Predicting non-canonical translation of protein-coding genes

Thesis Submitted for the degree "Doctor of Philosophy" by

Amit Alon

Submitted to the Senate of Tel-Aviv University:

27/10/2019

Predicting non-canonical translation of protein-coding genes

Thesis Submitted for the degree "Doctor Of Philosophy" by

Amit Alon

Submitted to the Senate of Tel-Aviv University:

27/10/2019

This work was carried out under the supervision of Prof. Oded Rechavi, TAU, in collaboration with Prof. Yitzhak Pilpel.

Prof. Oded Rechavi

Table of contents:

Abstract

The work presented in this thesis is focused on the possibility of having errors during protein translation. It is now clear that the original idea that “one gene encodes for one protein” was naïve, as multiple processes, for example, alternative splicing, allow multiple protein versions resulting from a single locus in the genome (Modrek and Lee, 2002). In my thesis work, I developed new computational tools for the discovery of non-canonical proteins and peptides. These tools were applied to predict non-canonical protein translation that can arise due to stop-codon read-through and frameshifting. Further, I demonstrated that these tools enable predicting prionicity (another example of non-canonical protein function). The underlying assumption was that in all of these cases, one mRNA molecule can be translated to produce multiple protein products owing to the alteration of the process of translation.

The tools designed revealed new proteins with prion-forming potential, abundant stop-coding read-through, and, most interestingly, many cases of apparent ribosomal frameshifting. Specifically, I found more than 200 novel human protein versions, hypothesized as resulting from ribosomal frameshift, and evidence for dozens on *Saccharomyces Cerevisiae* proteins where STOP codon read-through may significantly alter the protein’s original function. Furthermore, as a proof of principle, I also discovered two *Caenorhabditis Elegans* proteins with a high potential to act as prions.

To further investigate non-canonical behavior, I developed computational tools that can assist in systematically finding non-canonical translation of protein-coding genes, using the raw sequence of the protein or its encoding gene.

The first tool developed for this thesis was based on analyses of a set of DNA sequence properties that characterize protein-coding genes and was aimed at detecting deviation from a specific pattern that is expected to appear for these kinds of genes. The most important property is a periodic pattern in nucleotide variability. This is because the genetic code is redundant, allowing plasticity in creating a protein-coding sequence while still maintaining a desired amino-acid sequence.

This property can be used to examine a DNA sequence for its potential to encode a protein, based on the “extent” of periodic patterns that it holds. Given

a set of sequences from different species along with the frequency of codons coding for the same amino acid, assuming that amino acid identity should be mostly conserved to serve as an active protein, one can have a theoretical estimation of the level of conservation expected from a sequence, thus how many patterns for a period are expected to appear. I also used this property to find alternative translational frames in previously annotated protein-coding genes, revealing new candidate protein versions within already annotated genes, to be predicted.

The second tool used was a previously developed algorithm that would predict a protein's potential to have prion-forming domains. This original tool was mainly based on observed sequences that hold prion domains in the yeast *Saccharomyces Cerevisiae*. Once a computational basis was set, I moved on to find biological evidence and significance to our computational findings: (1) I found that the top-scoring protein in *C. elegans* (highest potential to have a prion forming domain) is a stress-related protein that might be involved in early pathogen exposure and protection. (2) I found that some of the genes predicted to undergo STOP codon read-through in *S. cerevisiae* gain a transmembrane domain due to translation of the 3' UTR and show evidence of change in their cellular location due to environmental changes. (3) I obtained evolutionary and experimental (via ribosome profiling) evidence of alternative translation options for some frame-shifted gene predictions from the human genome, strengthening the likelihood that these may be functioning units that could be produced by cells.

In my thesis, I focused mainly on the findings obtained from the non-canonical frame detection in humans and mammals in general, as these are the most elaborate and exciting results that I obtained. This challenge increased my understanding of a generally poorly understood subject. From the findings, I concluded that frameshifting might dramatically alter the translated product, generating major truncations, functional domain loss, binding properties alterations, protein folding changes, and more. The implications of these findings could be substantial; the results revealed a potential for an extra level of diversity in the proteome that may be programmed. If these phenomena are adaptive, then a strong regulation mechanism is likely to be in place. Such a mechanism is known to exist in organisms possessing compact genomes, such as in viruses, and the

identification of other mechanisms in mammals could be extremely important in the future.

I hypothesized that some of the frameshift events observed in humans may be adaptations to recent viral infections, which would utilize regulation mechanisms that would help generate new proteins when needed.

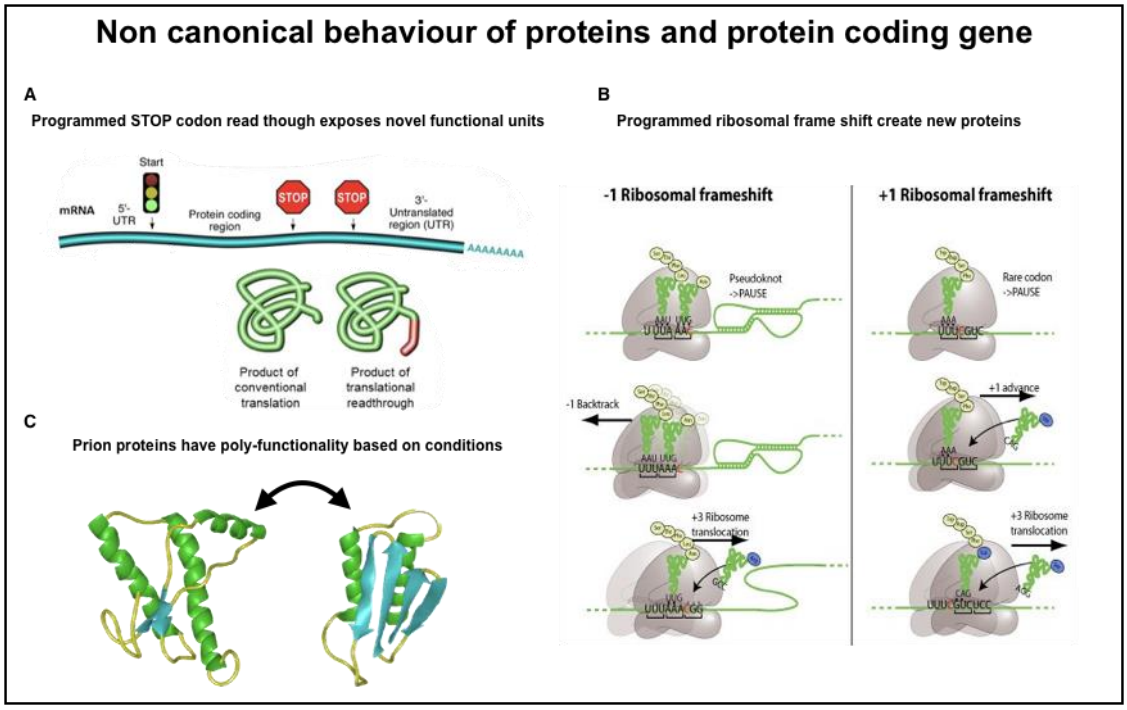


Figure 1: Graphical abstract of thesis layout. This thesis is divided into 3 chapters: **(A)** Non-canonical protein translation detection, via STOP codon read-through, in *Saccharomyces Cerevisiae*. **(B)** Non-canonical protein translation via frame-shifting detection in the human genome. **(C)** Predicting prion protein candidates in *C. Elegans*

Structure of the thesis

This thesis is divided into three chapters, each describing a different process forming non-canonical translation and activation proteins. The first part will set the basis for non-canonical protein translation by examining STOP codon read-through potential in the relatively simple genome of the yeast *Saccharomyces Cerevisiae*. The second and major part describes non-canonical protein translation analysis of the human genome. This part explores evidence of STOP codon read-through and ribosomal frameshifts in human protein-coding genes, based on the conservation profile that the DNA sequences hold when aligned against orthologs in other mammals. After delineating the mathematical procedure, I discuss the implications of the results and potential mechanisms of action. The third and last part focuses on the post-translational aspect of non-canonical protein function. I describe how I characterize prion-forming proteins in the nematode Caenorhabditis *Elegans* based on empirical evidence from experiments in yeasts.

Introduction

In this thesis, I focused on different aspects of protein translation. To date, when examining a species genome and proteome using existing databases, most of the sequences were derived using computational tools. Annotating an open reading frame (ORF) using computational methods requires some assumptions (Fickett, 1996):

1. Some sequence similarities (e.g., repeats) are less likely to appear in protein-coding genes
2. Some sequence similarities to other known protein-coding genes
3. Codon bias measures that correlate with other known biases from proven ORFs (of the same or similar species)
4. Template patterns matching known functional sites

There are many methods for identifying protein-coding regions (Brent, 2005, 2007; Frith et al., 2006; Mathé et al., 2002; Rogic et al., 2001; Yandell and Ence, 2012; Zhang, 2002), which all rely on the basic assumptions stated. If no experimental data is available, some methods can be applied to give an initial prediction (Dunham et al., 2000; Majoros et al., 2004; Salamov and Solovyev, 2000; Yada et al., 2003). Finding characteristics of large cDNAs, RNA-seq, and expressed sequence tags (ESTs) collections have also been developed to specify CDS sequences within them (Exploration et al., 2001; Furuno et al., 2003; Hatzigeorgiou et al., 2001; Lin et al., 2011; Min et al., 2005; Ota et al., 2004). When taking these assumptions to mind and examining the annotation of a genome, one must consider that these might narrow down the actual translational potential of a gene or a sequence since it had no previous evidence for it. With that in mind, The purpose of this thesis was to investigate vast genomic databases and lay the basis for systematic computational analyses that may uncover hidden, yet unreported, ORFs or translation deviations from the classical dogma (Harte et al., 2010).

The classical dogma of a typical eukaryotic mature mRNA is that of a monocistronic molecule with a tripartite structure: 5’ and 3’ untranslated regions (UTRs) are surrounding (or flanking) a single ORF or coding sequence (CDS). In eukaryotic mRNA, the model for scanning for initiation sites is based on the cap-dependent process of ribosome binding. Simplified, a 43S preinitiation complex binds to the cap structure at the 5’ end and scans the 5’

UTR to arrest at a translation initiation site (TIS). The large 60S subunit then joins to form a fully functional 80S ribosome, and polypeptide synthesis starts. This contrasts with bacterial ribosomes, which can also bind to internal binding sites in polycistronic mRNAs. This model cannot agree with a notion other than that the mRNA is monocistronic and is translated into a single polypeptide (Kozak, 1999).

The classical dogma for determining the annotated CDS usually chooses the longest sequence found between an initiation and an elongation site. Alternative (non-canonical) protein translation that is traditionally proposed (Klemke et al., 2001; Mouilleron et al., 2016):

- Alternative initiation sites may define non-canonical CDS translating different protein sequences. These initiation sites can be found within the annotated CDS and in the 5' UTR. Initiation sites residing in the 5' UTR will define upstream ORFs and are believed to translate translational regulatory elements (Calvo et al., 2009; Wethmar, 2014; Wethmar et al., 2014).
- Alternative initiation sites within the annotated CDS, whether in the canonical frame or a different one, would generate either a shorter ORF with a similar amino acid sequence (if in the canonical frame) or a completely different protein product (when in a different translational frame). The latter could also expand beyond the canonical STOP codon into the 3' UTR (Klemke et al., 2001).

Several studies conducting genome-wide bioinformatic analyses on mammalian genomes predicted candidates of alternative ORFs originating from translation initiation sites (TIS) within the annotated CDS (Chung et al., 2007; Ribrioux et al., 2008; Xu et al., 2010). Using different filters on predicted TIS and ORF (length, conservation, signals), these studies failed to predict several known alternative ORFs. Subsequent works used a less stringent strategy and came up with a catalog of ~17,000 alternative ORFs within annotated CDS in the human transcriptome (Vanderperre et al., 2012), and were later applied in other eukaryotes (Vanderperre et al., 2013).

Large-scale experimental studies also attempt to show the presence of non-canonical ORFs. These center around two main approaches:

The first approach is based on mass spectrometry (MS) experiments, where peptides are identified using a spectrometer fed with digested proteins and mapped to a given proteome. These rely heavily on a given sequence database. The primary databases used are UniProt Knowledgebase (Magrane and Consortium, 2011) and the NCBI Reference Sequence collection (Pruitt et al., 2014). The alternative sequences are not a part of these collections and thus cannot be identified when using them. To detect them, one must generate a predicted alternatively translated sequences database. Vanderperre et al., Did just that and reported 1259 novel human peptides in different cell lines, tissues, and fluids (Vanderperre et al., 2013). Another challenge posed in searching for alternatively translated proteins using mass spectrometry is that these would usually be short proteins. Short proteins, in general, are more difficult to detect by MS, and studies performed on human K 562 cells were able to detect ~200 alternative peptides (Ma et al., 2014; Oyama et al., 2004; Slavoff et al., 2013).

The second experimental approach for large-scale detection of translated sequences is ribosome profiling (Brar and Weissman, 2015; Ingolia, 2014; Ingolia et al., 2009a). In these experiments, mRNA fragments protected by ribosomes are isolated and treated with a nuclease. These fragments are later sequenced and mapped to some reference sequence database. Using this technique, translated fragments within transcripts, as well as TIS, can be identified (Ingolia, 2014; Ingolia et al., 2009a). In these experiments, the initial prediction of the alternative peptide sequence is not as crucial as apriori, but rather the transcriptomes are assumed to be alternatively translated. The detection efficiency here is much higher but still depends on having the right setting for the alternative translation to take action.

Considering all of this, It would seem that there are many alternative ORFs, and the complexity of translation might seem incredibly high. That said, certain translation mechanisms must exist to facilitate the different levels of alternative translation. For TIS other than the annotated one, more conditions or signals must exist, for example, Kozak sequences that mediate translation initiation in eukaryotic cells (Kozak, 1987, 1991, 1999). Programmed ribosomal frameshifting (PRF) is also a means of alternative translation that usually happen around specific sequence signatures “instructing” the ribosome to change its current reading frame (Jacobs et al., 2007). Lastly, stop codon

readthrough, which would elongate the canonical ORF, would require silencing termination factors or perhaps leaky sequences around the stop codon (Namy et al., 2001; True and Lindquist, 2000).

These phenomena may carry vast implications for both the functional and structural investigation and the interpretation of omics data. In many cases where an alternative ORF exists, it lies within the canonically annotated ORF (Vanderperre et al., 2013). Another implication is when considering synonymous mutations in sequences where an out-of-frame overlapping alternative ORF exists. These mutations may be synonymous in one frame but would most probably have stringer effects on another and may implicate pathological manifestations (Hunt et al., 2014). These would be difficult to test empirically since the mRNA sequence would likely be the same, but the translation mechanism deviates from canonical translation. For that reason, it is important to have a proper systematic prediction tool that would hint into areas where alternative translation would cause functional changes and eventually implicate its environment.

Chapter 1: Non-canonical protein translation detection via STOP codon read-through in *Saccharomyces Cerevisiae*

Overview of Chapter 1:

Protein-coding sequences have specific properties that differentiate them from other sequences. I used these properties to computationally explore the 3' *Un-Translated Regions* ('*3' UTRome'*) of protein-coding genes. This chapter rests on the notion that the proteome has an extra degree of diversity, which can be unveiled when the read-through of STOP codons occurs. I hypothesized that organisms can control such read-throughs and that under certain conditions, the STOP codon can be read through, allowing translation elongation. The concatenated parts, added to the otherwise shorter peptide, can affect the protein's function, binding properties, localization, secondary structure, and more. I developed a computational approach to scan whole genomes for protein candidates with probability potential for translation beyond the STOP codon. This approach considered many sequence-based properties from the possibly translated *UTRome*, such as the codon usage and amino acid composition profiles that match the species proteome. It also tested translation efficiency profiles throughout the original gene sequence and the sequence beyond the STOP codon. I examined evolutionary evidence of the possibility of translation beyond the STOP codon, using multiple sequence alignment profiles. Whenever such data was available, I analyzed ribosome profiling data and RNA sequencing data to find further evidence that the sequences have an actual potential to be translated.

Scientific background for non-canonical protein translation detection in *S. cerevisiae*

While usually, the three STOP codons (UAA, UAG, UGA) lead to detachment of the ribosome from the mRNA molecule (Jackson et al., 2012), there are some cases where this fails to happen; therefore, a STOP codon read-through event occurs. This can happen due to failure in detecting the STOP codon, causing insertion of a different amino acid instead of the STOP codon, so that translation continues in the original frame of the protein; or due to frame-shifts generating an altered sequence to be translated (Namy et al., 2004; Schueren et al., 2014). Either way, the translation will theoretically continue until the ribosome encounters another STOP codon, detaching from the mRNA. The resulting protein would possess an extra peptide that initially should not have been translated and thus might gain a new functioning unit (Jungreis et al., 2011). In *S. cerevisiae*, there is evidence of STOP codon read-throughs from ribosome profiling experiments, mRNA sequencing, and mass spectroscopy essays under various conditions (Baudin-Baillieu et al., 2014; Dunn et al., 2013). The most studied condition that leads to STOP codon read-through is manipulations of the translation termination factor SUP-35. Under normal conditions, this protein acts as a release factor of the ribosome. Under specific environmental conditions, SUP-35 will convert into a prion state [PSI+], and translation will not be terminated in the STOP codons it must release, causing an overall read-through of potentially many proteins. These events can affect other proteins and many cellular functions. As a control, STOP codon readthrough can be eliminated using the [PSI-] strain in such experimental manipulations where the prion version of SUP-35 cannot exist (Torabi and Kruglyak, 2011). Ribosome profiling experiments conducted on [PSI+] and [PSI-] strains showed that ribosomes are populating the 3' *Untranslated Region* (*UTR*), in a density that is comparable to that of the original *Open Reading Frame* (*ORF*) (Baudin-Baillieu et al., 2014; Dunn et al., 2013), which indicates a high propensity of translation in these regions. Moreover, several proteins showed a state of 'translational ambiguity, where more than one translational frame is active in (Baudin-Baillieu et al., 2014). Such read-through events are exciting as they may provide relatively new means for diversifying the proteome with a given genome.

Studies have also shown that STOP-codon read-through can happen due to the existence of “leaky” nucleotide motifs (downstream and upstream) (Namy et al., 2001; Skuzeski et al., 1991). It has also been stated that a phenomenon termed *Programmed Ribosomal Frameshift* (*PRF*) occurs in some species. This means that certain RNA sequences can cause the ribosome to move 1 or 2 nucleotides (upstream or downstream), resulting in a frameshift (Farabaugh et al., 2006; Jonathan, 2012). The presence of these sequences in the area right before the STOP codon can result in frameshifts that lead to a STOP codon read-through (Mikl et al., 2018).

STOP codon read-through is an interesting regulatory mechanism for exposing additional C-terminal domains of a protein, even if at much lower expression levels than the original protein. Viruses use this mechanism to increase functional versatility in a compact genome and control the ratio of two protein isoforms (Jungreis et al., 2011). It has been suggested as an evolutionary facilitator in yeast, where it is epigenetically controlled via a prion [PSI+], enabling the adaptation of new domains translated at low rates during normal growth but at higher rates in periods of stress when they might provide a selective advantage (True and Lindquist, 2000). STOP codon read-through in eukaryotes is also known to happen in transposable elements that could be endogenous retroviruses (Jungreis et al., 2011).

Previous work had laid down the basis for this hypothesis by exploring 12 Drosophila genomes to find possible STOP codon read-through events (Chan et al., 2013). In their work, they used the aligned orthologous sequences to find areas downstream to the canonical STOP codon that preserves properties of translated sequences. They used patterns of translated and untranslated sequences to predict the probability of the presence of nucleotides in specific locations. To complete their study, Chan et al. also used GFP tagging and mass spectrometry data to confirm the translation and function of these regions.

Research goal for detection of STOP codon read through in *S. cerevisiae*

My goal was to find groups of proteins across many species that display high STOP codon read-through potential. Once such genes are defined, I attempted to find further evidence to support these hypotheses. I wanted to characterize the effect of the translation elongation in cells and entire organisms, as could be captured in protein function change, localization, interactions, etc.

My main objective was to define a set of rules (and numerical thresholds) that will apply to many species towards systematically finding many read-through candidates.

My first aim was to produce a computational tool for systematically identifying read-through candidates for any species. I examined nucleotide sequences for several properties (depicted in the following sections), creating a system to rank genes by their probability of having STOP codon read-through and a functioning protein with the extra peptide translated.

I started by focusing on the yeast *Saccharomyces Cerevisiae* as a model since its genome is relatively small, simple, well-sequenced, and has a detailed phylogenetic conservation database. Moreover, a genetic background in which STOP codon read-through ([PSI+]) occurs is widely used and established, giving a solid starting point for testing for translation beyond the STOP codon and how it affects the cell.

