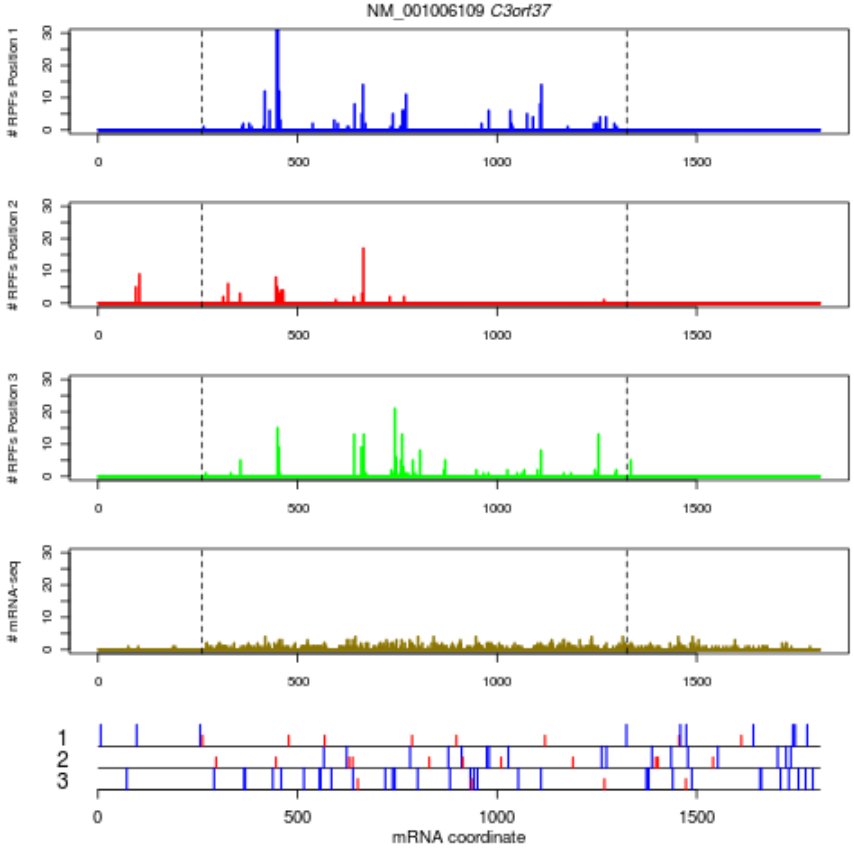


Supplemental Figure 25: Sub-codon profile, mRNA-Seq and ORF organization for

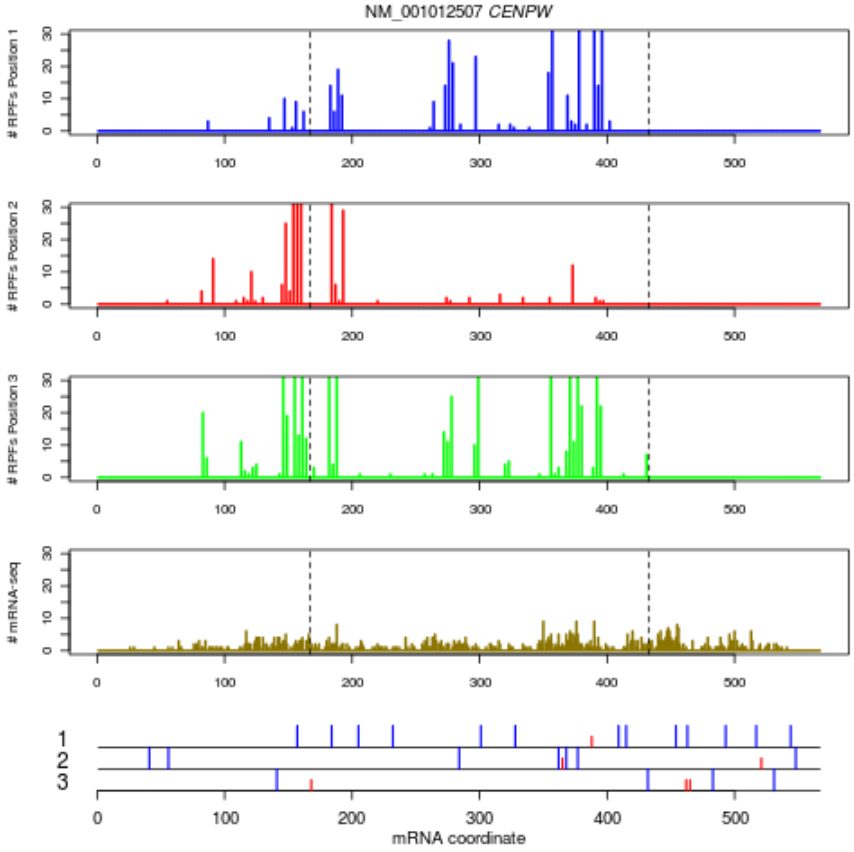


3. Upstream ORFs overlapping main protein coding ORFs (uORFs)

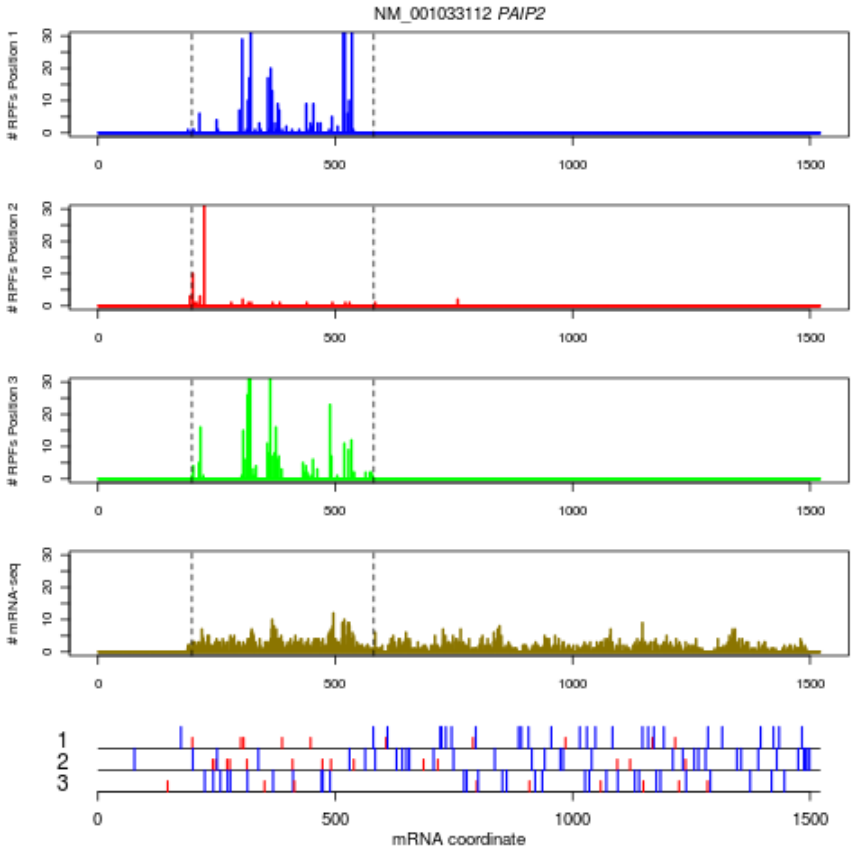
Supplemental Figure 26: Sub-codon profile, mRNA-Seq and ORF organization for



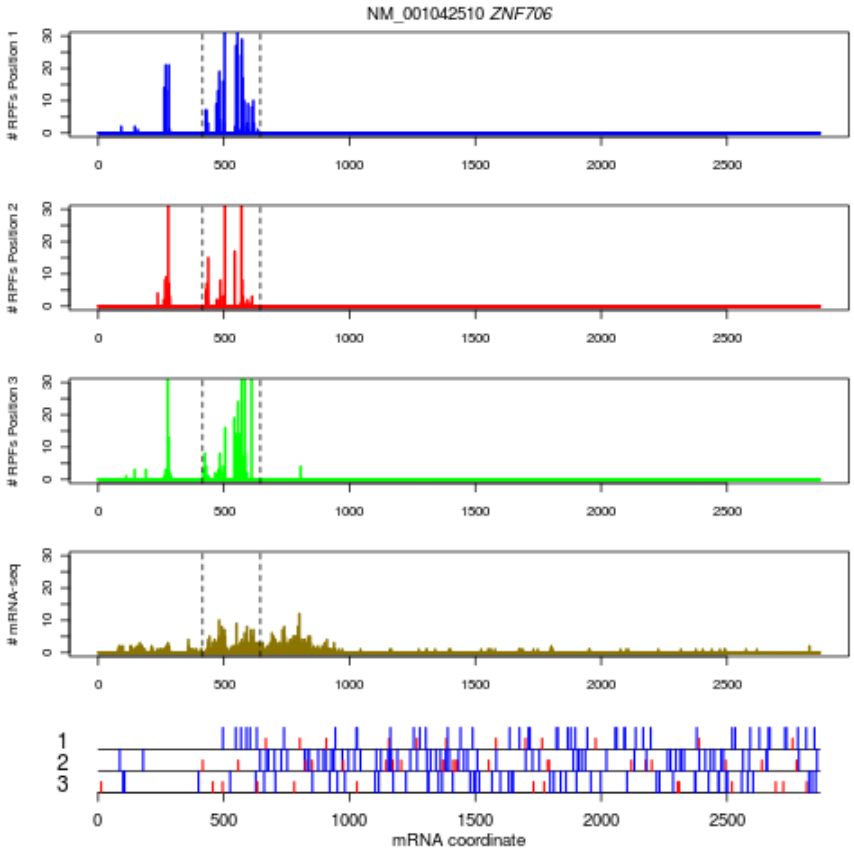
Supplemental Figure 27: Sub-codon profile, mRNA-Seq and ORF organization for



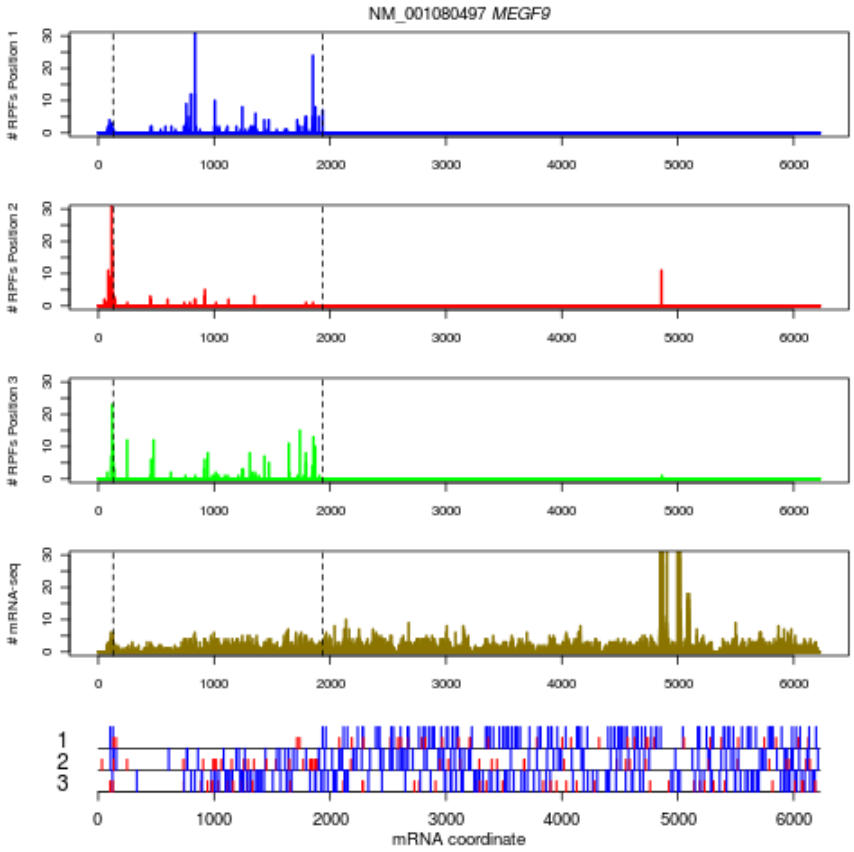
Supplemental Figure 28: Sub-codon profile, mRNA-Seq and ORF organization for



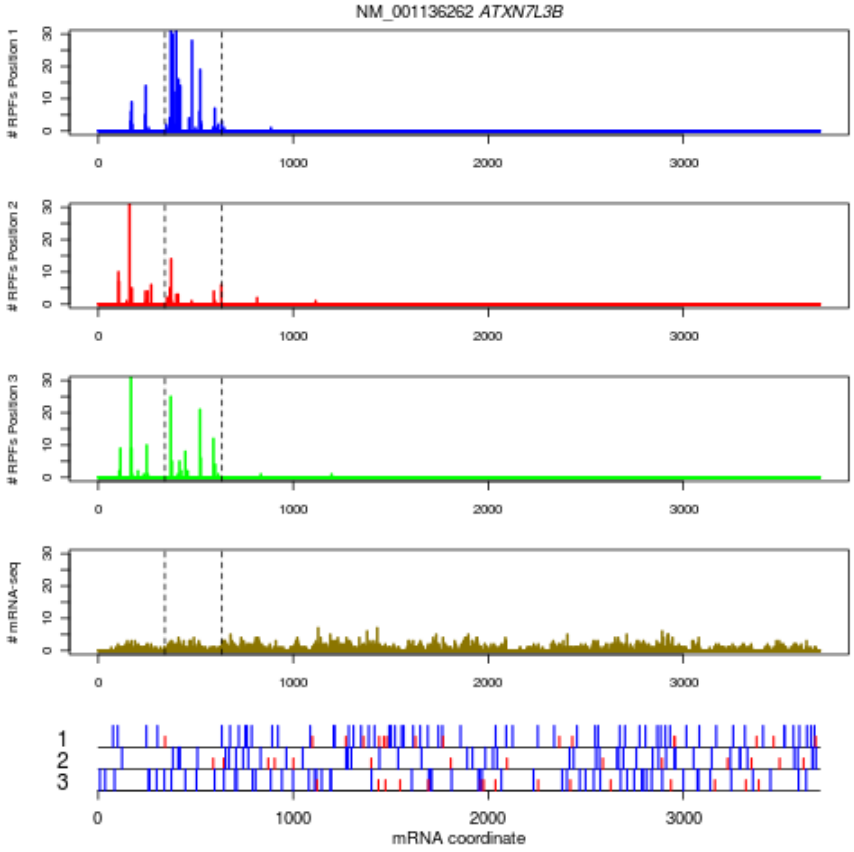
Supplemental Figure 29: Sub-codon profile, mRNA-Seq and ORF organization for



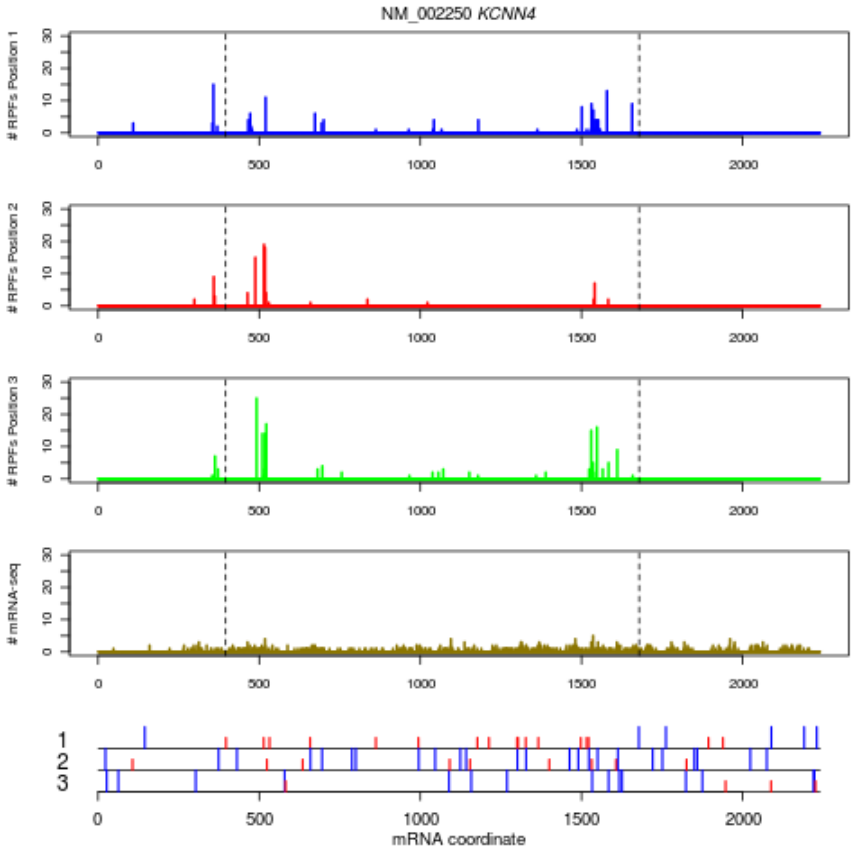
Supplemental Figure 30: Sub-codon profile, mRNA-Seq and ORF organization for



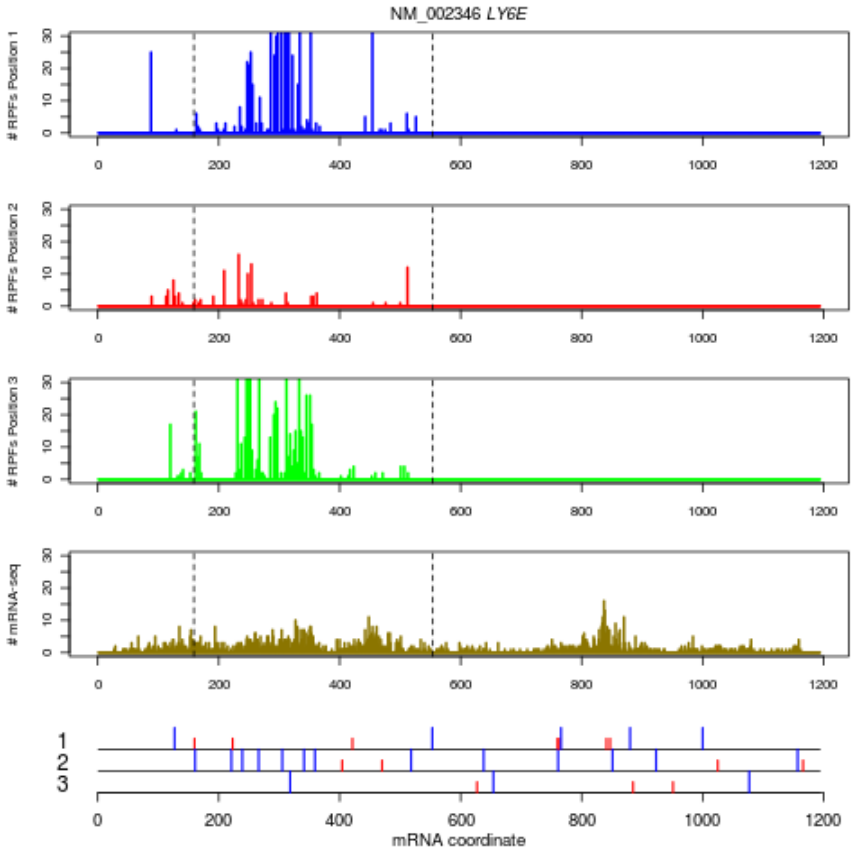
Supplemental Figure 31: Sub-codon profile, mRNA-Seq and ORF organization for



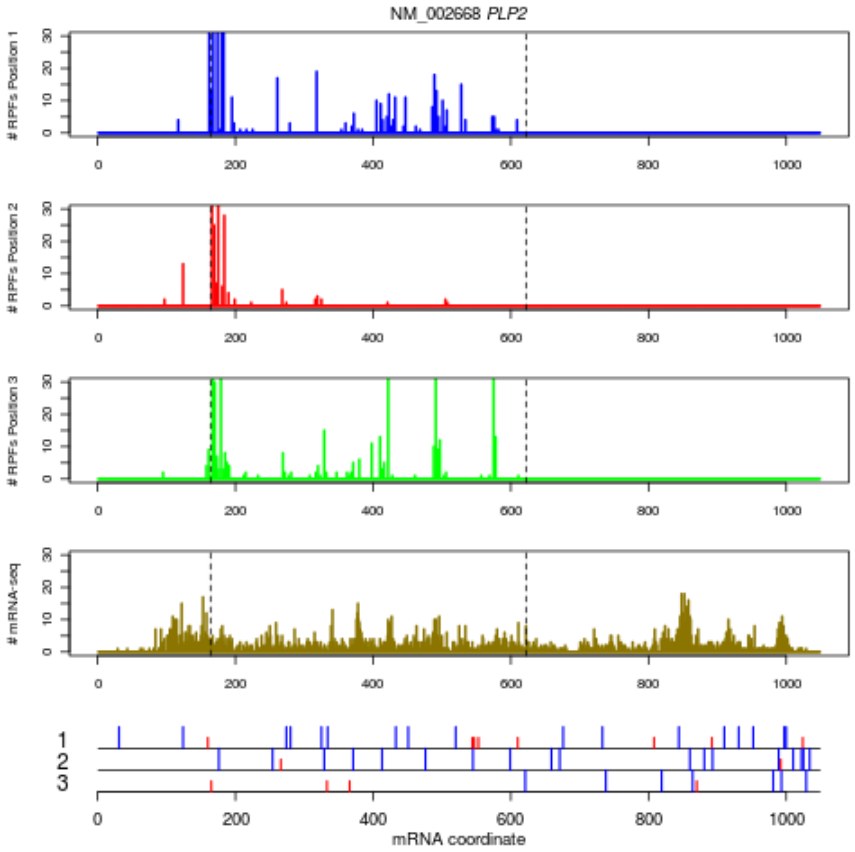
Supplemental Figure 32: Sub-codon profile, mRNA-Seq and ORF organization for



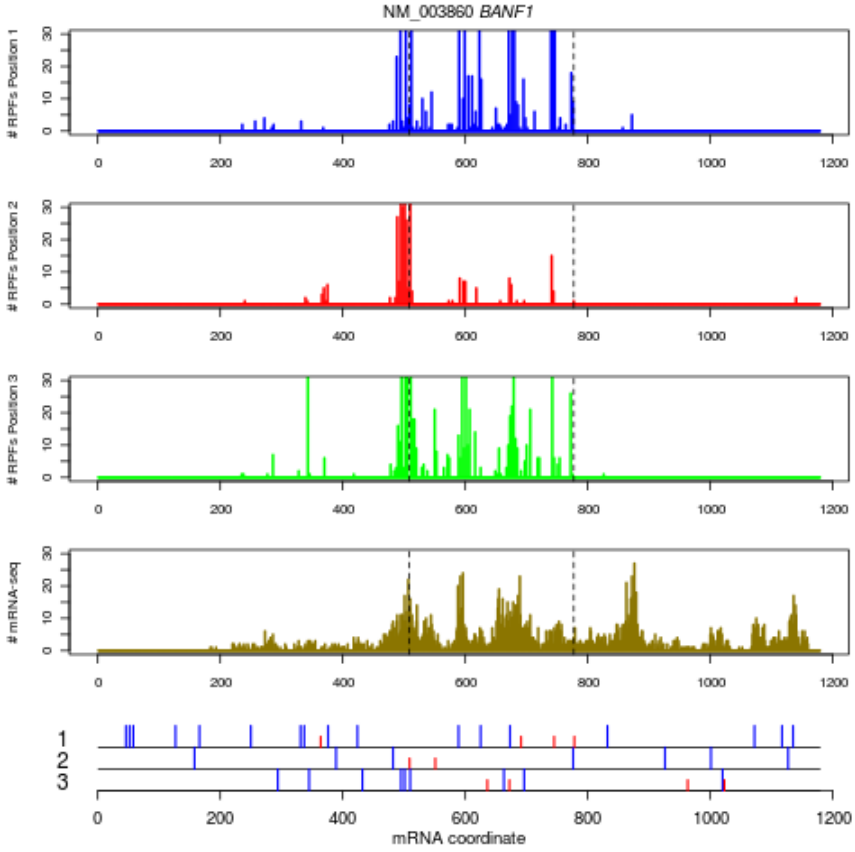
Supplemental Figure 33: Sub-codon profile, mRNA-Seq and ORF organization for



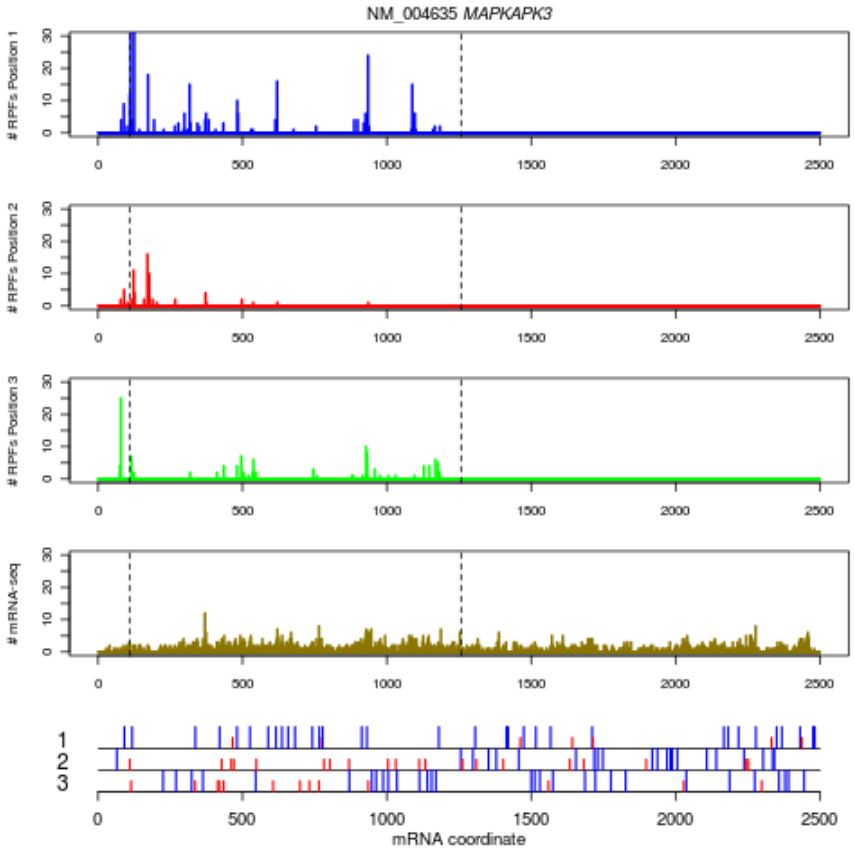
Supplemental Figure 34: Sub-codon profile, mRNA-Seq and ORF organization for



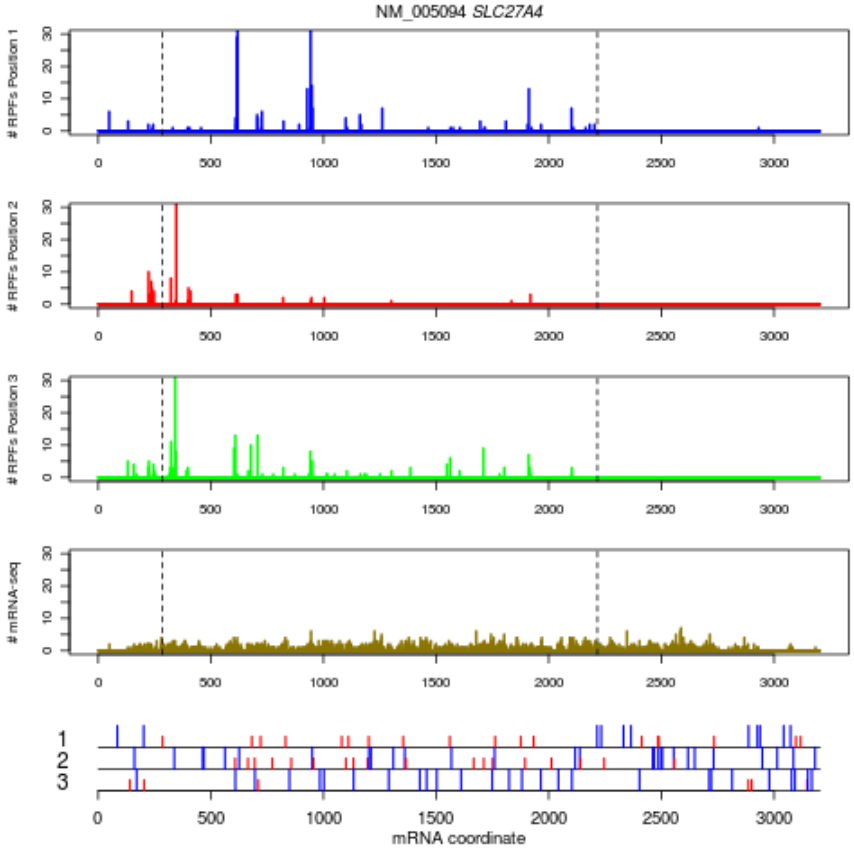
Supplemental Figure 35: Sub-codon profile, mRNA-Seq and ORF organization for



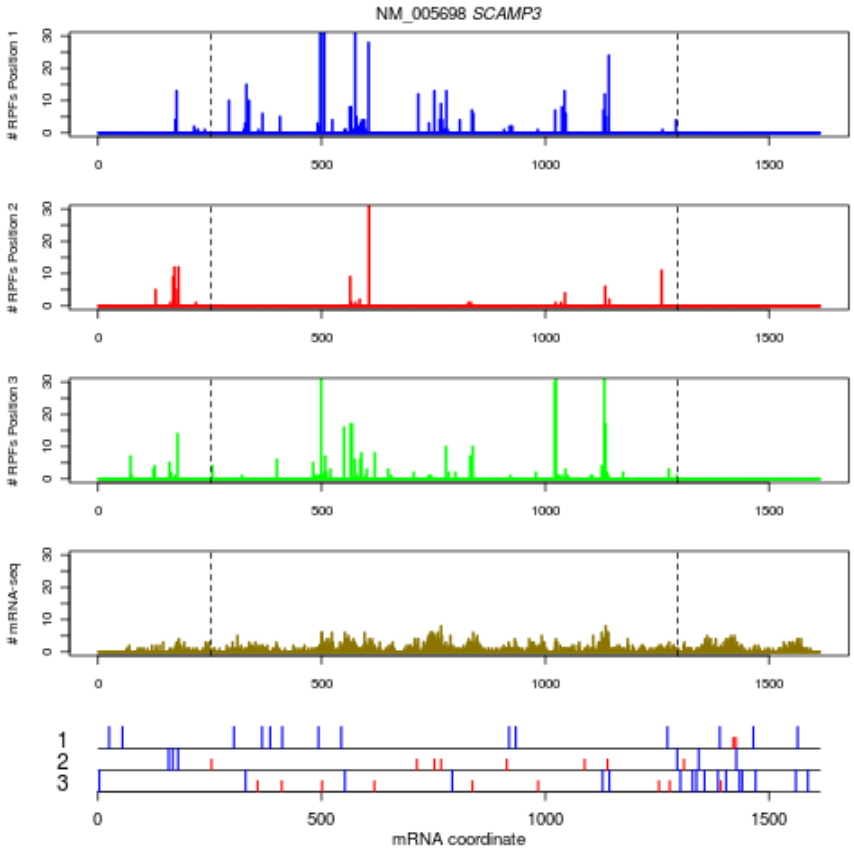
Supplemental Figure 36: Sub-codon profile, mRNA-Seq and ORF organization for



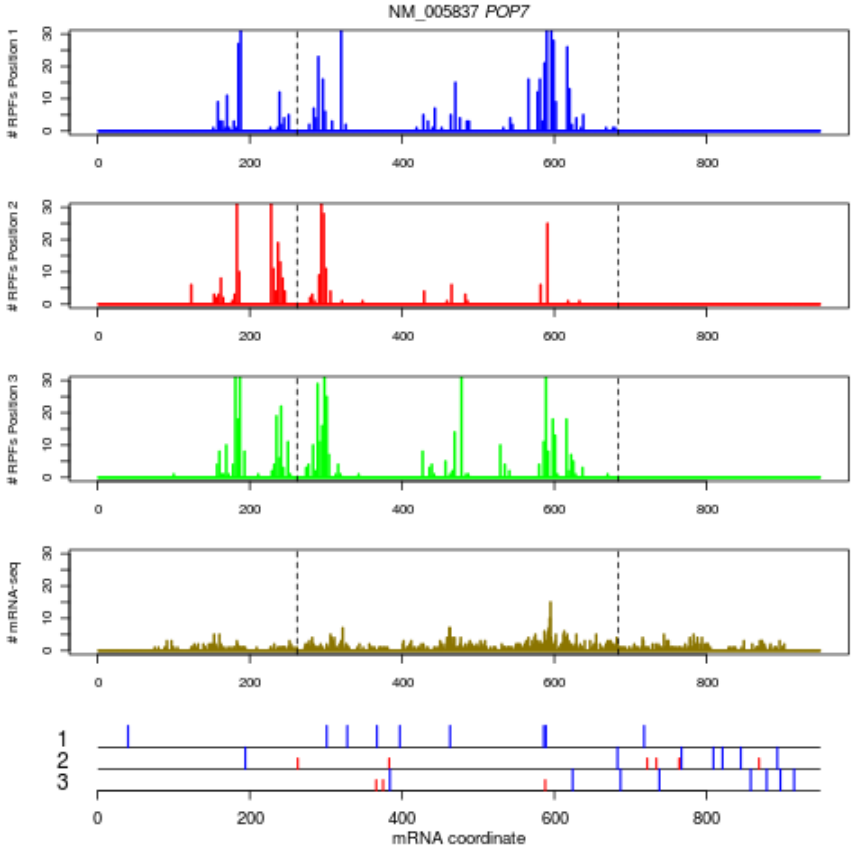
Supplemental Figure 37: Sub-codon profile, mRNA-Seq and ORF organization for



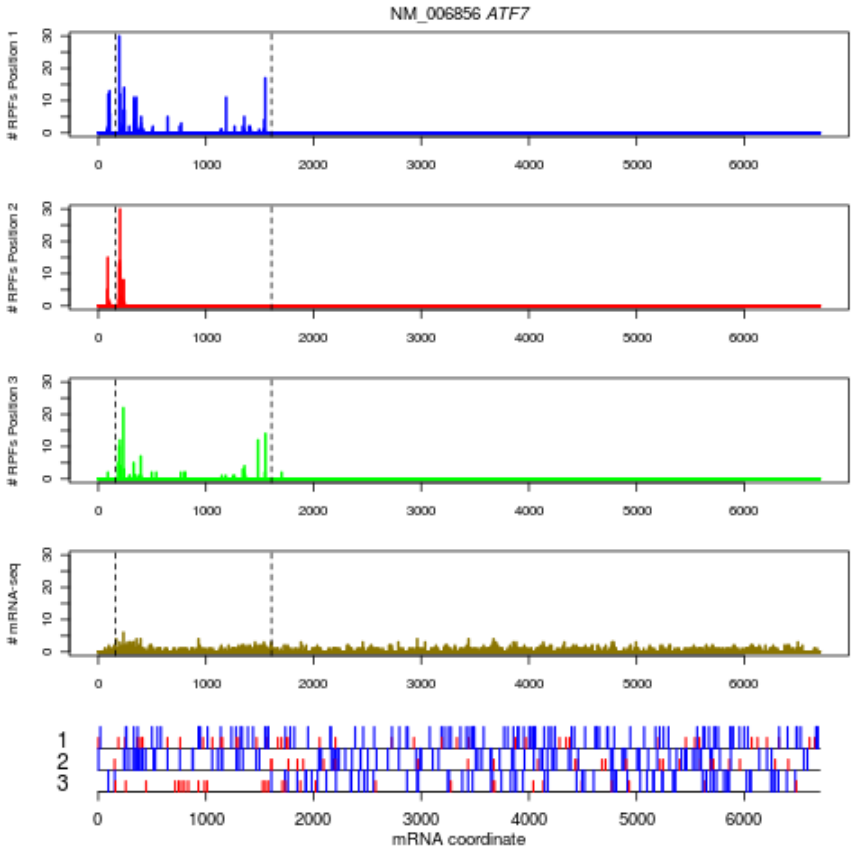
Supplemental Figure 38: Sub-codon profile, mRNA-Seq and ORF organization for



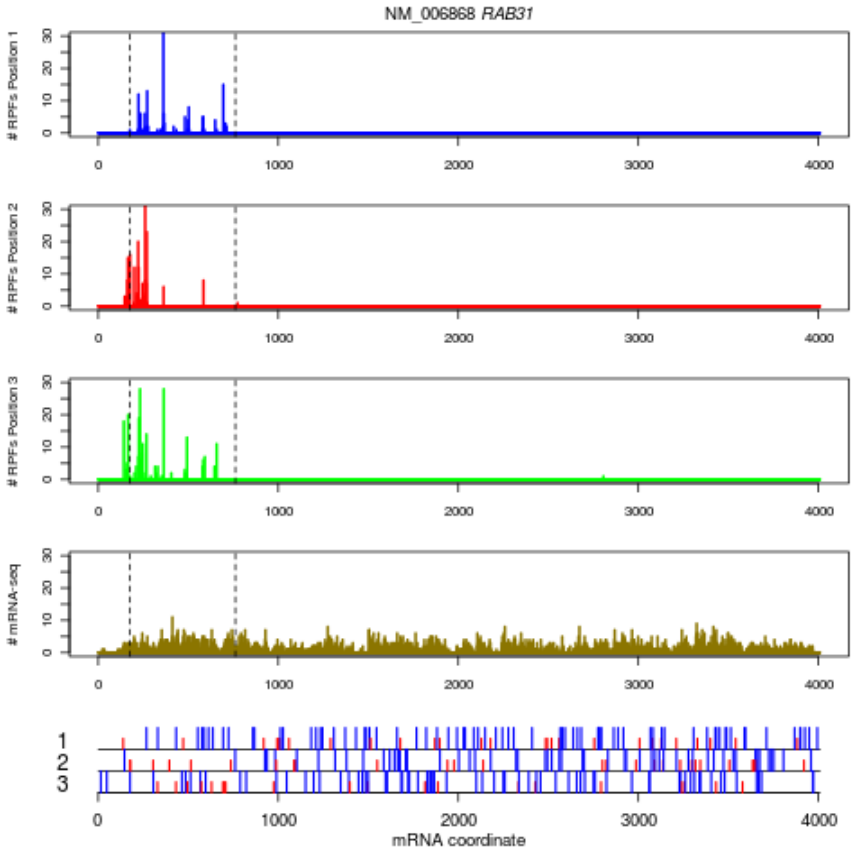
Supplemental Figure 39: Sub-codon profile, mRNA-Seq and ORF organization for



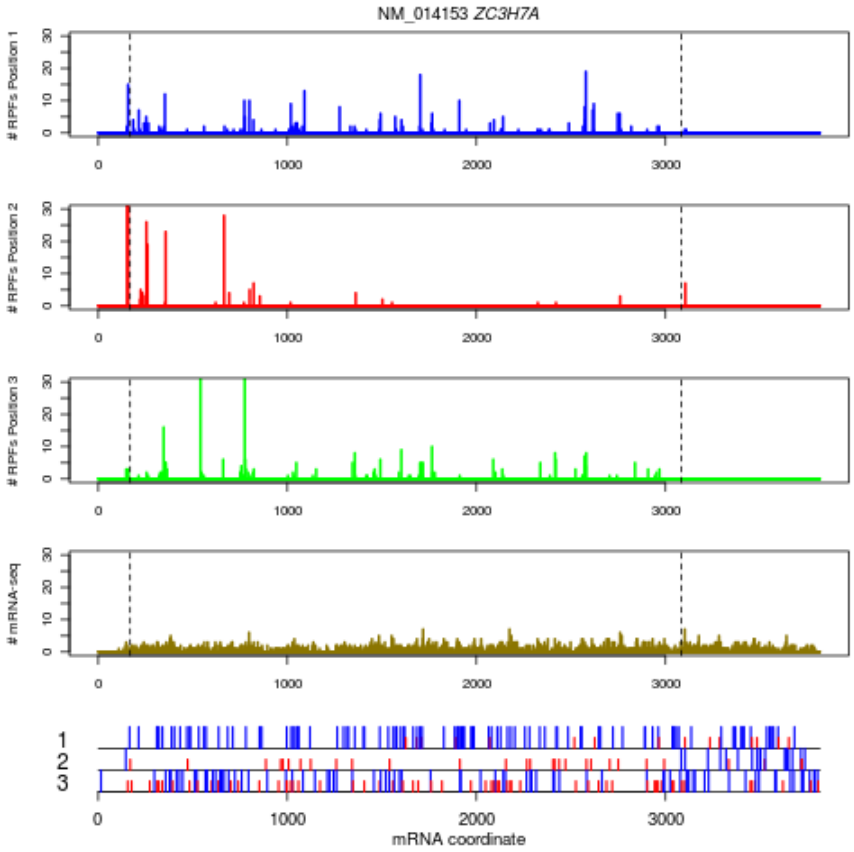
Supplemental Figure 40: Sub-codon profile, mRNA-Seq and ORF organization for



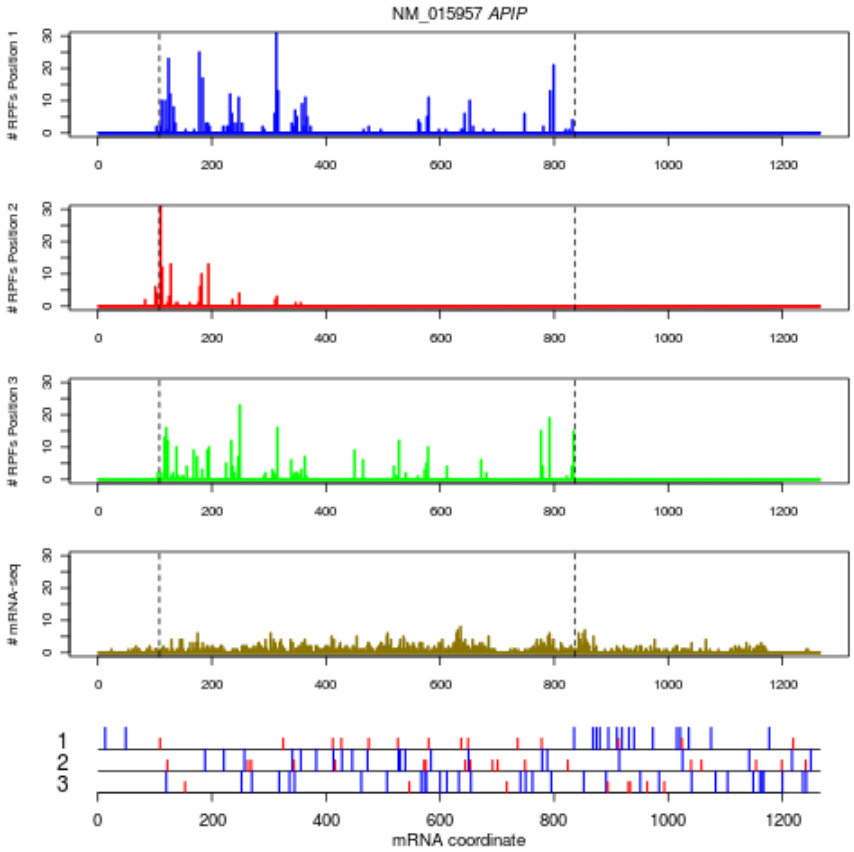
Supplemental Figure 41: Sub-codon profile, mRNA-Seq and ORF organization for



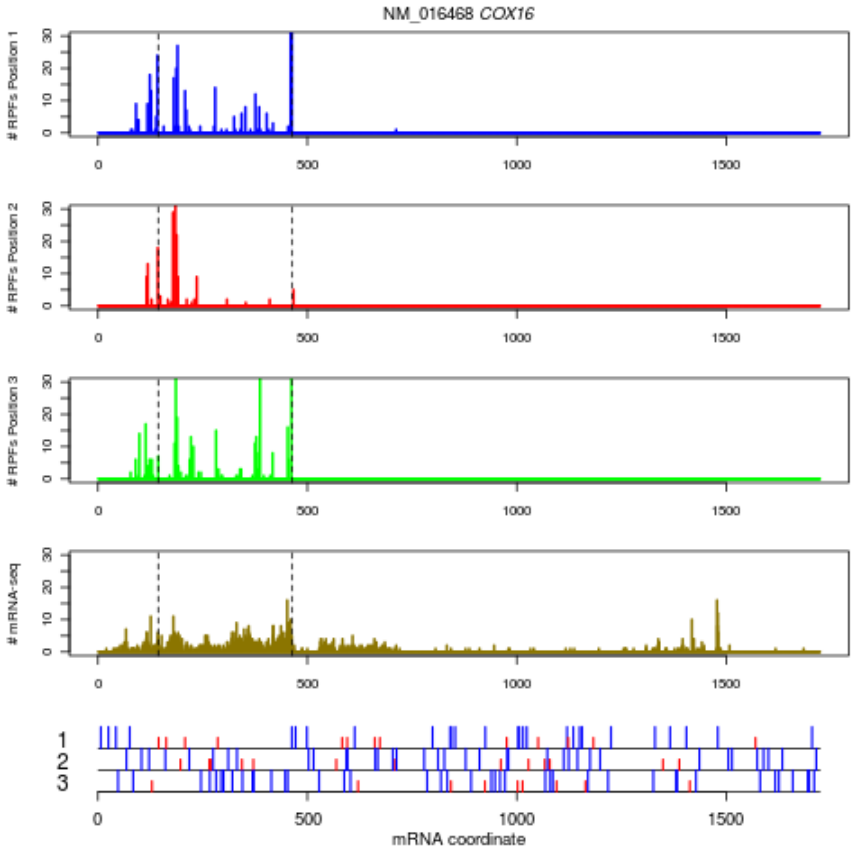
Supplemental Figure 42: Sub-codon profile, mRNA-Seq and ORF organization for



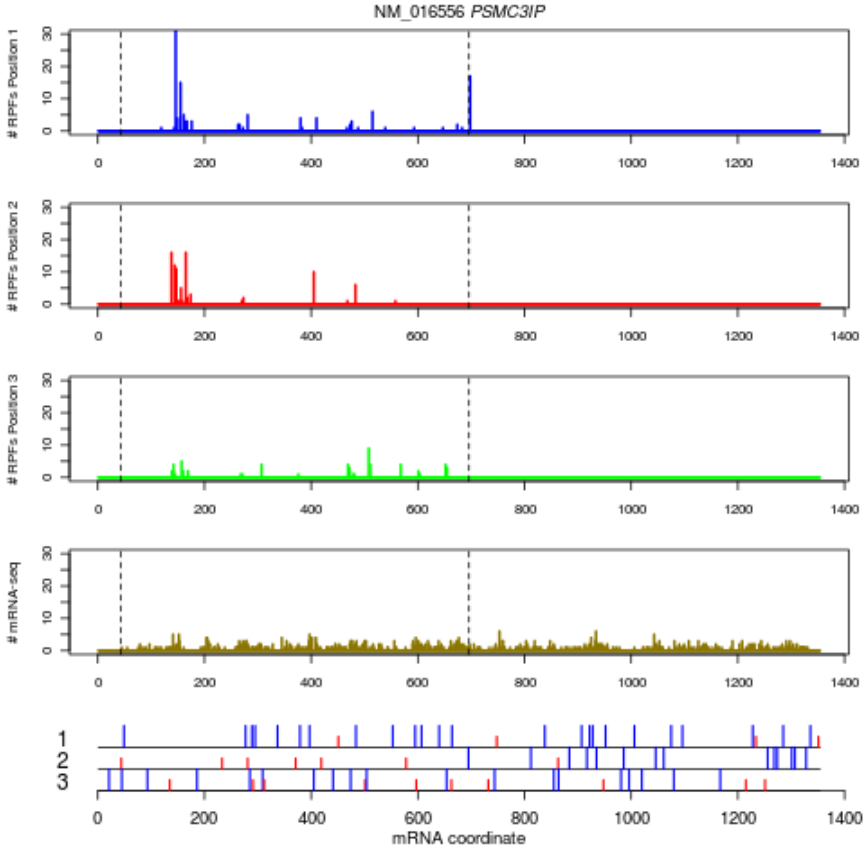
Supplemental Figure 43: Sub-codon profile, mRNA-Seq and ORF organization for



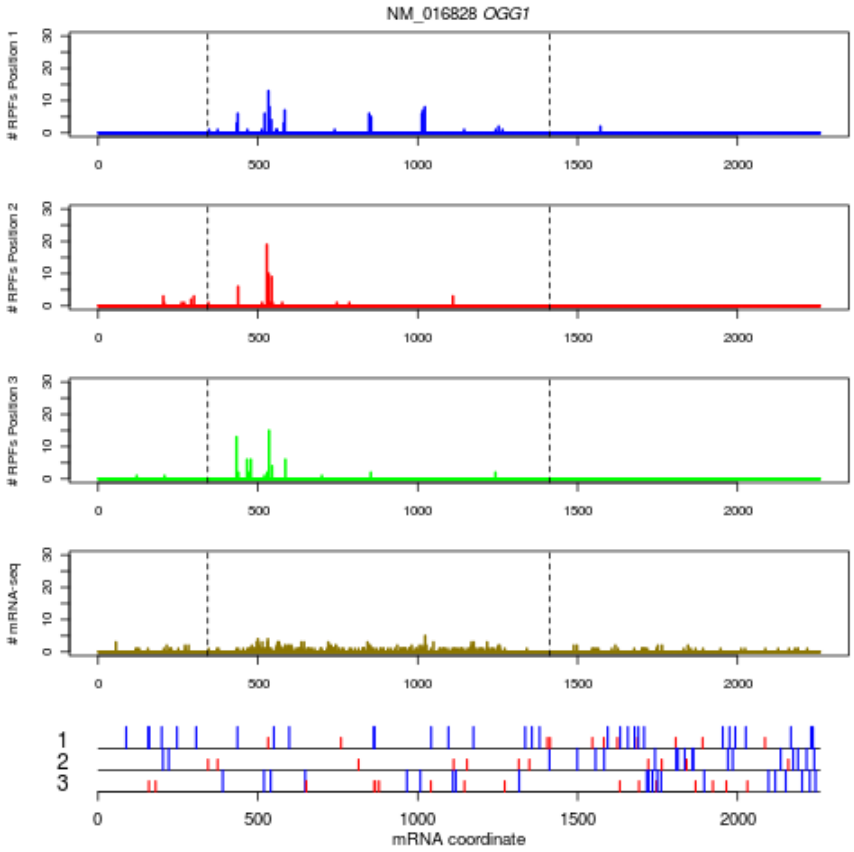
Supplemental Figure 44: Sub-codon profile, mRNA-Seq and ORF organization for



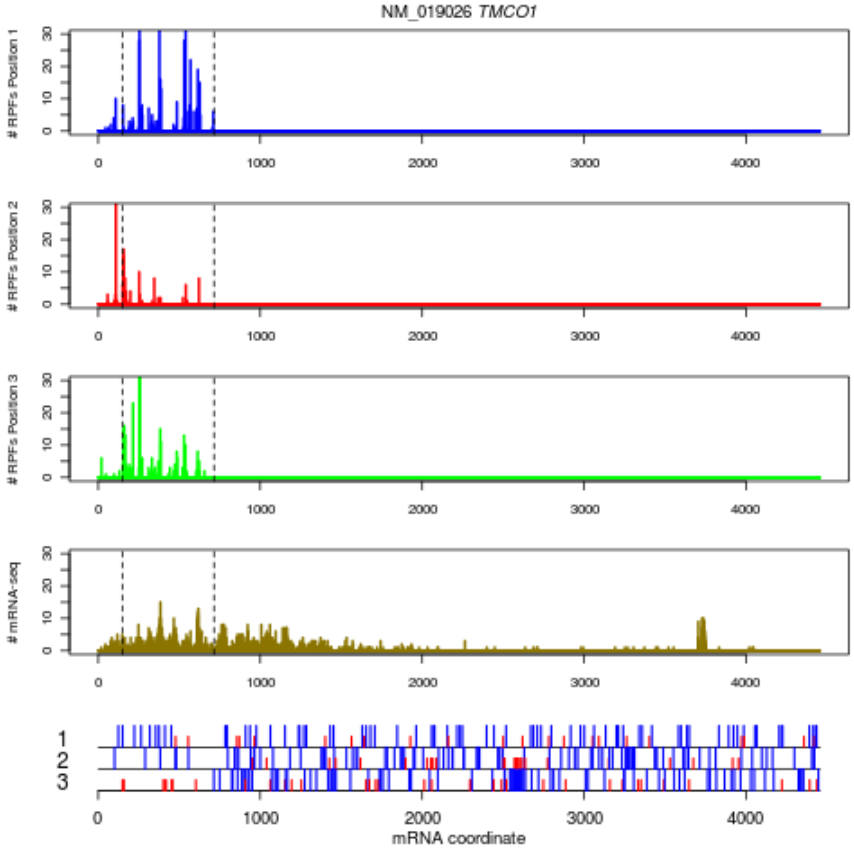
Supplemental Figure 45: Sub-codon profile, mRNA-Seq and ORF organization for



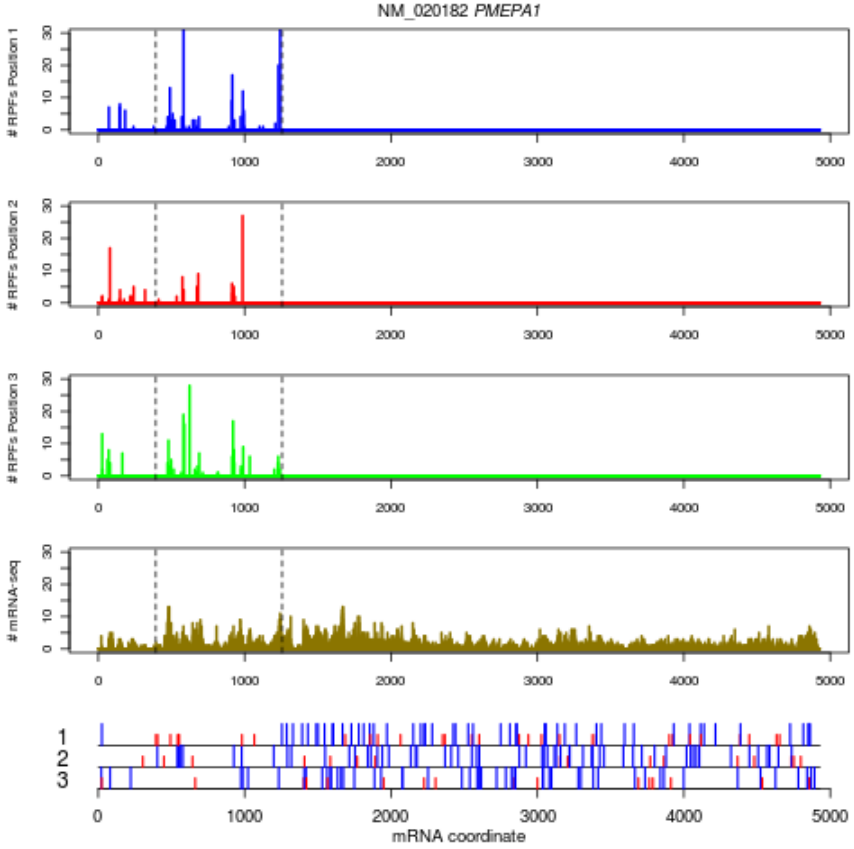
Supplemental Figure 46: Sub-codon profile, mRNA-Seq and ORF organization for



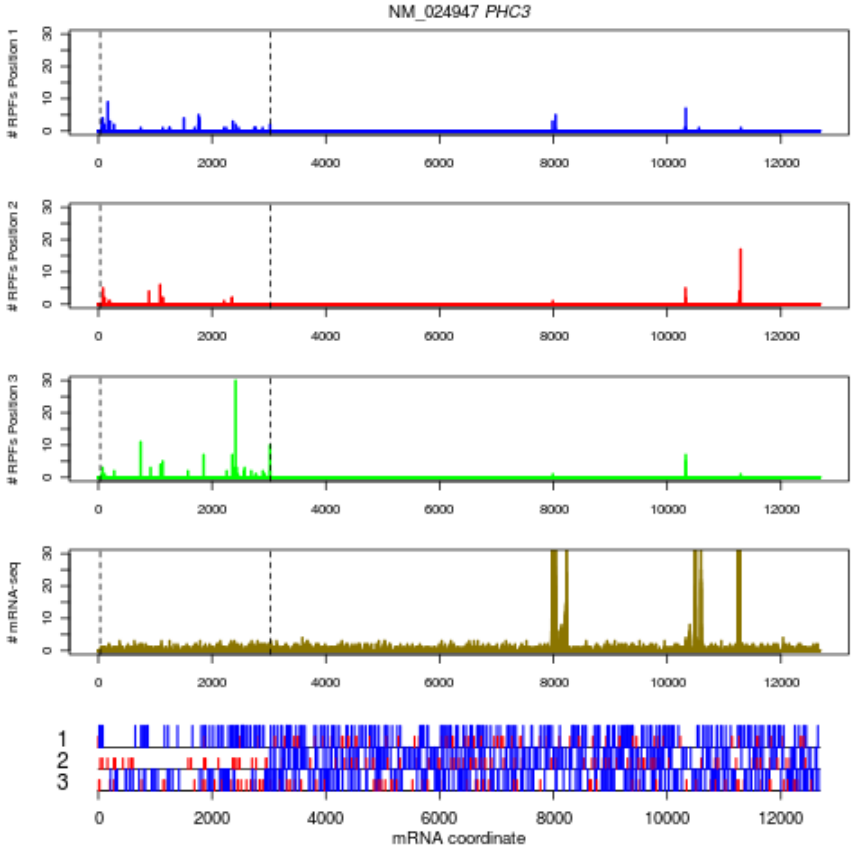
Supplemental Figure 47: Sub-codon profile, mRNA-Seq and ORF organization for



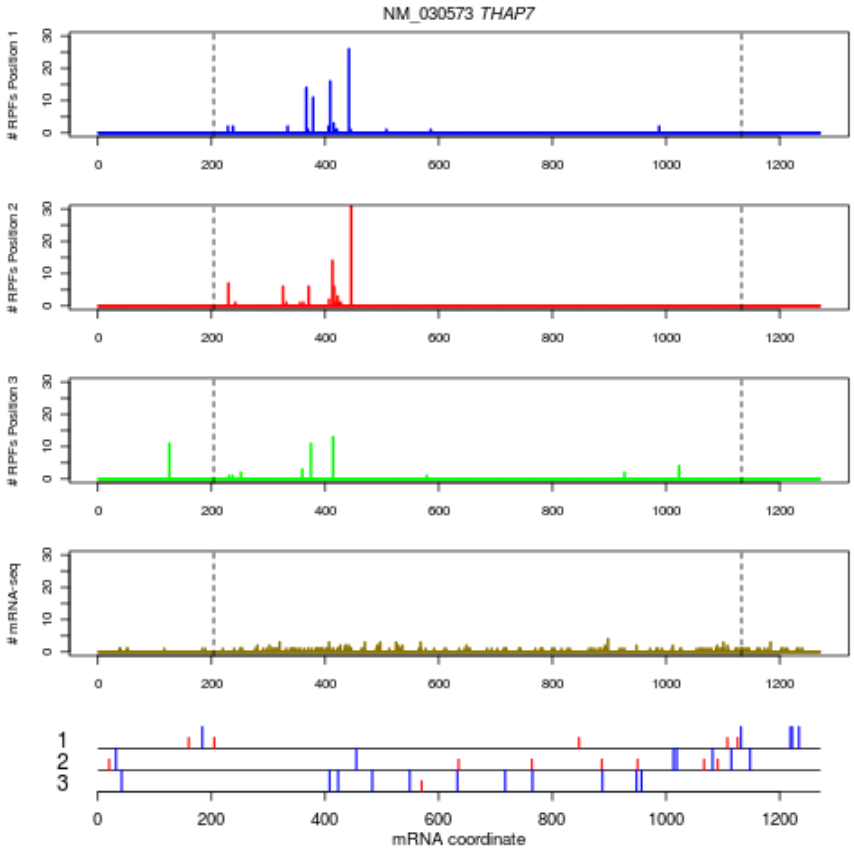
Supplemental Figure 48: Sub-codon profile, mRNA-Seq and ORF organization for



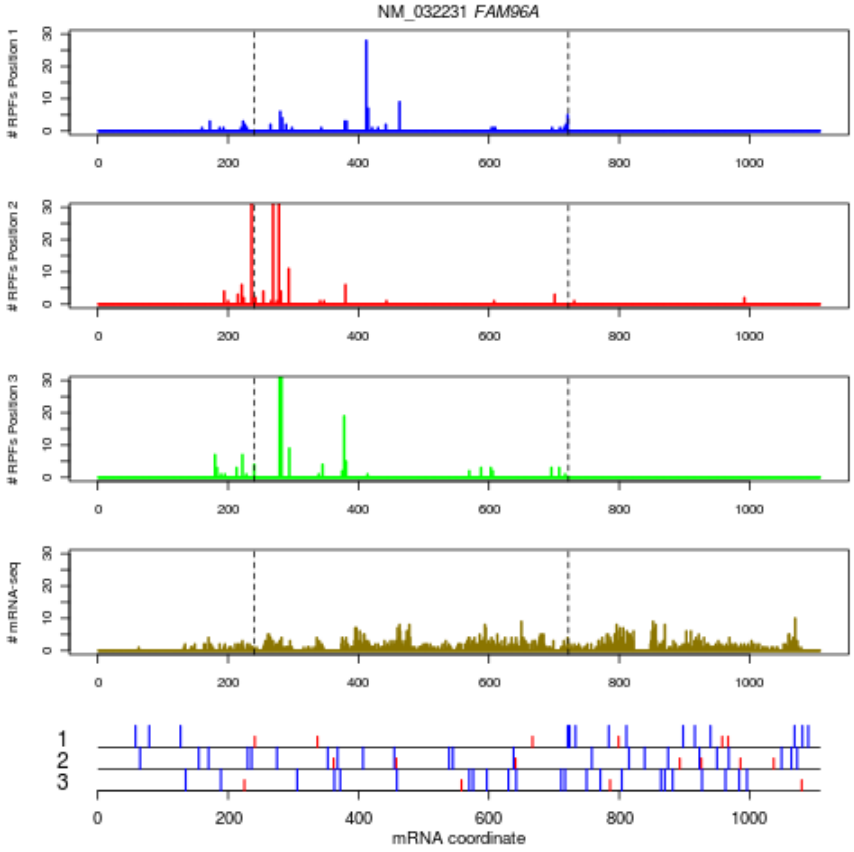
Supplemental Figure 49: Sub-codon profile, mRNA-Seq and ORF organization for