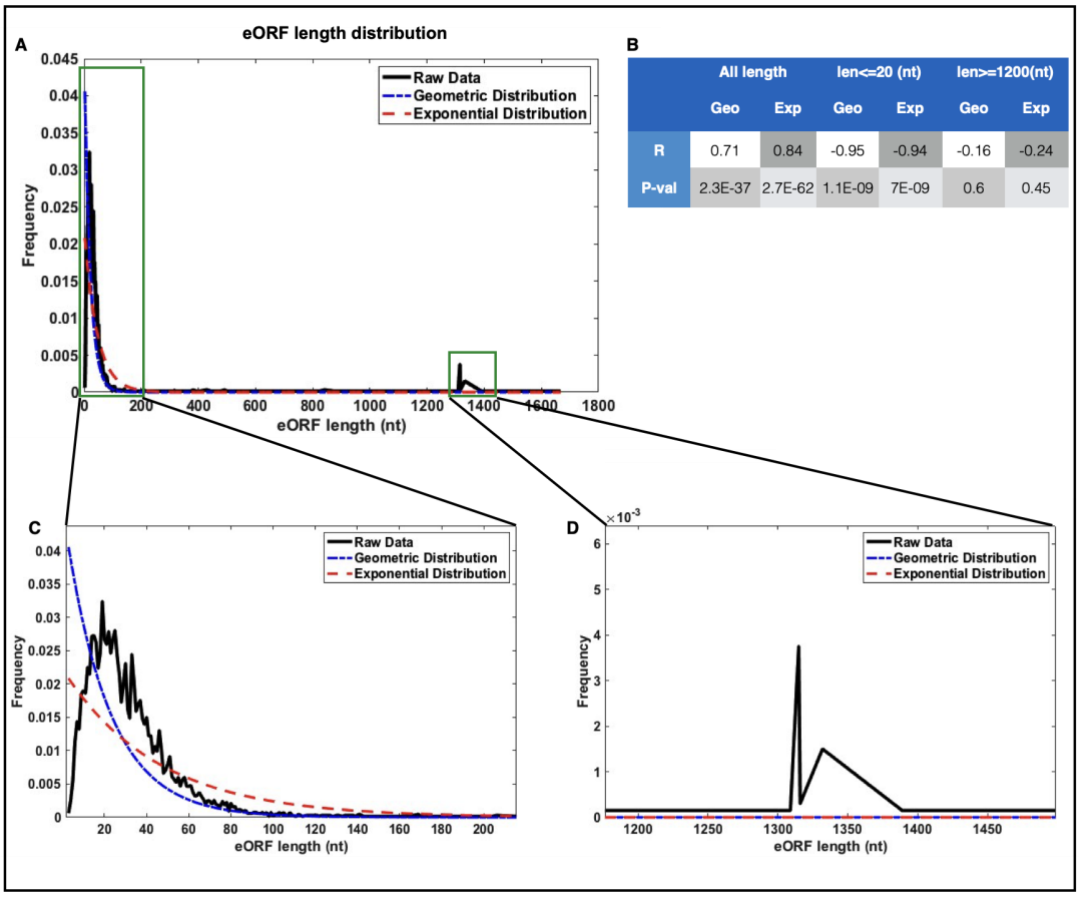
Results for detection of STOP codon read through in *S. cerevisiae*

1. *eORF* length

To examine what lies in the 3' UTR, I first recorded the number of codons downstream to the canonical STOP codon (until reaching a second STOP codon) in all three reading frames (0, +1, +2) for all protein-coding genes. The reading frame with the highest number of codons between the canonical STOP and the subsequent downstream STOP codon was termed the *eORF*. This was because having a STOP codon readthrough event that is programmed and intentional should, theoretically, provide an extended sequence that can be translated. The sequence is envisaged to be longer than expected at random in a null model of random sequences of nucleotides. After examining the distribution of lengths (figure 2A), it was clear that the distribution decays exponentially, as expected by random sequences. With that, it seems that the abundance of low-range lengths (up to ~20 codons) seemed somewhat higher than expected (figure 2B-C). Also, I found the distribution to have two distinct populations: one in the range of between 10-20 codons and another with a very high number of codons (over 400 codons). This can suggest that very long *eORFs* might have biological significance since they are not likely to appear by chance, contributing to the hypothesis that some regulation process was present. When examining Pearson's correlation between the curves representing *eORF* lengths distribution and the theoretical geometric distribution, I got high positive correlations, as calculated by linear Pearson's correlation (see figure 2B). The left tail of the distributions is negatively correlated as well as the right tail, although, for the latter, the correlation isn't statistically significant. For the decaying phase, there was a significant yet negative correlation, this is since this phase from the raw data is not purely decaying but represents a peak.

Often, a cutoff of 100 codons is used for the determination of the significance of a CDS, even though there is much evidence of peptides translated from short unannotated CDSs with important functions (Andrews and Rothnagel, 2014; Chng et al., 2013). Furthermore, excluding short CDSs is a part of annotation guidelines used by top publicly available gene sets, including GENCODE, RefSeq, Ensembl, and more (Mouilleron et al., 2016).

Due to all these, I decided to first look at ORFs with an *eORF* of >= 100 codons.



	All length		len<=20 (nt)		len>=1200(nt)	
	Geo	Exp	Geo	Exp	Geo	Exp
R	0.71	0.84	-0.95	-0.94	-0.16	-0.24
P-val	2.3E-37	2.7E-62	1.1E-09	7E-09	0.6	0.45

Figure 2: *eORF* length distribution. (A) *eORF* length distribution. Comparing *eORF* sequences to a randomly generated sequence, I calculated the theoretical length distribution using geometric distribution and its continuous version, the exponential distribution. I saw that the decay acts as predicted by these models, but the real *eORF* data has a peak at around 50 nts (**C**), and another at ~1350 nts (**D**). **(B)** The Pearson correlation coefficients and their p-values for the complete distribution (All lengths), short and long *eORF* lengths. (**C**) Zoom into the decaying phase. (**D**) The peak at very long *eORF* represents the yeast's transposable elements that are known to be fusion genes resulting from STOP codon read through at the first component. Green frames show the deviation from the theoretical distributions of lengths for short (<=20 nt) and long (>=1200 nt) *eORF* lengths.

In this group, if an ORF (protein) undergoes a read-through/frameshift, causing the *eORF* to be translated, it is more likely to have a functional effect. Such additions to a protein could code a new protein domain, a localization signal, create a new protein secondary structure, and so on. When testing the properties of genes in this group using Gene Ontology (*GO*) annotation, I found a strong enrichment for "transposition proteins" (p-value = 10^{-51}) which also have RNA binding function, as calculated by mHG model (Eden et al., 2009), that can identify, independently for each GO term, the threshold at which the most significant enrichment is obtained. The significance score is

corrected for multiple testing. These are the “gag proteins” of the retro-element that is abundant in the yeast genome (Clare et al., 1988) and are known to have an alternative translation form of fusion with the next ORF (at the translation level) (Clare et al., 1988). Since this isn’t news and considering the presence of very short protein sequences within the UTR (Nuclear Localization Signals (*NLS*) etc.), I decided that focusing on this particular group is too biased. I further decided to expand the analyses beyond detecting mere long extensions and generate a list of 3’ UTR sequence properties that could indicate STOP codon read-through.

Another interesting finding was a subset of genes having another STOP codon immediately after the canonical one. Some of them have more than one such STOP codon, creating a sequence of consecutive STOP codons. It might suggest that this signal acts as a tight regulator for translation termination in the STOP codon read-through induced condition. My hypothesis was that should the sequence after the stretch of consecutive STOP codons be translated, it might have a devastating effect on the cell, and evolution found a means of dealing with this. I termed these genes *Stop Means Stop* (*SMS*) and further investigated the implications of translation elongation of these genes. I examined their sequence properties, much like for *eORFs,* and continued looking for mutations that may cause the translation of these sequences and their links with phenotypes. I could not find significant enrichment with the essential genes or stress-related genes in the group where *SMS* was conserved. To determine if there was significance in sequence conservation of the adjacent STOP codon, I calculated conservation scores for the conserved SMS genes ortho-groups and a selected control group. The control group was designed to have sequences where the next in-frame codon after the canonical STOP codon was conserved such that at least one other species in the group had a codon coding for the same aa. Not all aa were selected to participate in building this control group. Instead, only those with at least three codons coding for an aa (to have fair comparability between STOP codons that are coded using three different codons). Since conservation scores were calculated using entropy, each comparison was made between orthogroups with the same number of species. In general, I found no correlation between the evolutionary conservation score of the canonical STOP codon and the next in-frame codon (Pearson’s correlation, R=0.02, p-value=0.05), and that the

next in-frame codon is usually much less conserved than the canonical stop codon (mean entropy score of 0.74 for the canonical score codon vs. 1.57 for the next in-frame codon). Comparing entropy scores between conserved SMS genes and the conserved aa coding control groups, there was no statistical significance for conserving the consecutive STOP codon (t-test for the entropy of codon identity, **table 1**).

# of sequences in ortho-group	Canonical STOP codon	Next in frame codon
2	p-val = 0.83	p-val = 0.64
3	p-val = 0.16	p-val = 0.64
4	p-val = 0.40	p-val = 0.18
5	p-val = 0.31	p-val = 0.24

Table 1: P-values for t-test on entropy scores of codon conservation in gene ortho-groups. The t-test was calculated for the entropy scores as calculated on the conserved SMS genes, and the control group as described in the Methods section. The null hypothesis was that the conservation scores are derived from the same distribution without assuming equal variance. As can be seen from the results, I could not reject this null hypothesis, thus concluded that this observed conservation could have happened by naturally occurring mutations. and there is no indication of

2. Sequence properties of *S. cerevisiae eORFs*

2.1. Amino acid composition correlation and codon usage correlation

When determining whether a sequence has the potential to be translated, important features to examine are amino acid composition and codon usage correlations. One might assume that a translated sequence should fit the trend that has already been set from the actual known proteome. For this, I used two control groups: The first was a random sequence consisting of 600 nucleotides (150 from each of the four nucleotides ACGT, preserving a constant GC content). The sequence was shuffled thousands of times (to fit the number of genes explored). The second control group was the group of the *eORFs* sequence randomly shuffled (conserving GC content).

For every group in this comparison, I calculated the aa composition and the codon usage correlation scores between the *eORF* sequences (natural, random, or shuffled) and the corresponding coding sequence for that protein, using Pearson's linear correlation measure. When simply observing the sequence correlation scores for codon usage, it is apparent that those scores for the actual *eORFs* are significantly higher relative to those obtained from the two control

groups. Applying a t-test to compare these scores distributions reveals that they are not derived from the same distribution, strengthening the assumption that these correlation scores were not random (p-value << 0.05). When testing if there is some relationship between the sequence correlation scores and the length of the *eORF* (assuming that longer sequences have a higher chance of having coding-like properties), I got positive Pearson correlation scores for both the actual *eORF* sequences and the two control groups. For codon usage correlation with the *eORF* length, I got R=[0.45, 0.34, 0.54] for the natural *eORF*, shuffled random sequences, and shuffled *eORF* sequences groups respectively, taking all genes into account. For aa composition correlation with *eORF* length, I got R=[0.33, 0.006, 0.46] for the natural *eORF*, shuffled random sequences, and shuffled *eORF* sequences groups respectively when all genes are taken into account. When calculating the same measure for the long *eORF* genes only (*eORF* length >= 100 nts), The drift between the natural *eORF* and the two control groups increases. For codon usage correlation with the *eORF* length, I got R=[0.70, 0.17, -0.19] for the natural *eORF*, shuffled random sequences, and shuffled eORF sequences groups respectively. For aa composition correlation with *eORF* length, I got R=[0.67, 0.004, 0.05] for the natural *eORF*, shuffled random sequences, and shuffled *eORF* sequences groups respectively. That is to say, *eORFs* with long extensions tend to have amino acid and codon composition that is typical to that of the entire proteome (figures 3 and 4). As expected, both control group sequences were not very correlative with the proteome composition for amino acid and codon. Although the t-test for comparison between sequence properties correlation scores revealed that the natural *eORF* sequences were not derived from the same distribution of the control groups, codon usage correlation scores appeared to be a better measure when taking the length of the *eORF* into account. This result probably stems from certain amino acids having up to 6 times more codons in the genetic table than others. Hence, they would appear to have a higher occurrence even in random nucleotide sequences.

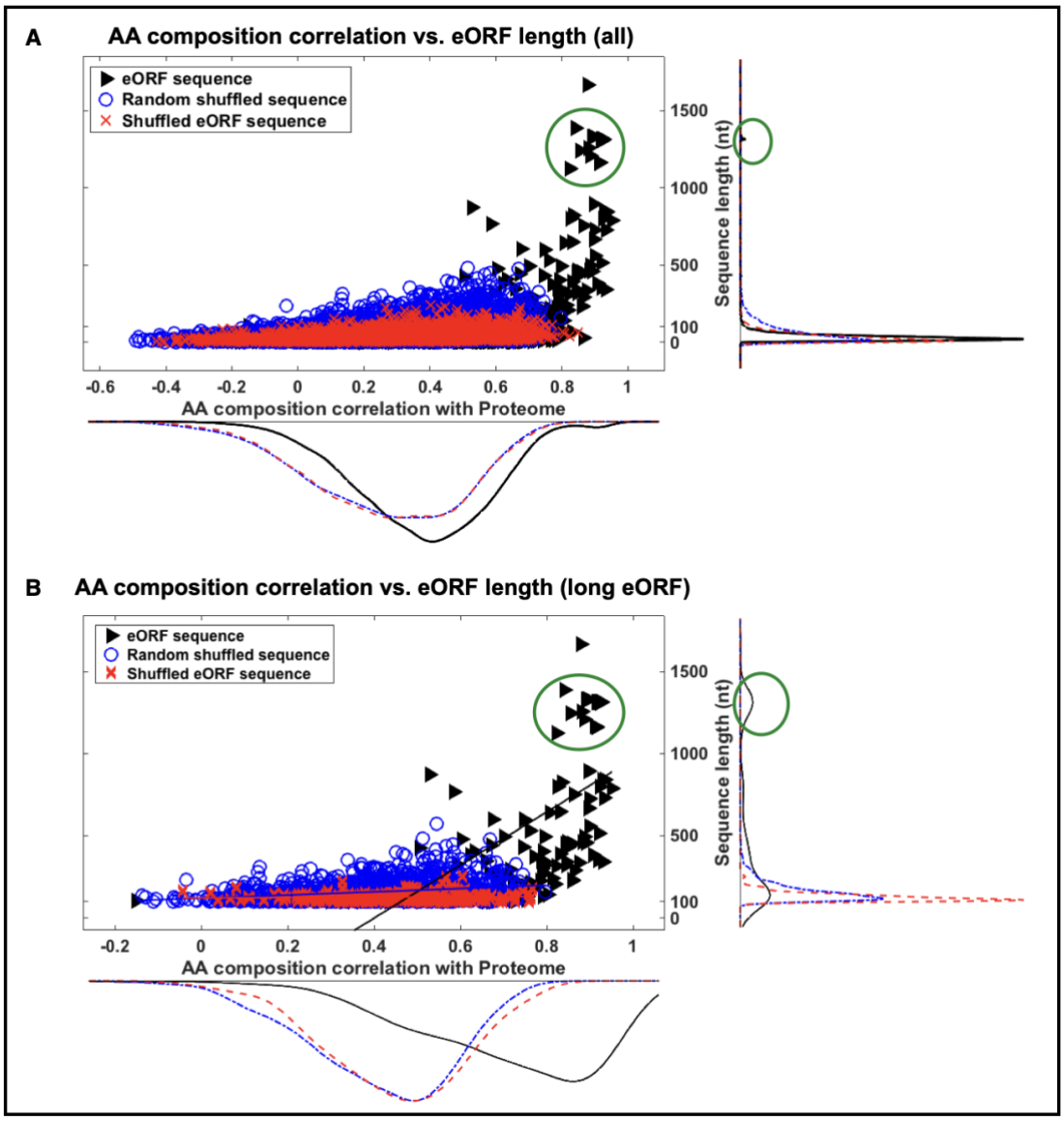


Figure 3: Amino Acid Composition Correlations. AA composition correlation scores between eORF and the proteome vs. *eORF* length (black). Two control sets were generated, the first (blue) are random sequences with 50% GC content. The second (red) was generated by shuffling nucleotides in the *eORF* itself (maintaining the GC content). **(A)** represents all *eORFs*. **(B)** are only *eORFs* longer than 100 nts. Green circles mark the transposable elements. Composition correlation scores were relatively higher for natural eORF sequences, but this trend became significant (t-test p-val<<0.05) when taking only longer sequences into account.

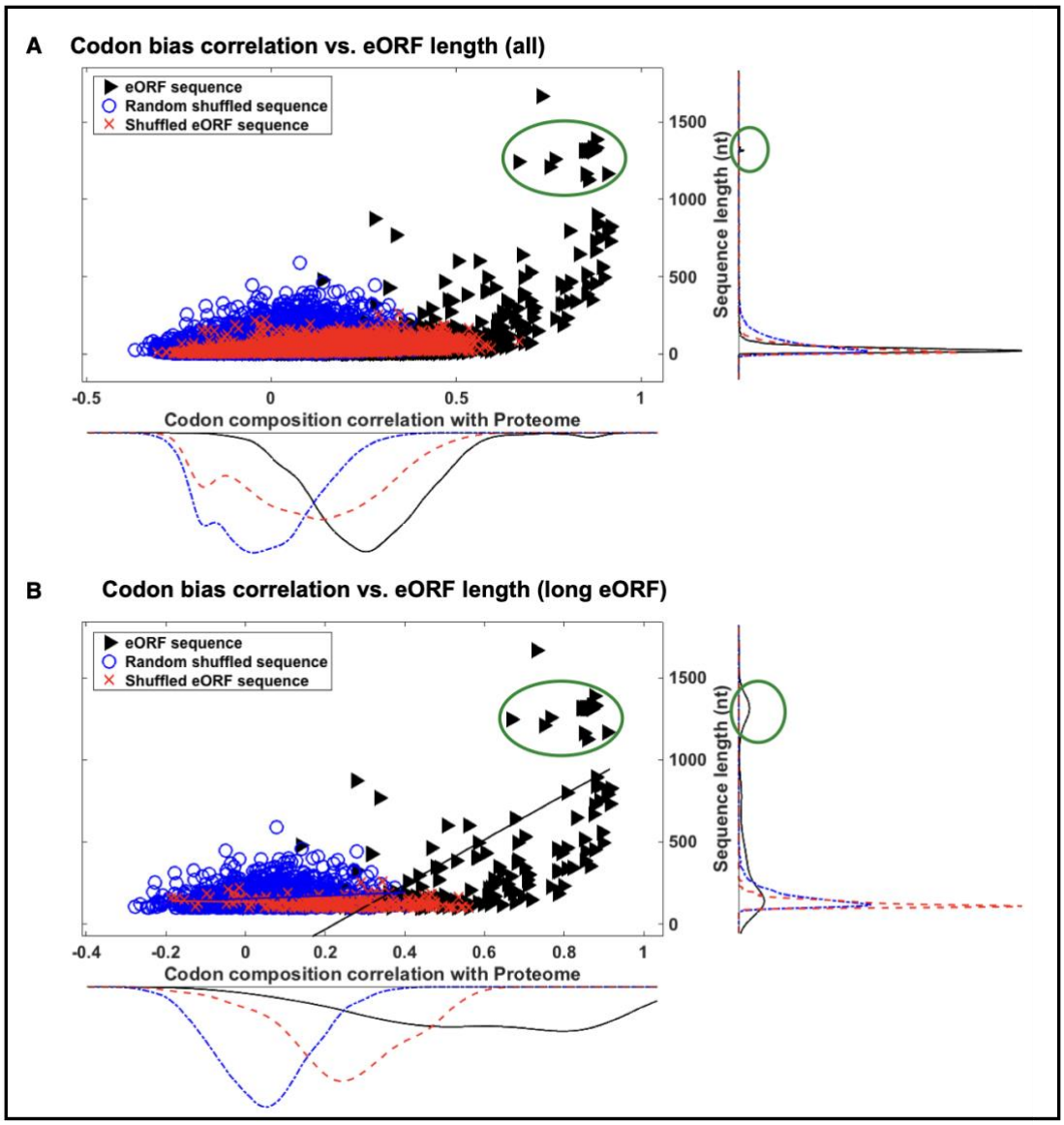


Figure 4: Codon bias in Composition Correlations. *eORF* Codon composition correlation to that of the *S. cerevisiae* proteome relative to *eORF* length (black). Two control sets were generated; the first (blue) are random sequences with 50% GC content. The second (red) was generated by shuffling nucleotides in the *eORF* (maintaining the GC content). **(A)** represents all *eORFs*. **(B)** are only *eORFs* longer than 100 nucleotides. Green circles mark the transposable elements. Composition correlation scores were relatively higher for natural eORF sequences, but this trend became significant (t-test p-val<<0.05) when taking only longer sequences into account.

3. Translation efficiency

3.1. tRNA Adaptation Index analysis

Another strong evidence for translation is the sequences' *tRNA Adaptation Index* (*tAI*). The tAI is a means of measuring translation efficiency. tAI was first computed as a measure of the codon bias in the composition (dos Reis et al., 2004). I found that coding regions hold higher local tAI scores than those observed in the UTRs (see figure

5A). Should some regions in the *eORF* have a higher tAI score than expected by random sequences, it might imply into translation of these sequences. I thus examined the local tAI scores (see methods section for details) for the last 100 codons (up-stream to the STOP codon), and the first 20 codons of the *eORFs*, as well as the last 20 codons of the *eORF,* to find an indication of high local tAI scores in the *eORF* comparable to those in the ORF.

The goal was to understand if I could identify possible read-through candidates where the local tAI score does not decrease to a shuffled sequence level once crossing the STOP codon. I compared local tAI profiles for the sequences with that of the same sequence after nucleotide shuffling to find significant increases in UTRs of genes that may be translated. As a positive control, I used the transposable elements genes (gag-pol) and the group of genes reported as having programmed read-through in (Namy et al., 2003). Since most genes from the latter positive control group revealed very short *eORFs* (mostly under 20 aa), I could not use them to estimate tAI properties. I had 45 transposable element sequences as a control group. These genes are particular in their sequence properties as they hold a specific codon usage pattern and code to similar sequences. Due to this, I could only try to estimate the deviation of local tAI scores from that of shuffled sequences. When assessing these differences, I could not find any sequence that holds a long stretch of local tAI scores that surpassed the 95 percentile of the shuffled sequence scores, neither in the coding sequence nor the *eORF*. Due to this, I was unable to decide on using this measure.

However, calculating the gene tAI scores of the ORFs and the *eORFs* separately (using formula (1) from the Methods section), I hoped to find some positive correlation between ORF and *eORF* tAI for genes that are being read-through. I used a control group of shuffled assignments of ORF and *eORF*. Referring to a group of genes that have been shown to have STOP codon read-through from ribosome profiling termed *RP-PRT* (Ribosome Profiling – Programmed Read-Through) (Dunn et al., 2013), they tend to have higher tAI scores for

the *eORF* compared with their ORF. After shuffling assignments, this relation seems to break (figure 5).

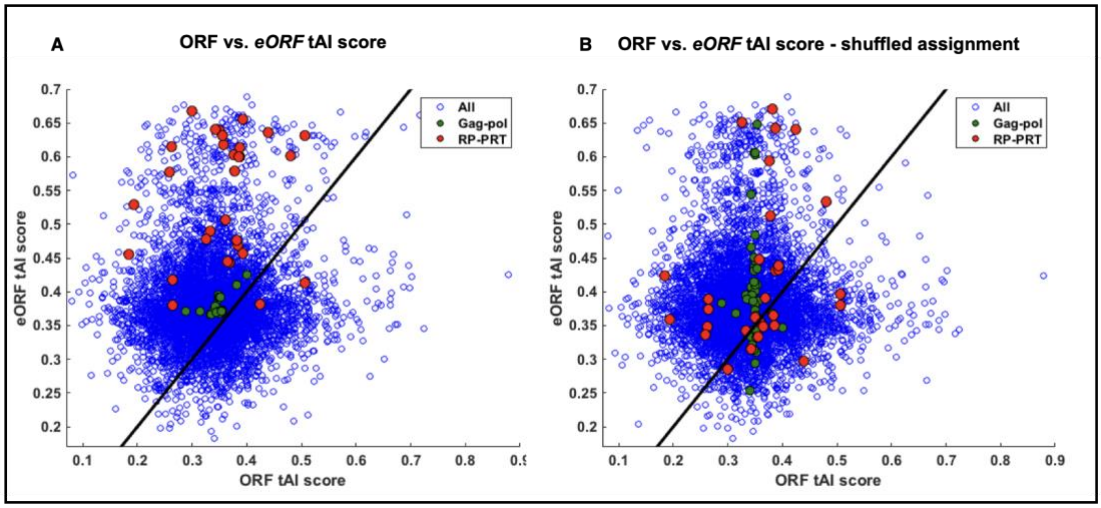


Figure 5: Translation adaptation index (*tAI*) of ORF vs. *eORF*. tAI is a measure of translation efficiency and each gene has a characteristic profile. **(A)** ORF tAI vs. *eORF* tAI. Black line marks the X=Y correlation. points below this line are points where gene tAI is higher in the *eORF*. Red dots are genes predicted to undergo STOP codon read-through from ribosome profiling data. Green dots show the gag-pol genes that have long translated 3' UTR. In general, no positive correlation was detected. **(B)** Same as **(A)** after *eORF* to ORF shuffled assignments. It is visible that the higher *eORF* tAI score that was present in the known read-through genes is broken.