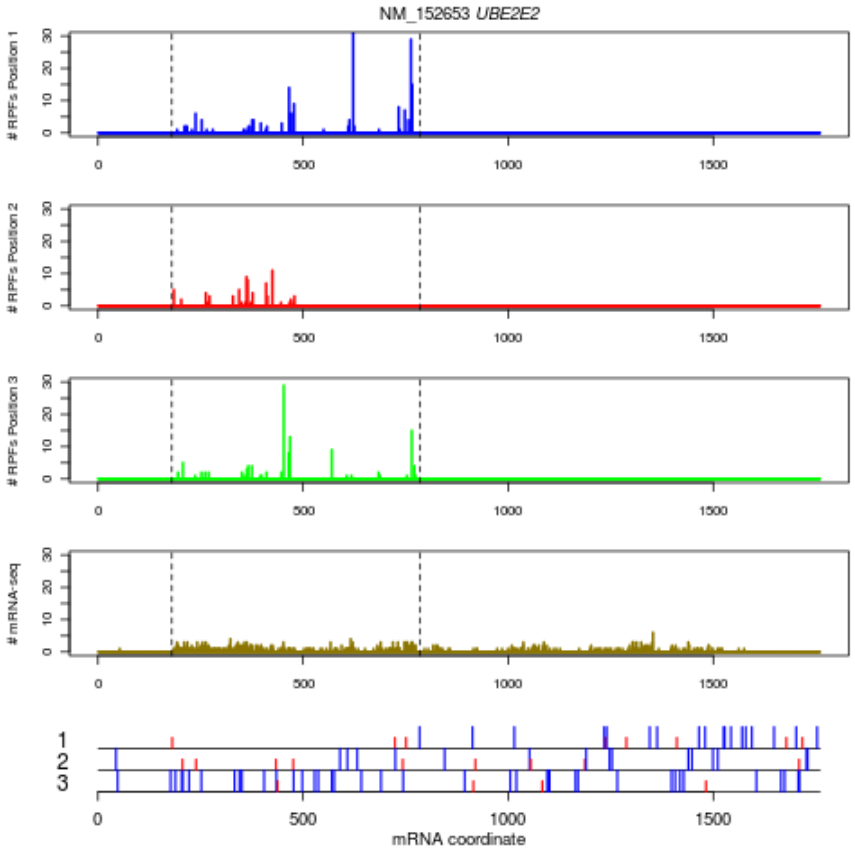
Supplemental Figure 50: Sub-codon profile, mRNA-Seq and ORF organization for



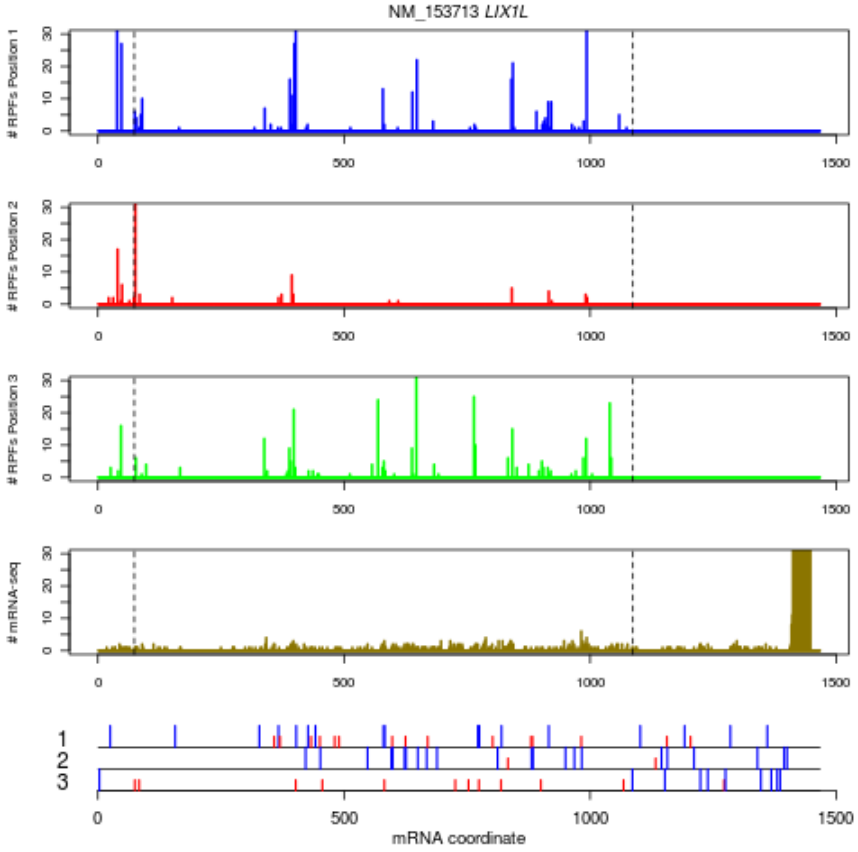
Supplemental Figure 51: Sub-codon profile, mRNA-Seq and ORF organization for



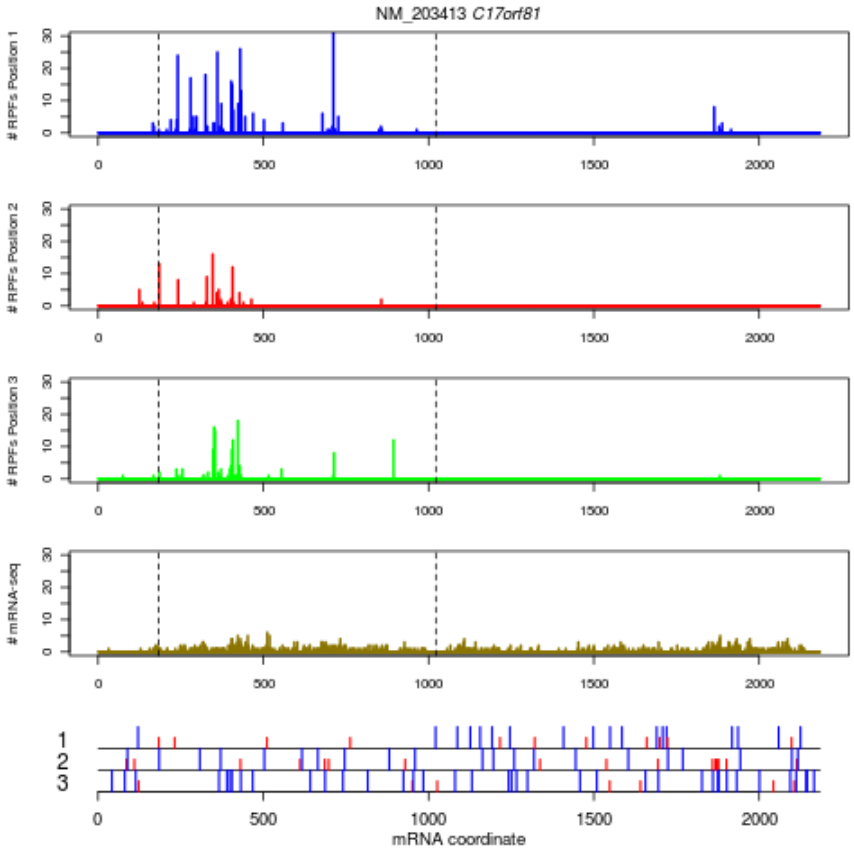
Supplemental Figure 52: Sub-codon profile, mRNA-Seq and ORF organization for



Supplemental Figure 53: Sub-codon profile, mRNA-Seq and ORF organization for

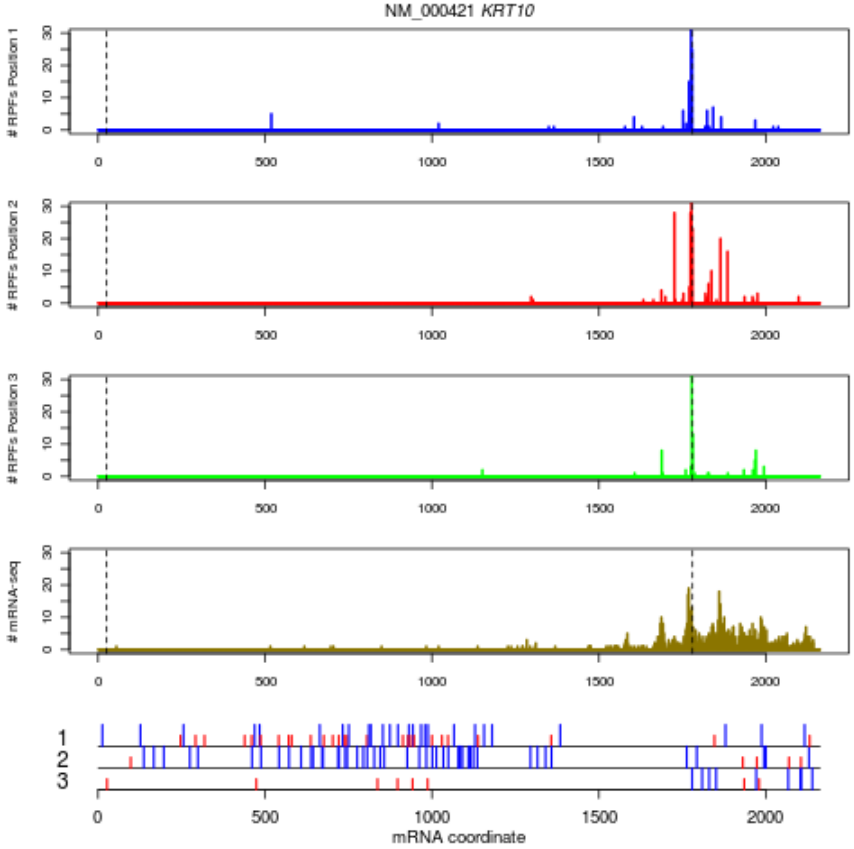


Supplemental Figure 54: Sub-codon profile, mRNA-Seq and ORF organization for

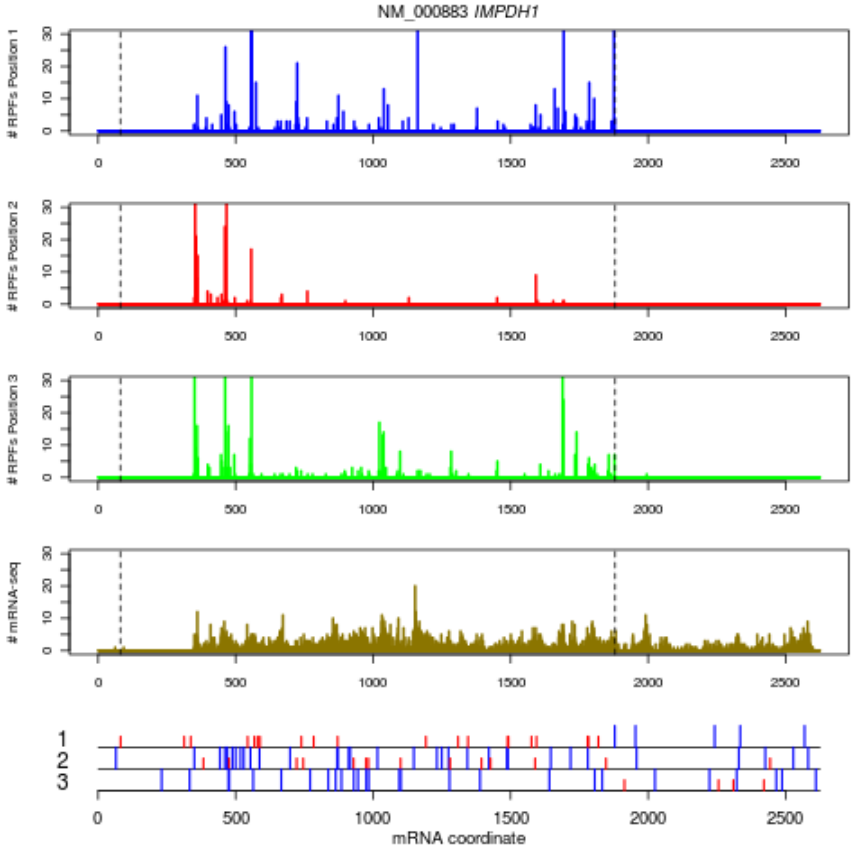


4. Alternative transcription initiation/alternative splicing cases.

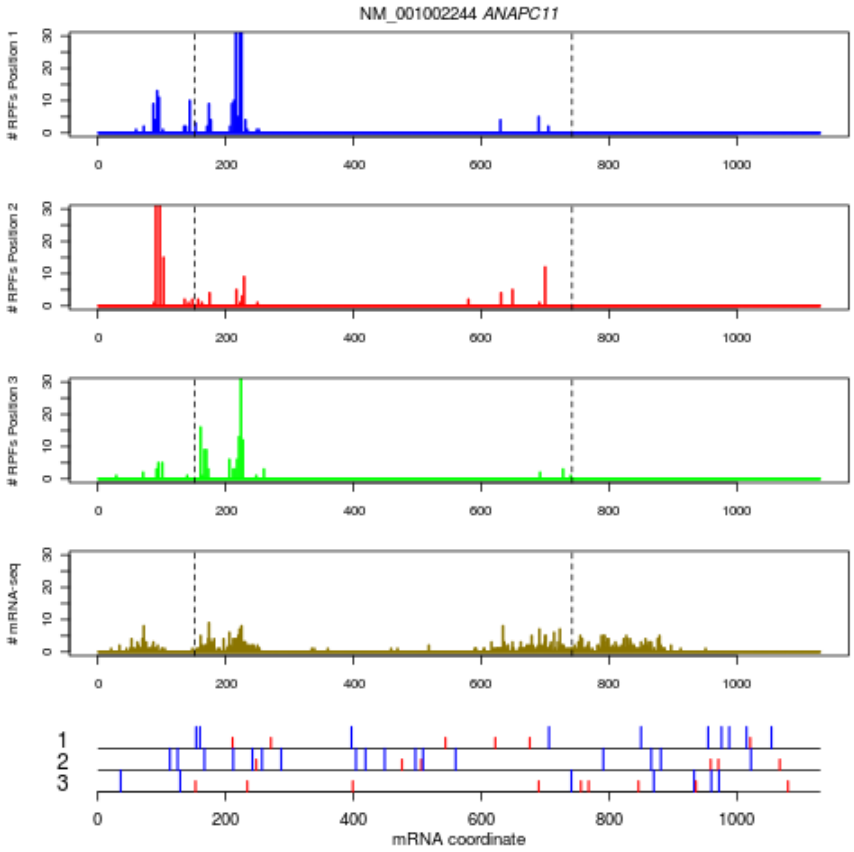
Supplemental Figure 55: Sub-codon profile, mRNA-Seq and ORF organization for



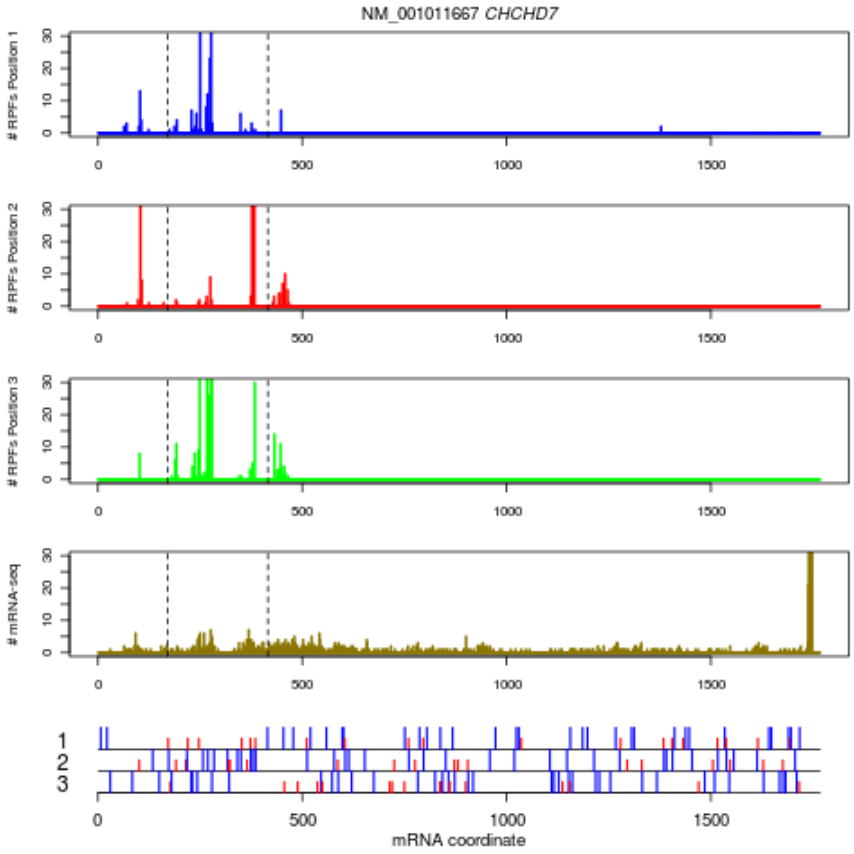
Supplemental Figure 56: Sub-codon profile, mRNA-Seq and ORF organization for



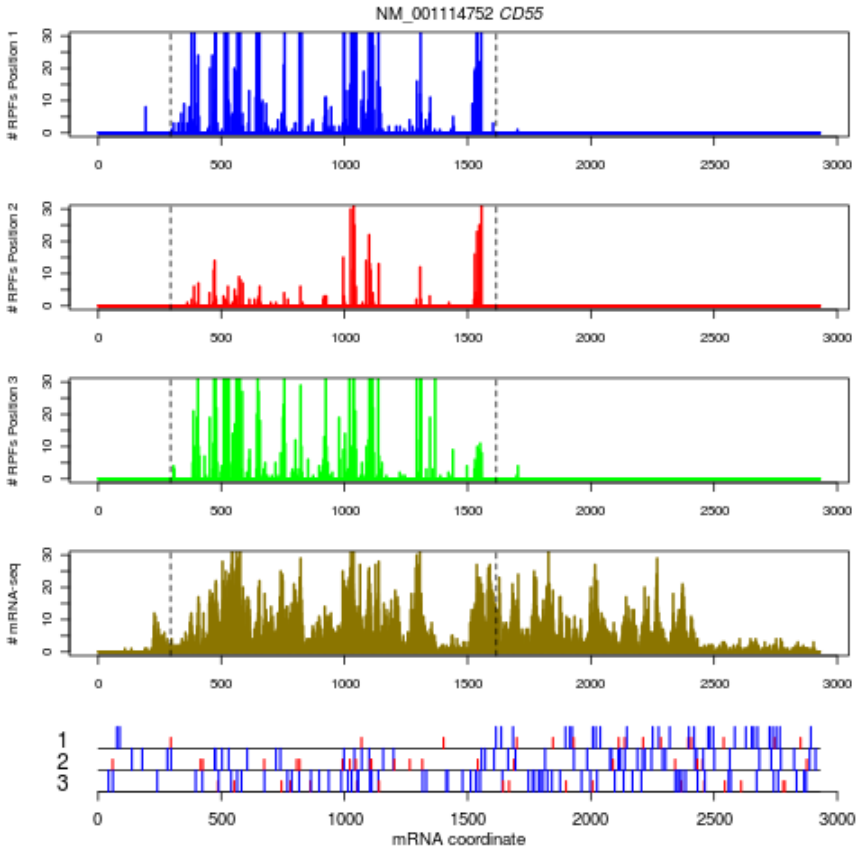
Supplemental Figure 57: Sub-codon profile, mRNA-Seq and ORF organization for



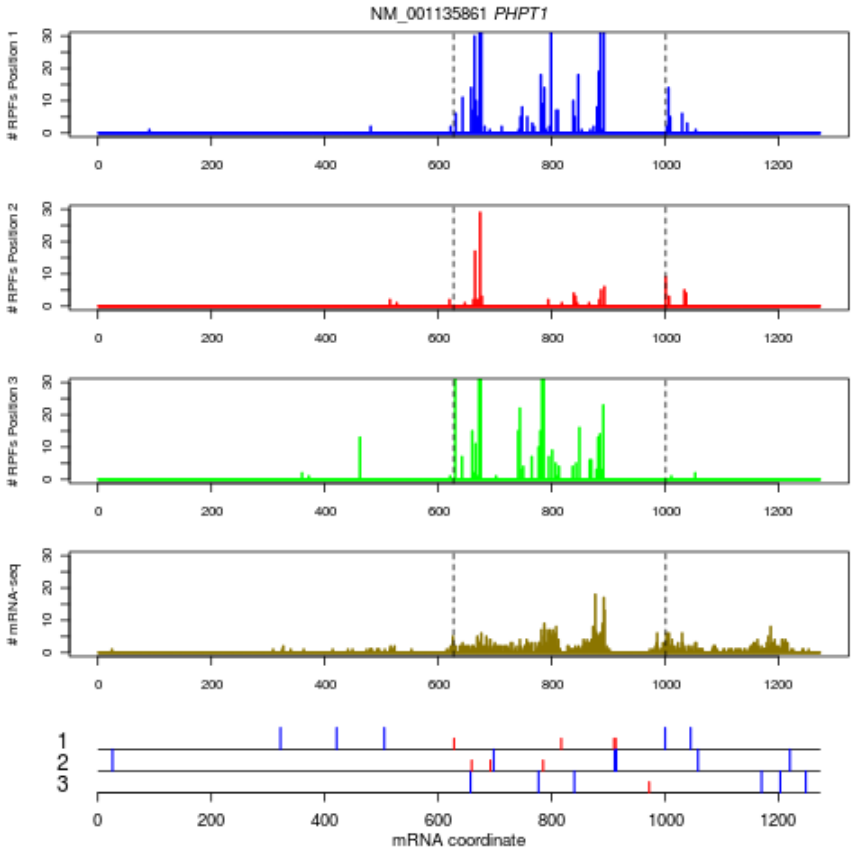
Supplemental Figure 58: Sub-codon profile, mRNA-Seq and ORF organization for



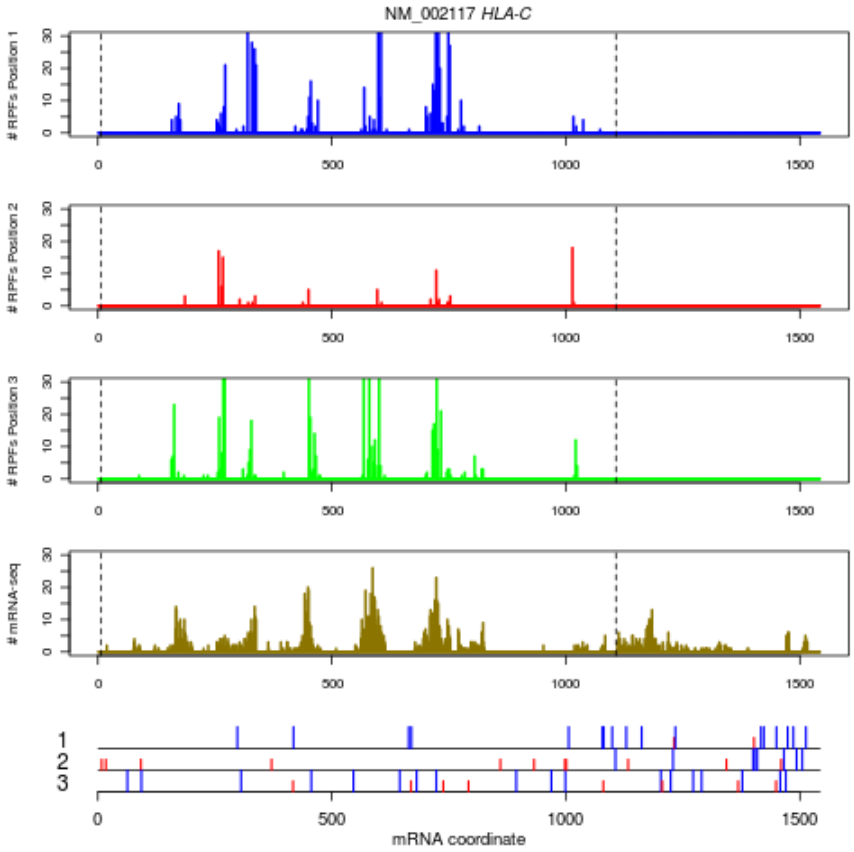
Supplemental Figure 59: Sub-codon profile, mRNA-Seq and ORF organization for



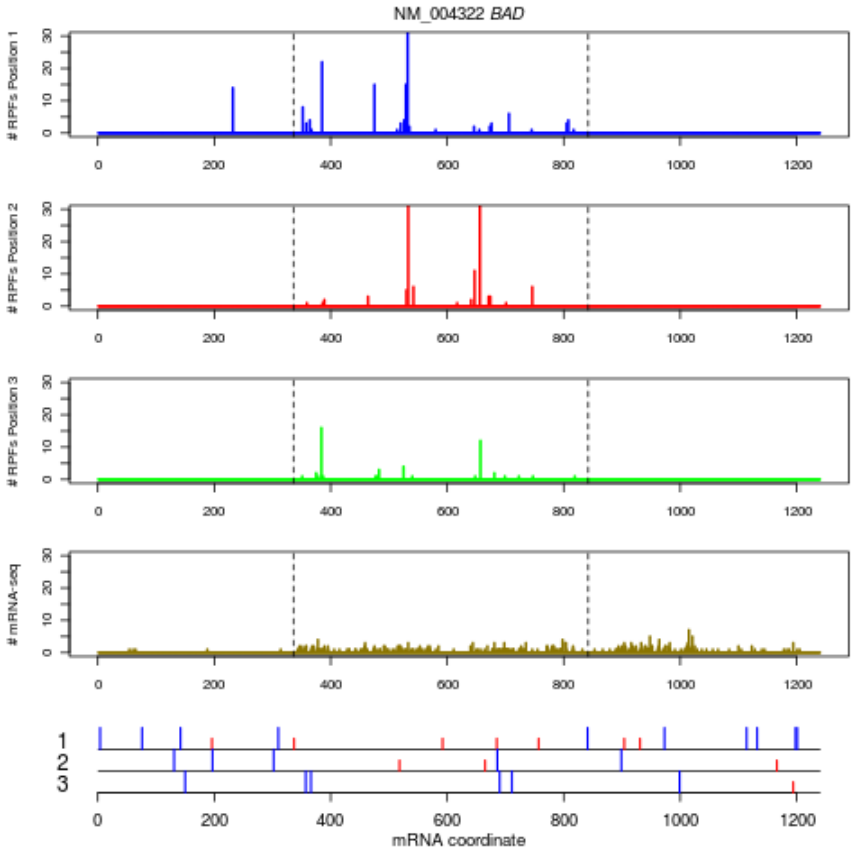
Supplemental Figure 60: Sub-codon profile, mRNA-Seq and ORF organization for



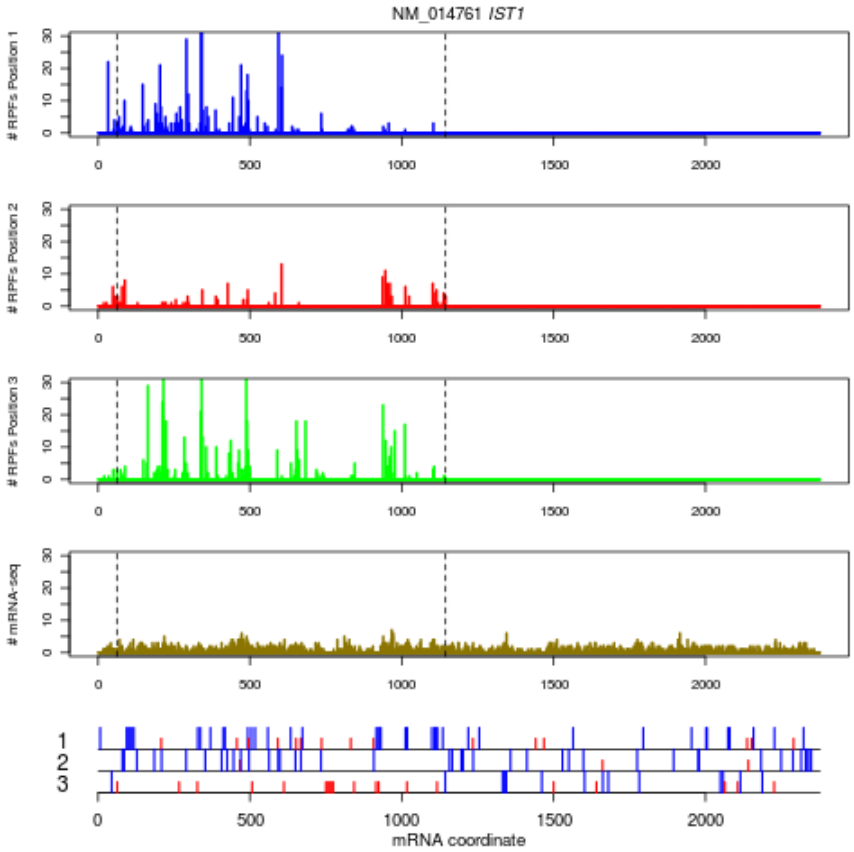
Supplemental Figure 61: Sub-codon profile, mRNA-Seq and ORF organization for



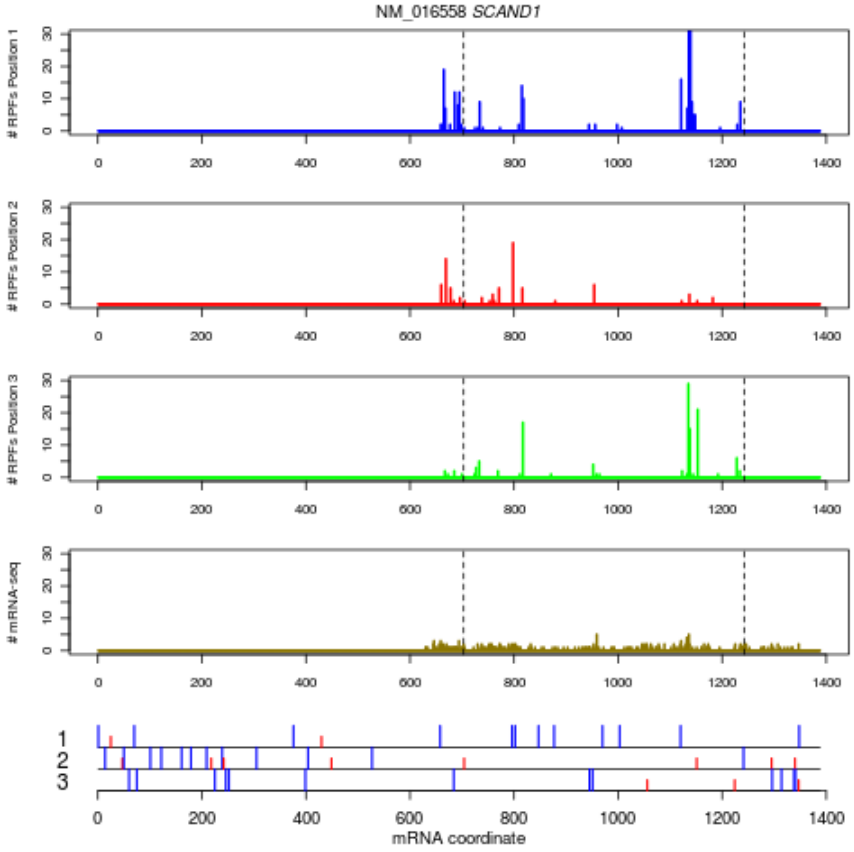
Supplemental Figure 62: Sub-codon profile, mRNA-Seq and ORF organization for



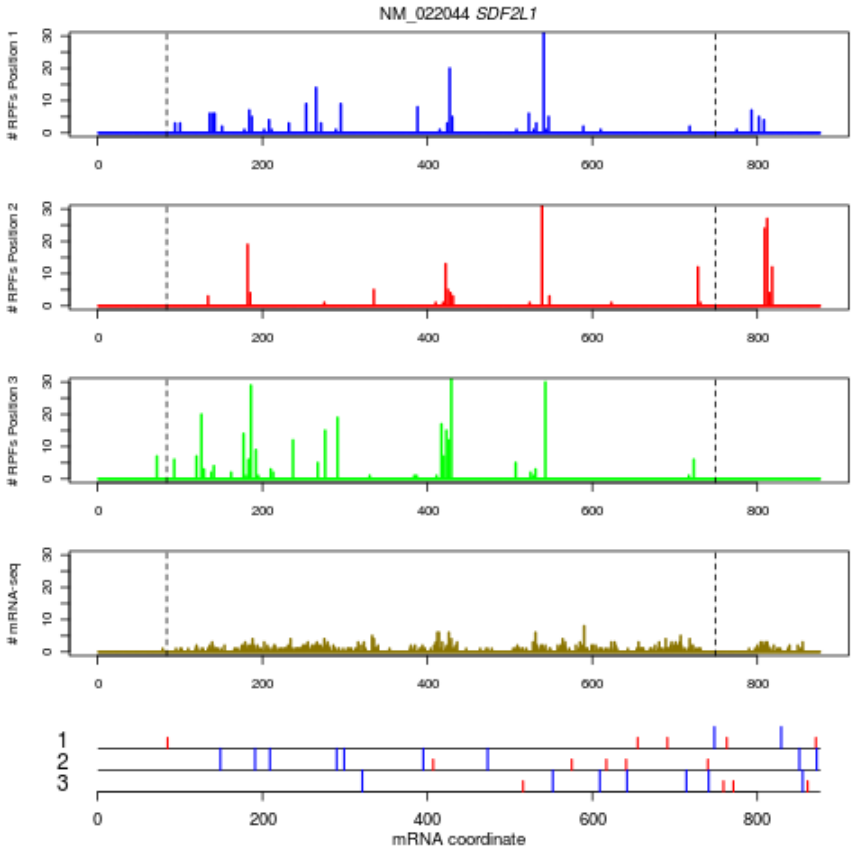
Supplemental Figure 63: Sub-codon profile, mRNA-Seq and ORF organization for



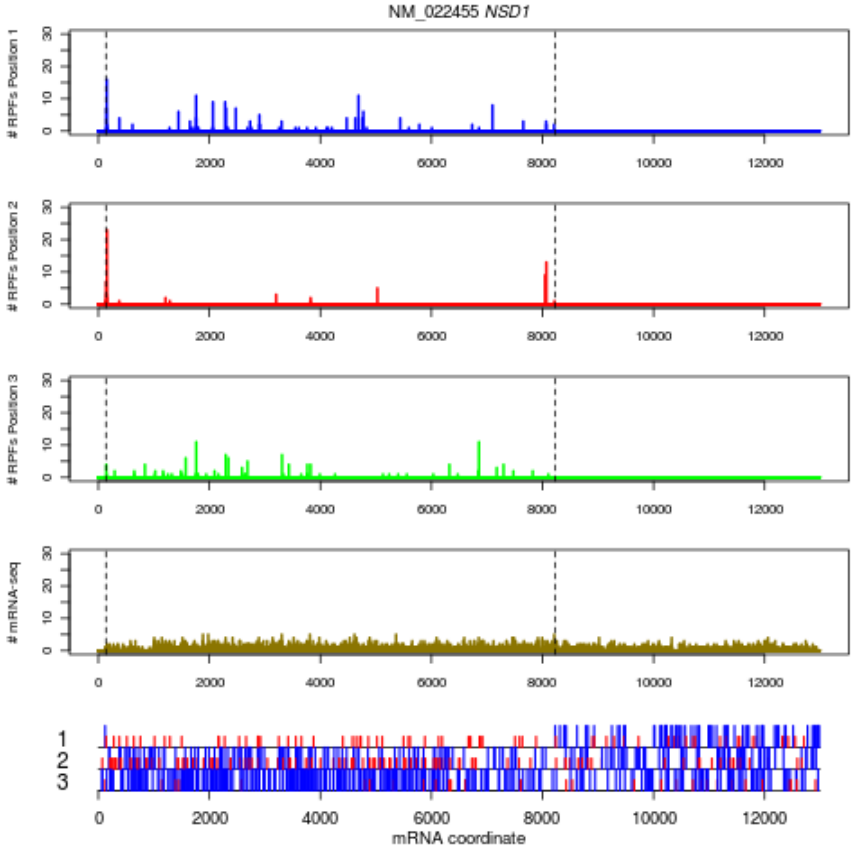
Supplemental Figure 64: Sub-codon profile, mRNA-Seq and ORF organization for



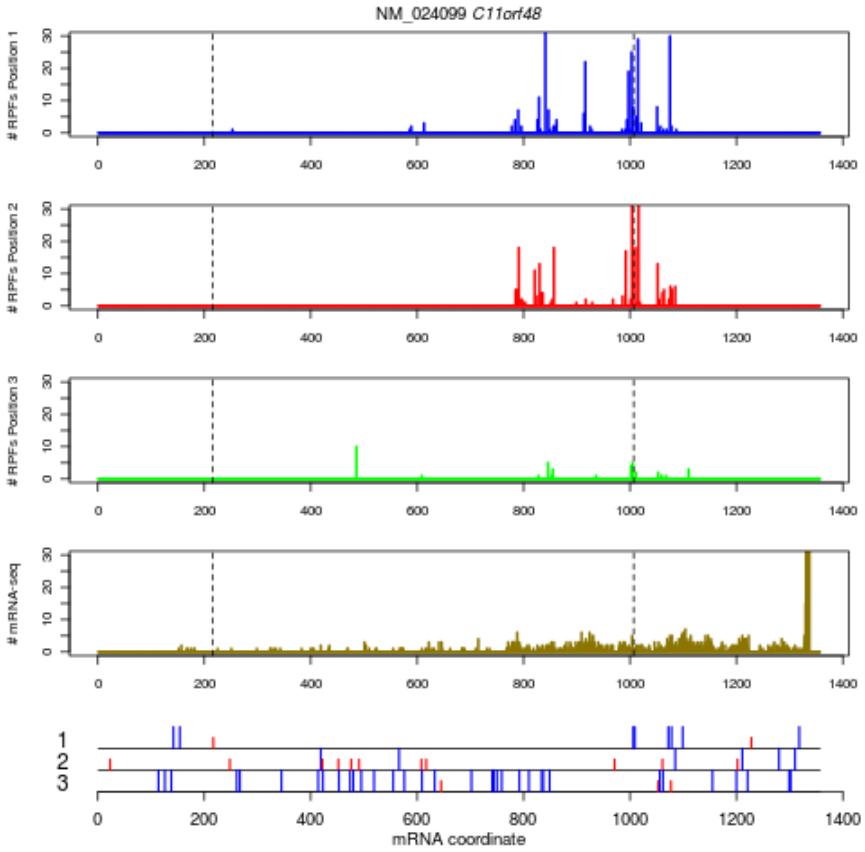
Supplemental Figure 65: Sub-codon profile, mRNA-Seq and ORF organization for



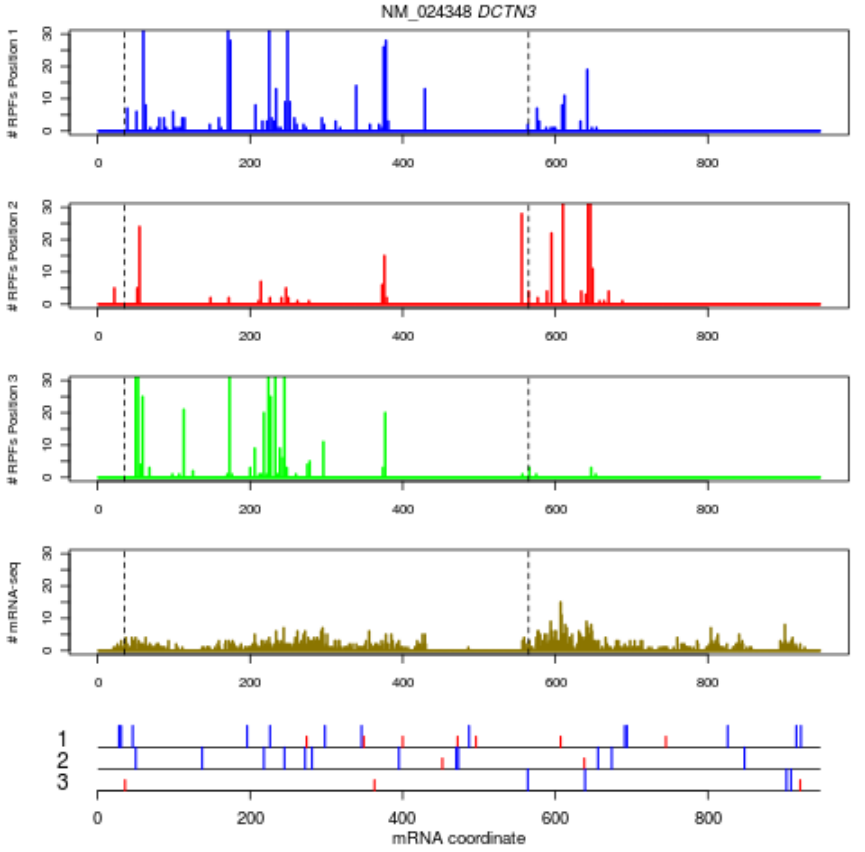
Supplemental Figure 66: Sub-codon profile, mRNA-Seq and ORF organization for



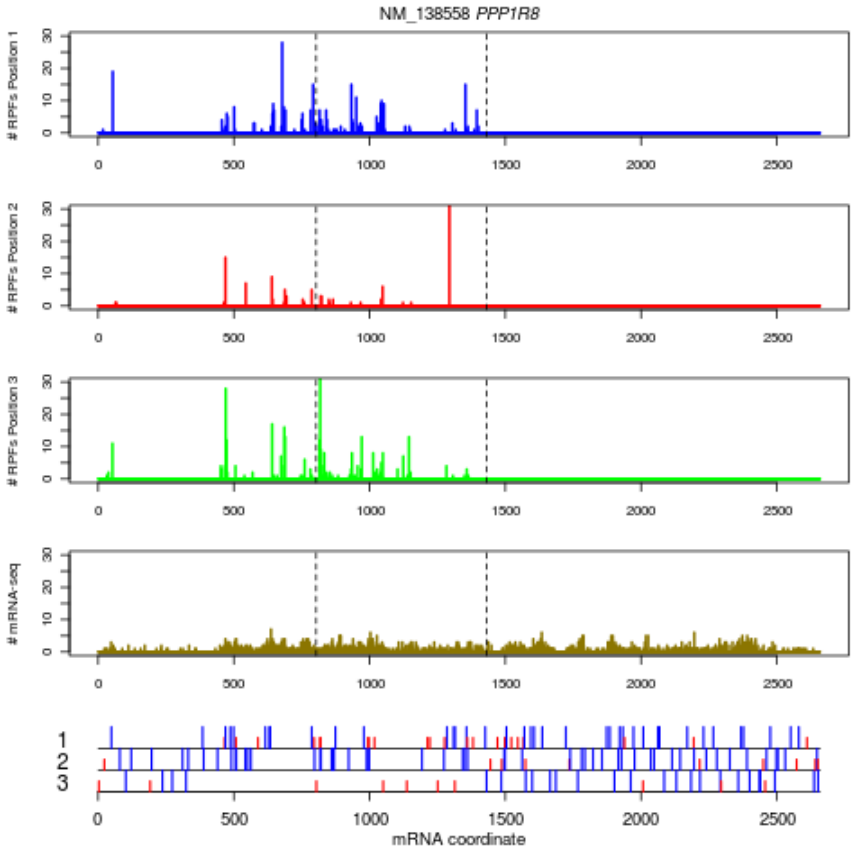
Supplemental Figure 67: Sub-codon profile, mRNA-Seq and ORF organization for



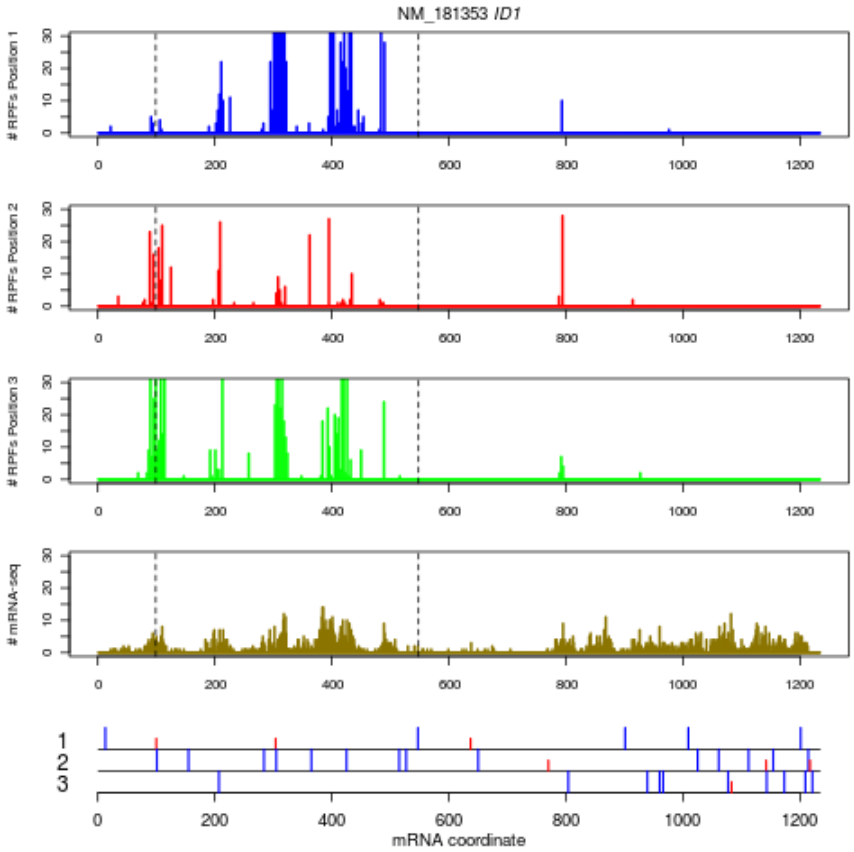
Supplemental Figure 68: Sub-codon profile, mRNA-Seq and ORF organization for



Supplemental Figure 69: Sub-codon profile, mRNA-Seq and ORF organization for

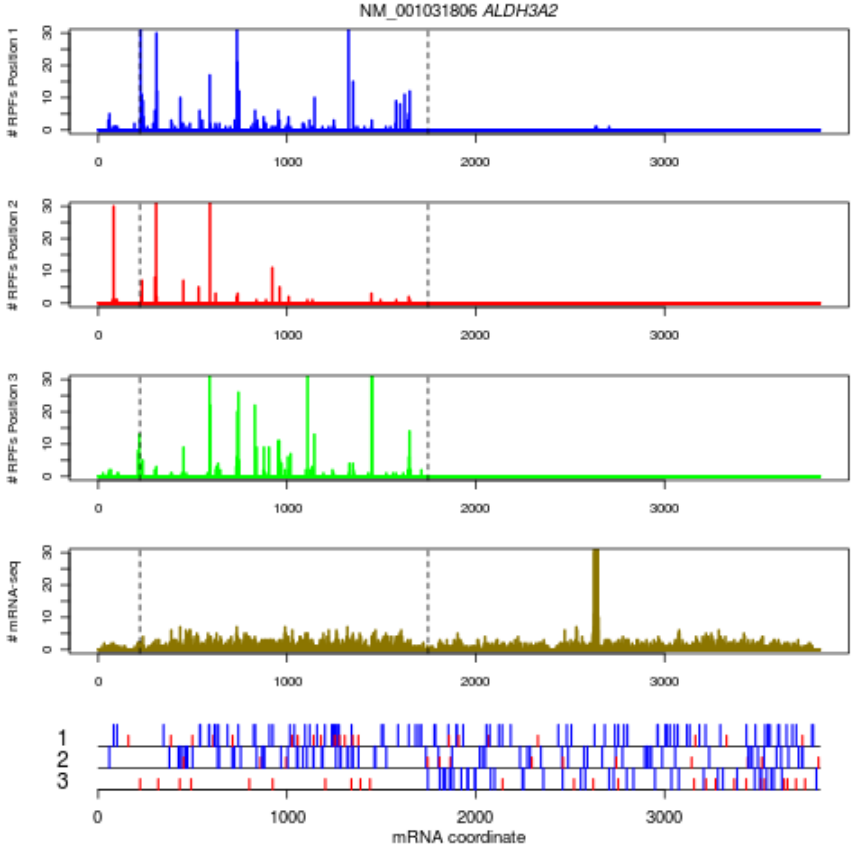


Supplemental Figure 70: Sub-codon profile, mRNA-Seq and ORF organization for

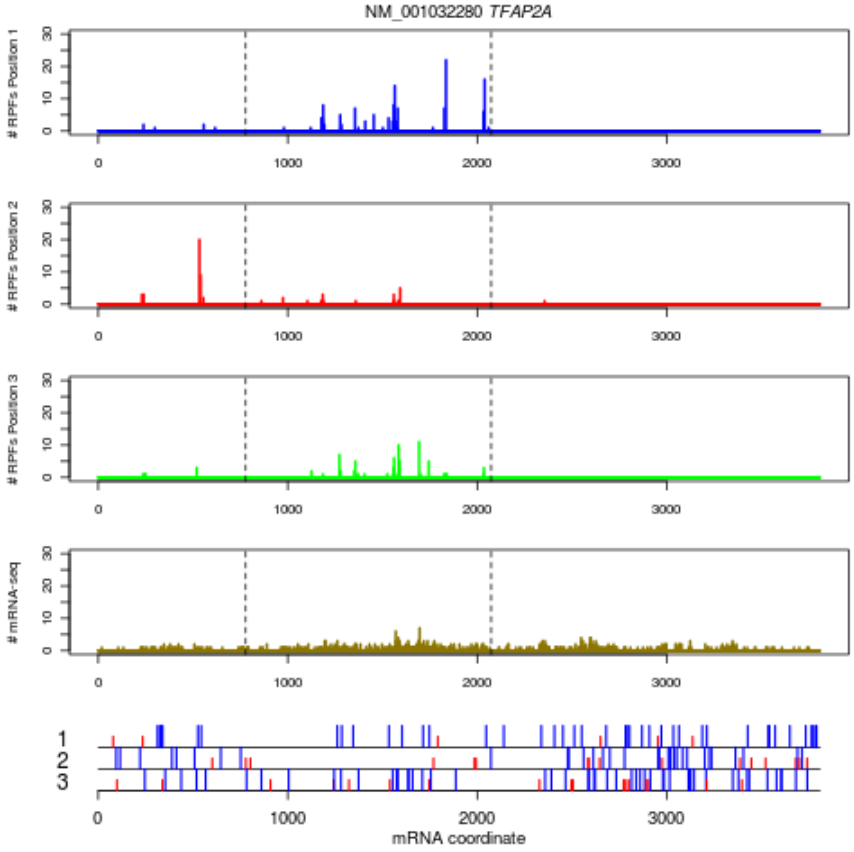


5. Unexplained cases.

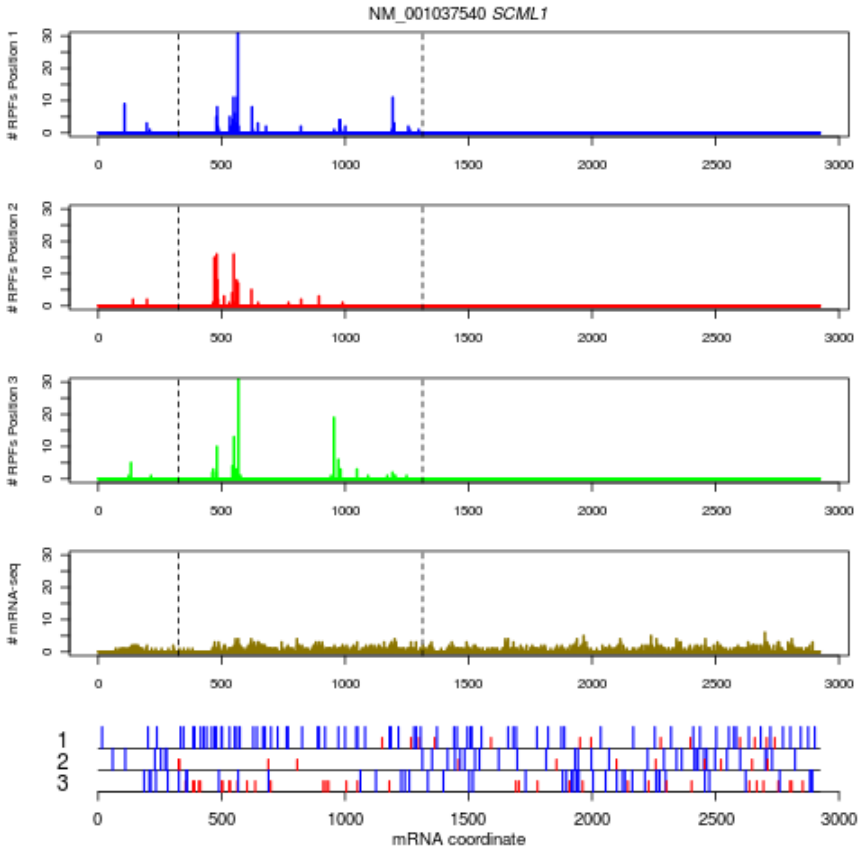
Supplemental Figure 71: Sub-codon profile, mRNA-Seq and ORF organization for



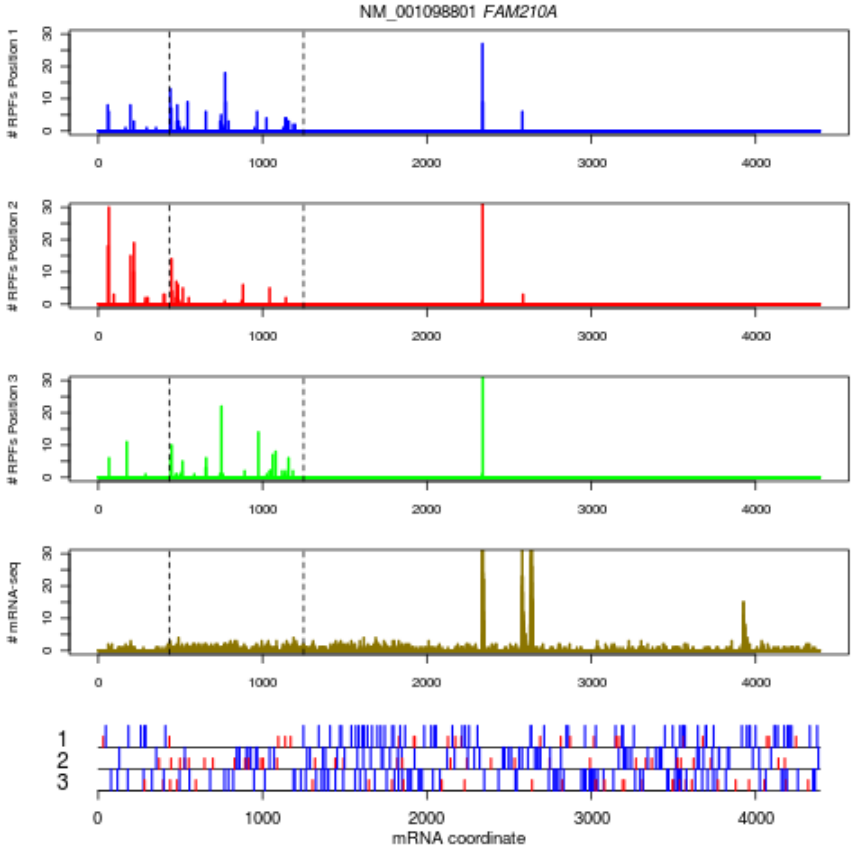
Supplemental Figure 72: Sub-codon profile, mRNA-Seq and ORF organization for



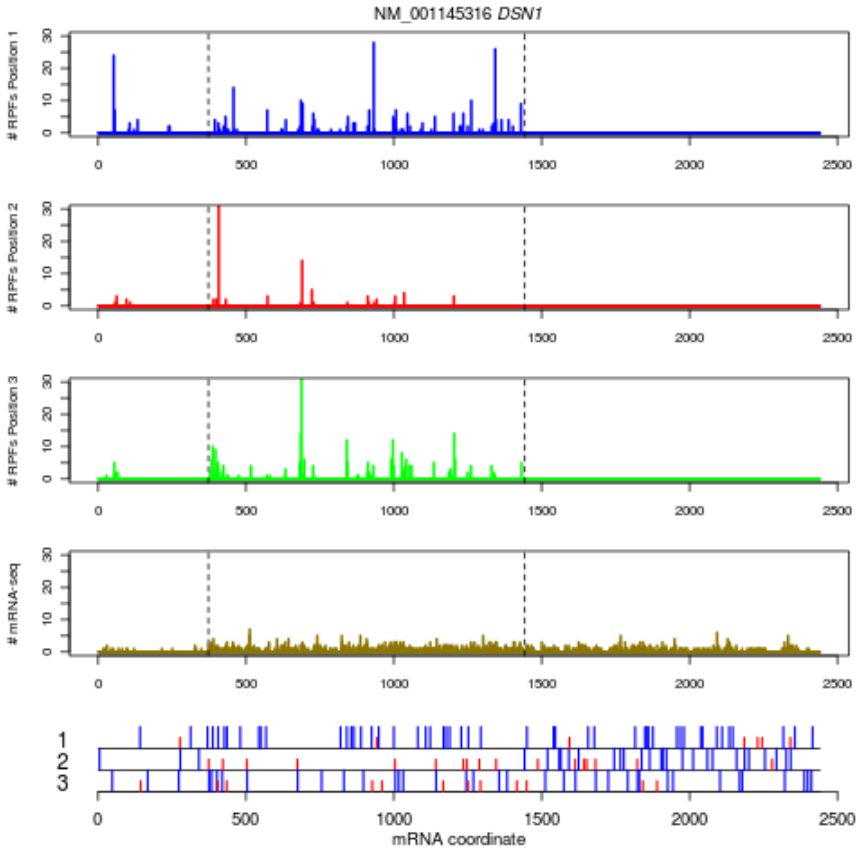
Supplemental Figure 73: Sub-codon profile, mRNA-Seq and ORF organization for



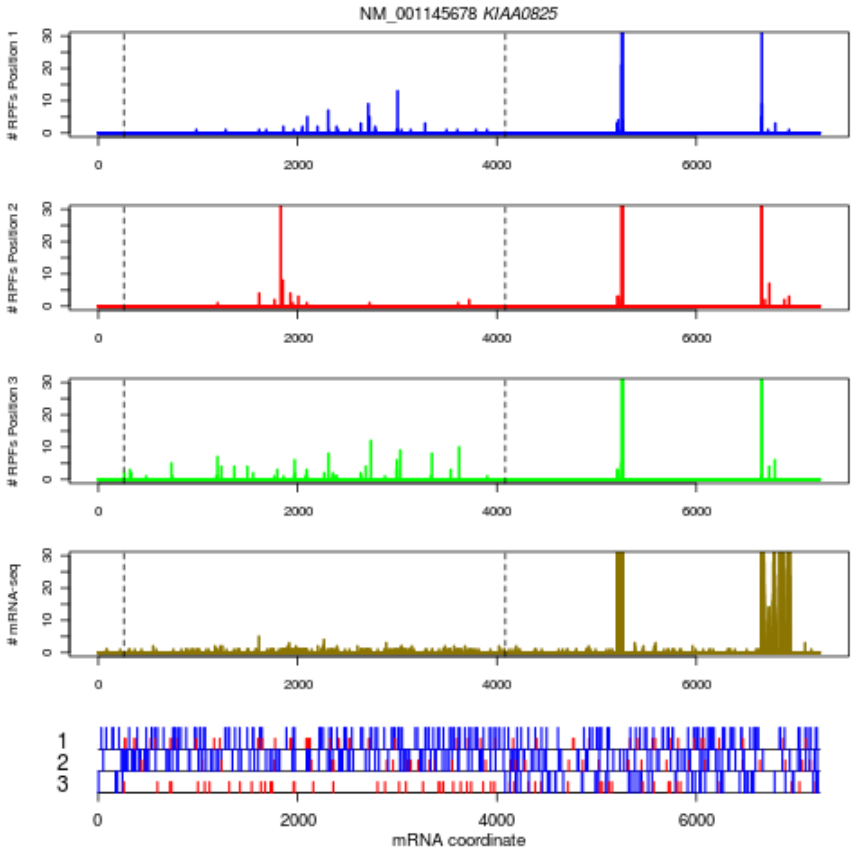
Supplemental Figure 74: Sub-codon profile, mRNA-Seq and ORF organization for



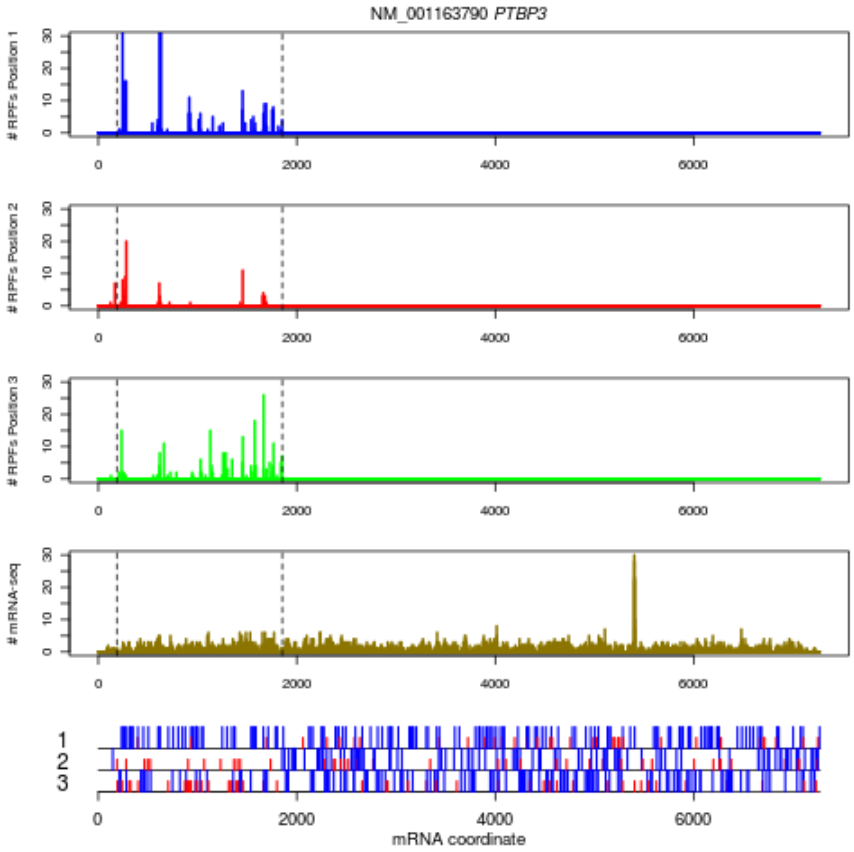
Supplemental Figure 75: Sub-codon profile, mRNA-Seq and ORF organization for



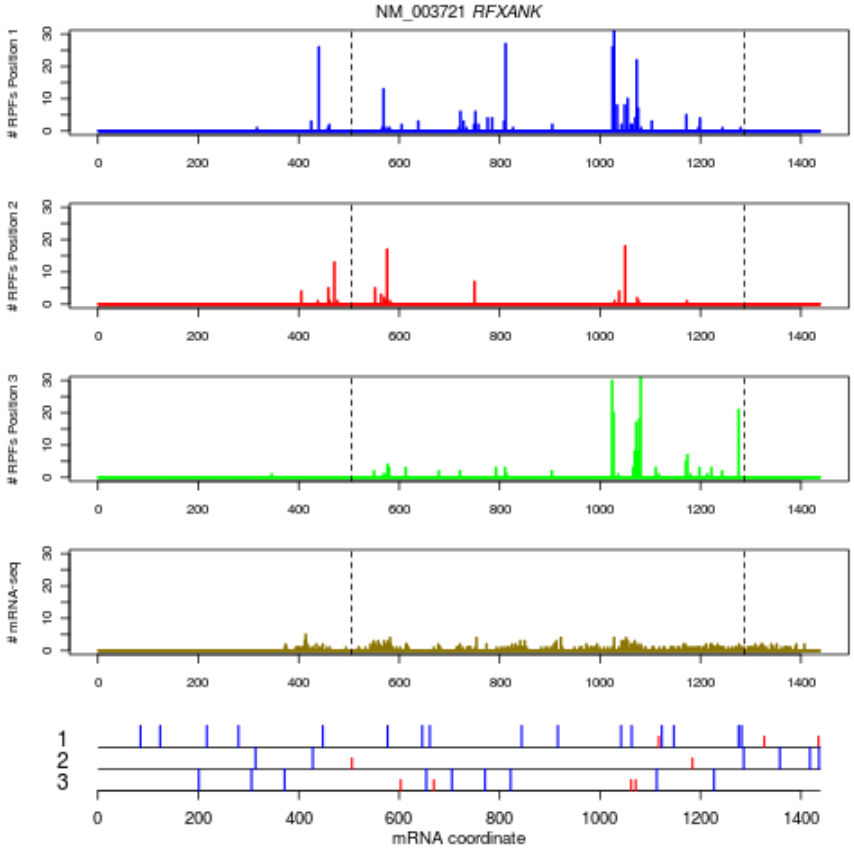
Supplemental Figure 76: Sub-codon profile, mRNA-Seq and ORF organization for