3.2. Ribosome profiling

From data gathered in (Dunn et al., 2013), I extracted lists of genes that have been suggested to have STOP codon read-through due to ribosome presence (as shown in supplement table 3 – link). I refer to these genes as *Positive for Read-Through using Ribosome Profiling* (*RP-PRT*). I used these genes as a positive control set or training set to extract important sequence features. As it turns out, these genes are not found in our predictions since they do not have exceptionally long *eORF*. Following this discovery, I understood that having a long *eORF* may not be a strong enough indicator for STOP codon read-through.

I did not find higher enrichment of ribosomes in the *3'* UTRs of genes having long *eORF* and did not detect a reading frame in the 3' UTRs. It appears as though the data we had from previous ribosome profiling experiments was too sparse to reach any insight. The sparseness was even worse when trying to do the same process for the genes predicted from the human genome. Not only was I having trouble

getting signals at all but determining the suitable tissue and conditions from which the data should be taken was too complex since the predictions did not hold any features in common.

4. Sequence signals indications

4.1. Frameshift signals and leaky STOP codon signals

I had to find a definite cause for the read-through event to strengthen the computational evidence for translational STOP codon read-through. Should a frameshift occur in the vicinity of the STOP codon (upstream to it), a read-through of the STOP codon will happen due to translation in the wrong frame. Previous works (Clare et al., 1988; Farabaugh et al., 2006; Jonathan, 2012) showed that some RNA sequences cause the ribosome to make translation errors resulting in the translation of the RNA molecule in the wrong frame. These errors can occur due to repetitive signals or mRNA secondary structure, which cause the ribosome to slide, or mRNA secondary structure to make translation errors resulting in a ribosomal frameshift.

I found only repetitive signals complying with frameshifts that result in shifting to the 3rd translation frame (+2). Out of 22 reported signals for the 3rd frame, only five were found in abundance.

In many other cases, I noticed different signals at different locations of the RNA. Theoretically, if a frameshift occurs, it can cause an early STOP codon to appear, activating the *Nonsense-Mediated Decay* (*NMD*) mechanism. I filtered The results to show only genes with a frameshift signal in the last 10% of the RNA molecule (close to the STOP codon). This way, I could reduce the probability of the newly appearing STOP codon (due to the frameshift) being more than 50 nucleotides upstream to any exon-exon junction, thus lowering the likelihood of NMD activation.

4.2. Protein motif signals (transmembrane, NLS, binding motifs)

Work by (Dunn et al., 2013) used various servers to predict functional protein motifs in the 3’ UTR. I used the same servers to find enrichment of motifs in genes predicted to show STOP codon read-through outputted from my analysis.

I analyzed eORFs with at least 20 codons for trans-membrane helix domain prediction since this is the minimal length of a transmembrane loop (Bowie, 1997). I found an enrichment of 1180/3254 genes that present a high probability for trans membrane helices in the *eORF*. This is in comparison to only 151/918 genes in a control set where the ***eORFs'*** nucleotide sequence was shuffled. This difference is significant using chi-square hypothesis testing with $p = 10^{-5}$.

I compared the number of transmembrane helices predicted in the *eORF* with ones found in the ORF. For *S. cerevisiae*, and found that for 14% of the genome, a transmembrane helix was predicted in the *eORF*, while none were predicted in the ORF. Of those genes with long *eORF*, 34% were predicted to have a trans-membrane helix in the *eORF,* while none were predicted in the ORF. This finding might suggest a change in location or even function of the protein due to *eORF* translation.

To test whether these proteins are found to be bound to a membrane or change their binding properties under stress, I used the work done by (Breker et al., 2014), where they used fluorescent tagging to determine the cellular localization of all *S. cerevisiae* proteins, under normal conditions and several stress conditions. I found that among the predicted *eORFs* to have a trans-membrane helix, six genes seem to locate to a membrane or change their location under different conditions, even though they do not seem to have any transmembrane domains in their ORF. It seems as though there is an activation under DNA replication stress for some of them. This may indicate a change in function caused due to the STOP codon read-through. The resulting proteins are described here:

4.2.1. YBL029-C: Protein of unknown function

This gene's *eORF* was predicted to reside in a +1 frameshift relative to the ORF frame of translation. Using a prediction engine for transmembrane helices in proteins, *TMHMM* (Krogh et al., 2001), no evident trans-membrane helices were found in the ORF, although it was reported to reside in the cell periphery. It is likely to have a trans-membrane helix in the *eORF*, which may explain its

cellular localization. *BLASTp* search for the *eORF* translation has a full match in a different *S. cerevisiae* strain (*FosterO*). When examining the sequence for this protein in the *FosterO* strain, it appeared to be a version of *S. cerevisiae* (*S288*) YBL029C-A with a ribosomal +1 frameshift in its *ORF* causing a STOP codon read-through and translation of the *eORF* in turn.

4.2.2. YGL208C: Protein of unknown function

The *eORF* for this gene seemed to be in the same frame of translation as the ORF. Transmembrane domain prediction predicts a domain in the *eORF* with no domains in the ORF. N-terminus GFP labeling showed localization to the cell periphery. No evidence of the extended peptide to other yeast strains was found.

4.2.3. YPL066W: Regulator of Rho1p, cofactor of Tus1p

The *eORF* for this gene seemed to be in the same frame of translation as the ORF. Transmembrane domain prediction predicts a domain in the *eORF* with no domains in the ORF. It was reported to locate in the cytoplasm and bud neck, which may suggest a dual role in the cell. No evidence of the extended peptide to other yeast strains was found.

4.2.4. YPR174C: Protein of unknown function

The *eORF* for this gene seemed to be in the same frame of translation as the ORF. Transmembrane domain prediction predicts a domain in the *eORF* with no domains in the *ORF*. It was reported to locate in the nucleus periphery and to have relative distribution to foci at the nuclear periphery increase upon DNA replication stress. No evidence of the extended peptide to other yeast strains was found.

4.2.5. YHR182W: Protein of unknown function

This gene's *eORF* was predicted to reside in a -1 frameshift relative to the ORF frame of translation. The TMHMM prediction engine found no evident trans-membrane helices in the ORF. It was reported to locate in the cytoplasm and cell periphery and was also re-localized from the bud neck to the cytoplasm upon DNA replication stress, suggesting a dual role in the cell. No evidence of the extended peptide to other yeast strains was found.

4.2.6. YML053C: Putative protein of unknown function

This gene's *eORF* was predicted to reside in a -1 frameshift relative to the ORF frame of translation. Transmembrane domain prediction predicted a domain in the *eORF* with no domains in the ORF. It was reported to locate in the cytoplasm and nucleus, which may suggest a dual role in the cell. No evidence of the extended peptide to other yeast strains was found.

5. Orthologous proteins analysis

5.1. *Multiple Sequence Alignment* (*MSA*) analysis for gene type classification according to STOP codon locations in the alignment

Translational STOP codon read-through can occur due to an evolutionary event causing a sequence to develop a premature STOP codon as a genetic diversity mechanism. If we examine aligned sequences of homologous proteins from different species, we can often find a shift in the STOP codon's location. This may suggest that a specific sequence was added/extracted from the final ORF along evolution. While this sequence will typically be translated and expressed in one species, it might only be translated under specific conditions in another. If the sequences are conserved along evolution, they most likely have a role. Generally aligned sequences of homologous proteins in different species show a certain degree of conservation inside the ORF and a much lower degree outside of it. Should a sequence beyond the ORF, and in our case downstream to the STOP codon, show a conservation degree close to that of the ORF, this sequence was perhaps selected for translation.

I used the raw data from (Wapinski et al., 2007), which includes multiple sequence alignments for 5594 *S. cerevisiae* genes with 23 other yeast strains. Starting with dividing each of the 5594 genes into one of four groups (as shown in figure 6):

1) Genes that were conserved only in the ORF. No significant conservation outside the ORF
2) Genes where the STOP codon of other strains appeared earlier than that of *S. cerevisiae* while keeping a high conservation level downstream to that STOP codon.

3) Genes where the STOP codon of *S. cerevisiae* appeared earlier than that of another strain while keeping a high conservation level downstream to that STOP codon.
4) Genes that had aligned STOP codons in other strains, as in *S. Cerevisiae* but kept a high level of conservation downstream to these STOP codons.

Since I searched for genes that present STOP codon read-through in *S. cerevisiae* with evidence of translation in other strains, I focused my efforts on further analysis of the last group. This group keeps a high conservation profile downstream of the STOP codon (as annotated for *S. cerevisiae*). It is conceived as probable for translational STOP codon read-through in *S. Cerevisiae*.

I found seven genes that show strong evidence of translation of the *eORF* since an orthologous gene has the *eORF* as part of its ORF. I suggested that these are genes where STOP codon read-through is highly probable, and no frameshift is needed to generate it. These genes are described here:

5.1.1. YNR069C: Protein of unknown function

Two closely related species, *S. Bayanus* and *S. Paradoxus* had orthologous versions for this *S. cerevisiae* gene, but their ORFs were extended to include the *eORF*. The genes *ORF* showed genomic organization compatible with STOP codon read-through, which contains the neighboring ORF YNR068C. This protein's shortened and read-through versions interact differently (in vitro) with another protein: Rsp5p (Namy et al., 2003; Novoselova et al., 2012).

5.1.2. YHL034W: Protein of unknown function

One closely related species, *S. Paradoxus* had an orthologous version of this *S. cerevisiae* gene, but its ORF was extended to include the *eORF*.

5.1.3. YLR030W: Putative protein of unknown function

One closely related species, *S. Paradoxus* had an orthologous version of this *S. Cerevisiae* gene, but its ORF was extended to include the *eORF*. The read-through would go on to include the next residing ORF YLR031W.

5.1.4. YLL052C: Water channel that mediates water transfer between membranes

Two closely related species, *S. Bayanus* and *S. Paradoxus*, and one farther species, *C. Glabrata*, had orthologous versions for this *S. Cerevisiae* gene. Still, their ORFs are extended to include the *eORF*. Had a slight overlap with the next ORF YLL053C.

5.1.5. YIL164C: Nitrilase

Three closely related species, *S. Bayanus, S. Paradoxus,* and *S. Kluvery*, and one farther species, *K. Waltii*, had orthologous versions for this *S. Cerevisiae* gene. Still, their ORFs were extended to include the *eORF*. In other *S. cerevisiae* strain backgrounds, This gene and the adjacent ORF, YIL165C, likely constituted a single ORF encoding a nitrilase gene (Brenner et al., 2016; Godard et al., 2007; Kellis et al., 2003).

5.1.6. YML003W: Putative protein of unknown function

Two species*, K. Waltii* and *K. Lactis* had orthologous versions for this *S. Cerevisiae* gene, but their ORFs were extended to include the *eORF*. Stop codon read-through would have resulted in a fusion with the neighboring ORF YML002W.

5.1.7. YMR084W: Putative protein of unknown function

In many related strains, the two adjacent ORFs, YMR084W and YMR085W, were merged, and together they were paralogous to GFA1 (Glutamine fructose six phosphate amidotransferase) (Byrne and Wolfe, 2005; Kellis et al., 2003).

5.1.8. YJL107C: Putative protein of unknown function

Three closely related species, *S. Pombe, S. Japonica,* and *S. Octosporus*, had orthologous versions for this *S. Cerevisiae* gene, but their ORFs were extended to include the *eORF*. In these species, the two slightly overlapping neighboring ORFs, YJL107C and YJL108C, were merged (Brachat et al., 2003; Harris et al., 2001; Sychrova et al., 2000; Yang et al., 2013).

After understanding the sequence features that could indicate and convince the possibility of STOP codon readthrough, I felt that the model is mature enough to explore the human genome, searching for translation

beyond the canonical STOP codon or other translation anomalies that will be described in the following chapter.

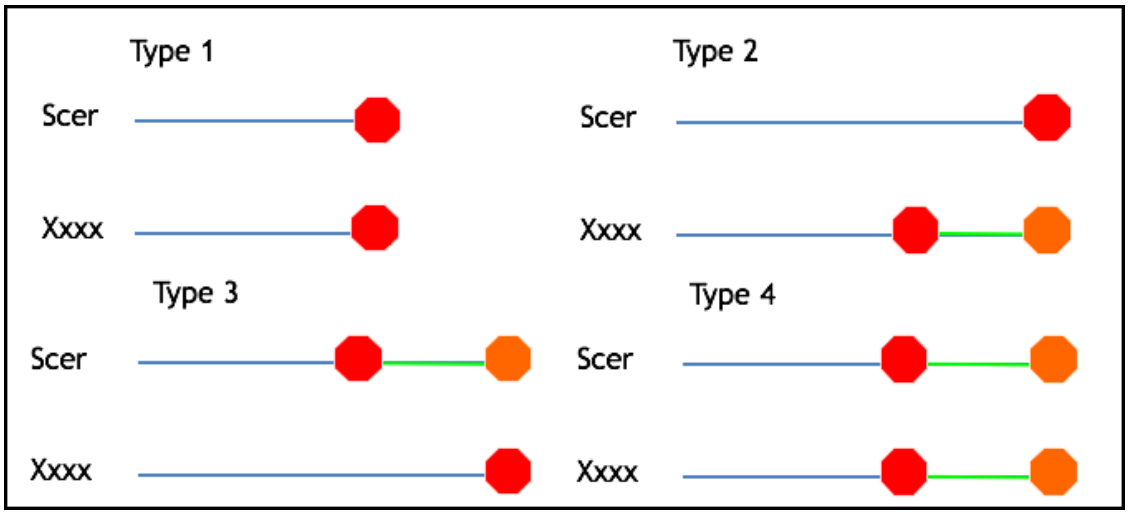


Figure 6. Dividing Multiple Sequence Alignments into Different Groups Based on the Relative STOP codon Location. Each gene was categorized into one of four groups (types) representing its alignment across evolution. Blue lines represent the CDS. Green lines represent the *eORF*. The canonical STOP codon is shown in red. STOP+1 is orange. I compared each yeast strain's sequence to that of *S. cerevisiae* to conclude which gene showed prominent evidence of translation beyond the STOP codon, and to which of the groups it fitted most.

Methods for detection of STOP codon read through in *S. cerevisiae*

1. 3' UTR sequence analysis

S. cerevisiae genome assembly version SACCEL3 was used in this work for both chromosome sequences and gene annotations. Only ORFs were analyzed (including putative and dubious ORFs). The mitochondrial chromosome was not part of the analysis. For each gene, the UTR was explored, ranging beyond the annotated end of the UTR. The codons were recorded until the closest STOP codon downstream for each of the three reading frames (+0, +1, +2) past the canonical STOP codon. The number of codons between the canonical STOP codon and the subsequent downstream stop codon in all three frames was counted, and the longest "stretch" was termed as *extended Open Reading Frame* (*eORF*) and recorded.

2. tAI calculation

Since tRNA copy number in the genome has a high positive correlation with the tRNA abundance in the cell, one can assume that it correlates with translation efficiency. With this information in hand, it is possible to rank a gene relative to its adaptation to the tRNA pool in the cell. An adaptiveness value (*tAI*) was calculated for each codon (dos Reis et al., 2004), which considered the tRNA copy number and the number of isoacceptors that recognize the tRNA and the anti-codon coupling strength due to wobble rules. Each organism has its own set of tAI values due to changes in the criteria mentioned above. The tRNA copy numbers were downloaded from the Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/) (Lowe and Eddy, 1996). Local tAI relates to the value calculated for each codon along the mRNA sequence and is also called the absolute adaptiveness index. The formula depicting it is:

(1) $tAI_{local,i} = \sum_{j=1}^{n_j}(1 - s_{ij})tGCN_{ij}$

where n_i is the number of tRNA isoacceptors that recognize the i-th codon, $tGCN_{ij}$ is the gene copy number of the j-th tRNA that recognizes the i-th codon, and s_{ij} is a selective constraint on the efficiency of the codon-anticodon coupling. This constraint was calculated by performing hill-climbing optimization of the

Spearman correlation between protein abundance and translation efficiency in *S. cerevisiae* (Tuller et al., 2010).

I then normalized each codon score by the maximal score for the maximal value of all local tAI values.

The tAI value of a gene is defined as the geometric average of all the tAI scores of the codons in the gene's sequence.

In general, tAI takes a geometric average form, but this can significantly reduce the value as the length of a sequence grows. For this reason, the averaged value was calculated using the codon frequency of the sequence using formula 2:

(2) $tAI_{AVG} = \prod_{1}^{61} tAI_{local}^{CF}$

I multiplied all local tAI values after raising them to the power of their codon frequency (CF) in the sequence.

3. <u>Orthologous proteins sequence analysis</u>

All ortho-groups from (Wapinski et al., 2007) presenting more than two species were analyzed for 3-way periodicity (3rd un-conserved nucleotide). UTRs for more species were extracted and aligned inside documented ortho-groups using data sources from their work. I examine the 3' UTRs' conceptual translation in the original canonical frame (0), the +1 frame, and the -1 frame. The number of codons until reaching a STOP codon was counted for each of these. The longest sequence was then termed the *eORF,* and the number of nucleotides added was measured and recorded for comparison. Once the sequences were re-aligned, 3-way nucleotide periodicity was tested. The genes were also divided into 1 of the four alignment groups based on the relative location of the STOP codon in each pair of organisms, as shown in figure 6.

Summary of Chapter 1:

In this chapter, I explored the *S. cerevisiae UTRome* to find exceptional sequence properties that may reveal non-canonical translation in the form of STOP codon read-through. I noticed that when looking at the sequence downstream to the canonical STOP codon, for some genes, there would be a very long sequence until another STOP codon is encountered. Should a STOP codon read-through occur, and translation elongates beyond the canonical STOP, a new large peptide could be appended to the protein, potentially affecting its functionality. I sought to devise rules defining the sequence properties typical to a sequence with translation potential and further aimed to find the biological implication of such events. I discovered that simply having a long sequence lacking STOP codons is not enough to create a criterion for potential translation. More sequence features should also be considered to provide a high indication for STOP codon readthrough. When looking at the composition and signals of *eORF* sequences, I could find some evidence of possible functional units.

When comparing predicted *eORF*s with a group of species in varying levels of evolutionary distance from *S. cerevisiae*, I could find evidence of some *eORFs* that are coded as part of the originating ORF or translated as fusion proteins in other species. This may indicate an evolutionary process developed to enrich translation options without changing the genome.

Due to a lack of sufficiently high density and quality data, I was not able to show experimental evidence of translation either from Ribosomal profiling data that was available to me or other sources to complete the full set of properties. However, from the evidence I was able to gather, I believe these approaches can help in providing systematic analysis and raising probable candidates for further exploration in experimental research.

Appendix A – Long eORF predictions in S. cerevisiae

Gene name	eORF length	eORF frame
'YAL037C-B'	170	1
'YAL031W-A'	105	1
'YAL019W-A'	1164	1
'YAR010C'	173	2
'YAR066W'	427	1
'YAR073W'	120	2
'YEL076C'	125	3
'YEL076C'	125	3
'YEL045C'	402	3
'YEL018C-A'	318	1
'YER039C-A'	268	1
'YER097W'	752	3
'YER137C-A'	1314	2
'YER159C-A'	1314	2
'YER189W'	1666	2
'YJL169W'	151	1
'YJL156C'	101	1
'YJL150W'	647	2
'YJL135W'	406	3
'YJL114W'	186	2
'YJL107C'	381	3
'YJL097W'	242	3
'YJL086C'	824	2
'YJL075C'	133	1
'YJL028W'	137	1
'YJR026W'	1314	2
'YJR028W'	1314	2
'YJR086W'	115	3
'YJR146W'	343	3
'YBL112C'	766	2
'YBL111C'	107	2
'YBL100W-A'	1331	2
'YBL077W'	107	1
'YBL039C-A'	318	3
'YBL008W-A'	841	2
'YBL005W-A'	1314	2
'YBR012W-A'	1315	2
'YBR027C'	381	2
'YBR121C-A'	319	3
'YBR156C'	104	3
'YBR159W'	110	2

'YBR201C-A'	129	2
'YBR224W'	895	2
'YBR227C'	138	1
'YDL195W'	113	3
'YDL162C'	789	3
'YDL050C'	125	1
'YDL037C'	871	1
'YDL022C-A'	451	2
'YDR034C-C'	1331	2
'YDR073W'	111	2
'YDR081C'	101	2
'YDR082W'	441	1
'YDR094W'	104	2
'YDR098C-A'	1314	2
'YDR156W'	130	2
'YDR157W'	373	1
'YDR170W-A'	389	2
'YDR210W-A'	1331	2
'YDR210C-C'	1314	2
'YDR261W-A'	1331	2
'YDR261C-C'	1163	2
'YDR304C'	108	1
'YDR316W-A'	1314	2
'YDR320C-A'	727	2
'YDR344C'	121	3
'YDR365W-A'	1314	2
'YDR505C'	127	1
'YDR509W'	115	1
'YIL164C'	122	1
'YIL141W'	797	2
'YIL086C'	168	1
'YIL085C'	111	2
'YIL082W'	1207	2
'YIL064W'	126	1
'YIL063C'	148	1
'YIL042C'	344	2
'YIL028W'	140	3
'YIL024C'	132	3
'YIL020C'	106	2
'YIR021W-A'	169	1
'YIR023C-A'	155	1
'YIR036W-A'	193	2
'YFL066C'	478	2

'YFL056C'	161	3
'YFL032W'	225	2
'YFL002W-B'	1331	2
'YFR009W-A'	203	3
'YFR012W'	270	2
'YFR054C'	124	3
'YKL023C-A'	205	3
'YKR103W'	339	1
'YKL145W-A'	187	3
'YKL115C'	515	2
'YKL084W'	224	2
'YKL031W'	205	1
'YKL020C'	495	1
'YKR032W'	291	1
'YKR077W'	104	3
'YOL163W'	219	1
'YOL159C-A'	152	2
'YOL103W-A'	1314	2
'YOL028C'	271	3
'YOR024W'	463	1
'YOR030W'	123	1
'YOR051C'	191	1
'YOR142W-A'	1314	2
'YOR158W'	163	1
'YOR192C-A'	1331	2
'YOR199W'	100	3
'YOR202W'	102	2
'YOR203W'	601	2
'YOR302W'	429	3
'YOR332W'	108	1
'YOR343W-A'	1331	2
'YCL076W'	142	3
'YCL067C'	293	2
'YCL066W'	102	2
'YCL057C-A'	168	3
'YCL041C'	492	2
'YCL042W'	464	3
'YCL033C'	103	2
'YCL022C'	189	1
'YCL020W'	1331	2
'YCR039C'	640	3
'YCR045W-A'	100	3
'YCR086W'	174	3