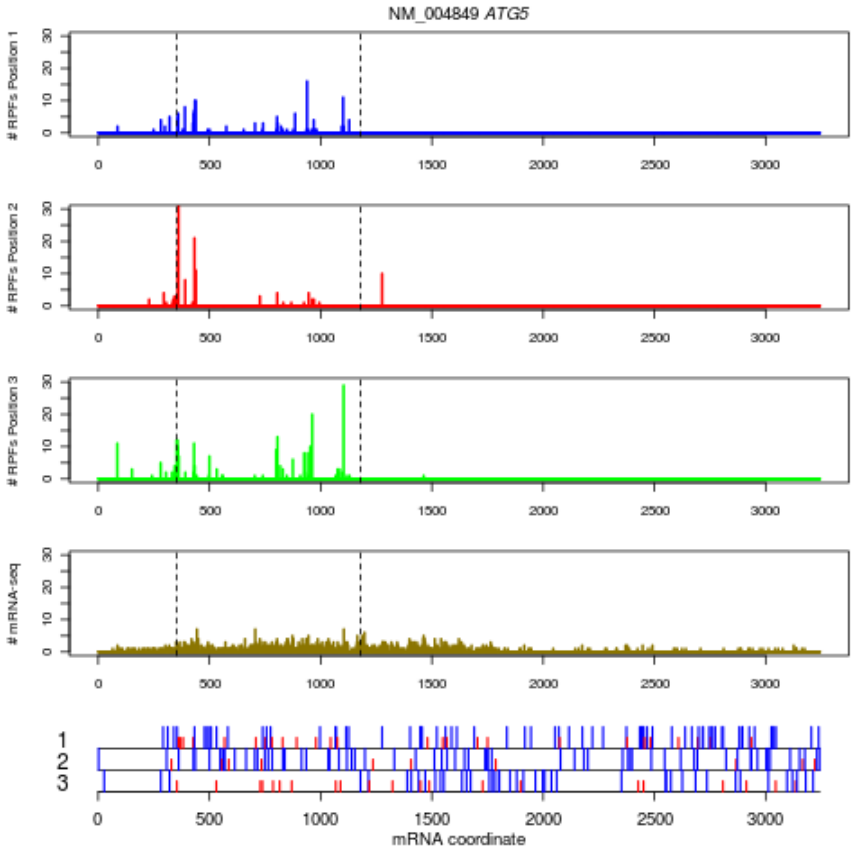
Supplemental Figure 77: Sub-codon profile, mRNA-Seq and ORF organization for



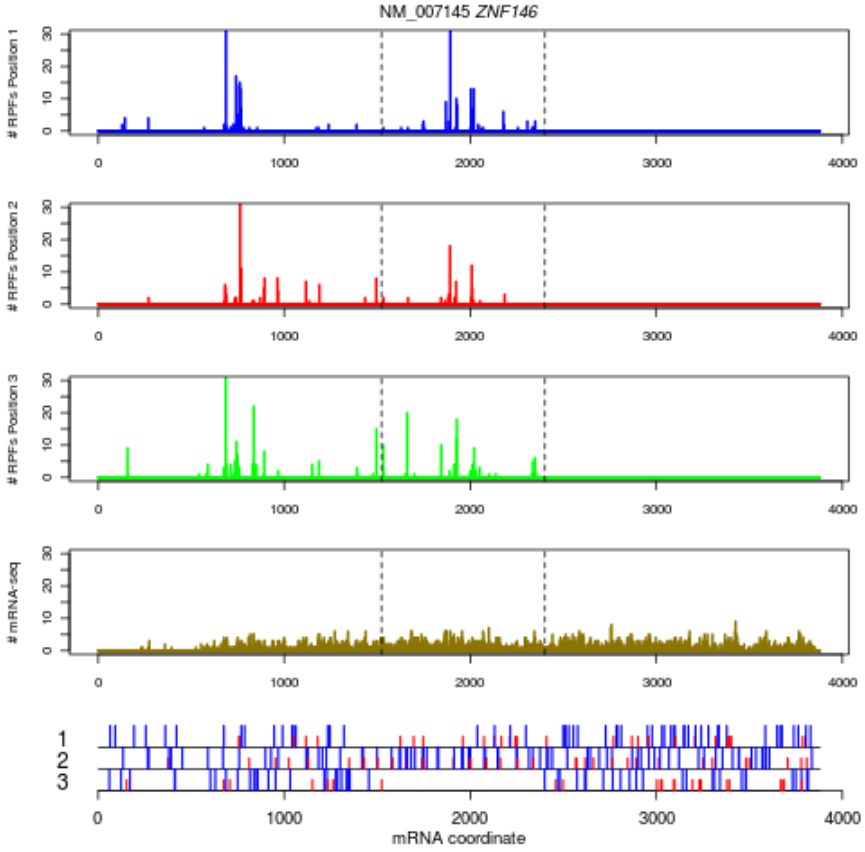
Supplemental Figure 78: Sub-codon profile, mRNA-Seq and ORF organization for



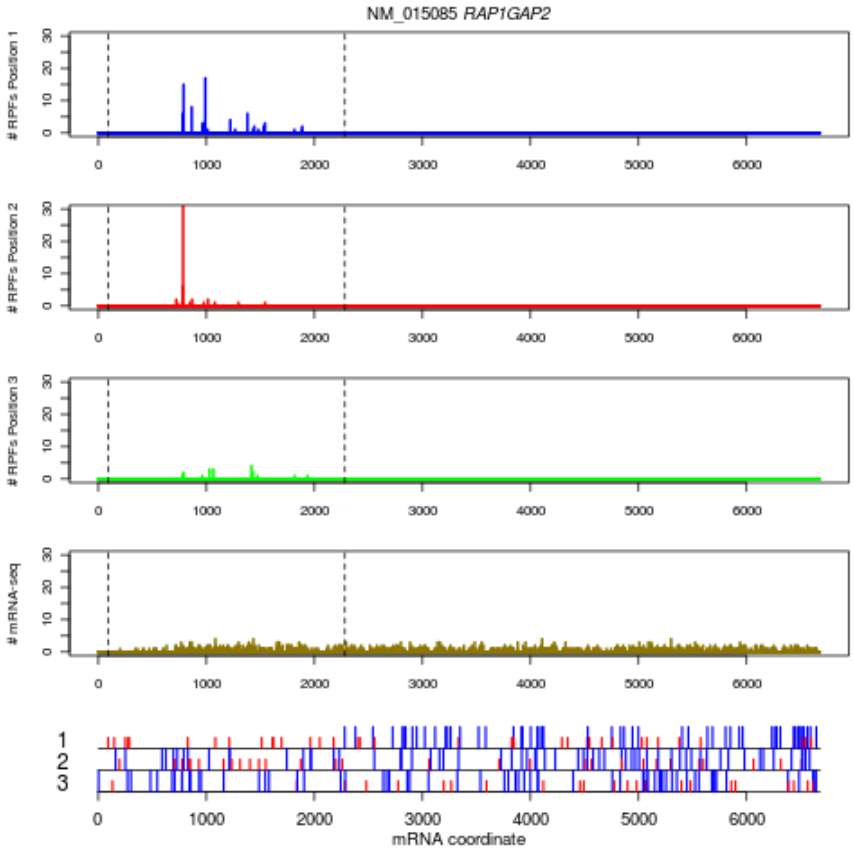
Supplemental Figure 79: Sub-codon profile, mRNA-Seq and ORF organization for



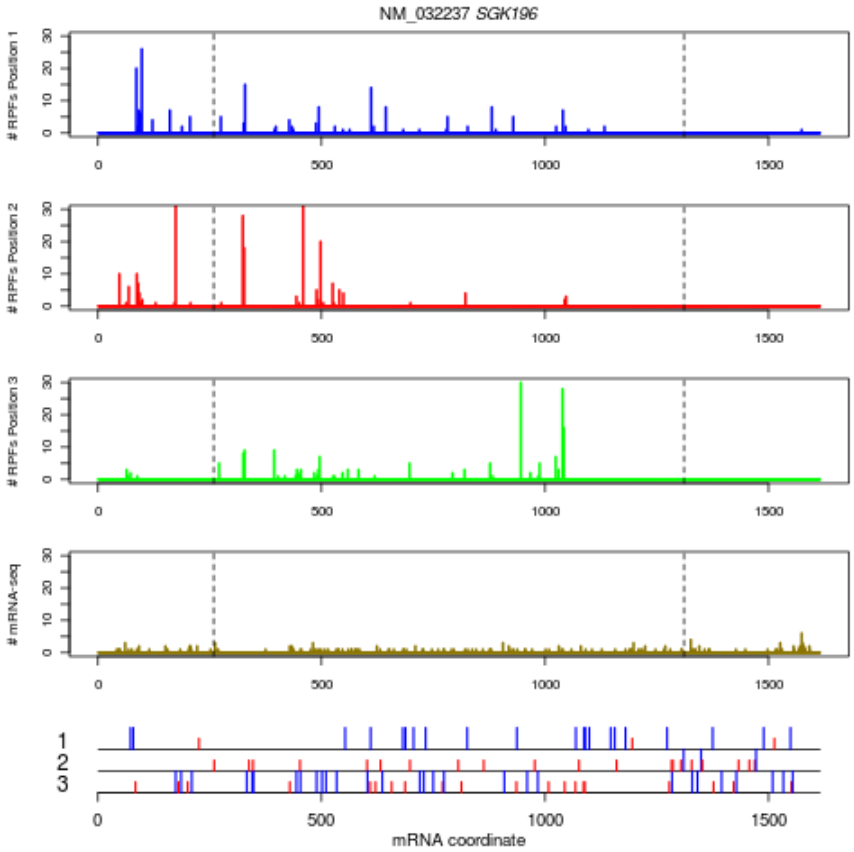
Supplemental Figure 80: Sub-codon profile, mRNA-Seq and ORF organization for



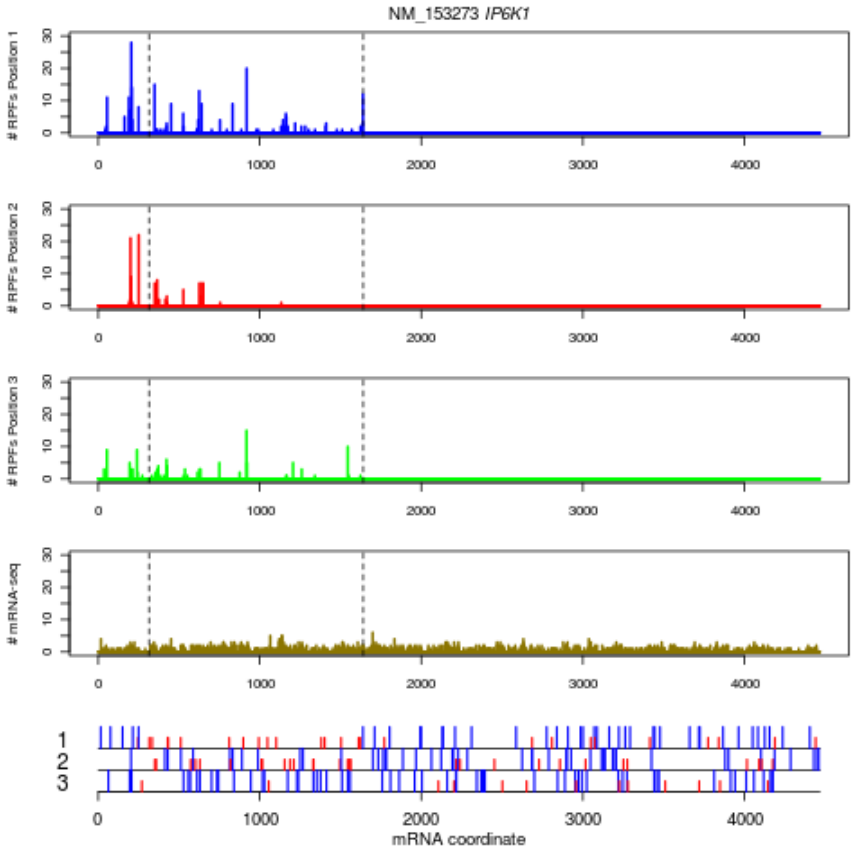
Supplemental Figure 81: Sub-codon profile, mRNA-Seq and ORF organization for



Supplemental Figure 82: Sub-codon profile, mRNA-Seq and ORF organization for

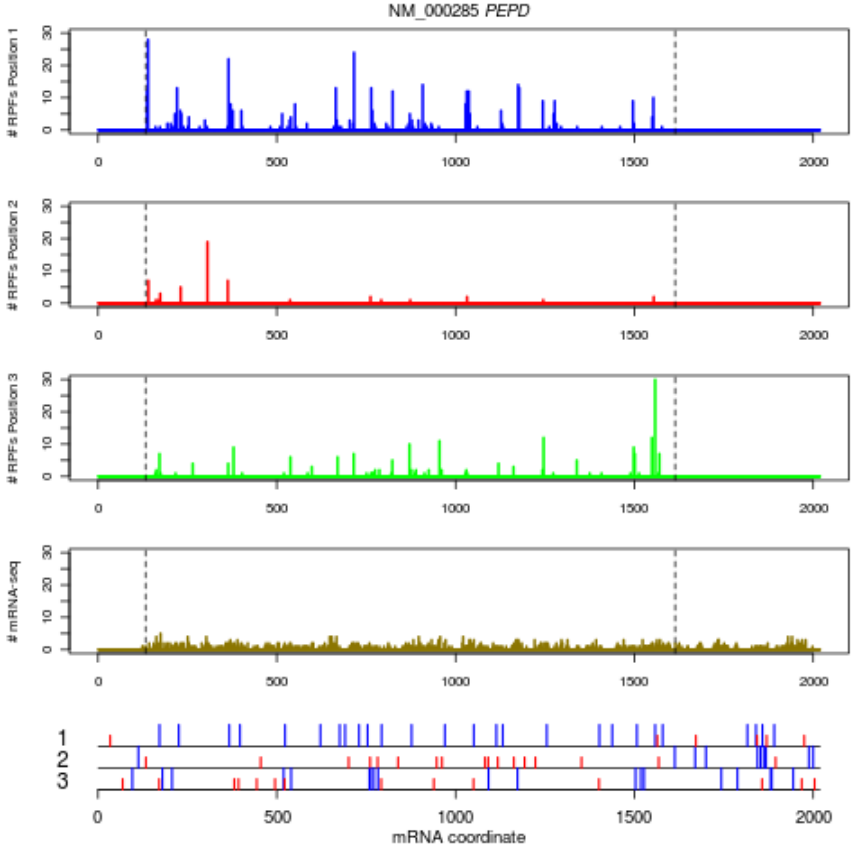


Supplemental Figure 83: Sub-codon profile, mRNA-Seq and ORF organization for

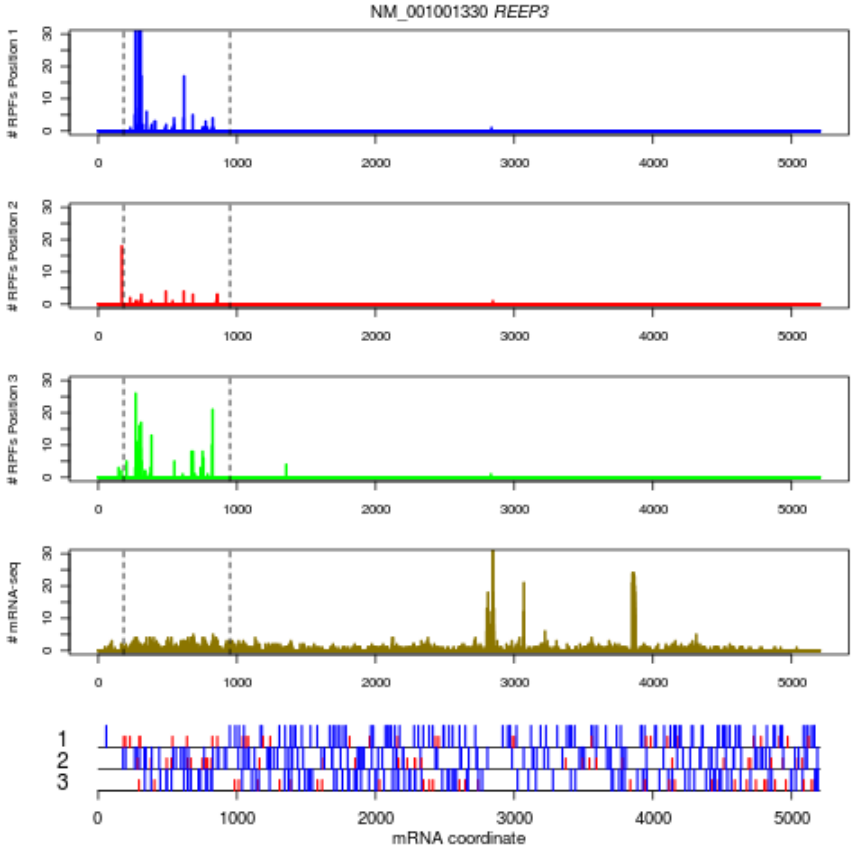


6. False positives.

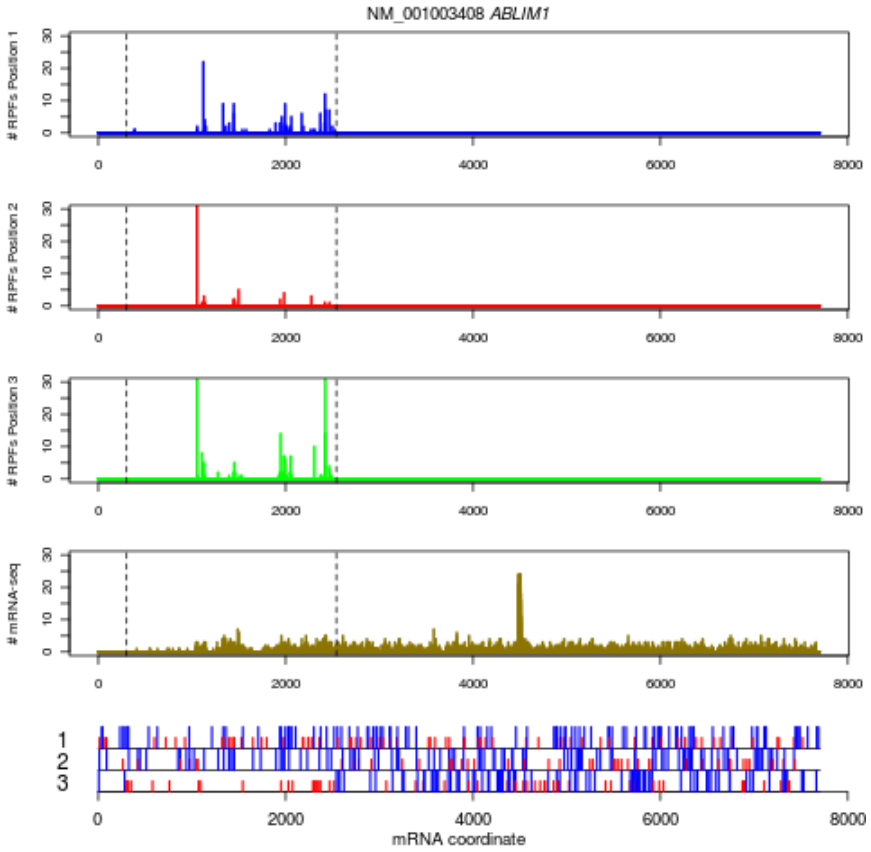
Supplemental Figure 84: Sub-codon profile, mRNA-Seq and ORF organization for



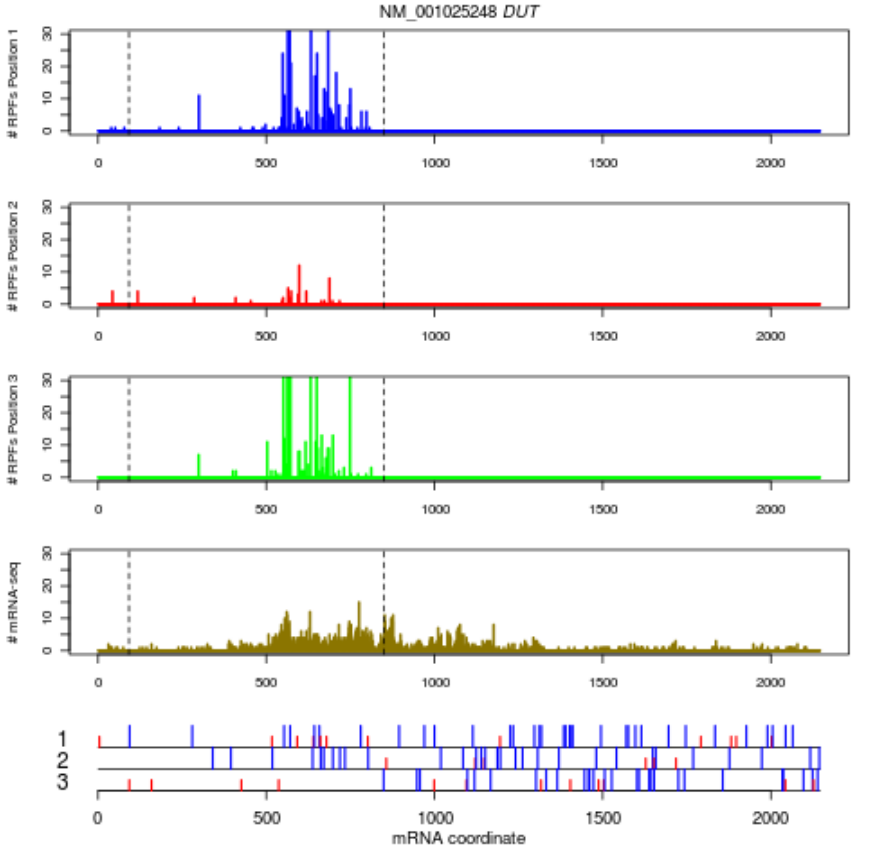
Supplemental Figure 85: Sub-codon profile, mRNA-Seq and ORF organization for



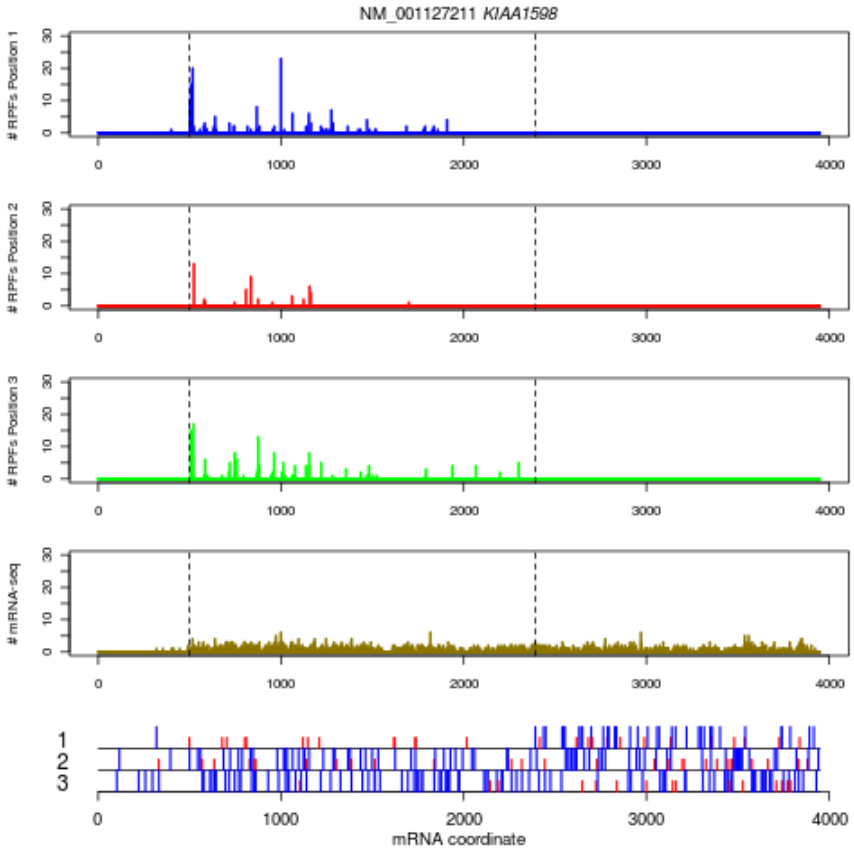
Supplemental Figure 86: Sub-codon profile, mRNA-Seq and ORF organization for



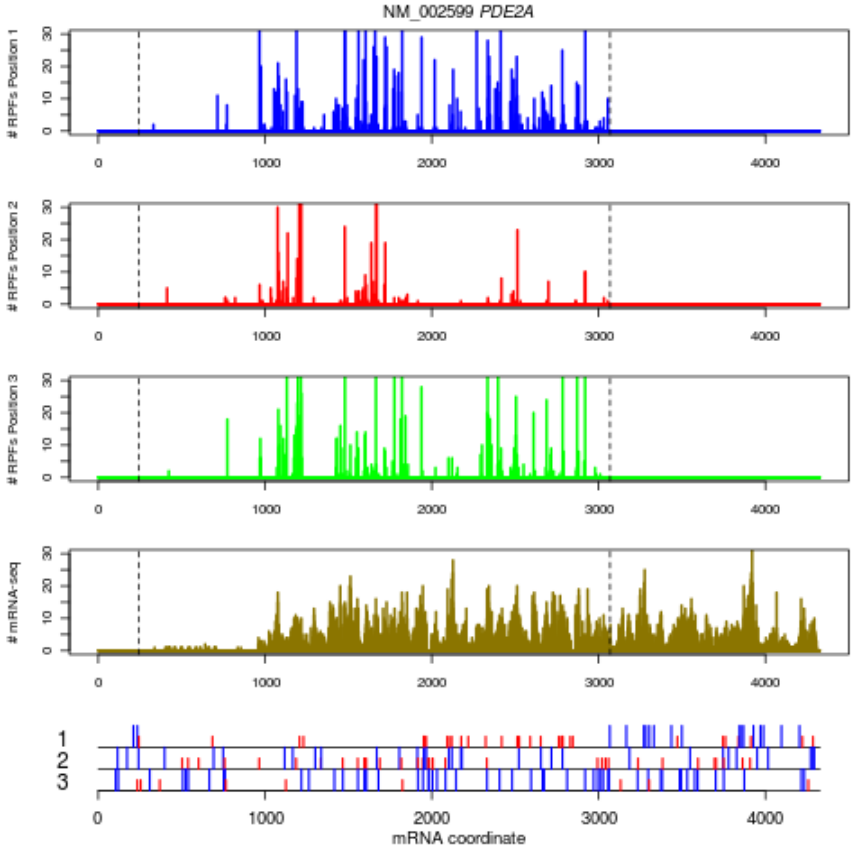
Supplemental Figure 87: Sub-codon profile, mRNA-Seq and ORF organization for



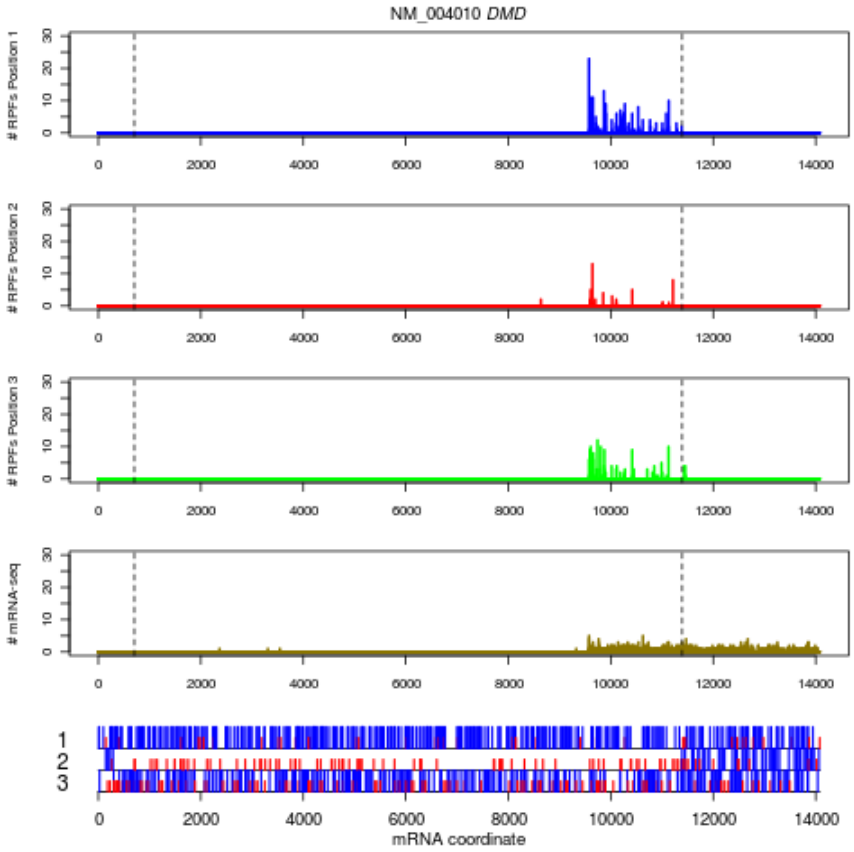
Supplemental Figure 88: Sub-codon profile, mRNA-Seq and ORF organization for



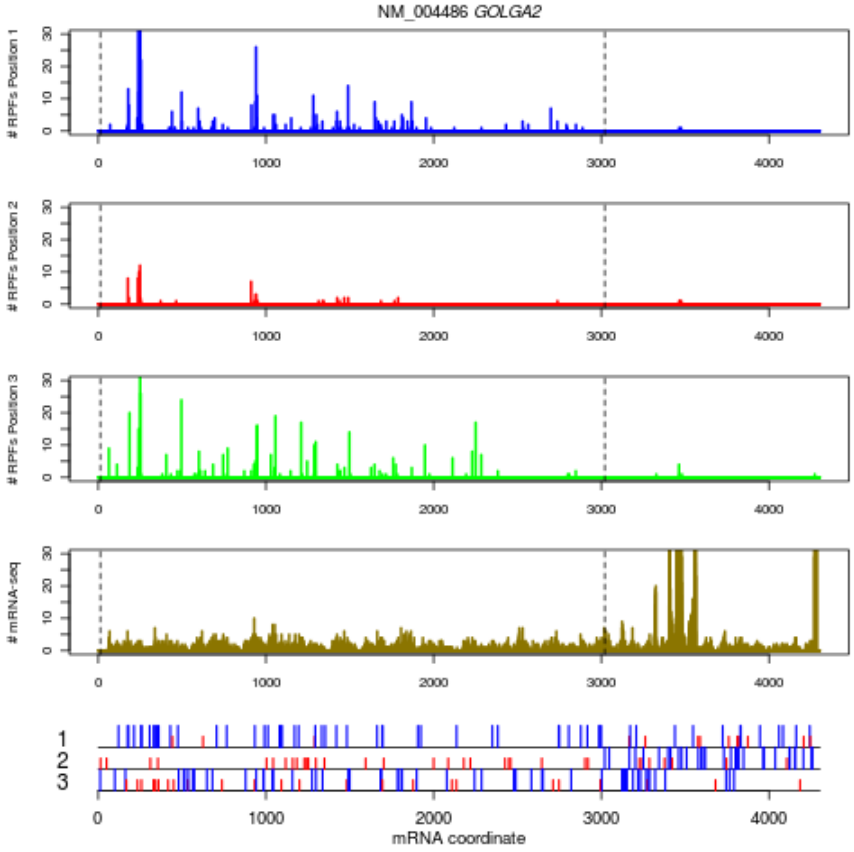
Supplemental Figure 89: Sub-codon profile, mRNA-Seq and ORF organization for



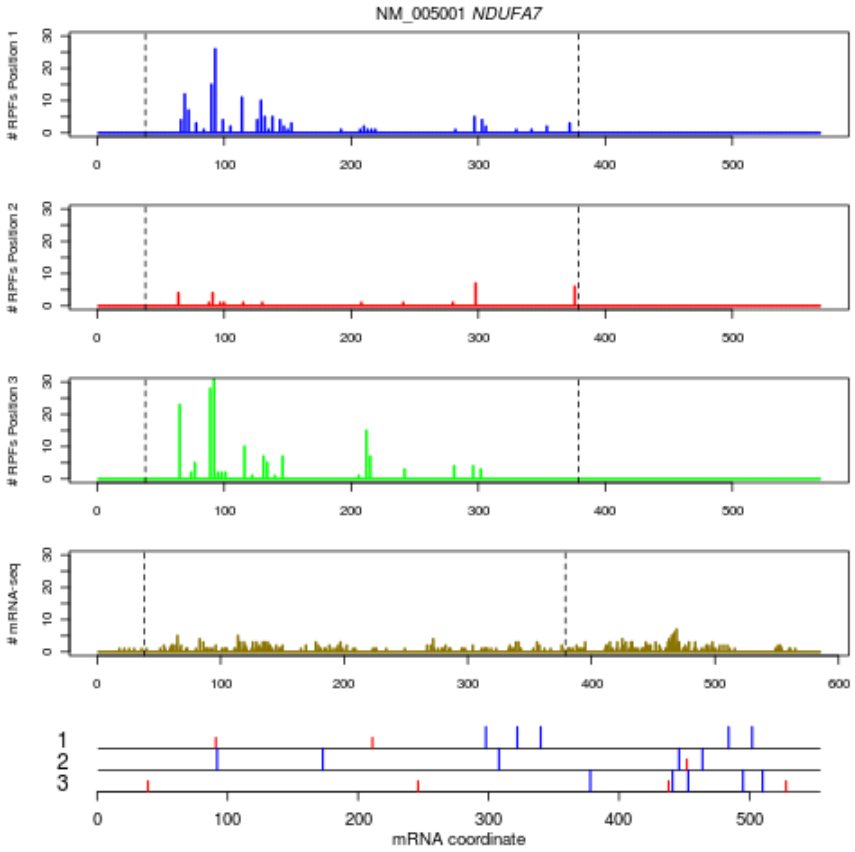
Supplemental Figure 90: Sub-codon profile, mRNA-Seq and ORF organization for



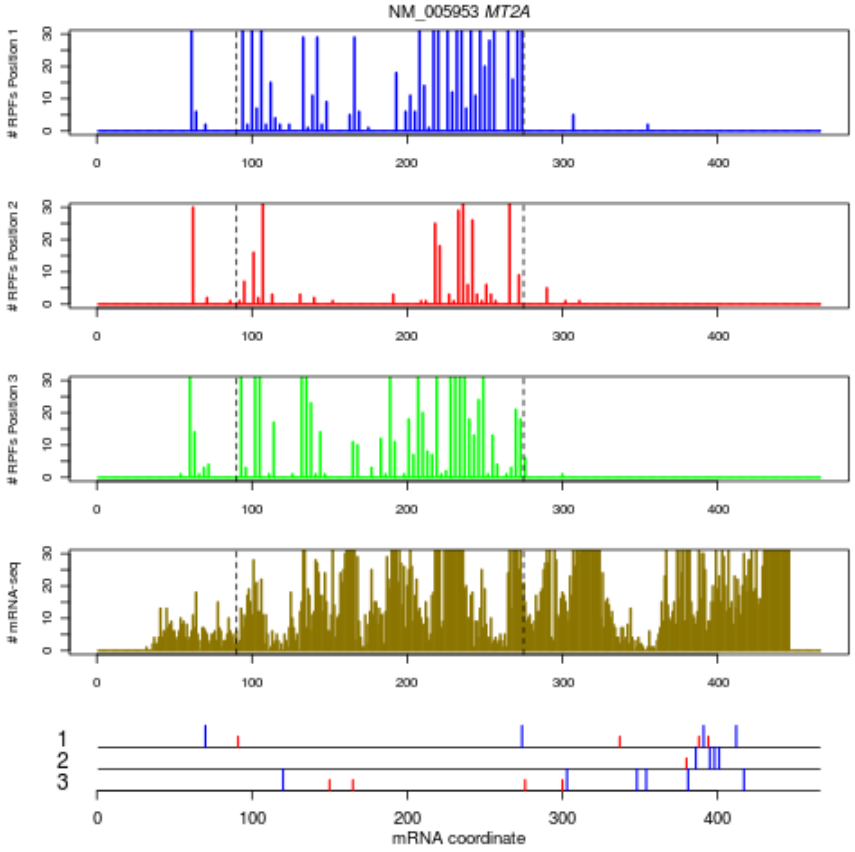
Supplemental Figure 91: Sub-codon profile, mRNA-Seq and ORF organization for



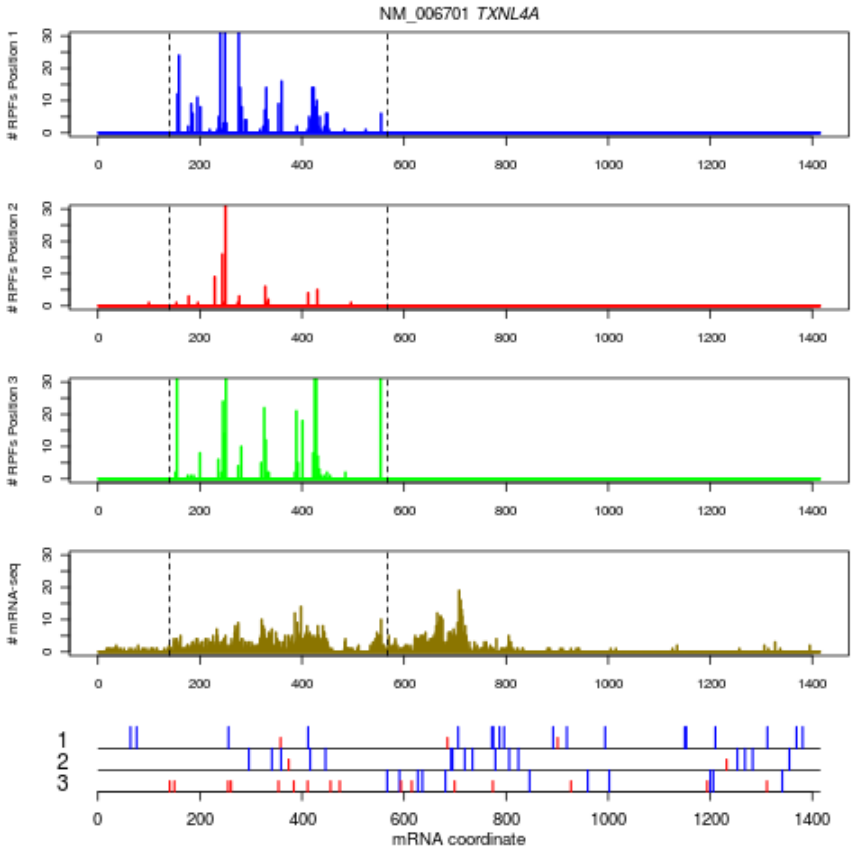
Supplemental Figure 92: Sub-codon profile, mRNA-Seq and ORF organization for



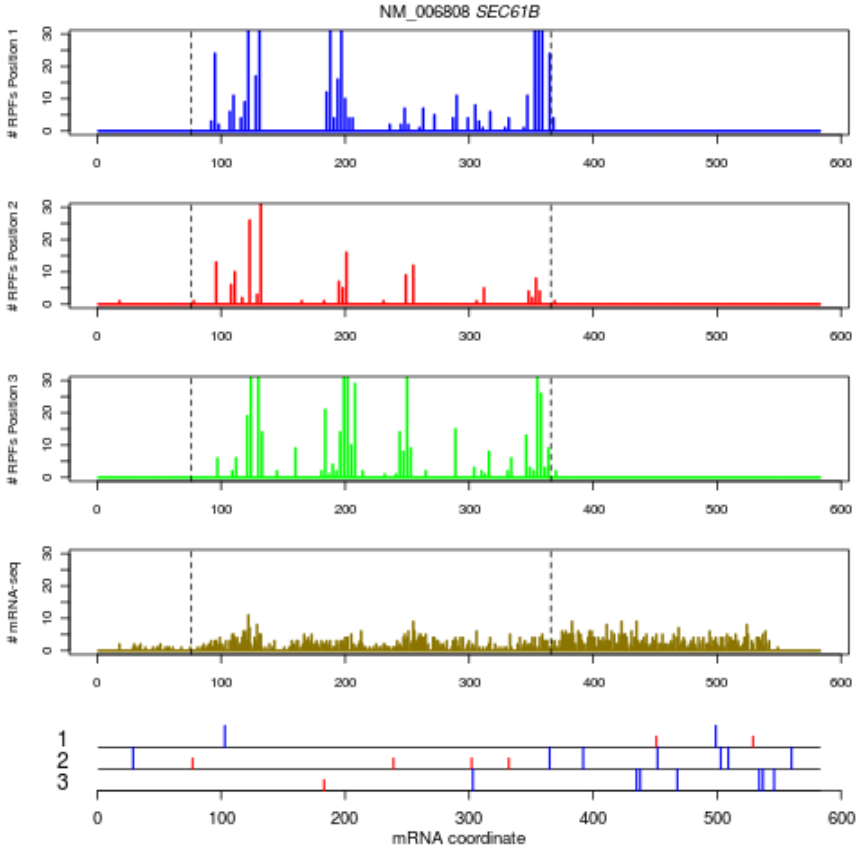
Supplemental Figure 93: Sub-codon profile, mRNA-Seq and ORF organization for



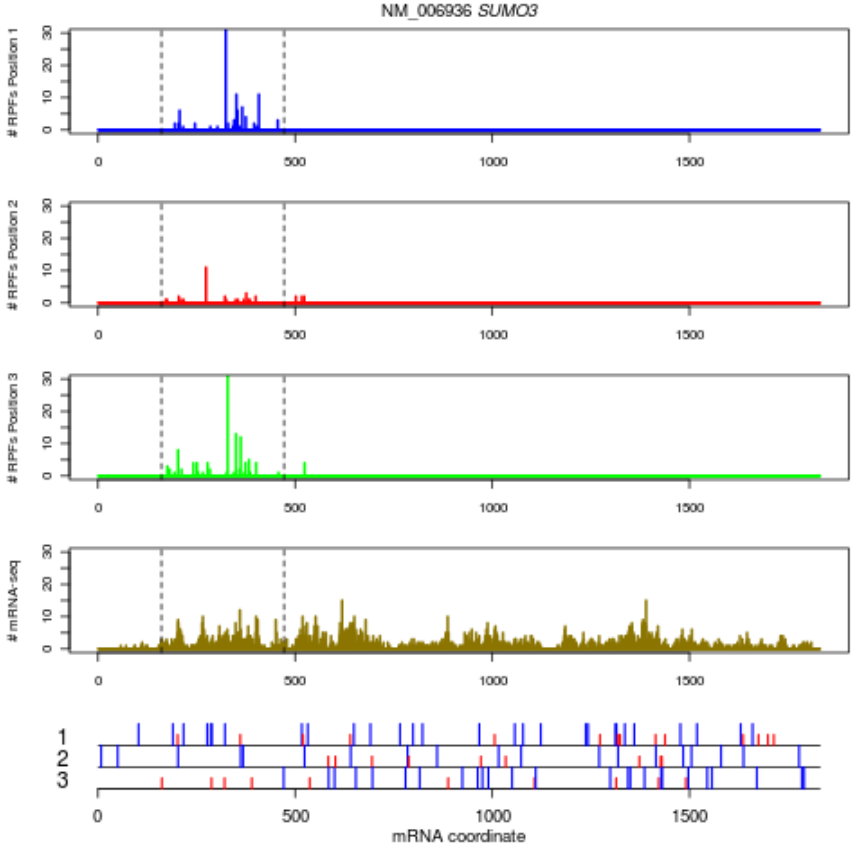
Supplemental Figure 94: Sub-codon profile, mRNA-Seq and ORF organization for



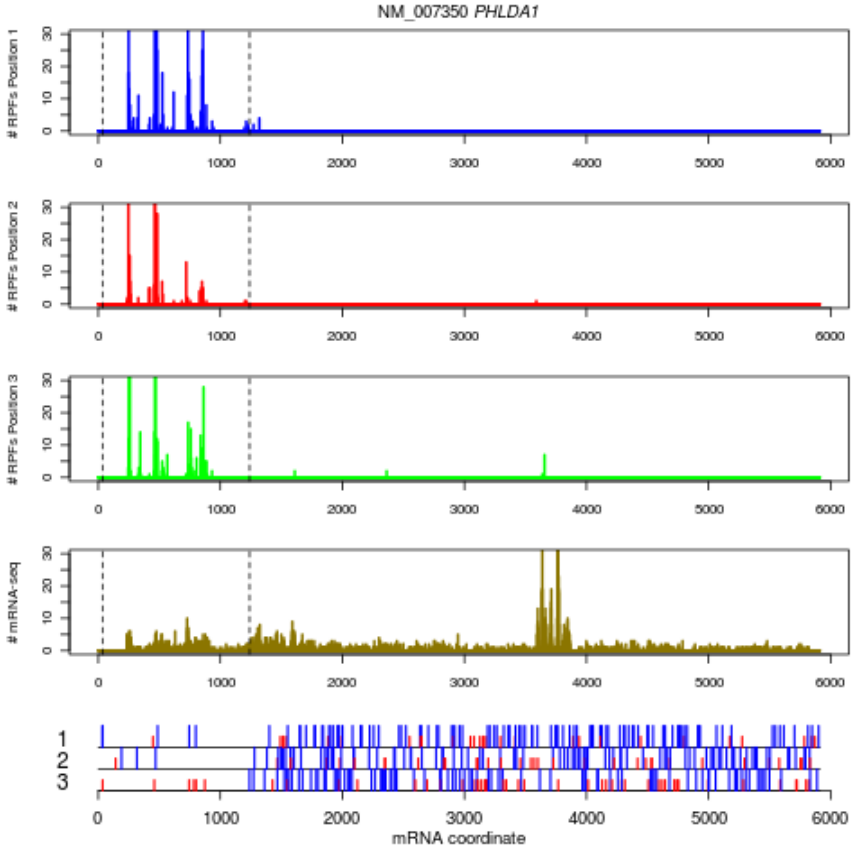
Supplemental Figure 95: Sub-codon profile, mRNA-Seq and ORF organization for



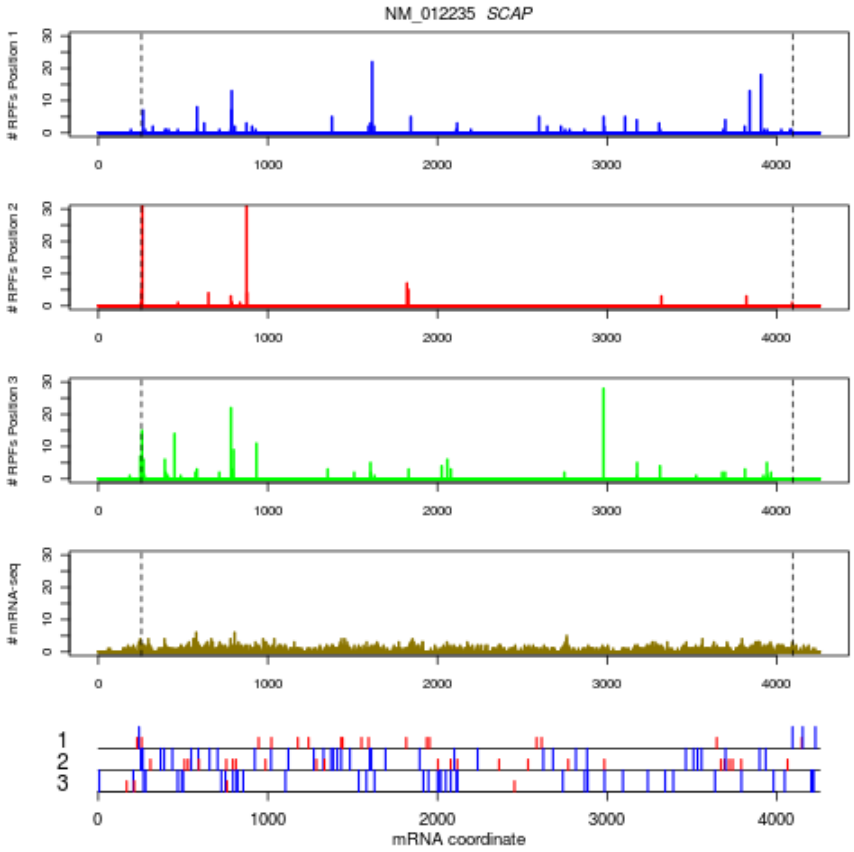
Supplemental Figure 96: Sub-codon profile, mRNA-Seq and ORF organization for



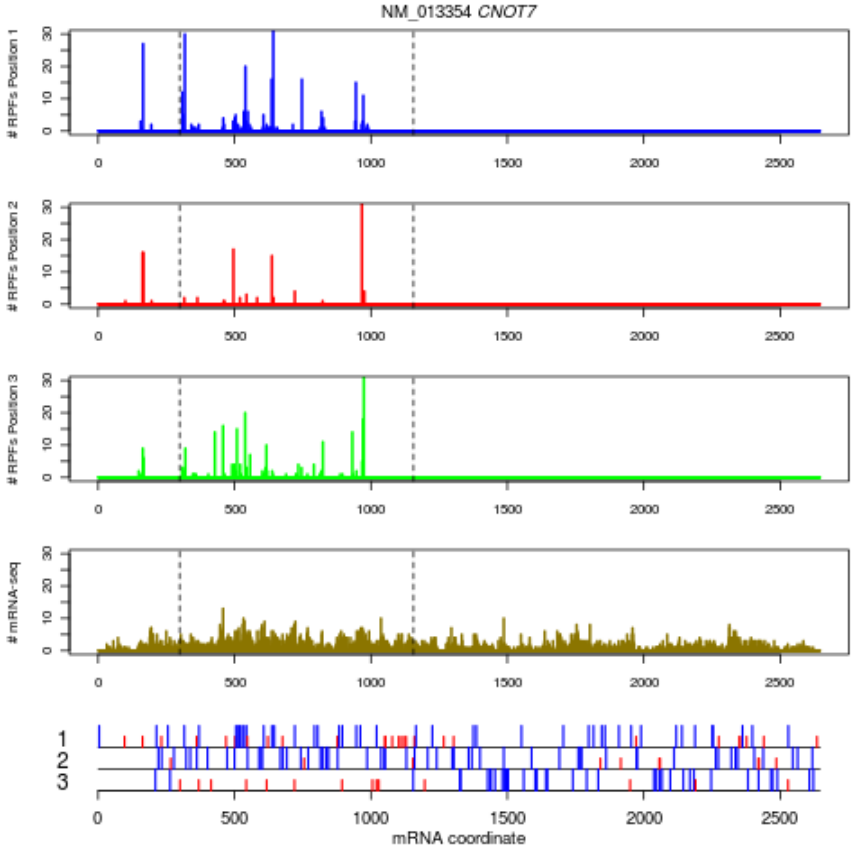
Supplemental Figure 97: Sub-codon profile, mRNA-Seq and ORF organization for



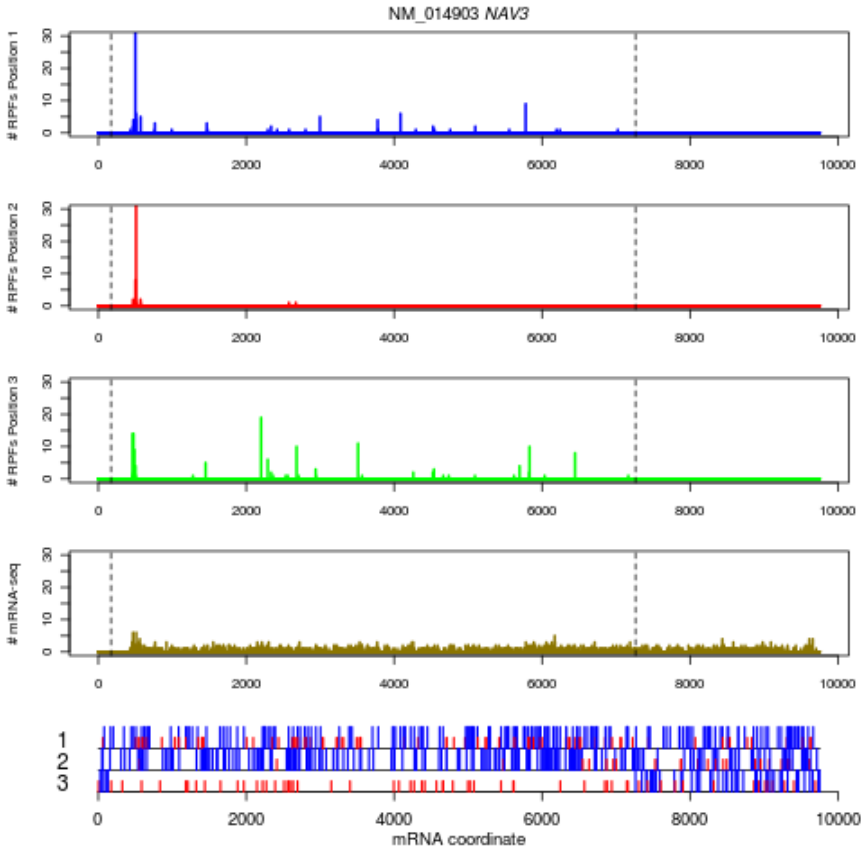
Supplemental Figure 98: Sub-codon profile, mRNA-Seq and ORF organization for



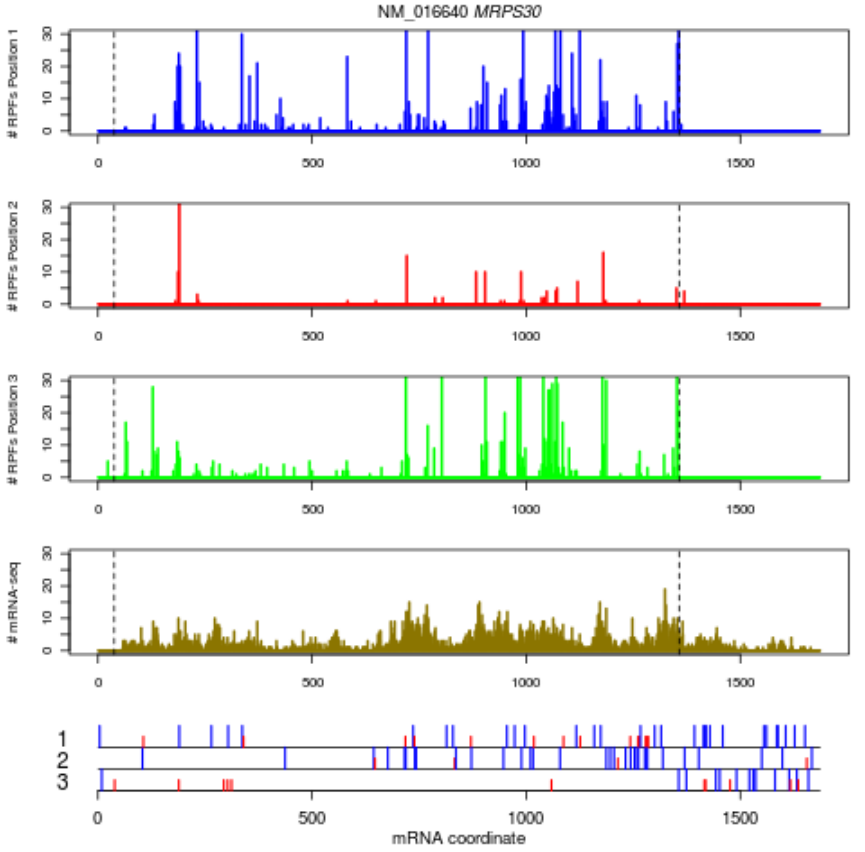
Supplemental Figure 99: Sub-codon profile, mRNA-Seq and ORF organization for



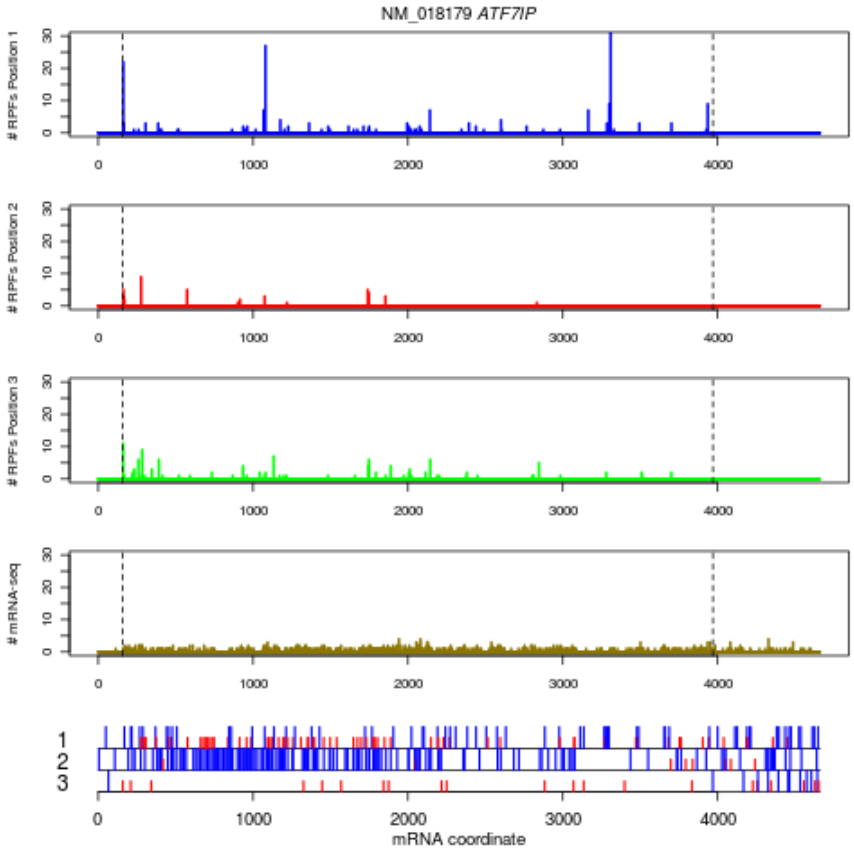
Supplemental Figure 100: Sub-codon profile, mRNA-Seq and ORF organization for



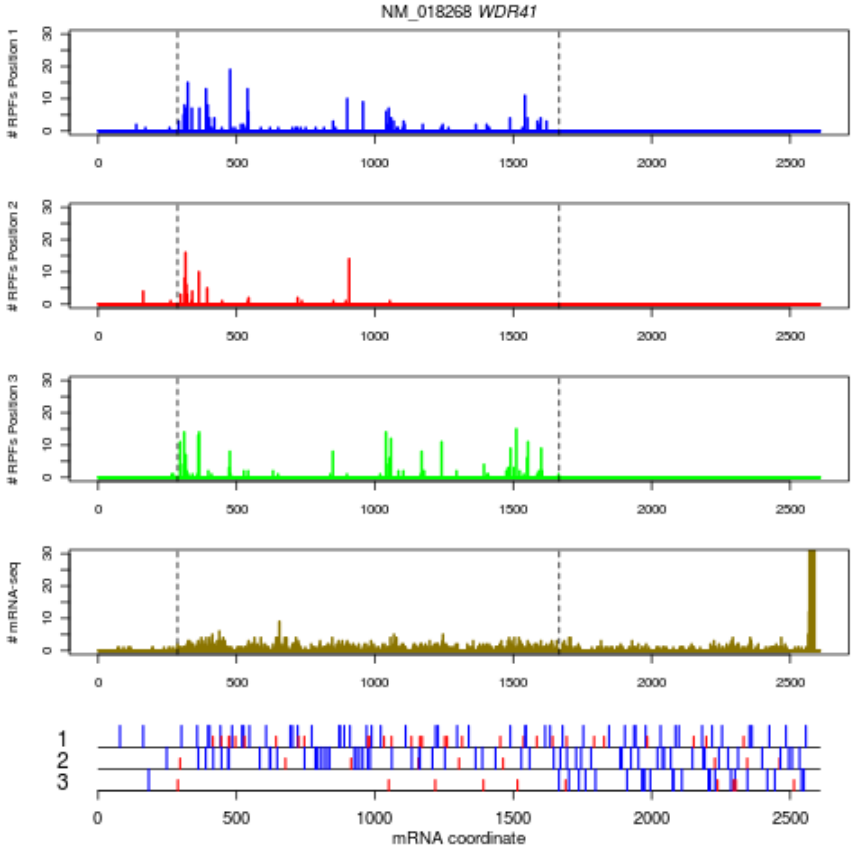
Supplemental Figure 101: Sub-codon profile, mRNA-Seq and ORF organization for



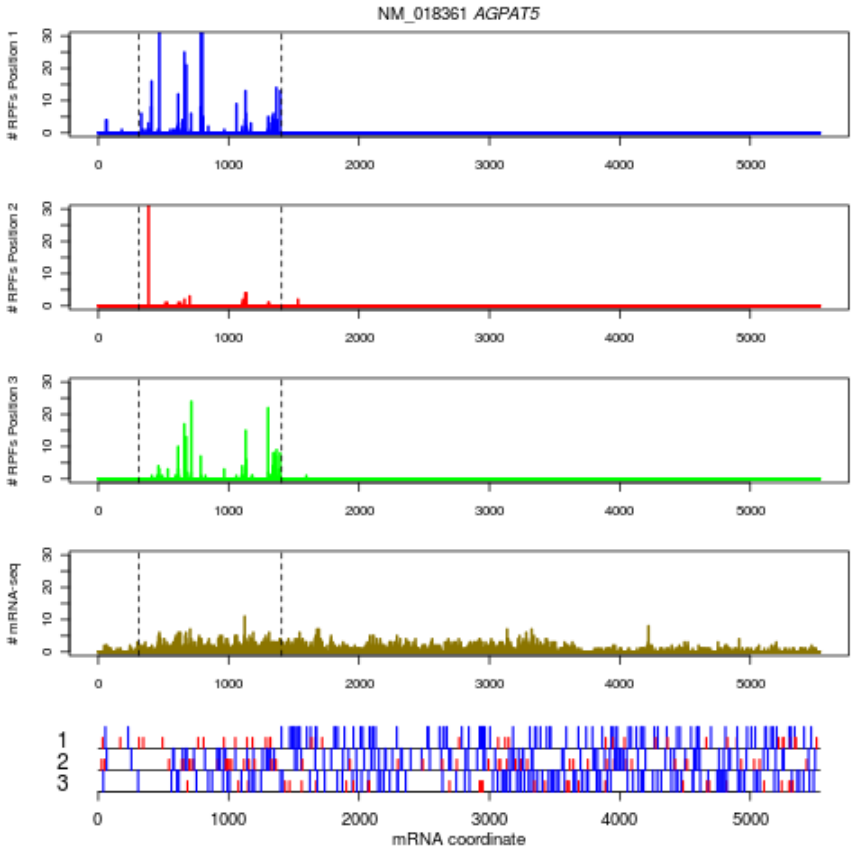
Supplemental Figure 102: Sub-codon profile, mRNA-Seq and ORF organization for