'YCR100C'	145	2
'YGL241W'	301	1
'YGL164C'	192	2
'YGL052W'	215	2
'YGR027W-A'	1314	2
'YGR038C-A'	1314	2
'YGR046W'	116	1
'YGR051C'	103	3
'YGR096W'	137	1
'YGR109W-A'	1256	2
'YGR161W-A'	1331	2
'YGR161C-C'	1314	2
'YGR163W'	126	3
'YGR226C'	129	2
'YLL052C'	135	2
'YLR030W'	233	3
'YLR102C'	133	3
'YLR140W'	323	3
'YLR157C-A'	1314	2
'YLR202C'	226	3
'YLR210W'	102	2
'YLR227W-A'	1314	2
'YLR256W-A'	314	2
'YLR313C'	130	3
'YLR349W'	100	1
'YLR365W'	103	1
'YLR368W'	728	3
'YLR393W'	492	2
'YLR410W-A'	1331	2
'YLR418C'	102	2
'YLR434C'	558	2
'YLR464W'	1243	3
'YLR465C'	259	1
'YNL305C'	102	1
'YNL304W'	101	2
'YNL284C-A'	1314	2
'YNL269W'	106	1
'YNL265C'	105	1
'YNL235C'	140	1
'YNL205C'	454	3
'YNL198C'	528	2
'YNL184C'	120	2
'YNL179C'	336	2

'YNL091W'	104	1
'YNL054W-A'	1308	2
'YNR014W'	177	1
'YNR066C'	1124	1
'YNR069C'	351	1
'YPL277C'	116	2
'YPL257W-A'	1314	2
'YPL149W'	143	3
'YPL076W'	135	1
'YPR039W'	341	3
'YPR137C-A'	1314	2
'YPR158W-A'	1315	2
'YPR158C-C'	1314	2
'YPR160W-A'	665	3
'YHL043W'	234	2
'YHL009W-A'	1388	2
'YHL005C'	160	3
'YHR058C'	256	1
'YHR073W-A'	841	3
'YHR213W-A'	104	3
'YHR214W'	427	1
'YHR214C-C'	1314	2
'YHR217C'	107	1
'YHR218W'	107	2
'YHR218W-A'	598	2
'YML132W'	111	3
'YML101C'	107	3
'YML045W-A'	1314	2
'YML040W'	1314	2
'YML009C-A'	122	3
'YML003W'	798	3
'YMR013C-A'	429	3
'YMR046C'	1314	2
'YMR051C'	1314	2
'YMR057C'	347	1
'YMR076C'	168	2
'YMR084W'	456	3
'YMR135W-A'	123	3
'YMR154C'	140	3
'YMR273C'	128	1
'YMR307C-A'	124	3

Chapter 2: An algorithm for prediction of potential frame-shifting during translation applied to protein-coding genes in the human genome

Overview of Chapter 2:

Protein translation, like other processes in the cell, can be altered under different conditions. Programmed *Ribosomal Frame-Shifting* (*PRF*), stop codon read-through, and translation of non-coding RNA has been empirically shown to occur under physiological conditions and be tightly regulated (McGillivray et al., 2018; Molina-García and Giraldo, 2017; Wills et al., 2006). I have developed a computational method for predicting translation potential and divergence from the canonical open read frame, using only DNA sequences. This method is based on DNA multiple sequence alignments of orthologous genes, taking into account the conservation patterns of protein-coding genes. These patterns are identified by a less conserved 3rd sub-codon position in the coding region of protein-coding sequences. Using *Fourier transformation,* I can quantify the potential for translation and recover hidden translation opportunities, either by STOP codon read-through or ribosomal frameshifting. Analysis of the human genome and comparison to 19 other mammalian genomes revealed exciting patterns of translation potential in 400 genes (I suspected that these genes have more than one active reading frame). I found that many of the predicted frameshifts reside close to the start or stop codons, suggesting the presence of upstream *Open reading frames* (*uORFs*) or stop codon read through. Using these predictions, I looked for new motifs that could regulate frameshifts and RNA secondary structures that could mediate them and the effects these non-canonical translated proteins may have.

Scientific background for non-canonical protein translation detection in the human genome

The classical notion of translation of an mRNA molecule is that the ribosome starts translation upon detecting a start motif (Kozak, IRES) (Kozak, 1987; Pelletier and Sonenberg, 1988), moves three nucleotides downstream, and attaches tRNAs based on the codons it reads. This process will continue until the first STOP codon is identified when the ribosome detaches, releasing the newly translated protein. Cases where (1) translation is initiated up-stream of the annotated AUG start site (McGillivray et al., 2018), or (2) when translation fails to terminate at STOP codons (Jungreis et al., 2011; Loughran et al., 2018; Molina-García and Giraldo, 2017) are well documented and heavily studied. Such events can happen due to changing conditions in the cell, affecting the ribosome and creating different proteins that may have different functions. In some cases, these processes are programmed and hypothesized to help the cell deal with changing environments or stress (Molina-García and Giraldo, 2017). Another level of complexity arises when the ribosome performs a programmed frameshifting (Dinman, 2006; Jonathan, 2012; Ketteler, 2012). In such cases, the ribosome does not move three nucleotides downstream to where it was, and the "normal" translation dynamics are interrupted. When a ribosomal frameshift happens due to an error, the probability of a newly functioning protein being translated is low, and the product will most likely be degraded (Atkins et al., 2016; Li and Zhang, 2015). In some cases, these alterations are regulated to allow the translation of new proteins from the same mRNA molecule (without splicing).

For example, *Ornithine decarboxylase Antizyme 1 (OAZ1)* is an enzyme that catalyzes the rate-limiting step in polyamine biosynthesis. It regulates the synthesis of polyamines by binding and inhibiting ornithine decarboxylase. The expression of antizyme is regulated by polyamine-enhanced frameshifting, creating a +1 ribosomal frameshift and activating the enzyme (Jonathan, 2012)

When looking at *protein-coding sequences* (*CDS*) across evolution, we see that a certain level of plasticity is allowed. Due to the redundancy of the genetic code, I observed variation in specific positions across CDSs, which maintain the amino acid sequence, and, thus, the functional translated protein. The variation will appear on average on the third sub codon position along the

gene, giving rise to a periodic signal holding a frequency of 3 (*3-way periodicity*). Examining this signal can help with the annotation of coding regions and allow for the detection of translation errors due to divergence from the expected pattern.

The non-canonical translation is a complex and exciting process that is yet to be systematically and thoroughly explored. In some cases, a non-canonical translation could have immense effects on a functioning protein that may impact entire cells, tissues, and much more.

Research goal for non-canonical protein translation detection in the human genome

I examined periodic variability signals of nucleotide sequences from protein-coding genes to find those that show divergence from the known canonical forms of translation. For example, I looked at the conservation patterns of untranslated regions (UTRs) (which shouldn't be translated, thus should not have a strong 3-way periodicity signal) to find an area outside of the CDS that may also be translated.

I also focused on detecting deviations from the periodic signal demonstrated within the CDS. These deviations are the ones that would maintain the frequency of this periodic signal but will change its phase. Changing the phase of this signal is equivalent to changing the sub codon position that is most variable, thus implying a change of translational frame. I could locate patterns consistent with translational ambiguities, specifically with frameshifts, laying down a set of filtration steps on the periodicity signal. For genes that present convincing evidence of another layer of translation, I sought to understand its implication in affecting protein levels and function and find the related environment to make this regulatory layer appear.

Results of non-canonical protein translation detection in the human genome

1. STOP codon read-through evidence in the human genome

Like in the case of the *S. Cerevisiae* genome, I also wanted to find evidence of translation beyond the canonical STOP codon for the human genome. I started by examining the same signals tested for the *S. Cerevisiae* genome: *eORF* length, composition (amino-acids and codons), known motifs, secondary structures, and orthology to other species (Fig 7). I found that, like for the *S. cerevisiae* case, the distribution of the eORFs lengths follows the theoretical geometric and exponential distributions (see figure 7B), so having a long *eORF* could happen at random and doesn't seem to be a unique property to hint upon translation (Pearson's correlation p-value << 0.05).

With that, I found that 35% (491 genes) of the long *eORF* set are highly conserved in the translation of the *eORF* with a set of species ranging from primates to rodents when compared to the average conversation observed in all 3' UTRs. Of these, I found enrichment for DNA-binding proteins, specifically homeobox-containing proteins (p-value = 3.3e-6 as calculated by mHG model (Eden et al., 2009)).

Previous work by (Jungreis et al., 2011) found evidence of abundant STOP codon read-through in *Drosophila Melangostar*. They found that genes undergoing STOP codon read-through show high conservation of the identity of the STOP codon across 12 different strains tested. Having done the same analysis for genes presenting long *eORF* and conservation of 3' UTR translation, I found that this conservation rate is the same for the entire genome, suggesting that many genes have a long *eORF* but do not undergo STOP codon readthrough.

Out of the genes having long conserved *eORF*, I found 80 genes associated with mutations causing the STOP codon to be altered. These genes do not hold any particular traits or form a distinct group (GO annotations, functional annotations, and such categories).

In summary, I concluded that although long *eORFs* do not seem to correlate with a higher probability of being translated to a functioning unit,

some genes that have long *eORF*s have some experimental evidence of possibly undergoing STOP codon readthrough. These cases are fascinating. The translation of such a long addition can affect the protein's function by changing its secondary structure, exposing localization signals, or adding additional functional units. While it would seem that these genes have no functional features in common on their own, they might change their destination after STOP codon readthrough, creating a new distribution of functional categories in the cell under different conditions. Not having a basic common ground may suggest surgical evolutionary processes aiding or disrupting many other processes or stress conditions in the cell and should be further investigated experimentally.

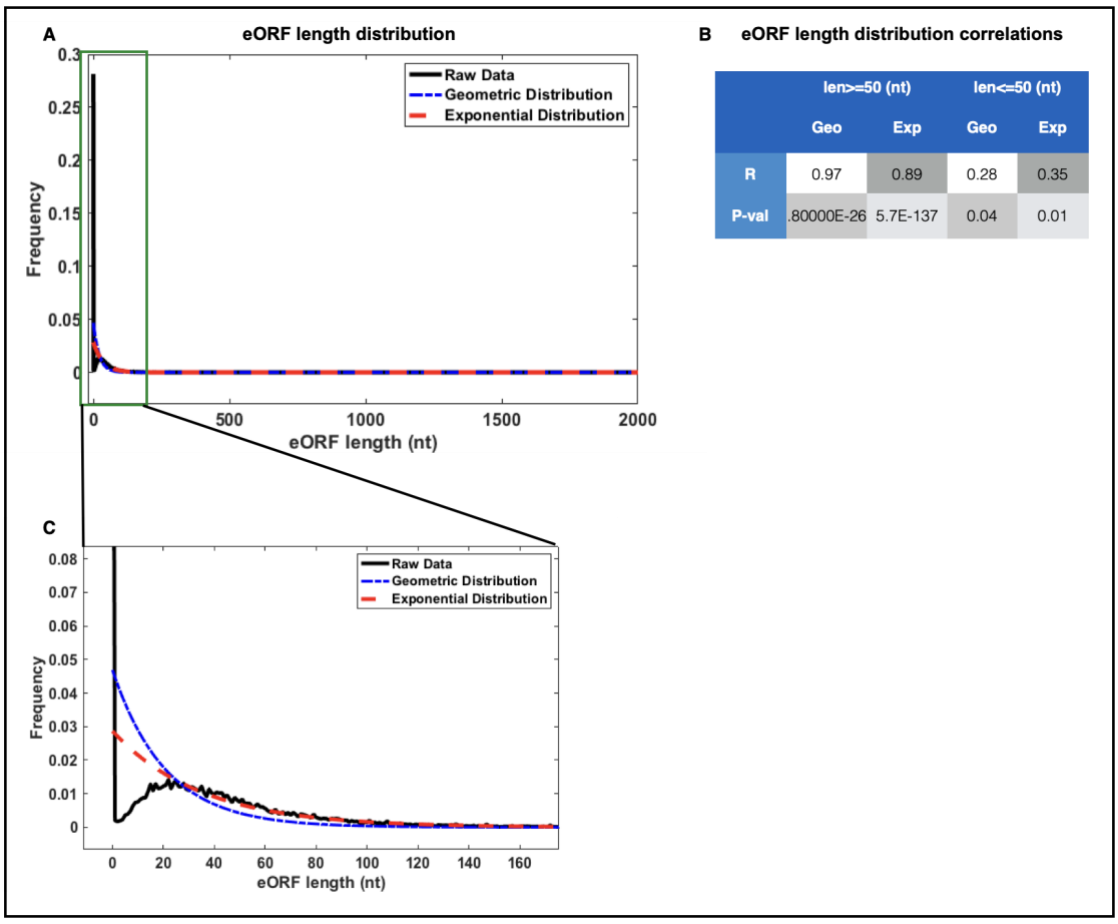


Figure 7: *eORF* length distribution. (A) *eORF* length distribution. Comparing the *eORF* sequence to a randomly generated sequence, I calculated the theoretical length distribution using geometric distribution and its continuous version, the exponential distribution. I saw that the decay acts as predicted by these models, but the real *eORF* data has a peak at around 30 nts (**C**). Green frame shows the deviation from the theoretical distributions of lengths for short *eORFs* (<=50 nt). **(B)** The Pearson correlation coefficients and their p-values for short and long *eORF* lengths. (**C**) Zoom in of the decaying phase to show the differences between the raw data and the theoretical distributions. It can be seen that the raw data did not simply decay, but rather presents a peak at around 30 nts.

1.1 Sequence properties for the predicted *eORF*

Like I did for *S. cerevisiae*, I wanted to see if I could find genes where there are extraordinary sequence properties, suggesting that should a STOP codon readthrough occur, translation is significantly more likely to take place compared to a random sequence. I used the same measure of aa and codon composition correlation between the ORF and the eORF. Also, I wanted to test whether there is some relationship between the sequence correlation scores and the eORFs' length.

There was a positive correlation between aa composition correlation scores and eORF length, similar to the one found for *S. cerevisiae* genes. Unlike the case for *S. cerevisiae*, hypothesis testing by t-test for the distribution of correlation scores showed that these scores were unlikely to appear at random compared to a shuffled sequences control set (p-value << 0.05). Upon further investigation, this does not apply when removing short sequences from the analysis, suggesting once more that aa composition correlation is not a strong enough measure (Fig 8).

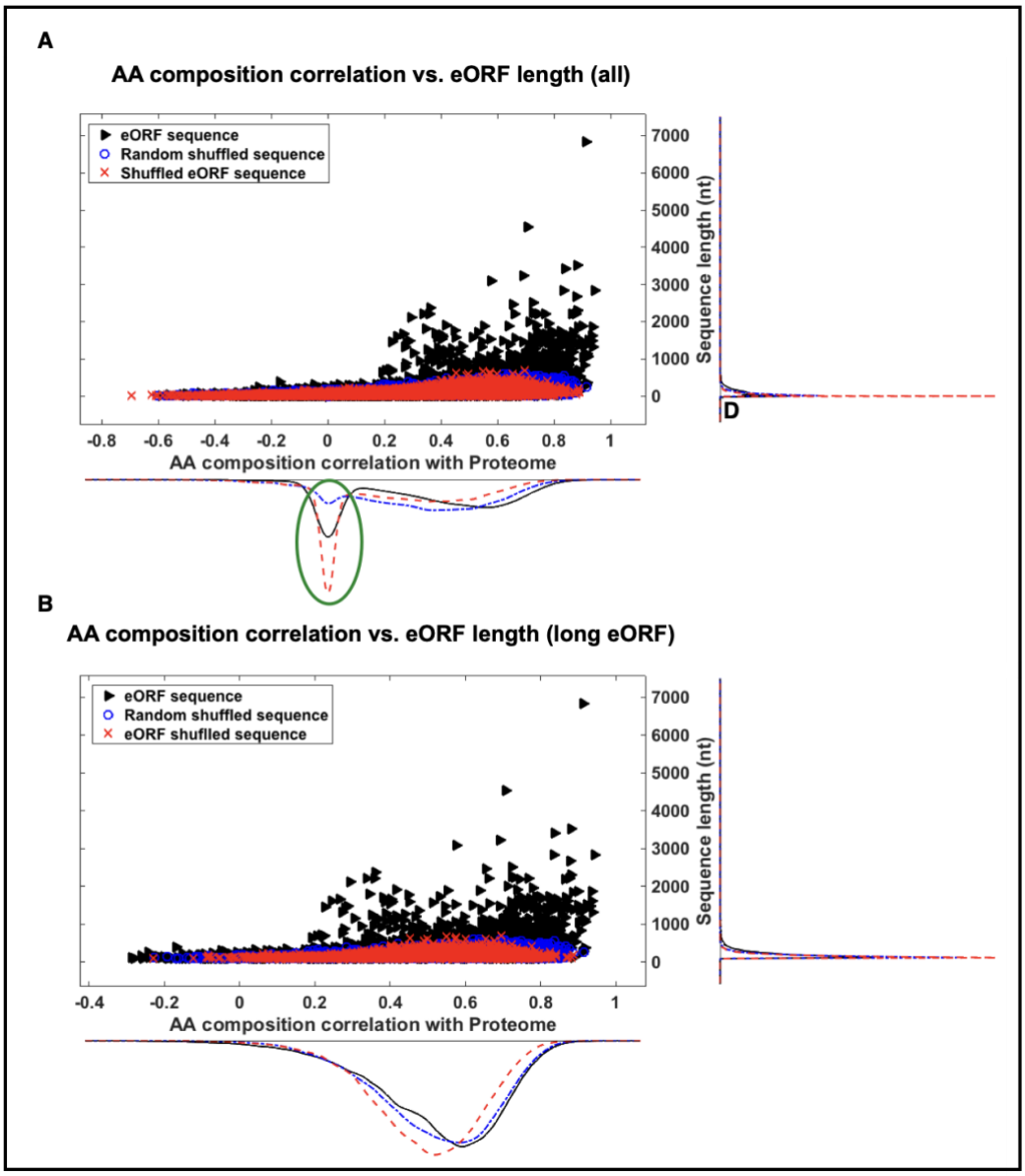


Figure 8: *eORF* sequence properties – AA composition correlation relative to *eORF* length. *eORF* AA composition correlation scores vs. *eORF* length (black). Two control sets were generated; the first (blue) is a random sequence with 50% GC content. the second (red) was generated by shuffling nucleotides in the *eORF* (maintaining the GC content). **(A)** represents all *eORFs*. **(B)** are only *eORFs* longer than 100 nucleotides. Longer *eORF* did not show higher positive correlations to the *eORF;* however, the non-correlative genes seemed not to be a part of this group (disappearing peaks marked in green circles). For aa composition correlation especially, it seemed that there was a subpopulation with a higher positive correlation between *eORF* and ORF which remained dominant when only considering genes with long *eORFs*.

For codon composition correlation scores between the ORF and the eORF relative to the length of the eORF, I saw an overall lower correlation (as expected); however, much like the *S. cerevisiae* case, the deviation of the scores from that of a random sequence seems greater and is also significant for both control groups under t-test (Fig. 9, p-val<<0.05).

That being said, there appeared to be a group that has a significantly higher correlation score (relative to random and shuffled sequences) exceeding the 95% percentile. These represent long enough eORF to be potentially effective in changing a protein's function should there be a STOP codon readthrough.

Using the reported read-through candidates in *D. Melanogaster* from (Jungreis et al., 2011) and seeing if their human orthologs extraordinary sequence properties, such as high composition correlations, did not reveal much. Although it seems there may be promising candidates in their work, most of them have very low sequence similarity with their human orthologs, and they have relatively short *eORF*s (< 100 nts).

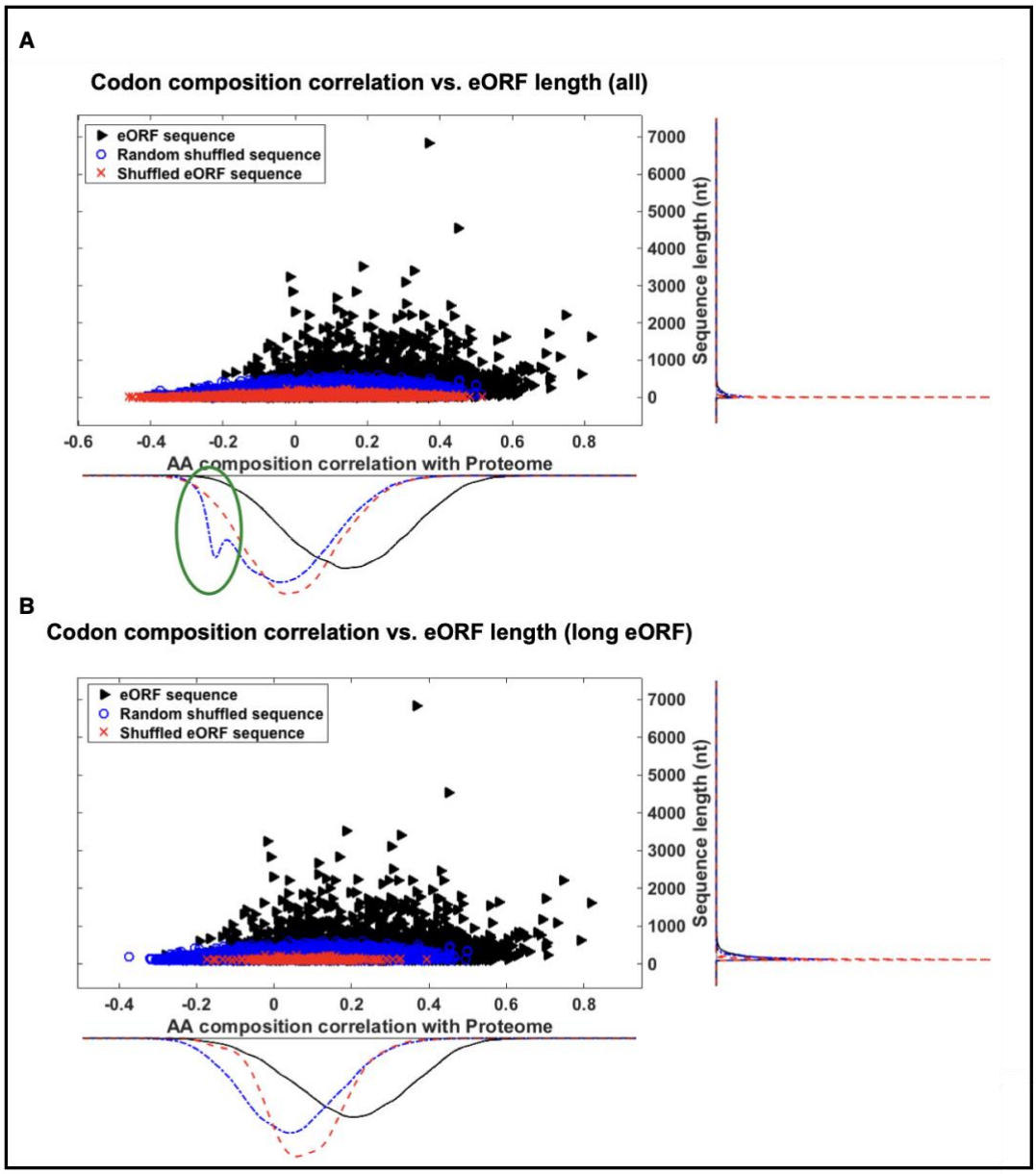


Figure 9: eORF sequence properties – codon composition correlation relative to *eORF* length. *eORF* codon composition correlation scores relative to *eORF* length (black). Two control sets were generated, the first (blue) is a random sequence with 50% GC content. the second (red) was generated by shuffling nucleotides in the *eORF* itself (maintaining the GC content). **(A)** represents all *eORFs*. **(B)** are only *eORFs* longer than 100 nts. longer *eORF* did not show higher positive correlations to the *eORF*, however, the non-correlative genes seemed to not be a part of this group (disappearing peaks marked in green circles).

1.2 SMS – Stop Means Stop

My next step was to find genes that had a sequence of multiple consecutive STOP codons. I suspected that should this sequence of STOPs be conserved in evolution; it may suggest a defense mechanism against STOP codon read-through. I found that out of ~1200 genes having

more than two consecutive STOP codons, ~850 genes conserve this property in evolution reaching as far as a mouse. Of 51 genes having more than one consecutive STOP codon, 16 conserved this property across evolution. This alone does not show any significance since it already has been shown that the first few nucleotides in the 3' UTR are highly conserved within mammals (first codons of the 3' UTRs hold 70% conservation) (Xie et al., 2010).

In addition, I looked at the resulting translation of the 3' UTR should the sequence of STOPs be read through. Although the probability of having a STOP codon read-through event is low, more so is the probability of reading through a sequence of STOP codons. Thus UTR translation would become more likely should a frameshift occur or a release factor become non-functional. I examined the 3' UTR for translatability in terms of aa composition, codon usage, known protein domains, and protein conservation. I could not find any enrichment specific to this group of genes.

Finally, I attempted to find evidence of the association between these genes and diseases. Specifically, I tested if there are documented genetic variations (insertions/deletions (indels)/*Single Nucleotide Polymorphism (SNPs)*) that may cause the stop codons to be read through. I found that ~200 genes hold a mutation that causes the canonical STOP codon to be lost and are associated with diseases. Since the annotation is for the canonical STOP codon alone, it could be that the second STOP codon could still provide termination unless the mutation creates a frameshift, then most likely, all STOP codons will be lost altogether. I found 15 genes with a frame-shift mutation right on the stop codon that are also associated with diseases. This does not mean that this mutation is the cause of the disorder, but further investigation may reveal new phenomena. The description of these results is summarized in table 2.

Gene ID	Symbol	Associated syndrome	Mutations	Mutation ID	Mutation Location	Mutation Description
ENSG00000063515	GSC2	velocardiofacial_syndrome	C/	rs781955300	19,148,992(-)	frameshift, stop_lost, terminator_codon_variant
ENSG00000064601	CTSA	deafness_autosomal_dominant_69	C/CATTTCTTTTTC	rs772734023	45,898,447(+)	coding_sequence_variant, frameshift, non_coding_transcript_variant, stop_lost'
ENSG00000079385	CEACAM1	colorectal_cancer	TT/TTT	rs775213687	42,510,905(-)	coding_sequence_variant, frameshift, genic_downstream_transcript_variant, stop_lost
ENSG00000092330	TINF2	oral_leukoplakia	CACTCACTC/CACTC	rs767558801	24,240,412(-)	3_prime_UTR_variant, coding_sequence_variant, frameshift, intron_variant, splice_donor_variant, stop_lost, terminator_codon_variant
ENSG00000104044	OCA2	angelman_syndrome_due_to_maternal_15q11q13_deletion	AGTTTCCTTTAGTCTTCGAGCA/A	rs766291945	27,755,351(-)	3_prime_UTR_variant, genic_downstream_transcript_variant, inframe_indel, intron_variant, stop_lost, terminator_codon_variant
ENSG00000106153	CHCHD2	parkinson_disease_late_onset	A/AA	rs35957514	56,101,806(-)	3_prime_UTR_variant, frameshift, stop_lost, terminator_codon_variant
ENSG00000112139	MDGA1	schizophrenia	CACACA/CACA	rs1049022822	37,638,060(-)	downstream_transcript_variant, frameshift, genic_downstream_transcript_variant, intron_variant, stop_lost, terminator_codon_variant
ENSG00000133742	CA1	carotid_artery_occlusion	T/	rs771203200	85,328,560(-)	frameshift, stop_lost, terminator_codon_variant
ENSG00000134115	CNTN6	peripheral_nervous_system_neoplasm	TT/TTT	rs563966864	1,358,599(+)	frameshift, intron_variant, stop_lost, terminator_codon_variant
ENSG00000135443	KRT85	hair_disease	GGCTCCATGACTCTACTAGGC/GGC	rs755977926	52,360,838(-)	3_prime_UTR_variant, inframe_deletion, stop_lost, terminator_codon_variant
ENSG00000143469	SYT14	spinocerebellar_ataxia_autosomal_recessive_11	AT/	rs756834000	210,161,042(+)	'3_prime_UTR_variant,' frameshift, non_coding_transcript_variant, stop_lost, terminator_codon_variant
ENSG00000154118	JPH3	leukodystrophy_hypomyelinating_2	TGT/	rs757987606	87,604,328(+)	3_prime_UTR_variant, coding_sequence_variant, inframe_deletion, inframe_indel, intron_variant, stop_lost, terminator_codon_variant
ENSG00000154485	MMP21	dextrocardia_with_situs_inversus	A/AAA	rs769800658	125,766,664(-)	frameshift, stop_lost, terminator_codon_variant
ENSG00000159197	KCNE2	cardiac_arrhythmia	CCCCC/CCCCCC	rs756561888	34,370,843(+)	coding_sequence_variant, frameshift, stop_lost
ENSG00000168124	OR1F1	hirschsprung_disease_1	CTGTC/CTGTCCTGTC	rs771267012	3,205,178(+)	coding_sequence_variant, frameshift, stop_lost

Table 2: List of frame-shift mutation on the canonical STOP codon that can generate multiple STOP codon read-through for a sequence of consecutive STOP codons and are associated with diseases.

2. Developing a novel algorithm for predicting non-canonical translation using conservation gene profiles

Non-canonical gene translation can be interpreted as many different deviations from the known canonical dogma of translation. In the previous section, I discussed the possibility of STOP codon readthrough (either programmed or due to mutations) and the possible effects of events such as these. A more elusive form of non-canonical translation is ribosomal frameshift. This process can have massive effects since it can completely alter a protein's sequence, thus generating a completely new functioning protein while eliminating the original. Predicting frameshift events is extremely hard using only the coding sequence. Programmed ribosomal frameshifts are rarely discovered in the human genome, primarily by experimental evidence.

For this reason, I developed a computational tool that could aid in producing lists of genes that possess properties and hints of translation in frames other than the annotated canonical one. I utilized the fact that the codon wobble position allows a certain degree of plasticity when examining protein-coding genes across evolution. Since different species have different codon usage patterns, highly conserved proteins show less conservation when exploring their coding sequence. This lack of conservation is not distributed randomly along the sequence but instead mainly appears in the third codon position, generating a periodic conservation signal when looking and CDS multiple sequence alignments.

I designed two computational models to achieve this task. Since I did not have enough positive samples of genes presenting this behavior (two known human genes undergo programmed ribosomal frameshift, and a handful more in viruses and such), training a model using positive samples was not an option. I built one rule-based model for translational frame determination using theoretical assumptions made by the design of the features (as detailed in the methods section) and another model trained on synthetic data produced by me. The frameshift site prediction model whose results are presented in this thesis, was rule-based and was designed by the theoretical assumptions captured in the model's features.

Both models were evaluated on the simulated data I created for assumption validations, alongside another synthetic database created from different sources than that of the original one, which holds the alignments' evolutionary properties later used for novel frameshift gene candidates.

The full details of the computational models are described in the methods section. I applied the models to the entire human transcriptome, and the results are reported henceforth.

3. Simulated MSAs

To best understand the algorithm's behavior under different conditions (different frameshifts, mixtures of frames, and gene conservation levels), I simulated hundreds of sequences deviating from the human collagen coding gene (COL1A – NM_000088) under different evolutionary constraints. I used a range of evolutionary distances between the sequences (from 0.5% up to 50%) and a range for the number of sequences (5-19), then mutated nucleotide positions in a manner where the amino acid change follows substitution matrices (BLOSUM). I only allowed mutations where the substitution score is non-negative.

The simulation was designed to include no evident frameshifts, a single frameshift (-1 or +1), and a region of frame mixture (0/+1). From simulation results, I deduced the constraints of conservation that may affect the algorithm's outcomes—these insights aided in building a rule system for frameshift prediction. Figure 10 shows the phase distribution calculated for all windows in the synthetic MSA dataset. The phase value gives an indication of which codon is the most variable within the triplets that are the codons – a way to determine the frame of translation. The phase calculation was not conducted on each codon on its own, but rather on a series of consecutive codons. Thus, every measurement was actually an “average” phase for all the codons representing a window analyzed. It can be seen clearly that for the different coding frames, the phase distribution was non-overlapping (Fig. 10A). In contrast, the frame mixture phases showed undeniable overlaps, making it harder to use this measure for frame differentiation (Fig. 10B-C). As detailed in the methods section, the

frame determination of a sequence is greatly affected by its phase as computed by the model, yet it is not likely that phase alone can help in differentiating frame mixtures from pure frame transitions.

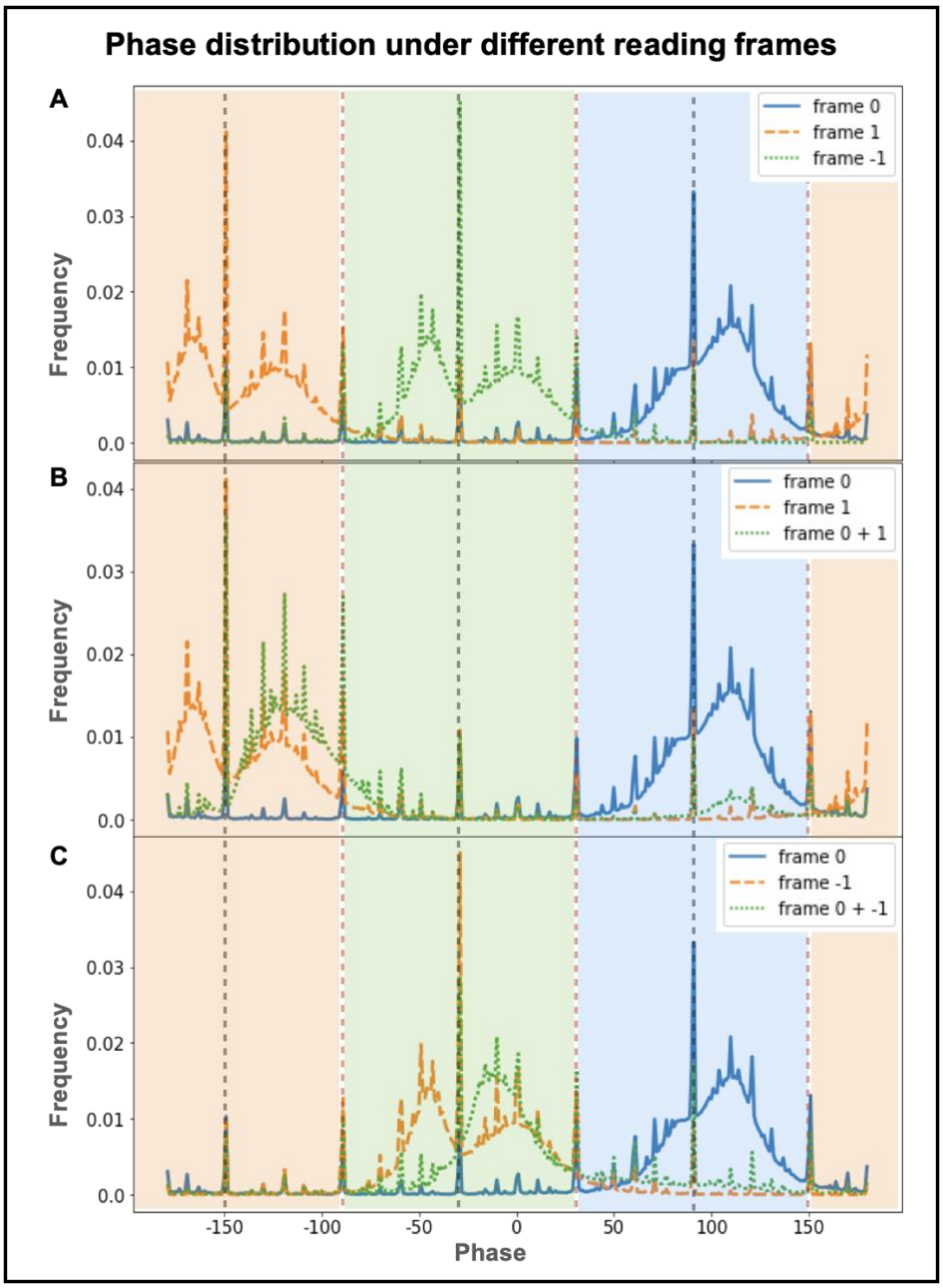


Figure 10: Phase distribution under different frames of translation. Using the synthetic MSA dataset generated for this work, I plotted the distribution of phases for every generated scenario. The sectors used for each frame determination are marked across this figure with a colored background. Gray dashed lines represent the center of each section, and red dashed lines represent the limits of each section. **(A)** Pure frames phase distribution. Every translational frame shows a distinct peak (-150, -30, and 90 for frames +1, -1, and 0, respectively), marking the dominant phase for pure frame translation. This frame matches the theoretical design. Aside from the prominent peak for every frame, two more distinct peaks surround it. These are likely phases representing transition windows where part of the sequence is in the previous frame during the transition. **(B)** Comparing phase distribution for +1 pure frame translation and mixture of 0 and +1 frames. The spectrum for the mixture of frames overlaps that of the pure frame making it harder to separate them. However, the mixture of frames has only one additional peak other than the prominent peak. **(C)** Same as **(B)** for -1 pure translation and the 0 and -1 translational frames mixture.

I also used this dataset of simulated MSAs to train a gradient boosting model (classifier) for the frame prediction. The features I selected for model training were the phase, magnitude, positive magnitude, and overall genetic variation computed for each window across the entire synthetic MSA dataset. Each window was a sample, and the labels were the window's translational frame.

Figure 11 shows success rates (and false-positive rates) of different evolutionary distances (mutation rate and the number of sequences) in each one of the cases simulated. Since the evolutionary distance in the database I used for later predictions was on average 10% in mutation rate, including overall 19 species (marked in black frames in figure 11), and the evolutionary distance within the primate group (which makes up most of the data) was 7%, I concluded that I am successful in detecting frameshifts in this data set with enough confidence.

To see if there is a good generalization for the models, I simulated evolutionary divergence using another gene sequence as the base sequence for mutations. I used the human hemoglobin subunit alpha 1 coding gene (HBA1 – NM_000558). I wanted to simulate sequences that can represent the actual genomic data I had and had used to generate candidate frameshift events. For this reason, I simulated the sequence with an evolutionary rate similar to the real data I had (10%) and having 19 species in every gene group. Using the rule-based model I get a 0.22 true positive rate (TPR) for dual coding (0/+1) frameshift events detections, 0.3 TPR for -1 frameshift events detections and 0.34 TPR for +1 frameshift events detections. When applying the second model where frame inference is calculated using a model trained on the synthetic MSA database, I get 0.45 TPR for dual coding (0/+1) frameshift events detection, 0.78 TPR for -1 frameshift events detections, and 0.91 TPR for +1 frameshift events detections. This alone does not necessarily imply that the latter model is more accurate but rather might be overfitted for the synthetic data. With that, the fact that the numbers are almost 50% lower than the ones calculated when computing TPR for the MSA dataset used for assumption making might suggest that the rule-based model is under-fitted and not completely generalized. Thus, I can assume that aside from

the predictions shown in this work, there might be many more candidates in the human genome.

Figure 11: Simulation success rates and false detection rates. Simulating different cases (single +1 frameshift, single -1 frameshift, dual coding regions and no frameshifts), I could estimate my success rate by counting the number of times where my algorithm was successful in detecting the right transition in the correct site (given some error space). I simulated sequences with varying levels of evolutionary distances, which are affected through mutation rate and number of sequences. Black boxes (and purple box) mark the success rates (and false detection rate) for an evolutionary distance as representing the data analyzed in this work. It is evident that dual coding regions have much lower success rate as this is a more complex scenario.

4. The models detect known genes in which frameshifts occur

4.1. Human OAZ1 and PEG-10

After analyzing the entire human genome, ~1200 genes were predicted to have non-canonical translation via alternative frame translation. Two of these genes are well known and characterized as ribosomal frameshifting incidents. The first is the Ornithine decarboxylase antizyme (OAZ1). OAZ is an enzyme that catalyzes the rate-limiting step in polyamine biosynthesis. It regulates the synthesis of polyamines by binding and inhibiting ornithine decarboxylase. Antizyme expression is regulated by polyamine-enhanced frameshifting, creating a +1 ribosomal frameshift and making the enzyme

active (Jonathan, 2012). OAZ1 is highly conserved in vertebrates, conserving its sequence and the mechanism of action. As shown in figure 12A, two active frames, one after the other, constructed the profile of the translational frame of OAZ1. The location of the predicted frameshift matched that of the reported frameshift up to 6 AA.

The second known case of ribosomal frameshift is Retro transposon-derives protein (PEG-10). This gene includes two overlapping reading frames of the same transcript encoding distinct isoforms. The shorter isoform has a CCHC-type zinc finger motif containing a sequence characteristic of gag proteins of most retroviruses and some retrotransposons. It functions in part by interacting with members of the TGF-beta receptor family. The longer isoform has the active site DSG consensus sequence of the protease domain of pol proteins. The longer isoform results from -1 translational frameshifting which is also seen in some retroviruses (Wills et al., 2006). As seen in figure 12B, two active frames, one after the other, constructed the profile of the translational frame of PEG-10. The location of the predicted frameshift matched that of the reported frameshift up to 8 AA.

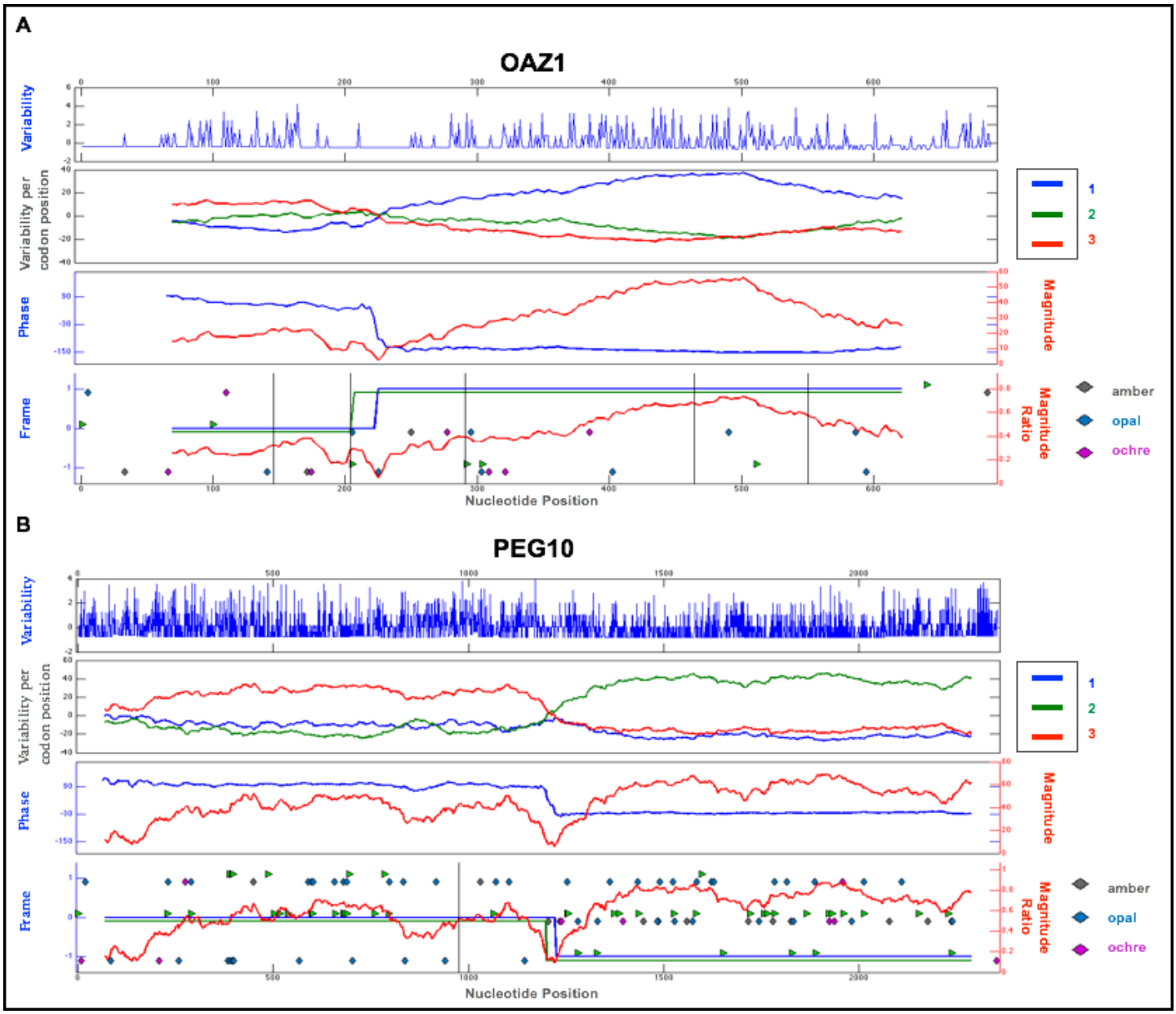


Figure 12: The known cases of ribosomal frameshifts as detected from my algorithm. Top panel shows the variability scores for each nucleotide position. Second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). Third panel shows the raw output from the analysis – magnitude and phase. Bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account (see the methods section for details). Green line shows the annotated frames from the literature. **(A)** OAZ1 is known to have a +1 frameshift resulting in an elongated protein version revealing functional domains. The error of estimation of the frameshift site was 6 amino acids. **(B)** PEG10 is a fusion protein translated via -1 ribosomal frameshift. The error in estimation of the frameshift site was 8 amino acids.