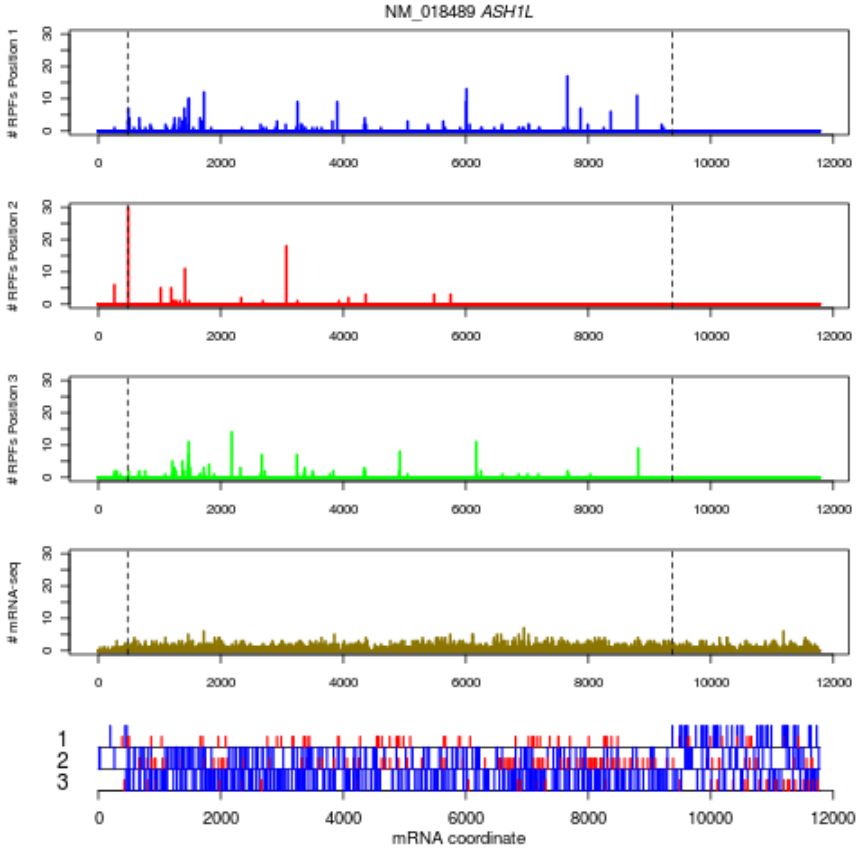
Supplemental Figure 103: Sub-codon profile, mRNA-Seq and ORF organization for



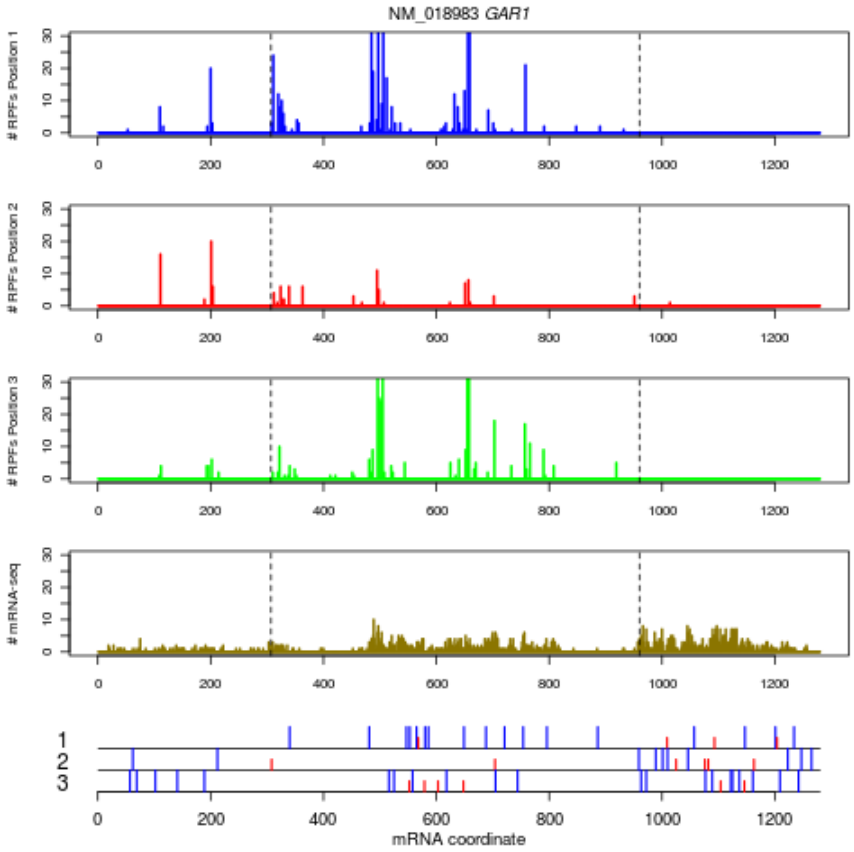
Supplemental Figure 104: Sub-codon profile, mRNA-Seq and ORF organization for



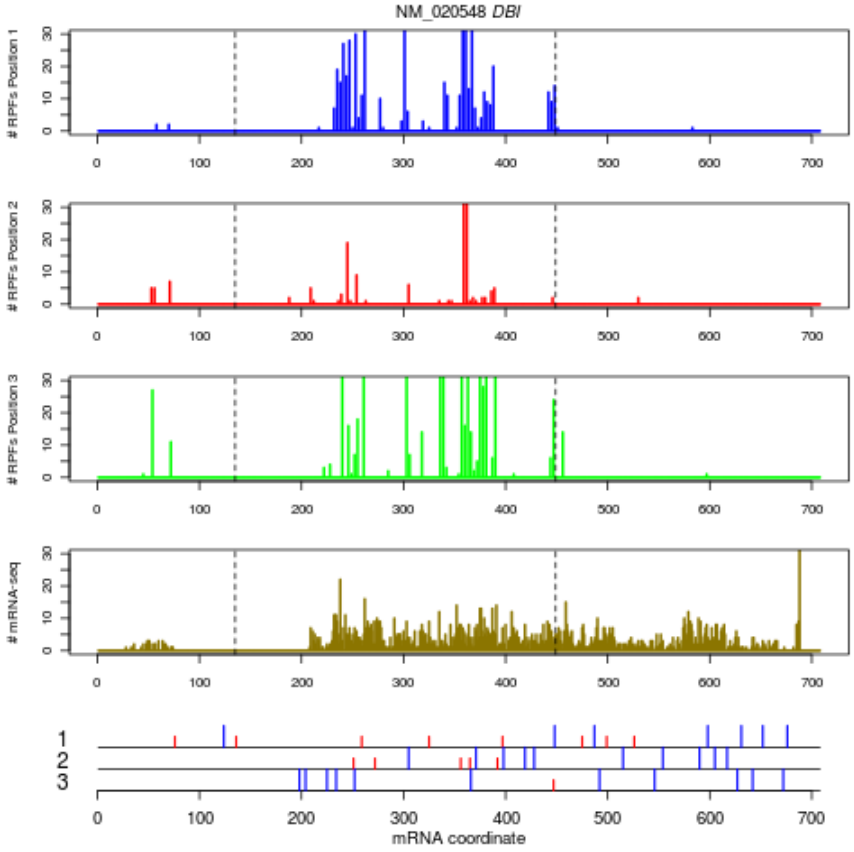
Supplemental Figure 105: Sub-codon profile, mRNA-Seq and ORF organization for



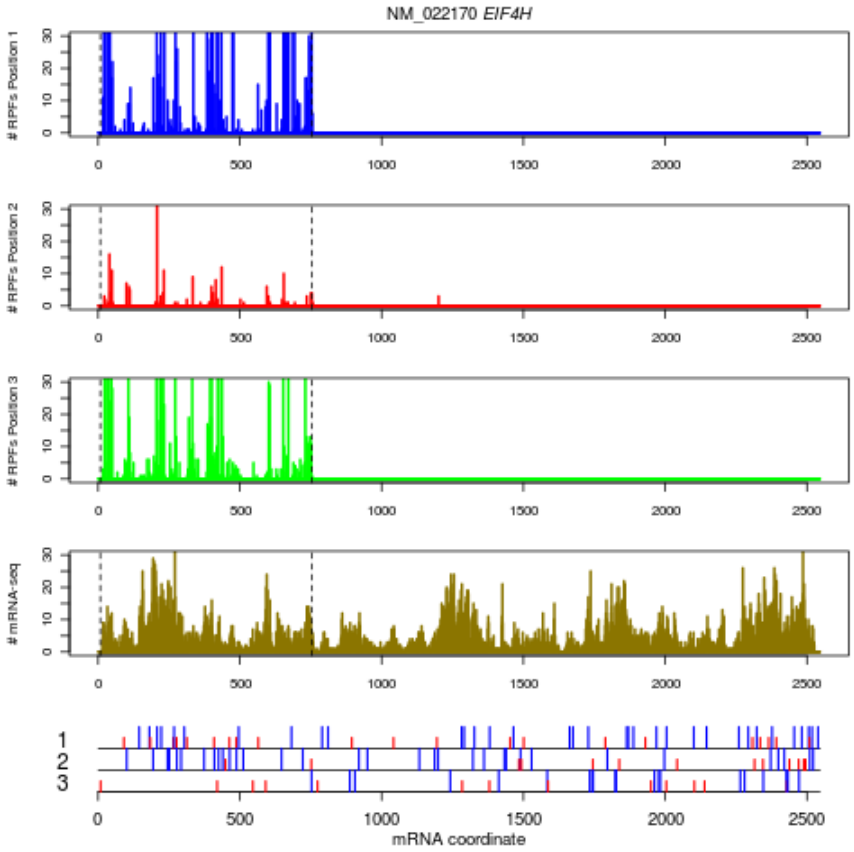
Supplemental Figure 106: Sub-codon profile, mRNA-Seq and ORF organization for



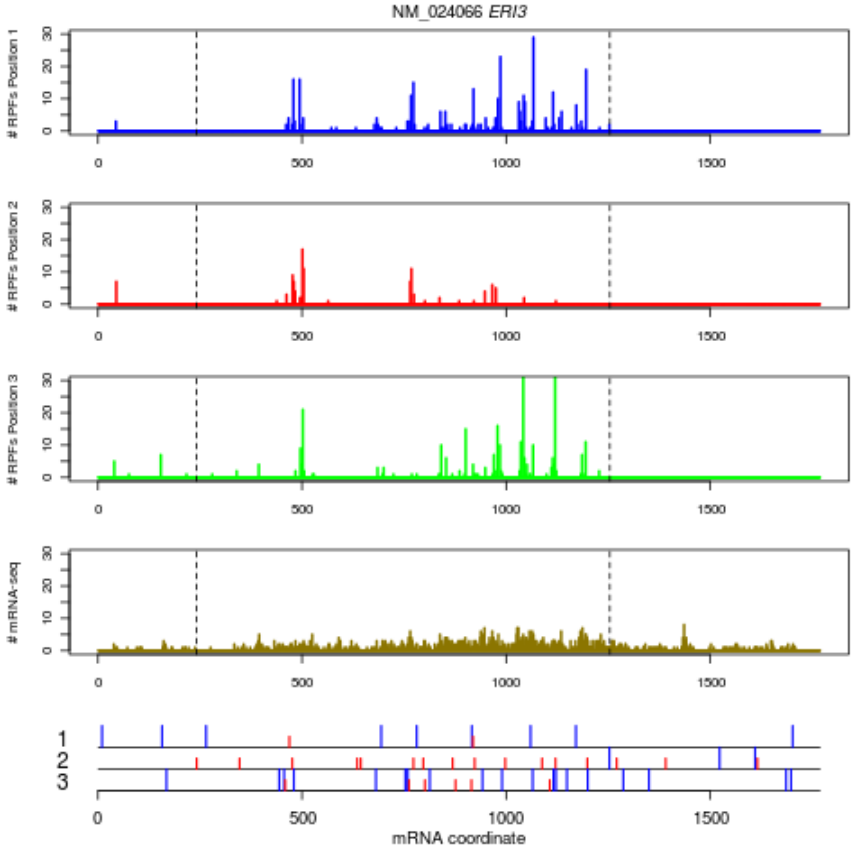
Supplemental Figure 107: Sub-codon profile, mRNA-Seq and ORF organization for



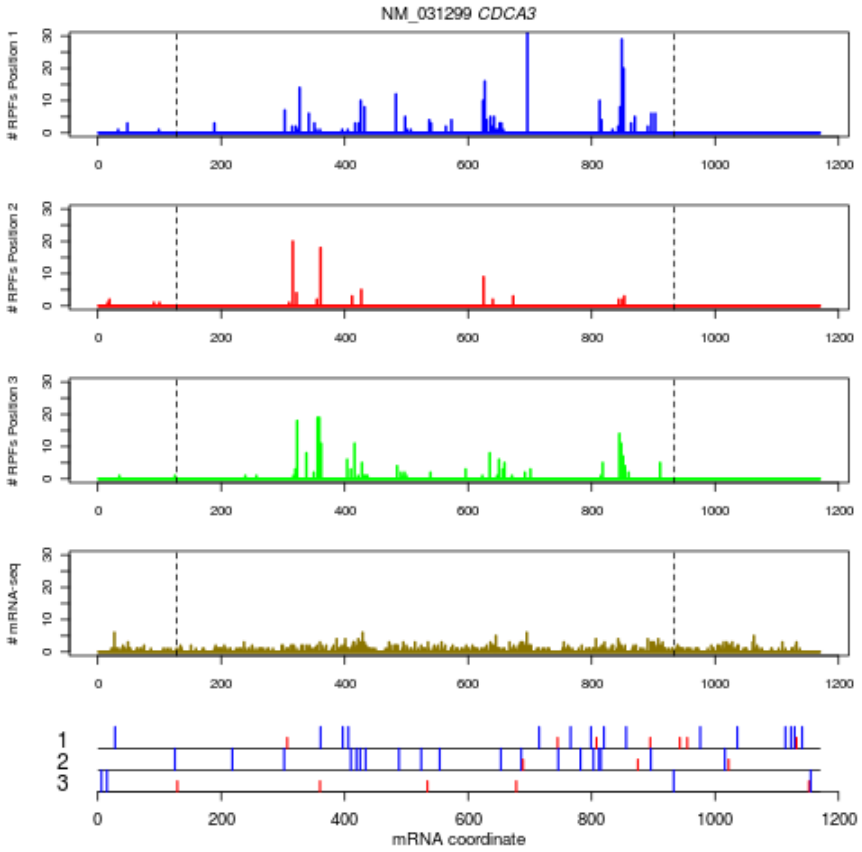
Supplemental Figure 108: Sub-codon profile, mRNA-Seq and ORF organization for



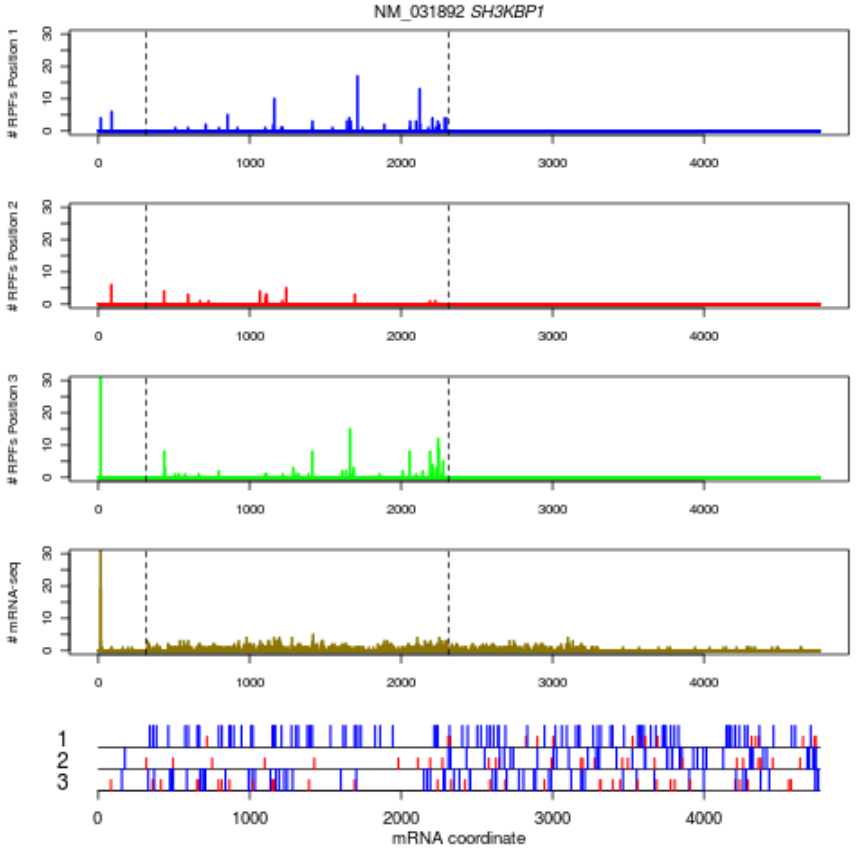
Supplemental Figure 109: Sub-codon profile, mRNA-Seq and ORF organization for



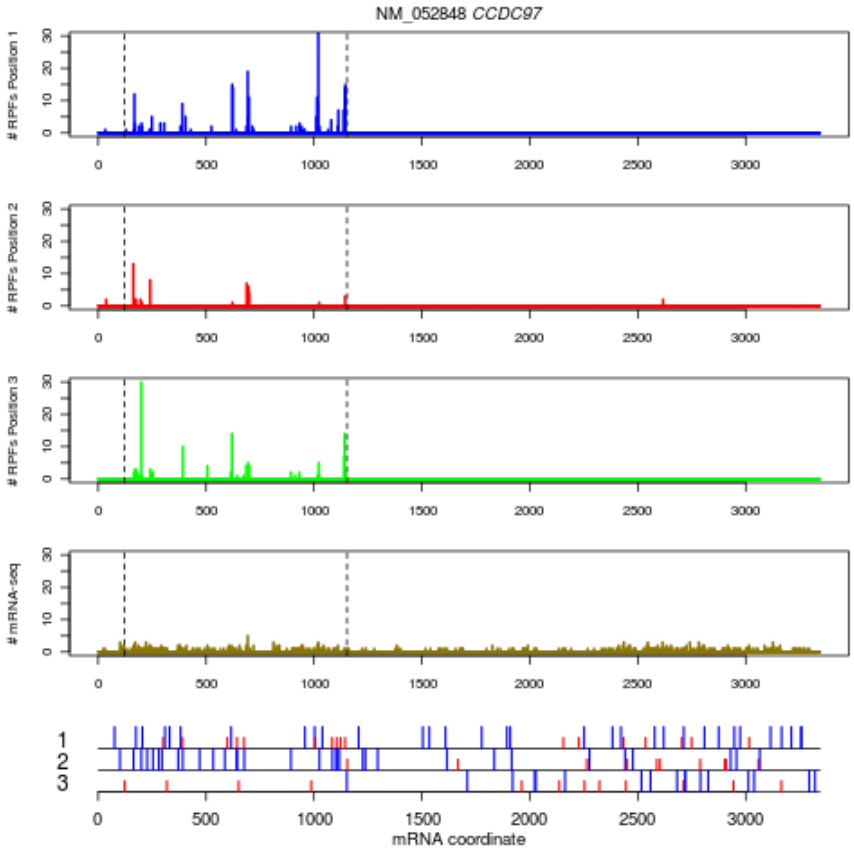
Supplemental Figure 110: Sub-codon profile, mRNA-Seq and ORF organization for



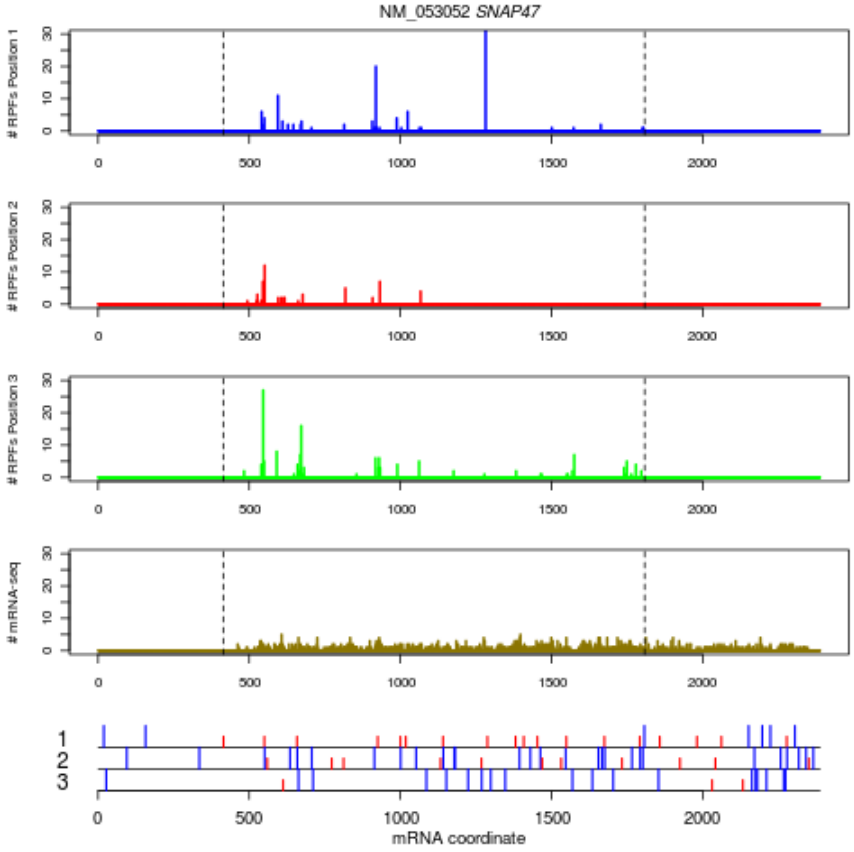
Supplemental Figure 111: Sub-codon profile, mRNA-Seq and ORF organization for



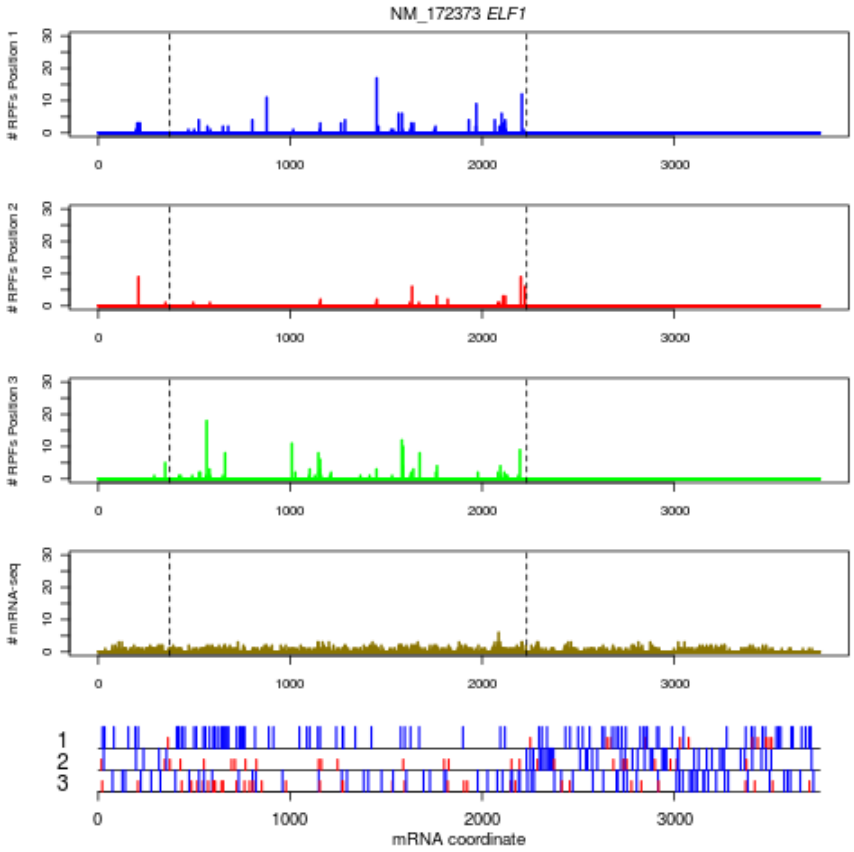
Supplemental Figure 112: Sub-codon profile, mRNA-Seq and ORF organization for



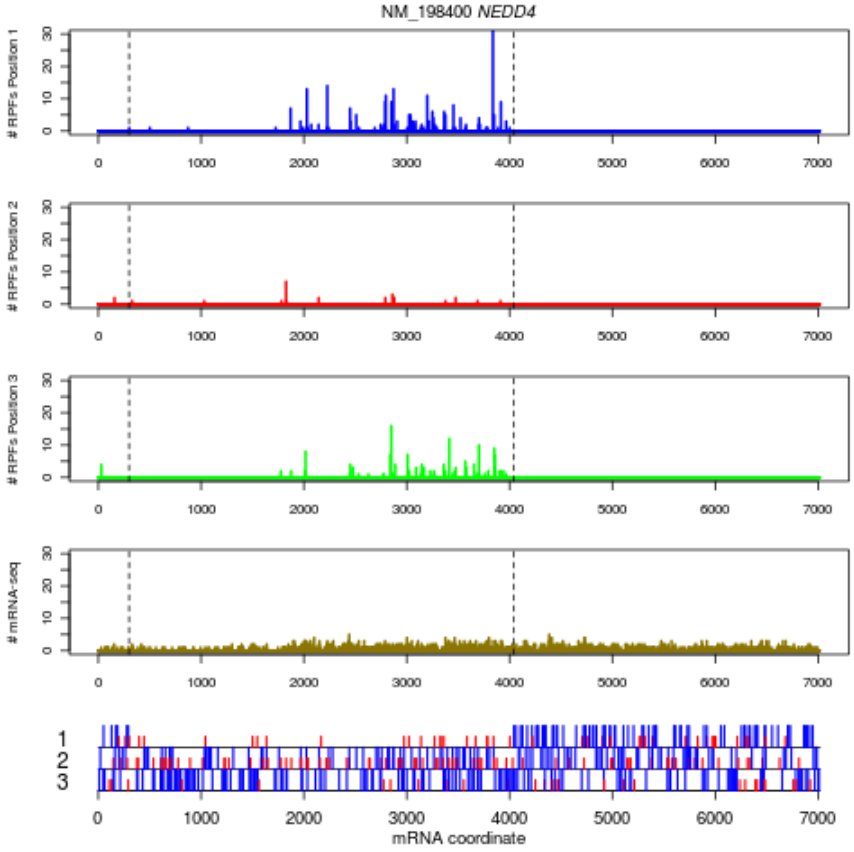
Supplemental Figure 113: Sub-codon profile, mRNA-Seq and ORF organization for



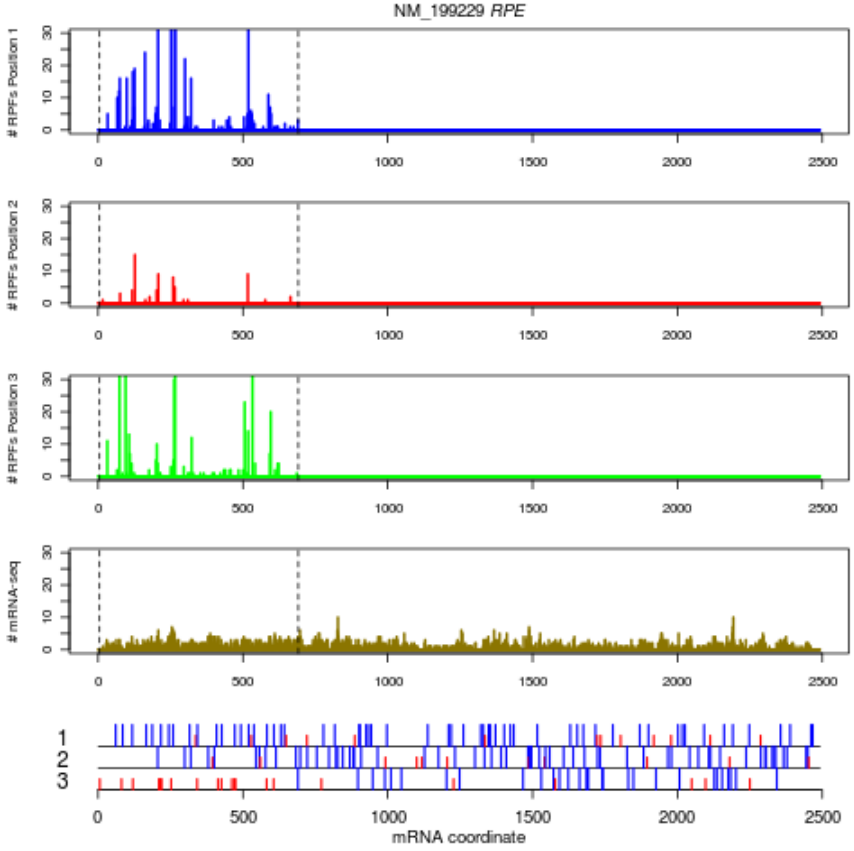
Supplemental Figure 114: Sub-codon profile, mRNA-Seq and ORF organization for