4.2. HIV gag-pol fusion gene

Viruses are well known for utilizing their compact genome by using the same locus to translate different proteins using many mechanisms, including ribosomal frameshifts. One well-known case is for the HIV gag-pol polyprotein. The polyprotein translation is contingent on a -1 ribosomal frameshift revealing the pol ORF that is located downstream of the gag ORF with a slight overlap. The polyprotein will eventually be cleaved into the viral enzymatic proteins (Nikolaitchik and Hu, 2014). I analyzed ~1500 HIV variants in the locus of the gag-pol polyprotein to receive the variation signal. This case study gave me a case of the double encoding region and a -1 frameshift, and the algorithm detected both signals distinctively, allowing the proper interpretation (figure 13B).

4.3. HPV E2-E4 overlapping region

In HPV viruses, two ORFS overlap in locus but are programmed as coding in different frames (Tan et al., 2012). The E4 ORF resides entirely within the E2 coordinates but is annotated to translate in a different frame. In this case, E4 is not translated via ribosomal frameshifting, but rather this is a case of the dual coding region within a transcript. My model successfully located this “dual coding” region, showing unique phase properties (fig. 13A) that were later confirmed via simulation.

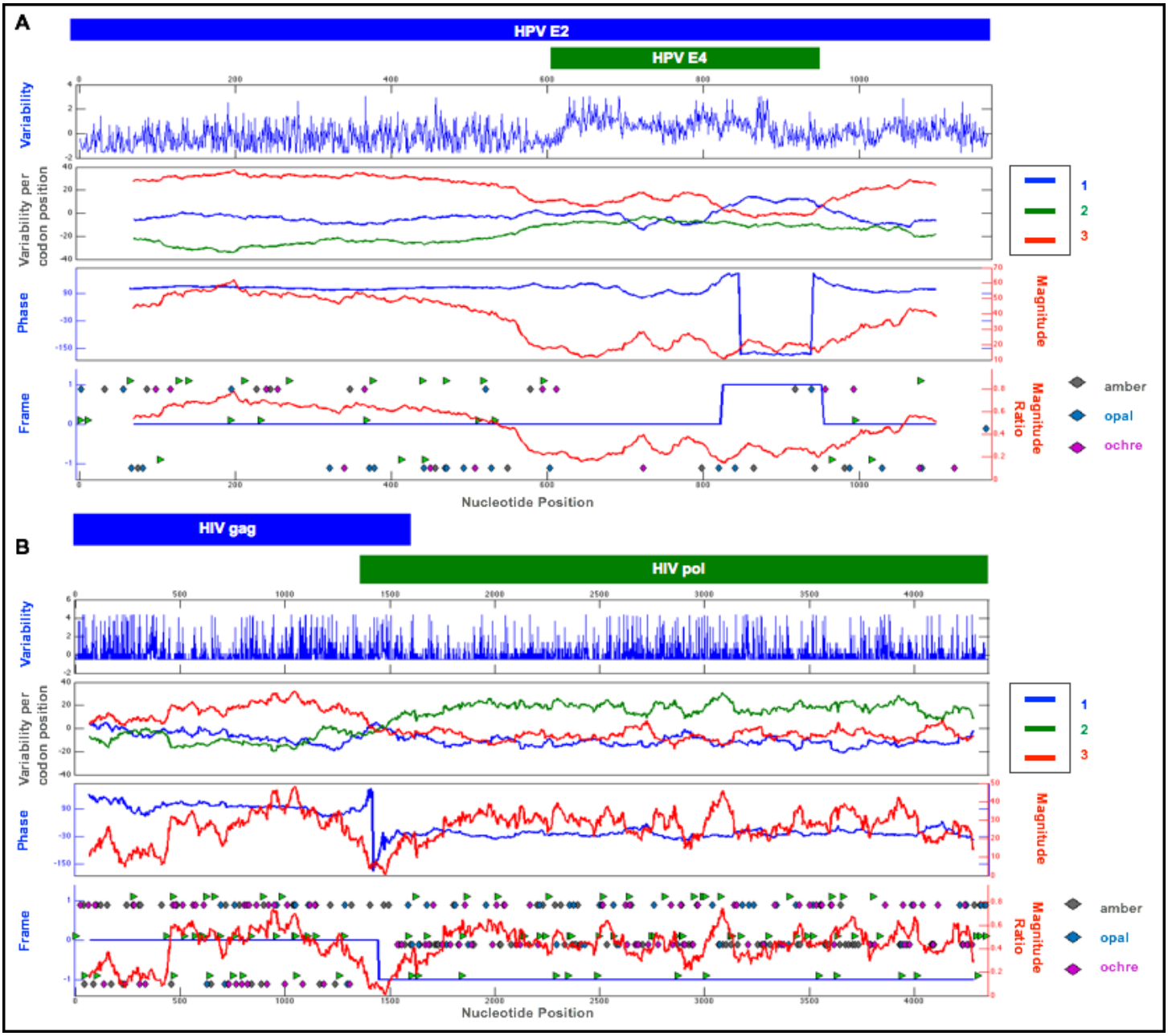


Figure 13: The known cases of viral ribosomal frameshifts as detected from my algorithm. The top panel shows the variability scores for each nucleotide position. The second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). The third panel shows the raw output from the analysis – magnitude and phase. The bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account (see methods section for more details). The green line shows the annotated frames from the literature. **(A)** HPV dual-coding region E2-E4. The effect on conservation is quite visible in the overlapping area, but my algorithm detected this mixture with great delay. **(B)** HIV gag-pol fusion protein is much like the mammalian PEG10 case, only with a slight overlap between the ORFs. This overlap is visible from the singularity observed in the frameshift site.

5. Detecting frameshifts on the novel COVID-19 genome

The Novel Corona Virus Disease (*COVID-19*), which upon infection causes Severe Acute Respiratory Syndrome Coronavirus 2 (*SARS-CoV-2*), is one of the major causes of respiratory diseases in the past two years.

Using multiple sequence alignments of *COVID-19* infected samples from humans and bats, I detected two highly probable frameshift events on annotated ORFs from the human sequence. One frameshift event predicted a -1 frameshift on ORF1a, resulting in the translation of a fusion polyprotein product between the two consecutive ORFs, ORF1a and ORF1b. The second was an out-of-frame alternative ORF (having a start and a STOP codon) within the ORF3a coding region. Both these events were reported and experimentally shown to happen (Firth, 2020; Jungreis et al., 2020).

5.1. ORF1a -1 frameshift generates a fusion polyprotein between consecutive ORFs.

As reported in (Firth, 2020), Most of the *SARS-CoV-2* genome encodes for ORF1ab, which includes within it a -1 programmed ribosomal frameshift, generating a polyprotein that should later be cleaved. This frameshift is triggered by a slippery site and a downstream pseudoknot, much like in the case of the HIV gag-pol polyprotein (Baranov et al., 2005). The ribosomal frameshift was reported to happen four codons before the annotated STOP codon for ORF1a at nucleotide position 13,190 (Jungreis et al., 2020).

Applying my model for frame prediction and frameshift location prediction revealed the same frameshift reported with a drift of four codons. Figure 14 shows my model's predicted profile for this gene, showing high periodic signals and high accuracy in location prediction.

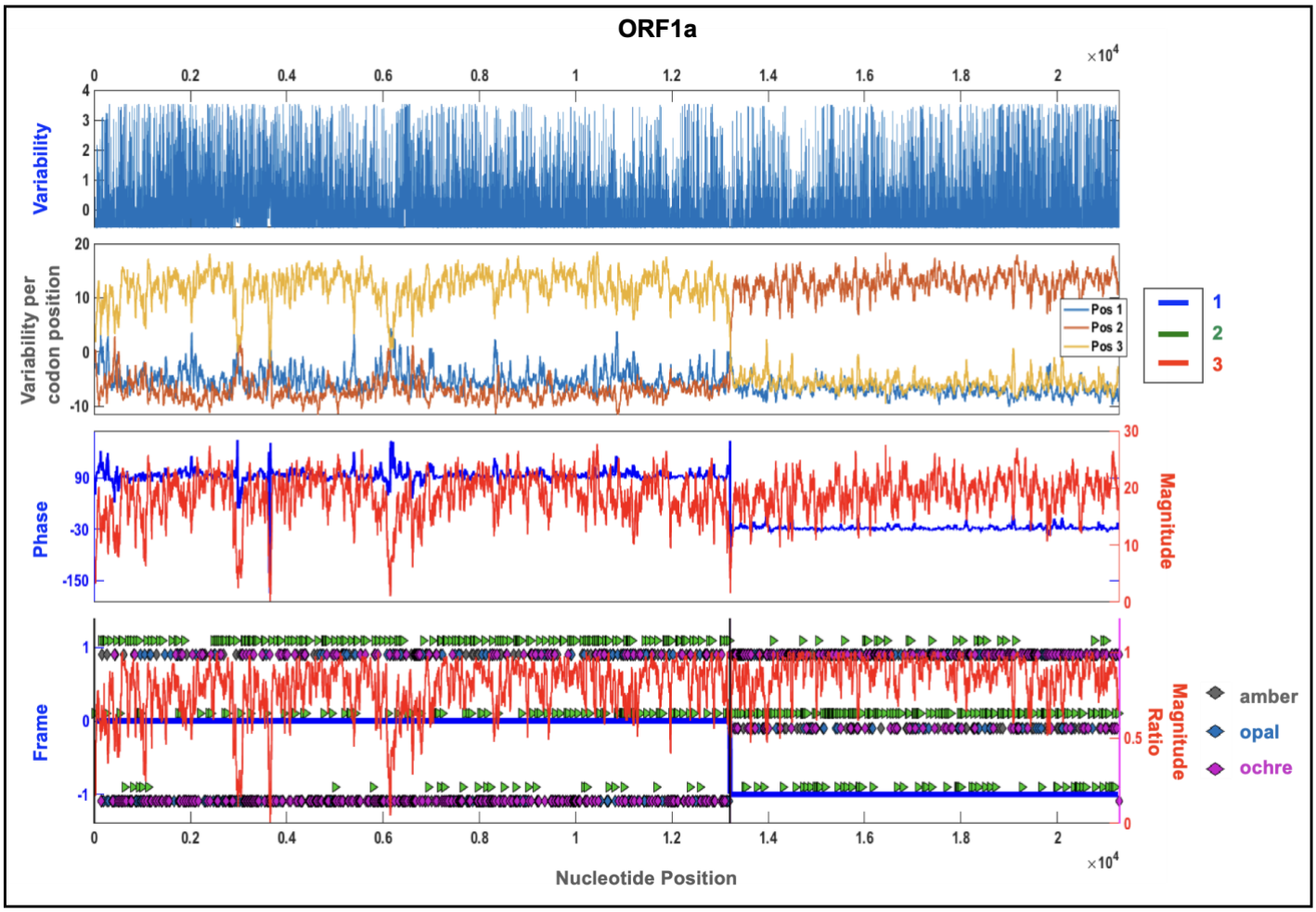


Figure 14: Frameshift in COVID-19 ORF1a. Analysis scheme as previously described where a -1 frameshift appears on nucleotide 13,178, just four codons upstream to the reported location. Top panel shows the variability scores for each nucleotide position. Second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). Third panel shows the raw output from the analysis – magnitude and phase. Bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account.

5.2. ORF3a encodes another non-canonical ORF within it, frameshifted +1 relative to itself.

When analyzing the profile produced by the model for ORF3a, there is an inset of a second non-canonical reading frame. The angle corresponding with that reading frame would fit a mixture of the 0 and +1 frames (much like in the HPV dual-coding region reported earlier). Since the alternative translational frame has a start and stop codon surrounding the predicted area of non-canonical translation, it is more probable that this is an alternative reading frame rather than two events of frameshifts on the same transcript.

Indeed, (Finkel et al., 2021) found experimental evidence from ribosome foot-printing for translation initiation in that region and that matches the

proposed alternative reading frame. Later, (Firth, 2020) provided further computational support for the existence of this putative ORF.

The results from my model can be seen in figure 15. The drift between the alternative initiation site and the predicted start site for the alternative frame was eight codons, and the drift between the alternative STOP codon and the predicted stop site for the alternative frame was ten codons. As in other cases of dual coding I analyzed, the errors in location prediction are greater than those made for complete phase transition.

Accumulating more positive control samples where it is experimentally proven that an alternative translational frame exists strengthens the validity of my model, and would provide me with higher confidence in producing further candidates for experimental testing.

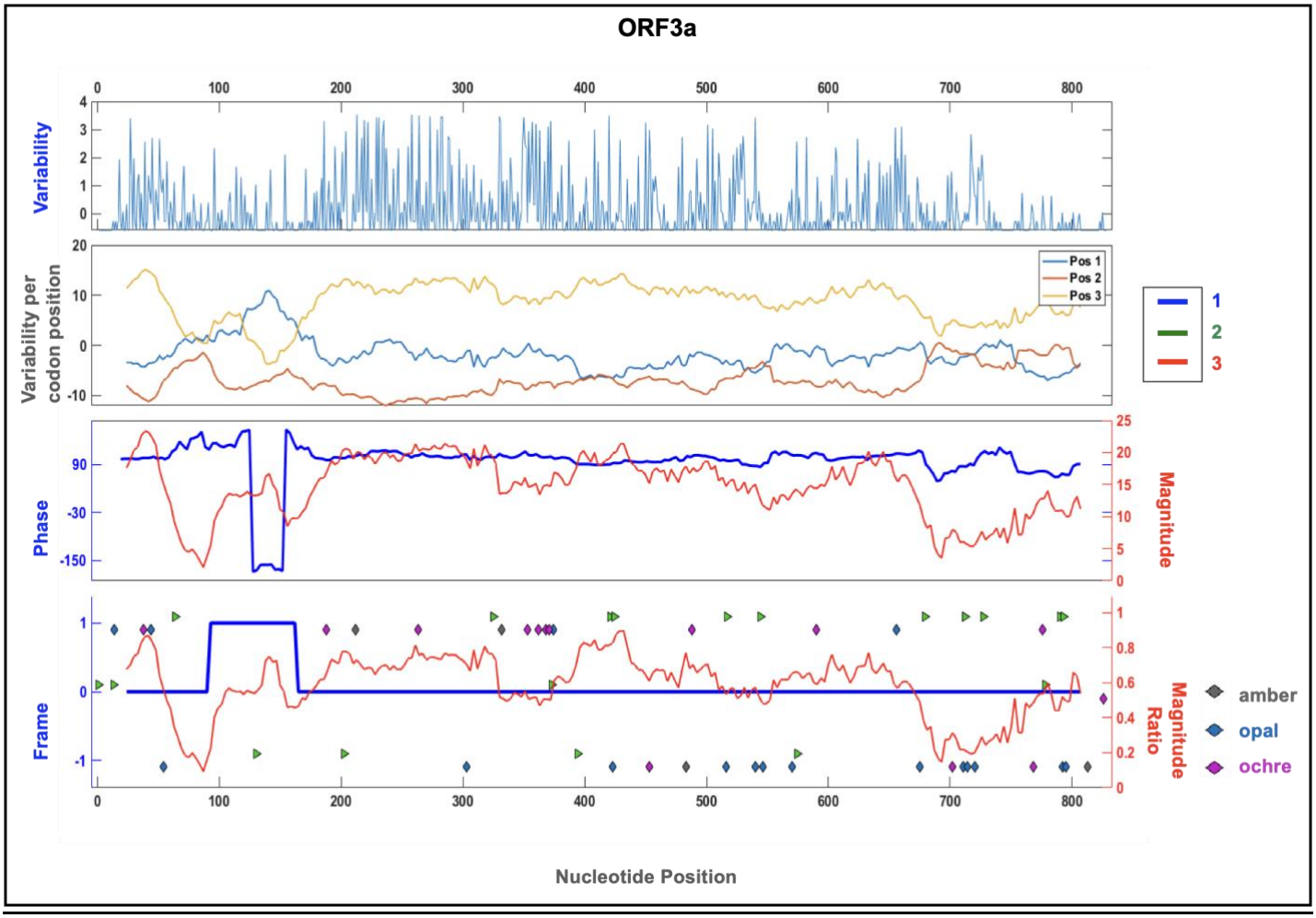


Figure 15: Frameshift in COVID-19 ORF3a. Analysis scheme as previously described where a +1 dual coding region is visible. Top panel shows the variability scores for each nucleotide position. Second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). Third panel shows the raw output from the analysis – magnitude and phase. Bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account.

6. The model predicted 400 novel frameshifting events in the human genome

I analyzed ~40,000 human transcripts to produce their 3-way periodicity profiles. Using the pipeline developed for predicting frameshifts described in the Methods section (figs. 27-28, 32), ~1200 genes were predicted to have two distinct translation frames. Of these predictions, 400 genes (coming from ~300 different loci) were labeled novel events because the newly predicted peptide was not found in any conceptual translation database of the human genome. This way, I eliminated any known case of ribosomal frameshift or any case of alternative splice variants that will create transcripts that appear frameshifted relative to the primary transcript of that gene. Out of these 300 loci, 86 were confirmed by both models as described in the methods section, presenting as top candidates for non-canonical translation in humans.

Of those, I could not, to date, tell if the signal arose due to them undergoing ribosomal frameshifting during translation or due to having an additional splice variant not yet annotated. But the potential for translation in another frame is represented by the variability profile of their genes.

Previous work (Michel et al., 2012) analyzing the periodic signal arising from ribosome profiling experiments aimed to also find areas in the genome where dual-frame encoding can happen. There, they utilized a triplet periodicity that arises from the ribosome-protected fragments (RPF) after alignments to the mRNA. The RPF aligned mostly to the first or third sub codon positions (Guo et al., 2010). This suggested that for ribosome profiling data, the phase of the periodic signal can be used to assess the reading frame. As reported later in this chapter, I have also used available RPF data to provide further evidence for my findings. The work reported by (Michel et al., 2012) detected 108 protein-coding genes where a duality in the coding frame was detected as calculated by their algorithms. They manually inspected the 108 profiles assumed to have dual coding and divided them into six subgroups. Out of the 108 reported dual-coding genes, 33 of them were later labeled as false positives, putting the error rate for their detection at over 30%. Two of their reported groups were the identification of upstream and non-upstream overlapping ORFs. Most of them are short sequences that are not easily detected by my methods, as

currently reported. The other relevant groups were the known cases of OAZ1 and PEG10, alternative splicing events, and cases that are otherwise unexplained – i.e., novel predictions.

My model's presented configuration for detection was for high confident, simple read-through cases. This means that some events that should be detected were filtered along the way. After manually inspecting the profiles presented in (Michel et al., 2012), I reported agreement on some of their novel results and the explainable cases (as can be seen in the supplemental figures). In (Michel et al., 2012), the data used for analysis was a mere 6000 genes that were those that had the highest RPF coverage from data collected from HeLa cells (Guo et al., 2010). After optimizing a threshold for determining periodicity transition score (PTS), only 800 genes were left for valid analysis. Of those, only 108 genes were predicted to have dual-coding regions and were later manually inspected (Michel et al., 2012). This sparsity of analyzed genes in that work and the fact that my model was optimized to analyze the simple, most probably cases left little to no chance of overlap between predictions. That said, the known cases of OAZ1 and PEG10 are successfully detected by both methods under many different constraints.

Of their 108 genes predicted to have more than one encoding frame, 44 are non-upstream ORFs (*nORFs*) and upstream ORFs (*uORFs*), which are mostly regulatory sequences and do not result in protein products. Thus I did not expect that my model would have been able to detect those. Moreover, the method presented in (Michel et al., 2012) reveals 16 instances of alternative splicing variants resulting in frameshifted sequences. It is assumed that the human genome has thousands of these events (Kovacs et al., 2010), and since I present an identification of ~1000 such events, it could be that my model is more sensitive for this task.

Examining those cases labeled "unexplained" by (Michel et al., 2012), 13 in total, show that five cases might have also been detected by my algorithm (see supplemental figures 1-7). The genes SCML1, DSN1 and KIAA0825 shows clear signs of non canonical translation, while the gene PTBP3 and FAM210A show ambiguous frame due to low significance in the magnitude, which may also suggest dual coding. Since (Michel et al., 2012) did not give exact coordinates, nor the expected frameshift, but

merely presented the RPF counts per sub-codon position, I could only qualitatively and manually examine the overlap between my results and theirs. It should also be noted that they self-reported 33 false-positive results, which is 30% of their entire predictive space.

The manually inspected comparison of profiles can be found at the end of this chapter under supplemental figures 1-7, alongside a description of the comparison between findings in this work and (Michel et al., 2012).

Coming back to the novel predictions presented in this work, there seemed to be nothing striking about this group of genes as they have many different functions, expression patterns, etc. Examining sequence properties for these predictions (amino acid compositions, codon usage) and comparing them to randomly picked groups of genes did not give rise to anything that may indicate a mechanism. That said, when examining the result of a frameshift in some of these genes, I saw some sub-groups emerging. For 34 genes, it seemed as though the frameshift might be causing the gene to be truncated by encountering a series of STOP codons right after the frameshift happens (an example can be seen in figure 16A). Cases like these may suggest that the translation in that frame must come to a halt.

In figure 16B, I showed a case where frameshifting caused the protein to lose a poly methionine sequence that may be important for the original gene's function. Moreover, the new frame of translation has no STOP codon, suggesting that translation, in this case, may continue into the 3' UTR due to the frameshift.

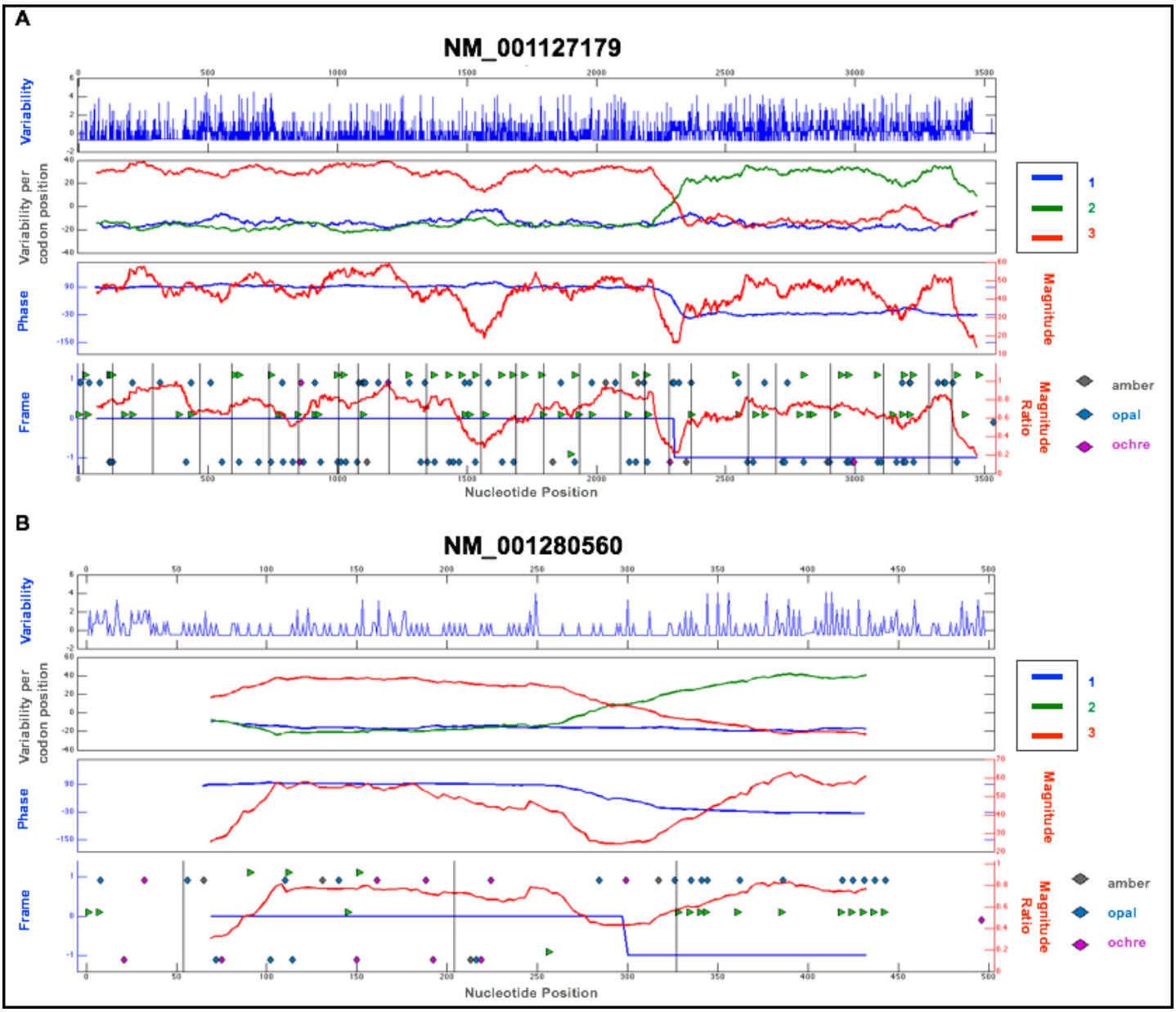


Figure 16: Examples of novel frameshift predictions. The top panel shows the variability scores for each nucleotide position. The second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). The third panel shows the raw output from the analysis – magnitude and phase. The bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio, taking optimum normalization into account. **(A)** Frameshifting would cause a significant part of the protein to be truncated in this gene. Translation doesn't appear to be active in the new frame, as multiple STOP codons would cause translation to end. **(B)** For this prediction, it seemed that a poly-Methionine signal would be lost if the frameshift occurred. Also, the new frame does not code for a STOP codon, so translation might proceed into the 3' UTR in the new frame.

6.1. Some of the predicted new peptides show a high correlation with the human proteome after frameshifting

After observing the distribution of amino acid (*aa*) composition of all the proteins in the cell, I could assess the correlation between the conceptual translation of the frame-shifted versions of the subset mentioned above of frame-shift genes predictions with their predicted

frameshift. It seemed as though for some genes that are predicted to have a frameshift, the aa composition after the frameshift had higher correlation scores with the proteome (using Pearson's linear correlation score calculated on the vectors representing aa frequencies for the entire proteome, and every version of the protein) than the original protein (fig 17A). This observation may suggest a real functional change in the protein after frameshifting. It could help determine which frameshift predictions would most probably have major effects and are less likely to be a product for degradation. It may also hint into frameshift events that cause the new protein version to acquire known and common protein motifs that were initially not in the sequence. Similar behavior was observed when I looked at the codon compositions of these genes (fig 17B).

Changes in codon composition could also have major effects on translation outcomes. For one, codon composition has a role in the translation efficiency (Tuller et al., 2010), implying protein folding, translation errors, and more. Another possible outcome could be due to changes in mRNA secondary structure. This could also affect the translation (Soemedi et al., 2018). Still, more importantly, it could alter the mRNA's stability (Mauger et al., 2019), which might generate more or less stable mRNAs after frameshifting.

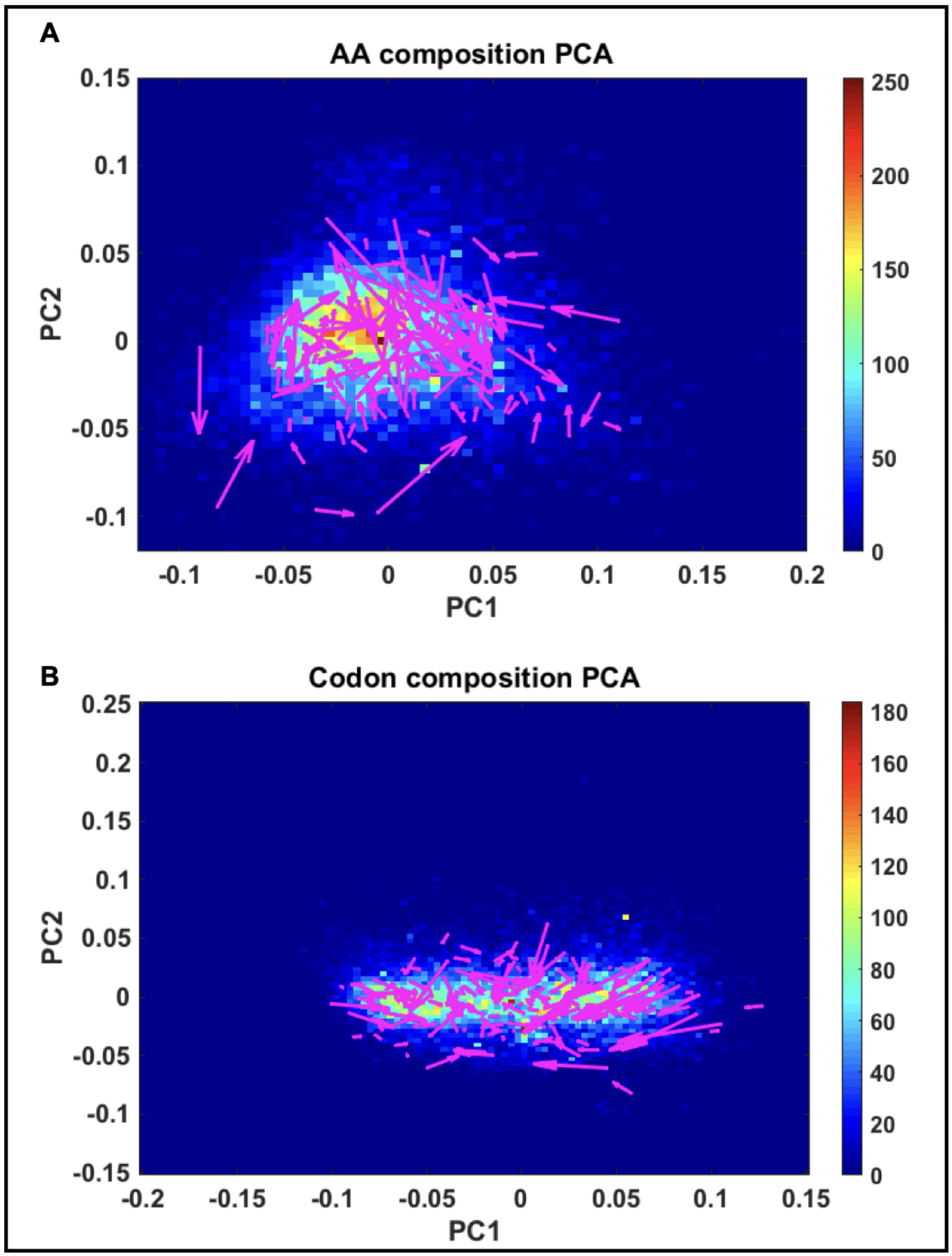


Figure 17: AA and codon composition changes for the original and the predicted frameshifted version of the protein. To understand if a peptide translated due to frameshifting could be functional, I analyzed the changes this would have had on the original protein as would be canonically translated. Purple arrows are all the frameshift candidates where the correlation score between the frameshifted version and the entire proteome was higher than the original protein or gene without frameshifting. The base of the arrow marks the original score, while its head marks the frameshifted score. Density plot shows the composition scores of the entire proteome. **(A)** Amino Acid composition PCA. **(B)** Same as **(A)** for protein-coding genes and codon composition PCA.

6.2. Genes that exhibit +1 frameshifts may carry a signal in mRNA secondary structure

Several databases for predicting ribosomal frameshifting events based on motif recognition after analysis of known frameshift sites exist (mostly from plants, yeasts, viruses, and bacteria) (Jacobs et al., 2007; Moon et al., 2007; Theis et al., 2008). These are based on observation and mostly show prediction and identification for -1 ribosomal frameshift events. I attempted to cross the area of frameshift as predicted from the algorithm in this thesis with each of the datasets but got poor matching. This could occur due to an error in the exact frameshift location estimation (motifs are relatively short, less than 25 nucleotides) or because mammals are barely analyzed to generate motifs in these data sets. When examining the properties of frameshift motifs, I noticed that mRNA secondary structure has an important role in generating ribosomal frameshifts, as well as the sequence composition (Dulude et al., 2002; Li et al., 2002). Previous work mainly described and characterized -1 ribosomal frameshift, and I sought to try and find a motif of such nature from my +1 predicted frameshift events.

To do so, I analyzed the mRNA secondary structure energy (see methods section for details) to find a signature of strong structures around the frameshift location. I found that a subset of the genes predicted to show +1 frameshifts that show a signature of strong secondary structure right around the predicted frameshift location (fig 18A). This was also observed when examining the energy signature for OAZ1s mRNA secondary structure, a gene known to undergo a +1 frameshift (fig 18B). This change in secondary structure exceeded the 95% confidence interval calculated for a shuffled sequences control group. I did not observe such signatures for the -1 frameshift prediction sequences. This finding could serve as a motif for identifying novel +1 frameshift events. It may also aid in distinguishing between ribosomal frame-shift events and frameshift that are observed due to un-annotated alternative splice variants or might even hint into the process under which a frameshift might occur.

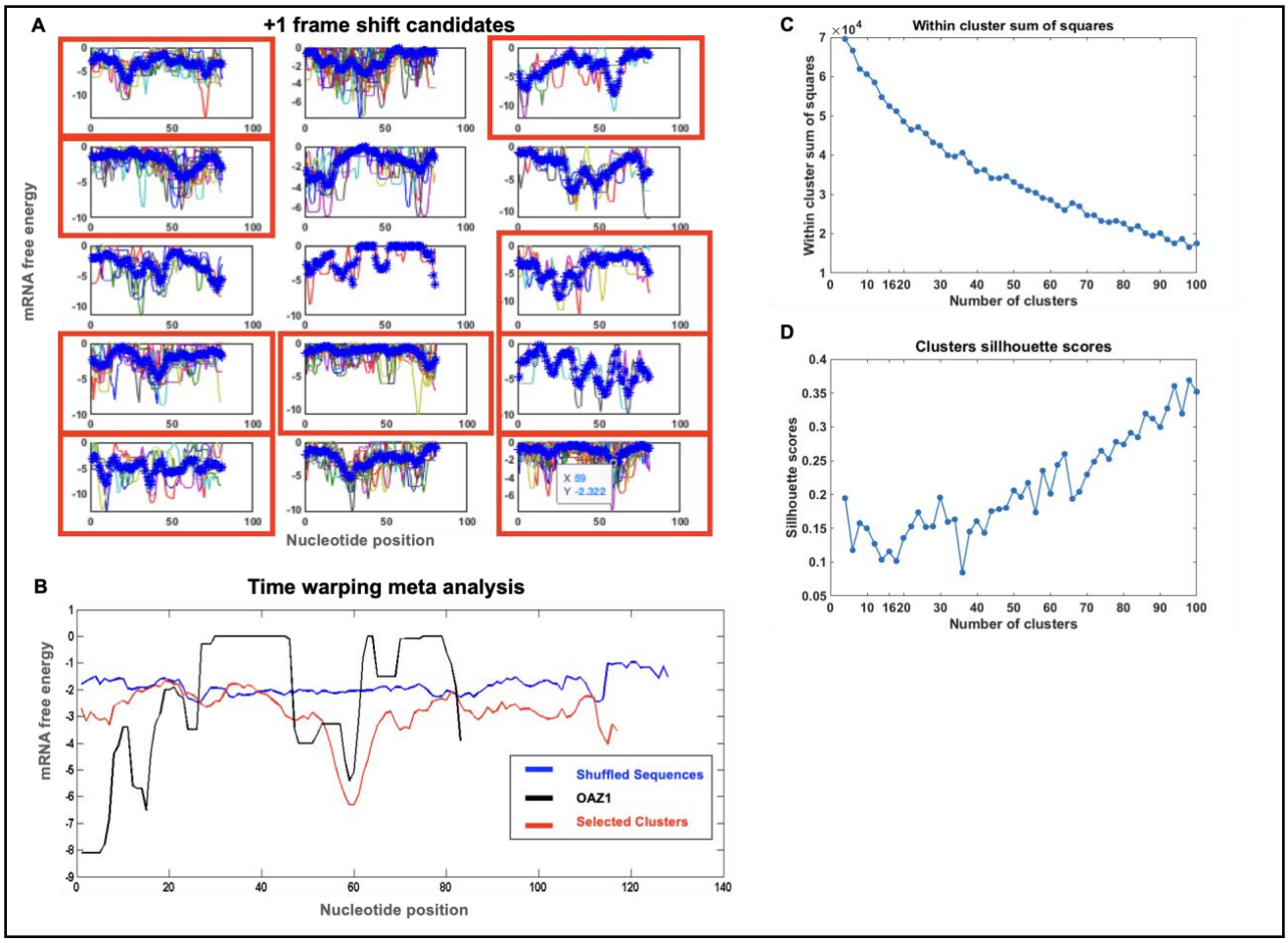


Figure 18: mRNA secondary structure energy signature. mRNA secondary structure showed a signature of low free energy (tight mRNA structure) for a subgroup of predictions, near the frameshift site.**(A)** Sunbgroups as clustered by their energy patterns around the predicted frameshift site. Red frames mark those clusters where a distinct low energy peak exists (suggesting tight mRNA structure). **(B)** Aligning the selected groups from (A) with the energy profile produced for the only known case of +1 ribosomal frameshift in mammals, OAZ1, gave great overlap. This was an indication that this may be involved as a mechanism for regulation. This was compared to the energy patterns that the shuffled sequences produced. The shuffling was done by only shuffling the codons and not at the nucleotide level. **(C-D)** Determining the number of clusters. I used two scoring functions to find the best number of clusters for this analysis. Both within clusters sum of squares and silhouette scores (which are common methods for cluster number determination) did not aid in having a distinct threshold. 16 clusters seem to be a fair tilting point for this analysis.

6.3. Genes that exhibit +1 frameshifts are “young” (recently evolved)

BLASTing protein sequences of the predicted 400 conceptual translations in the newly predicted frame, against *C. elegans*, *D. Melanogaster,* and *S. Cerevisiae* proteome databases, revealed that +1 frameshifted gene predictions tend to not have orthologs in any of these groups when compared to -1 frameshifted gene predictions, and the entire human proteome (fig. 19). -1 ribosomal frameshifts are well characterized in viruses. They usually happen on a slippery site that causes the ribosome to change its original reading frame, followed by an mRNA pseudoknot which re-stabilizes it (Dinman, 2006). +1 frameshifts are much less common. Not only that, but when they do come to be, no common sequence motif seems to apply. I thus believe that this type of frameshift might have evolved more recently. While the human proteome is very diverse, it could still benefit from plasticity derived perhaps from different stress conditions, under the regulation of programmed ribosomal frameshift.

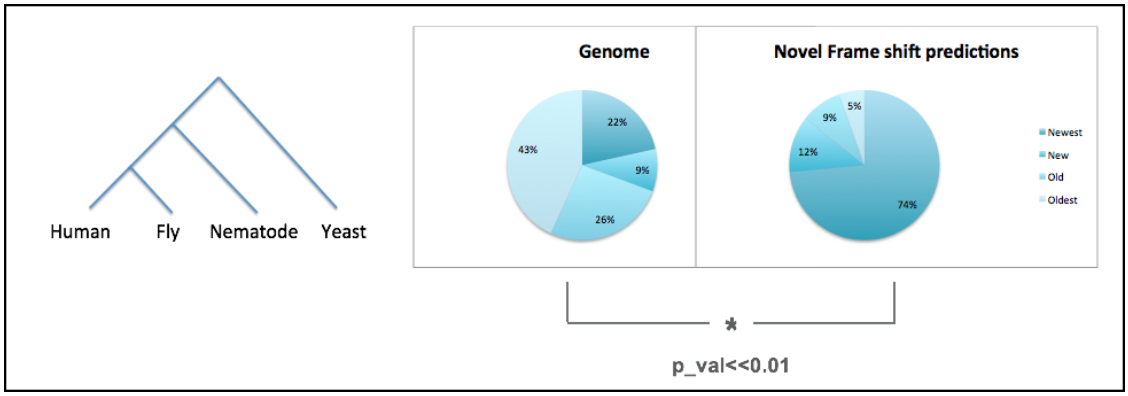


Figure 19: +1 frameshift predictions seemed newer in evolution. I described the “age” of a gene as the distance in evolution for which it has an ortholog. I tested orthologs in Fly, Nematode, and Yeast. The older a gene is, the farther it would go with having orthologs in evolution. I saw that genes predicted to have +1 frameshifts tend to not have orthologs in any of these organisms, marking them as newly evolved.

7. Confirmation by ribosome profiling P-site location analysis

The models could analyze periodicities that arise not only from variation in MSAs. A similar periodicity arises when observing the sub codon localization of ribosome P-sites from ribosome pull-down and sequencing experiments. In addition, I can utilize this to detect translation in non-canonical frames using newly acquired experimental data. Using the data

obtained from Ohler et al. (Calviello et al., 2016), I tested the strength of the model in detecting such events on ribosome profiling data for the known genes in the human genome that undergo ribosomal frame-shift, as well as the set of novel, predicted genes I gathered.

I did not find an overlap between the novel predicted frameshift genes as predicted from ribosomal profiling data and those predicted from MSA data. Most predictions were not convincing as they had very low ribosome counts. However, I detected the two known genes in the human genome that undergo ribosomal frameshift (OAZ1 and PEG10). In addition, I was also able to detect a few cases of frameshifts that occur due to splicing. This gave me enough confidence in the use of this kind of data and the motivation to find high-quality data from different tissues and under different conditions to validate the results from the MSA analysis.

Figure 20 shows the results of the rule-based model when analyzing the data of ribosome counts (P-site) in OAZ1 and PEG10. The results are consistent with those obtained from MSA, albeit with a drift of 16 codons between the predicted and annotated frameshift location in OAZ1 and ten codons in PEG10 (Figure 20).

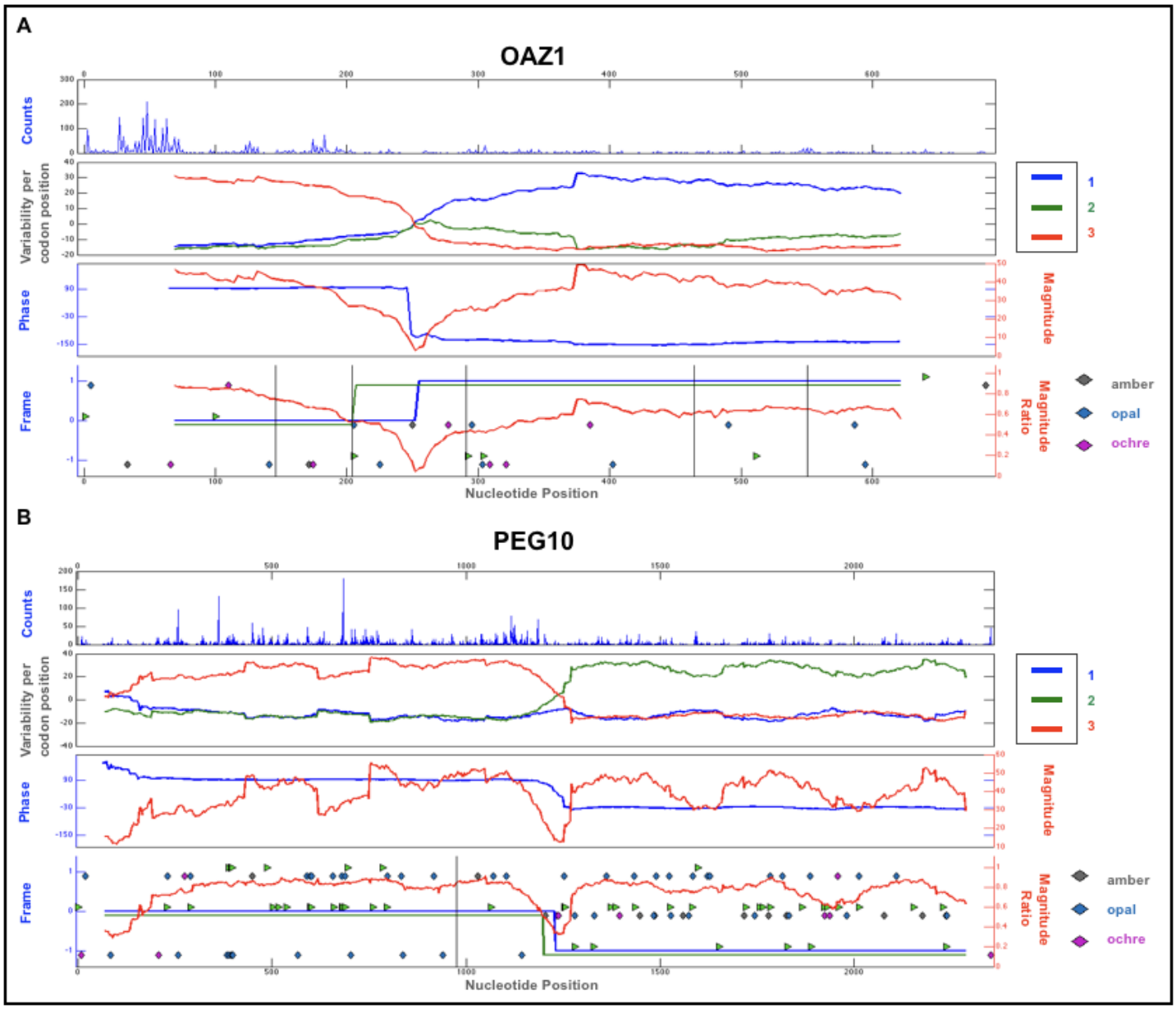


Figure 20: Ribosome profiling P-site location analysis on the known cases of ribosomal frameshift. Top panel shows the ribosome P-site counts for each nucleotide position. Second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). Third panel shows the raw output from the analysis – magnitude and phase. Bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account. Green line shows the annotated frames from the literature. **(A)** OAZ1 is known to have a +1 frameshift resulting in an elongated protein version revealing functional domains. The error of estimation of the frameshift site was 16 amino acids. **(B)** PEG10 is a fusion protein translated via -1 ribosomal frameshift. The error in estimation of the frameshift site was 10 amino acids.