Supplemental Figure 115: Sub-codon profile, mRNA-Seq and ORF organization for

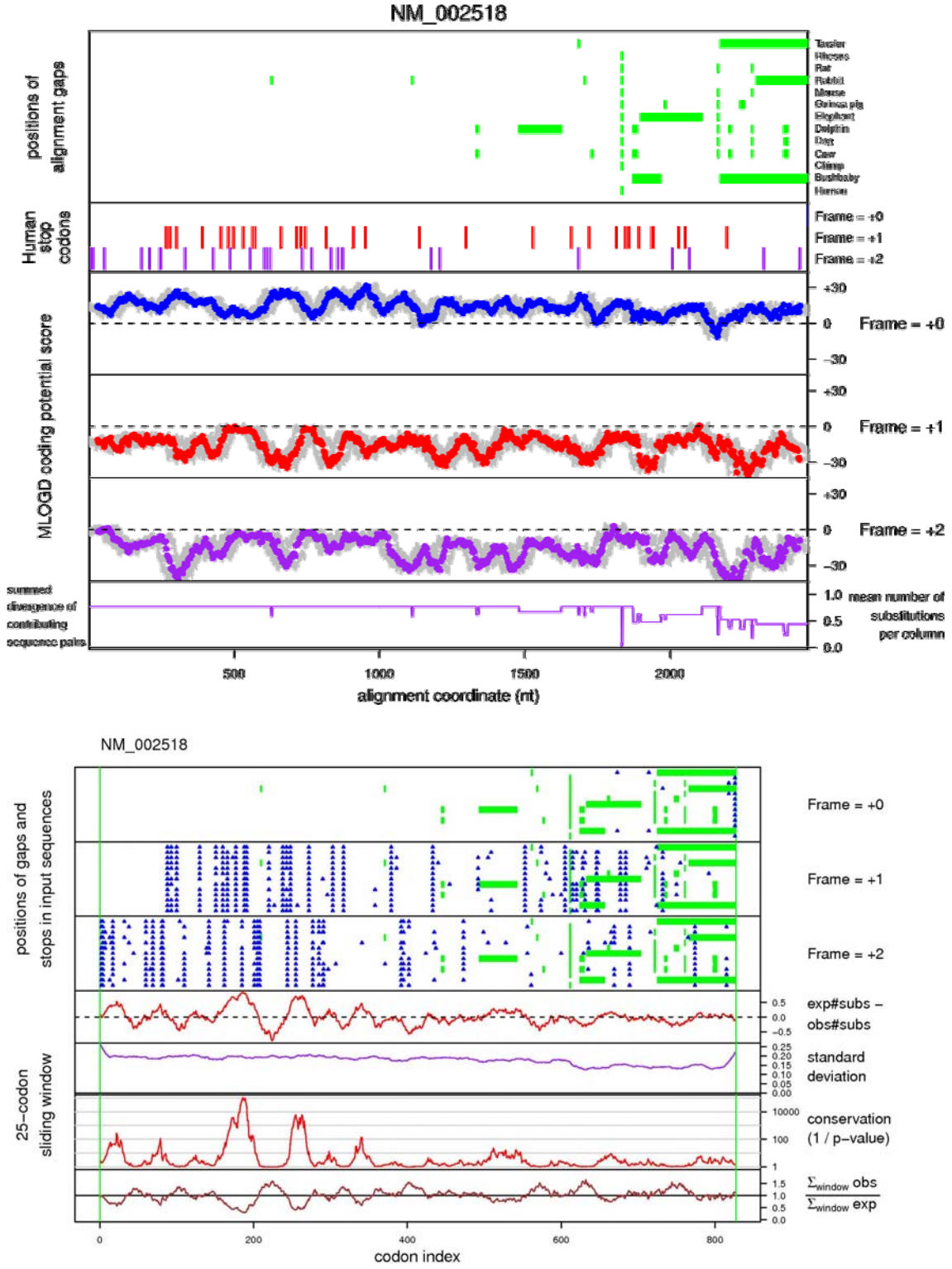


Supplemental Figure 116: Sub-codon profile, mRNA-Seq and ORF organization for



Comparative sequence analysis

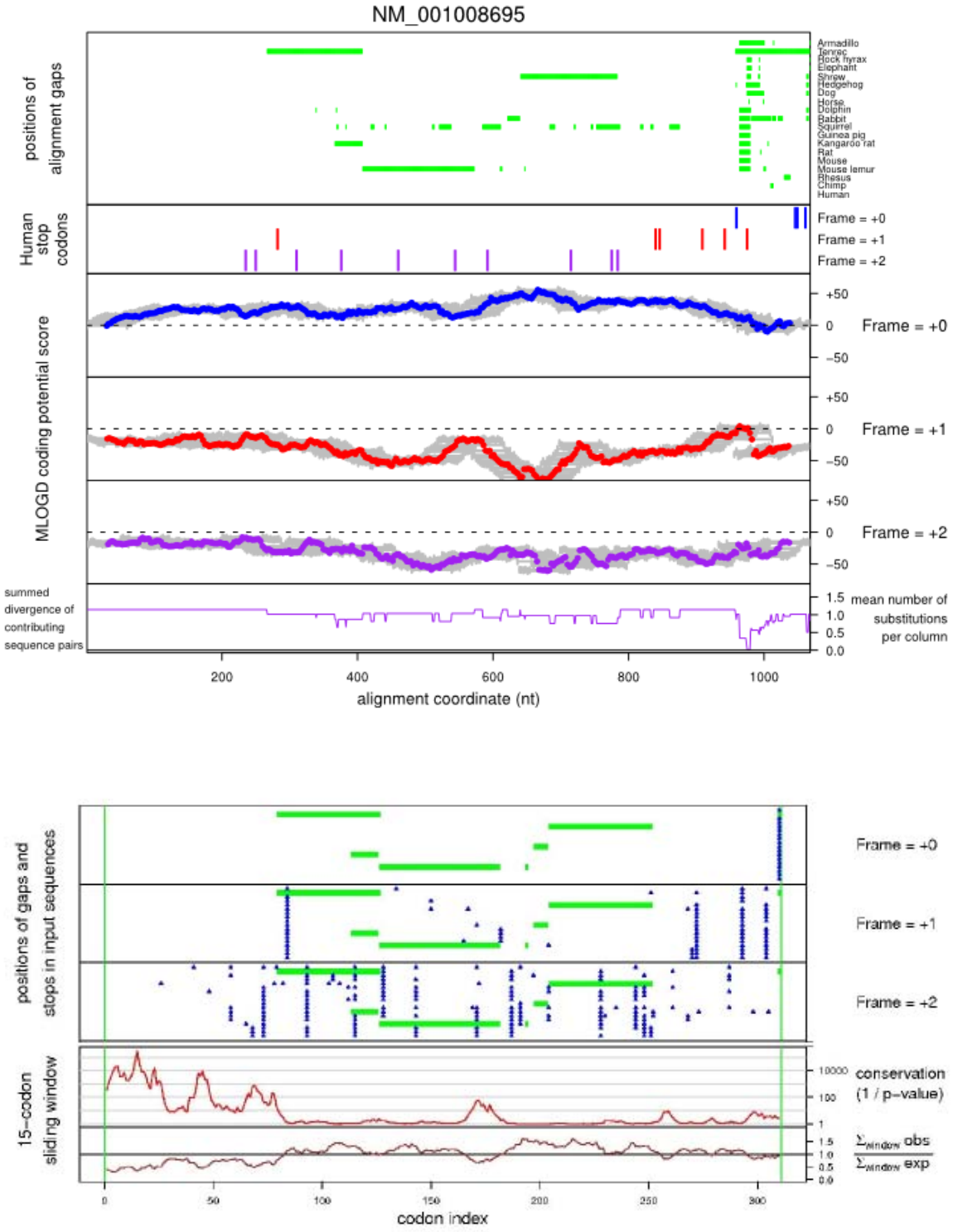
Supplemental Figure 117: Coding likelihood (calculated by MLOGD) and synonymous site conservation for *NPAS2* (NM_002518).



While the MLOGD analysis and synonymous site conservation do not indicate any purifying

selection acting on the region where the RPFs occur in the +1 frame, there are no in-frame stop codons within the same region.

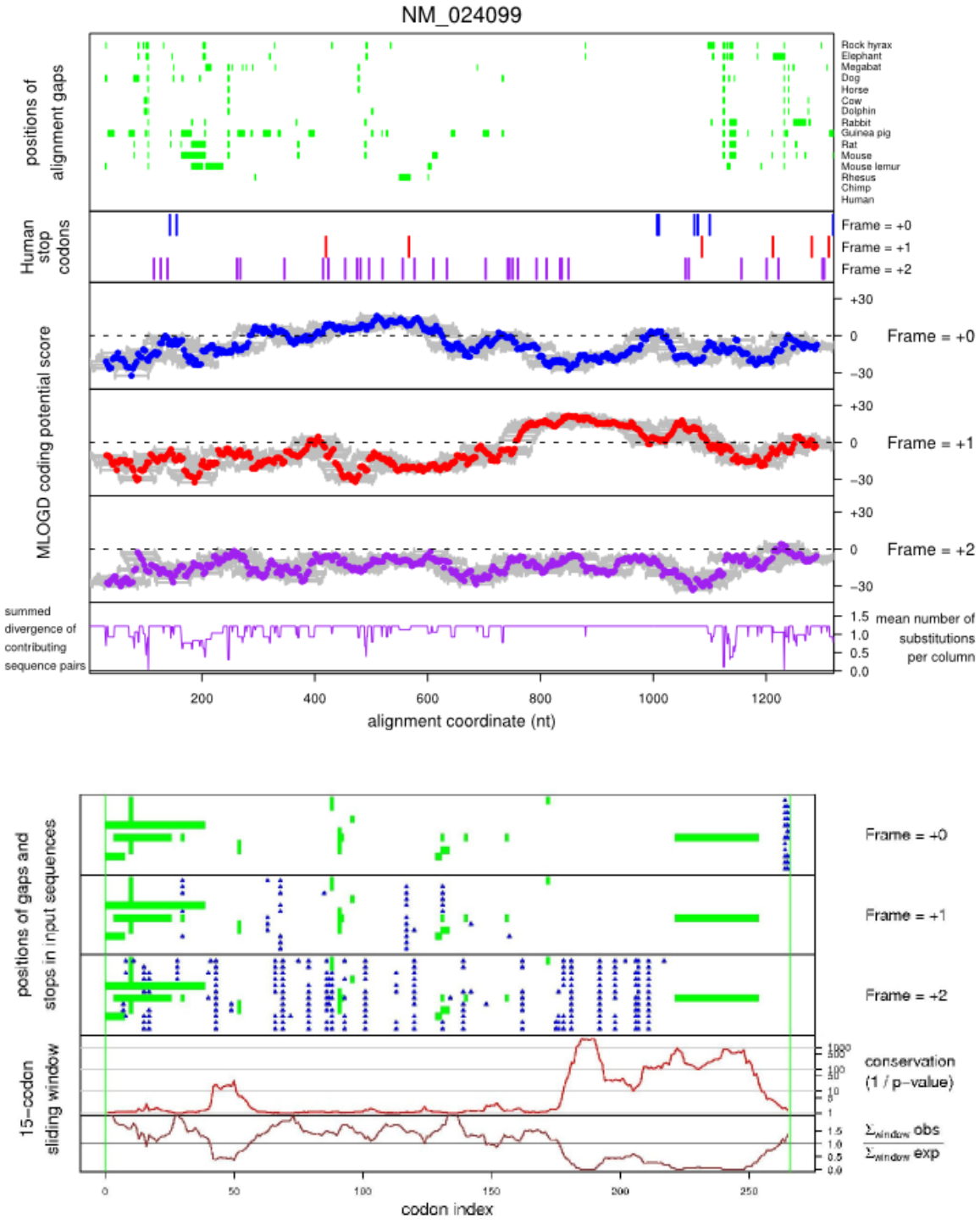
Supplemental Figure 118: Coding likelihood (calculated by MLOGD) and synonymous site conservation for *THAP7* (NM_001008695)



MLOGD analysis indicates that purifying selection operates mainly in the zero frame (CDS frame). However there are no in-frame stop codons within the human uORF region in any of the other orthologous genomic sequences. This indicates that while the protein sequence encoded

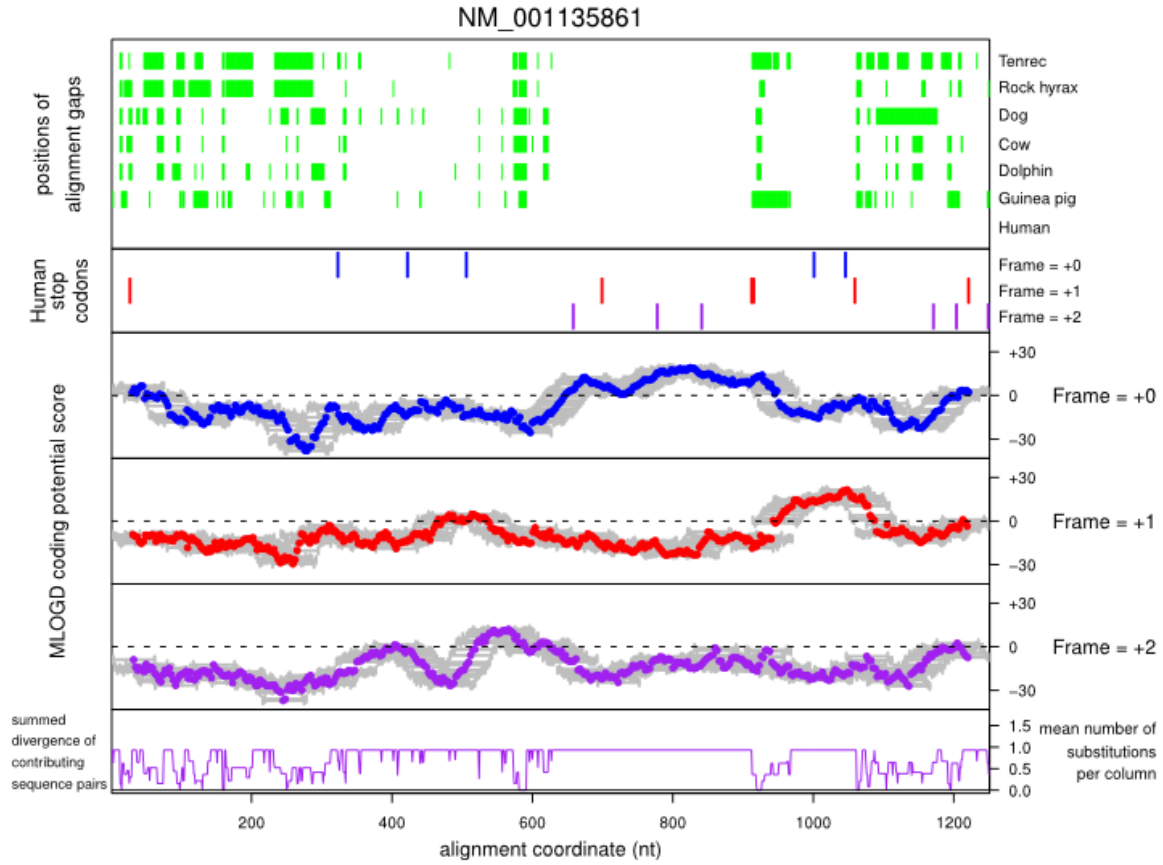
by the uORF is not highly conserved, and may have no functional role as a peptide product, the existence of the uORF is evolutionarily preserved and therefore its translation likely plays an important functional role. However, significant conservation of synonymous positions in the pORF indicates that the uORF constrains the evolution of the corresponding sequence more than what would be expected just by the avoidance of stop codons in the alternative frame.

Supplemental Figure 119: Coding likelihood (calculated by MLOGD) and synonymous site conservation for *C11orf48* (NM_024099)



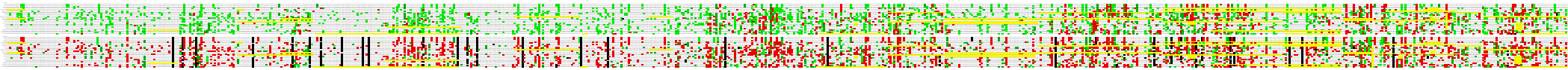
The +1 frame has a positive coding potential in the area of dual decoding. Also there is enhanced synonymous site conservation in this region of the pORF. This is consistent with the existence of an additional evolutionarily conserved feature (viz. protein coding in the alternative frame) in this region.

Supplemental Figure 120: Coding likelihood (calculated by MLOGD) for *PHPT1* (NM_001135861)



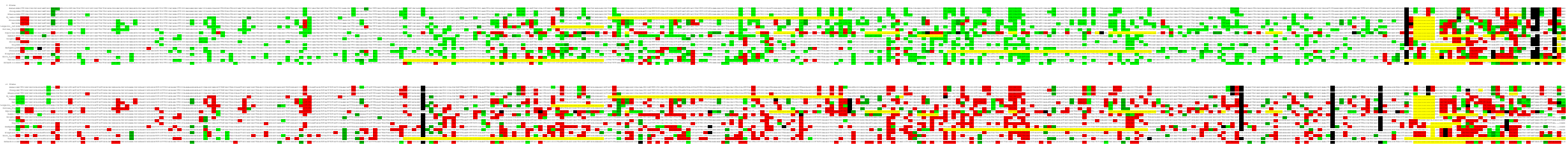
The coding likelihood statistics are consistent with dual decoding of the 5'-end of the fourth exon.

Supplemental Figure 121



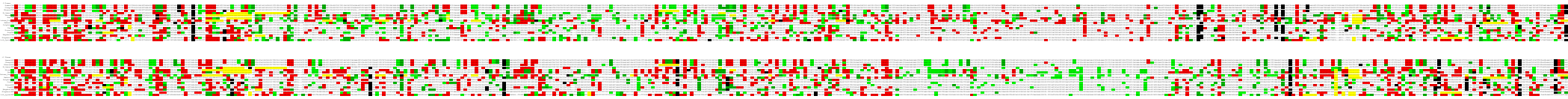
Genomic multiple codon alignments for *NPAS2* (NM_002518) from 23 vertebrate species.
Synonymous codon substitutions shows as bright green, positive (in BLOSUM62 matrix) non-synonymous codon substitutions are shown in dark green,
other non-synonymous codon substitutions are shown in red, gaps in yellow and stop codons are in black.

Supplemental Figure 122



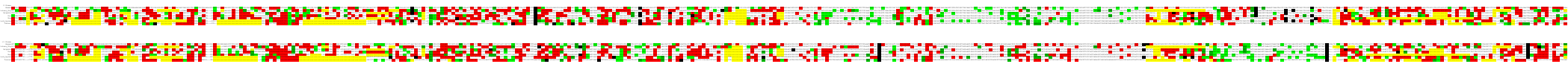
Genomic multiple codon alignments for *THAP7* (NM_001008695) from 19 vertebrate species.
Synonymous codon substitutions shows as bright green, positive (in BLOSUM62 matrix) non-synonymous codon substitutions are shown in dark green,
other non-synonymous codon substitutions are shown in red, gaps in yellow and stop codons are in black.

Supplemental Figure 123



Genomic multiple codon alignments for *C11orf48* (NM_024099) from 15 vertebrate species.
Synonymous codon substitutions shows as bright green, positive (in BLOSUM62 matrix) non-synonymous codon substitutions are shown in dark green,
other non-synonymous codon substitutions are shown in red, gaps in yellow and stop codons are in black.

Supplemental Figure 124



Genomic multiple codon alignments for *PHPT1* (NM_001135861) from 7 vertebrate species.
Synonymous codon substitutions shows as bright green, positive (in BLOSUM62 matrix) non-synonymous codon substitutions are shown in dark green,
other non-synonymous codon substitutions are shown in red, gaps in yellow and stop codons are in black.

Supplemental Discussion

False positives, unexplained profiles, staccato reads and potential ambiguity in the functional classification of dually coded genes.

Although we were able to identify many dually decoded regions, a number of mRNAs with a high PTS are false positives. Sub-codon profiles of about a third of all candidates have a high PTS for reasons other than dual translation. The most prominent example (4th in Supplemental Table 1) is a profile for dystrophin *DMD* mRNA (RefSeq NM_004010). As mentioned earlier the longest transcript isoforms were used for the generation of sub-codon profiles in our study. However, examination of the sub-codon profile and mRNA-seq (see Supplemental Fig. 90) reveals that RPFs and mRNA-seq fragments cover only a small part of the *DMD* pORF variant and therefore are most likely produced by a short mRNA variant. However, since the CSCPD value depends on both the relative number of RPFs in a particular CDS region and its length, a slight deviation from the expected RPF sub-codon distribution is amplified into a strong PTS signal by a prolonged CDS region lacking RPFs. This example demonstrates the drawbacks of our current approach - as it requires manual verification of the predicted candidates.

In addition to the 33 false positive candidates whose sub-codon profiles are inconsistent with dual coding, we found 13 instances where the profiles conform to the expected distribution for dual coding regions, but dual coding is inconsistent with the ORF organization of the available transcript variants. The highest scoring candidate in the 'unexplained' category (9th in Supplemental Table 1) is the mRNA for protein kinase-like protein *SGK196* (RefSeq NM_032237). Examination of the corresponding profile (see Supplemental Fig. 82) shows a relatively large area of the mRNA (from its 5'-end to about one third of CDS) where the profile is consistent with translation in the +1 frame

relative to the CDS. However, the corresponding region of mRNA is interrupted by multiple stop codons.

Such inexplicable profiles could be artifacts of the ribosome profiling method. Profiles with RPF sub-codon distributions that are inconsistent with the normal reading frame are likely to arise for statistical reasons in such a high-throughput study. Theoretically possible sequence or site-dependency of RPF length variability may also lead to incorrect frame detection. An additional source of errors could be ambiguous alignments, where the origin of RPFs is different from the mRNA for which the profiles are constructed. RPFs with the same sequence could originate from different translated mRNAs if they share significant sequence similarity. When a particular sequence occurs in several translated mRNAs, such RPFs are aggregated and result in an enhanced signal even in those locations of mRNAs that are not translated. We termed instances of such high peaks staccato reads. For example, staccato reads can be seen in the profile for NM_001145678 mRNA (see Supplemental Fig. 76) where high density peaks occur for a couple of 3'UTR locations in all sub-codon positions.

During our manual evaluation of sub-codon profiles, we examined whether particular high RPFs could originate from the translation of different mRNAs. However, we cannot exclude the possibility that particular RPFs may originate from splice junctions of mRNA variants that are not known in the RefSeq database.

In addition, it needs to be pointed that some candidates could be classified in more than one category. An example is NM_016556 (see Supplemental Fig. 45). The start codon of the alternatively translated frame is located downstream of the previously predicted pORF start. However, the observed CDS start site (second) is located downstream of the alternative start codon.

Thus we classified this case as an overlapping uORF rather than a non-upstream protein coding nORF.

Our classification is largely based on available data on RNA transcripts. Identification of an alternatively decoded ORF in a particular mRNA does not necessarily mean that this ORF is translated in the mRNA for which the sub-codon profile was built. For example, it is possible that the region corresponding to the high density of RPFs is part of an exonic region of two mRNAs, one of which has not been identified and therefore cannot be found in the available databases.

R scripts for calculating PTS

#For the top 1000 expressed genes, get the 95 percentiles for the 3 sub-codon positions and write results to csv file

```
gene_counter = 1
interpolation_mat1 = matrix(NA, nrow = 1000, ncol=100)
interpolation_mat2 = matrix(NA, nrow = 1000, ncol=100)
interpolation_mat3 = matrix(NA, nrow = 1000, ncol=100)


#Set the working directory to the folder containing the ribosome profiles of top 1000 expressed genes
setwd("/home/1000_Genes")


top_genes <-list.files()
for (gene in top_genes) {
  reads = read.table(gene, header=T, sep=",")
        attach(reads)

read_mat = matrix(NA, nrow = 1, ncol = length(Coordinate))

for (j in 1:length(Coordinate)){
        read_mat[,j] = Count[j]}

upprop = matrix(NA,nrow = 1,ncol = length(Coordinate))
downprop = matrix(NA,nrow = 1,ncol = length(Coordinate))
seq_length1 = (length(Coordinate)/3)-1
seq_length2 = seq_length1 - 1

#calculate the cumulative upstream proportions
denom = cumsum(read_mat[1,])
upprop[1,(1+3*(0:seq_length1))] = cumsum(read_mat[1,(1+3*(0:seq_length1))])/
denom[(3+3*(0:seq_length1))]
upprop[1,(2+3*(0:seq_length1))] = cumsum(read_mat[1,(2+3*(0:seq_length1))])/
denom[(3+3*(0:seq_length1))]
upprop[1,(3+3*(0:seq_length1))] = cumsum(read_mat[1,(3+3*(0:seq_length1))])/
denom[(3+3*(0:seq_length1))]

#calculate the cumulative downstream proportions
totalsDown = rep(NA, seq_length1)
for (i in 1:seq_length1){

totalsDown[i] = sum(read_mat[1+3*(i:seq_length1)]+ read_mat[2+3*(i:seq_length1)]+
read_mat[3+3*(i:seq_length1)])}
```

```
downprop[1,(1+3*(0:seq_length2))] = rev(cumsum(rev(read_mat[1,(1+3*(1:seq_length1))])))/
totalsDown[0:seq_length1]
downprop[1,(2+3*(0:seq_length2))] = rev(cumsum(rev(read_mat[1,(2+3*(1:seq_length1))])))/
totalsDown[0:seq_length1]
downprop[1,(3+3*(0:seq_length2))] = rev(cumsum(rev(read_mat[1,(3+3*(1:seq_length1))])))/
totalsDown[0:seq_length1]


#Calculate the CSCPD (absolute difference between cumulative upstream and downstream
proportions for sub-codon positions 1,2,3)
third_full_seq_minus1 = (length(Coordinate)/3)-1
y1 = abs(upprop-downprop)[1+3*(0:third_full_seq_minus1)]
y1_withoutNA = y1[which(y1 != "NA")] #omitting the last entries with possible NaN entries due to
division by zero and also Na's
y1_withoutNaN = y1_withoutNA[which(y1_withoutNA != "NaN")]
length_third_seq_withoutNaN = length(y1_withoutNaN)
length_third_seq_withoutNaN_minus1= length(y1_withoutNaN)-1


y2 = abs(upprop-downprop)[2+3*(0:third_full_seq_minus1)]
y2_withoutNA = y2[which(y2 != "NA")]
y2_withoutNaN = y2_withoutNA[which(y2_withoutNA != "NaN")]

y3 = abs(upprop-downprop)[3+3*(0:third_full_seq_minus1)]
y3_withoutNA = y3[which(y3 != "NA")]
y3_withoutNaN = y3_withoutNA[which(y3_withoutNA != "NaN")]


x1 = 1+3*(0:length_third_seq_withoutNaN_minus1)
x2 = 2+3*(0:length_third_seq_withoutNaN_minus1)
x3 = 3+3*(0:length_third_seq_withoutNaN_minus1)

#Converting all coordinates in coding region to relative values between 0 and 1 and using a
smoothing function
ys1 = smooth.spline(x1/(length_third_seq_withoutNaN*3),y1_withoutNaN)
ys2 = smooth.spline(x2/(length_third_seq_withoutNaN*3),y2_withoutNaN)
ys3 = smooth.spline(x3/(length_third_seq_withoutNaN*3),y3_withoutNaN)

#Sampling 100 equidistant CSCPD values between 0 and 1
xout = 0.01*(1:100)
yout1 = predict(ys1,xout)$y
yout2 = predict(ys2,xout)$y
yout3 = predict(ys3,xout)$y

interpolation_mat1[gene_counter,] = yout1
interpolation_mat2[gene_counter,] = yout2
interpolation_mat3[gene_counter,] = yout3

gene_counter = gene_counter+1
detach(reads)
```

```
}


#to get the 95th quantiles
cnames = c("Percentile_P1", "Percentile_P2", "Percentile_P3")
rnames = rep("",100)
quantiles_95 = matrix(0, nrow = 100, ncol=3,dimnames=list(rnames,cnames))

for(j in 1:100){
        quantiles_95[j,1] = quantile(interpolation_mat1[,j],0.95,na.rm = T)
        quantiles_95[j,2] = quantile(interpolation_mat2[,j],0.95,na.rm = T)
        quantiles_95[j,3] = quantile(interpolation_mat3[,j],0.95,na.rm = T)}

write.csv(quantiles_95, file = "/home/quantiles_95.csv",col.names = TRUE, row.names = FALSE)


#########################################################################

#To determine the PTS for batches of 1000 genes

#Read in the 100 95th quantile values obtained previously
percentiles=read.csv("/home/quantiles_95.csv", header=T)
attach(percentiles)
xout = 0.01*(1:100)


#Calculate the difference between the gene's CSCPDs and the 95th quantiles and cumulate to get the PTS for each sub-codon position and total PTS

gene_counter = 1
#Getting the PTS for a group of 1000 genes
difference_mat1 = matrix(NA, nrow = 1000, ncol=100)
difference_mat2 = matrix(NA, nrow = 1000, ncol=100)
difference_mat3 = matrix(NA, nrow = 1000, ncol=100)

#to get a vector first of all the gene names in the folder which can then be fed into the PTS matrix as row names
rnames = rep("",1000)
cnames = c("PTS1", "PTS2", "PTS3", "PTS")
PTS = matrix(0, nrow = 1000, ncol=4, dimnames=list(rnames,cnames))

setwd("/home/Genes/")
high_genes <-list.files()
for (gene in high_genes) {
        print (gene_counter)
        rnames[gene_counter] = gene
        reads = read.table(gene, header=T, sep=",")
        attach(reads)

read_mat = matrix(NA, nrow = 1, ncol = length(Coordinate))
```

```
for (j in 1:length(Coordinate)){
        read_mat[,j] = Count[j]}

upprop = matrix(NA,nrow = 1,ncol = length(Coordinate))
downprop = matrix(NA,nrow = 1,ncol = length(Coordinate))
seq_length1 = (length(Coordinate)/3)-1
seq_length2 = seq_length1 - 1

#calculate the cumulative upstream proportions
denom = cumsum(read_mat[1,])
upprop[1,(1+3*(0:seq_length1))] = cumsum(read_mat[1,(1+3*(0:seq_length1))])/
denom[(3+3*(0:seq_length1))]
upprop[1,(2+3*(0:seq_length1))] = cumsum(read_mat[1,(2+3*(0:seq_length1))])/
denom[(3+3*(0:seq_length1))]
upprop[1,(3+3*(0:seq_length1))] = cumsum(read_mat[1,(3+3*(0:seq_length1))])/
denom[(3+3*(0:seq_length1))]

#calculate the cumulative downstream proportions
totalsDown = rep(NA, seq_length1)
for (i in 1:seq_length1){
totalsDown[i] = sum(read_mat[1+3*(i:seq_length1)]+ read_mat[2+3*(i:seq_length1)]+
read_mat[3+3*(i:seq_length1)])}

downprop[1,(1+3*(0:seq_length2))] = rev(cumsum(rev(read_mat[1,(1+3*(1:seq_length1))])))/
totalsDown[0:seq_length1]
downprop[1,(2+3*(0:seq_length2))] = rev(cumsum(rev(read_mat[1,(2+3*(1:seq_length1))])))/
totalsDown[0:seq_length1]
downprop[1,(3+3*(0:seq_length2))] = rev(cumsum(rev(read_mat[1,(3+3*(1:seq_length1))])))/
totalsDown[0:seq_length1]

#Calculate the CSCPD (absolute difference between cumulative upstream and downstream
proportions for sub-codon positions 1,2,3)
third_full_seq_minus1 = (length(Coordinate)/3)-1

y1 = abs(upprop-downprop)[1+3*(0:third_full_seq_minus1)]
y1_withoutNA = y1[which(y1 != "NA")]
y1_withoutNaN = y1_withoutNA[which(y1_withoutNA != "NaN")]
length_third_seq_withoutNaN = length(y1_withoutNaN)
length_third_seq_withoutNaN_minus1= length(y1_withoutNaN)-1

y2 = abs(upprop-downprop)[2+3*(0:third_full_seq_minus1)]
y2_withoutNA = y2[which(y2 != "NA")]
y2_withoutNaN = y2_withoutNA[which(y2_withoutNA != "NaN")]

y3 = abs(upprop-downprop)[3+3*(0:third_full_seq_minus1)]
y3_withoutNA = y3[which(y3 != "NA")]
y3_withoutNaN = y3_withoutNA[which(y3_withoutNA != "NaN")]
```

```
x1 = 1+3*(0:length_third_seq_withoutNaN_minus1)
x2 = 2+3*(0:length_third_seq_withoutNaN_minus1)
x3 = 3+3*(0:length_third_seq_withoutNaN_minus1)

ys1 = smooth.spline(x1/(length_third_seq_withoutNaN*3),y1_withoutNaN)
ys2 = smooth.spline(x2/(length_third_seq_withoutNaN*3),y2_withoutNaN)
ys3 = smooth.spline(x3/(length_third_seq_withoutNaN*3),y3_withoutNaN)


xout = 0.01*(1:100)
yout1 = predict(ys1,xout)$y
yout2 = predict(ys2,xout)$y
yout3 = predict(ys3,xout)$y


difference_mat1[gene_counter,] = yout1 - Percentile_P1
difference_mat2[gene_counter,] = yout2 - Percentile_P2
difference_mat3[gene_counter,] = yout3 - Percentile_P3

for (j in 1:100)
{
        if (as.numeric(difference_mat1[gene_counter,j]) >= 0){
                PTS[gene_counter,1] = PTS[gene_counter,1] + difference_mat1[gene_counter,j]
                }
        if (difference_mat2[gene_counter,j] >= 0){
                PTS[gene_counter,2] = PTS[gene_counter,2] + difference_mat2[gene_counter,j]
        }
        if (difference_mat3[gene_counter,j] >= 0){
                PTS[gene_counter,3] = PTS[gene_counter,3] + difference_mat3[gene_counter,j]

        }

}

PTS[gene_counter,4] = PTS[gene_counter,1] + PTS[gene_counter,2] + PTS[gene_counter,3]
gene_counter = gene_counter+1

detach(reads)

}

write.csv(PTS, file = "/home/PTS.csv",col.names = TRUE, row.names = rnames)
```