8. Case Studies of frameshift prediction genes

8.1. NM_152559 – Methyltransferase like 27 (METTL27)

METTL27 is located in the *Williams-Beuren syndrome (WBS)* critical region. WBS results from a hemizygous deletion of several genes on chromosome 7q11.23, thought to arise due to unequal crossing over between highly homologous low copy repeat sequences flanking the deleted region. Haploinsufficiency of METTL27 may cause certain cardiovascular and musculoskeletal abnormalities observed in the disease (Fusco et al., 2014). My prediction model predicted a -1 frameshift in the proteins' C-terminus, causing it to both partially change its sequence and also read through the canonical STOP codon extending translation to the 3' UTR (fig 21). The signal picked up by the model originated from the fact that other than the human sequence for this gene, all other sequences in the alignment contained an 11 nucleotides gap, causing a -1 frameshift to be detected. This means that for all other species analyzed, the major transcript for this gene was spliced in a manner that is frameshifted relative to that of the human major transcript.

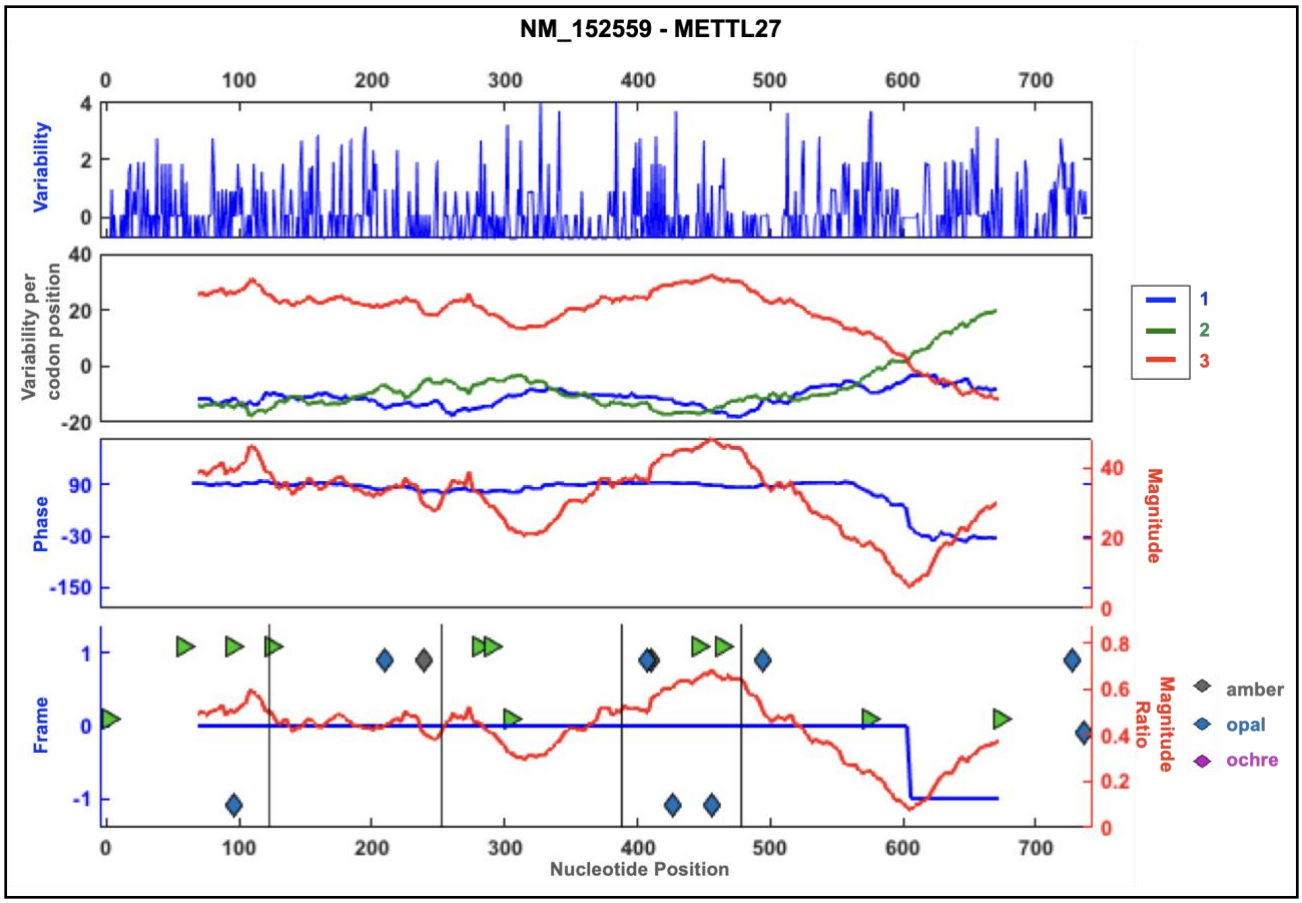


Figure 21: METTL27 frameshift prediction. Analysis scheme as previously described where a -1 frameshift appears on nucleotide 603. This correlated with the fact that the MSA holds an 11 nucleotides gap – creating this frameshift signal detected by my model. The top panel shows the variability scores for each nucleotide position. The second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). The third panel shows the raw output from the analysis – magnitude and phase. The bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio, taking optimum normalization into account.

Analyzing the conceptual translation of the new version of the protein, as predicted by the algorithm, resulted in a complete match to the major product in all other species included in the analyses (fig 22).

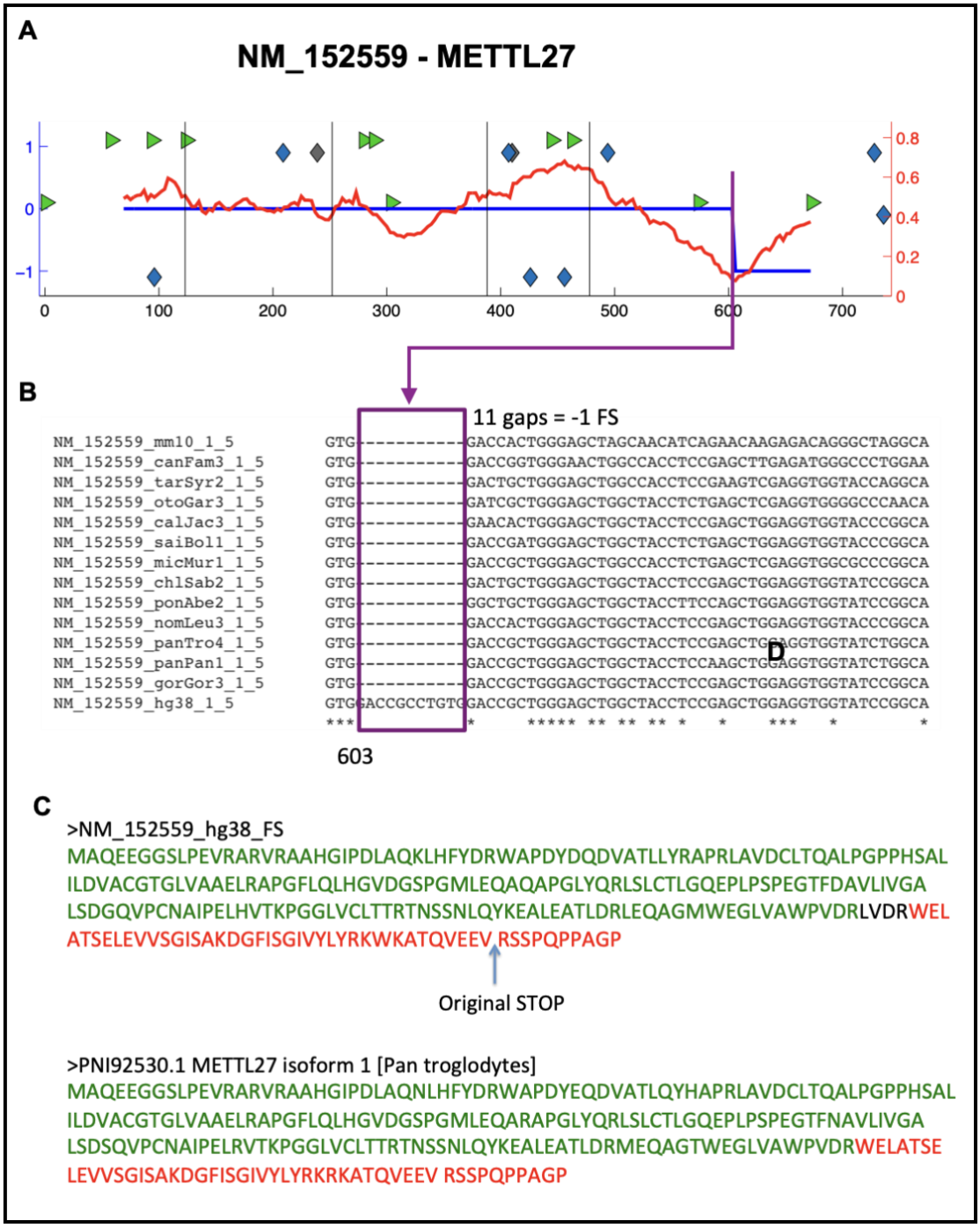


Figure 22: METTL27 frameshift would give higher similarity to the major transcript of this gene in other species. (A) The -1 frameshift appears on nucleotide 603, which correlates with the fact the MSA holds an 11 nucleotides gap – creating this frameshift signal detected by my model **(B)**. **(D)** BLASTp on the new version of the protein generated hits from many organisms where the protein is annotated as the version of the major transcript, strengthening the assumption that it could be that this transcript should have also been annotated in the human genome.

There seemed to be no common functional domains specifically mapped to the region changed. Still, when examining the proteins' secondary structure with the frameshift, it appeared that a beta-sheet replaced an initially present alpha helix, and an intrinsically disordered

region is gained in the new C-terminus. The addition of this region, along with the loss of an alpha helix, may suggest potential protein-protein interaction (fig 23).

Recent work studying genetic variation from more than 60K individuals with different genetic backgrounds and disease (Lek et al., 2016) reveals that for Non-European (non-Finnish) individuals, there was a high probability for a single nucleotide deletion generating a -1 frameshift creating the exact frameshift I predicted. No clinical implications were described for this SNP, nor was this further studied as an effector of some phenotype. Still, since the experimental evidence is present, it might be of interest for further studies.

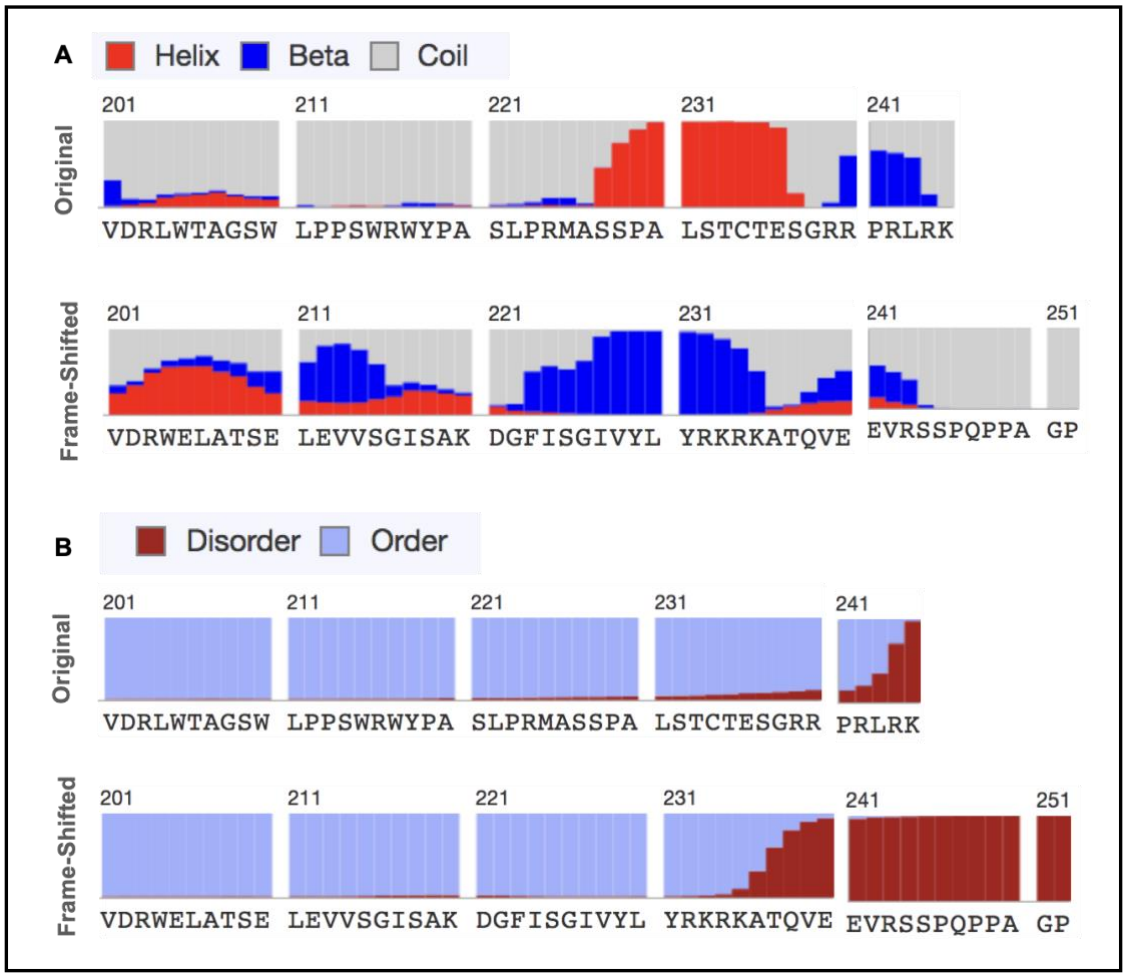


Figure 23: METTL27 frameshift replaces an alpha helix with a beta-sheet. (A) Analyzing the protein secondary structure and intrinsically disordered regions, showed that due to the frameshift, a structural change could take place that may affect the function and create protein-protein interactions. **(B)** The added peptide that would be appended should the frameshift occur creates an intrinsically disordered region that could also indicate protein-protein interaction in this new section of the protein.

8.2. NM_199243 – G Protein-Coupled Receptor 150 (GPR150)

This gene encodes an orphan member of the class A rhodopsin-like family of *G-protein-coupled receptors (GPCRs)*. Within the rhodopsin-like family, this gene is a member of the vasopressin-like subfamily, including vasopressin and oxytocin receptors. The silencing of this gene, due to promoter methylation, is associated with ovarian cancer progression. All GPCRs have a transmembrane domain that includes seven transmembrane alpha-helices. A general feature of GPCR signaling is the agonist-induced conformational change in the receptor, leading to activation of the heterotrimeric G protein. The activated G protein then binds to and activates numerous downstream effector proteins, which generate second messengers that mediate a broad range of cellular and physiological processes. From my prediction, it seemed that if there is a frameshift in the middle of the gene then the translation will resume its original frame (fig 24). Since this gene is encoded by one exon, this is a good candidate for an actual ribosomal frame-shift event since there were no known splice sites. While the chances of having two ribosomal frameshifts are low, it might still happen, especially if there is a strong mRNA secondary structure that could mediate this.

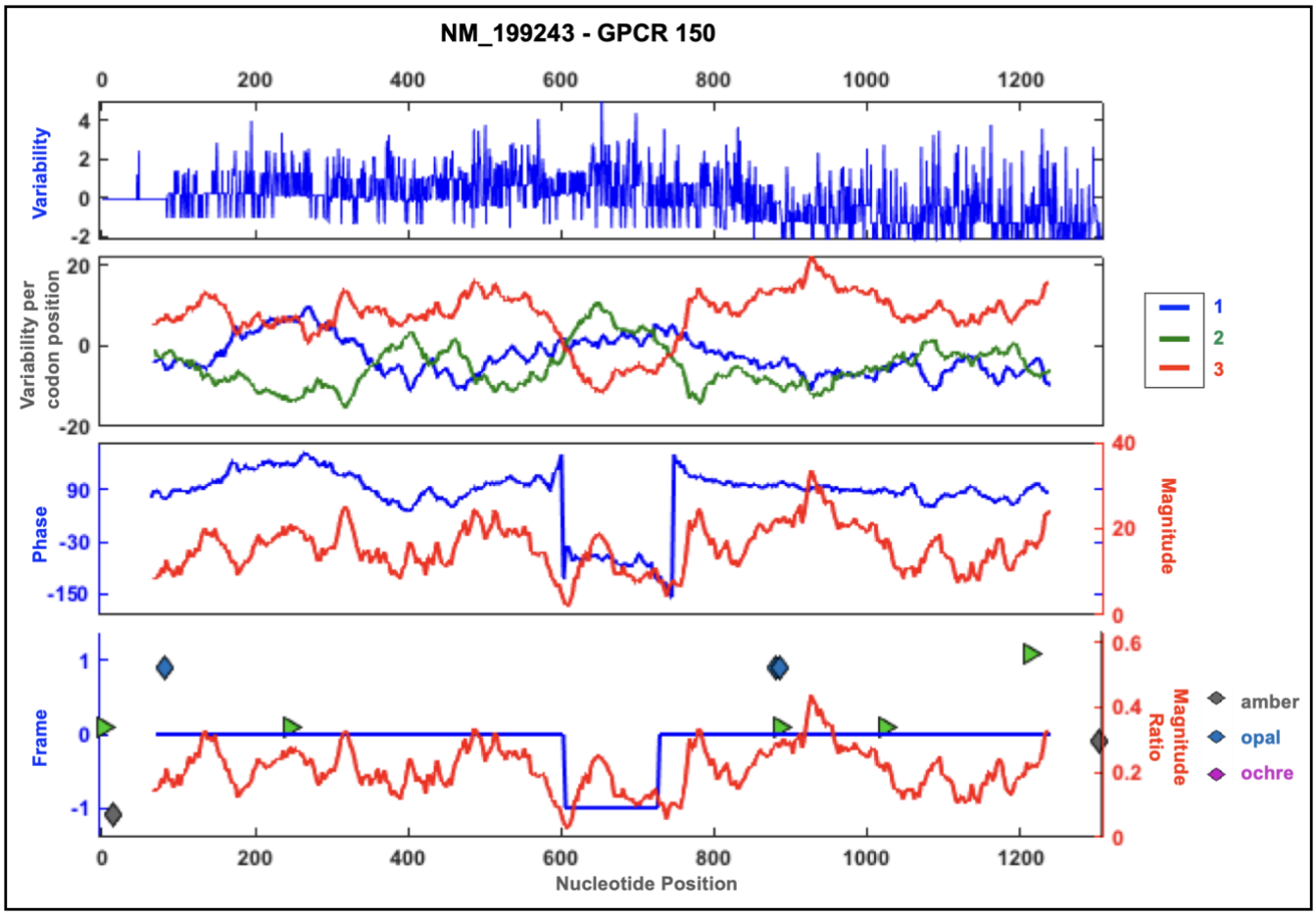


Figure 24: GPCR 150 frameshift prediction. (A) Analysis scheme as previously described where a -1 frameshift is evident but later the 0 frame is resumed. Top panel shows the variability scores for each nucleotide position. Second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). Third panel shows the raw output from the analysis – magnitude and phase. Bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account.

Examining the gene's free energy of mRNA secondary structure in sliding windows shows low energy areas in the predicted locations of frameshifts (fig 25).

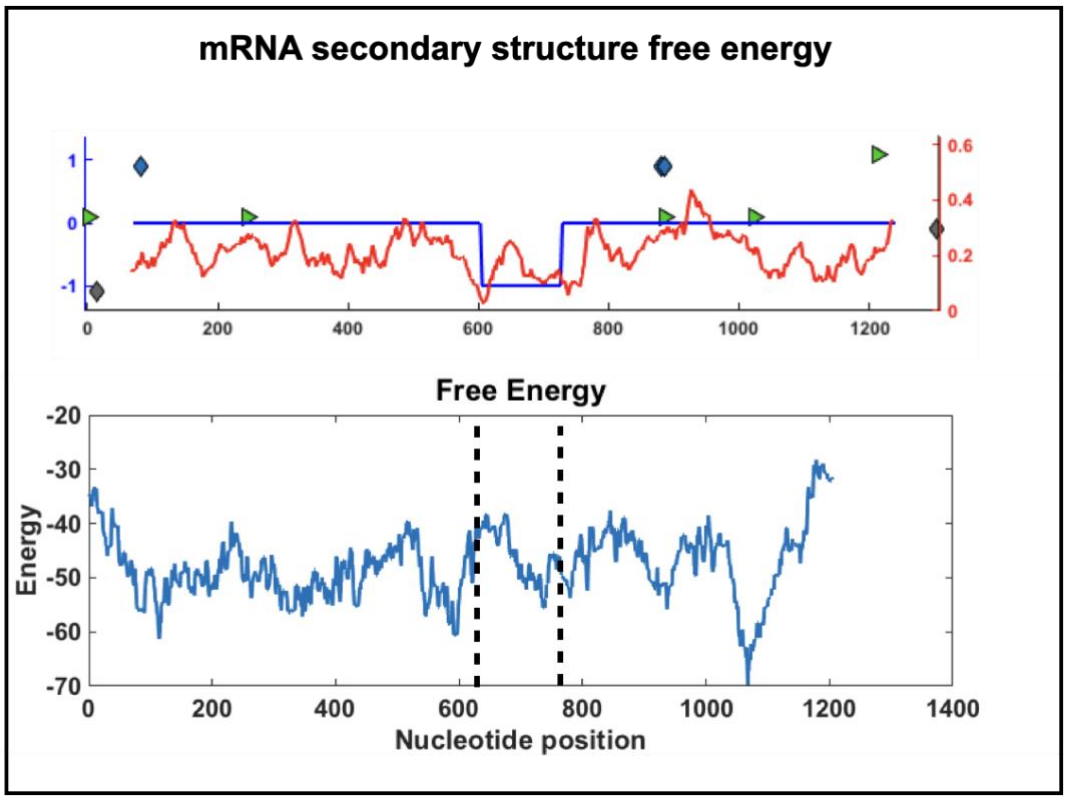


Figure 25: GPCR 150 frameshiftshown low mRNA secondary structure energy. mRNA secondary structure free energy showed that the regions of frameshift (both to and from the new frame) are inside low energy wells, indicating a strong secondary structure that could mediate the process. Black horizontal dashed lines mark the locations of predicted frameshifts.

Generally, GPCRs have seven transmembrane domains, an outer cellular binding domain, and a cytoplasmic signaling pathway. The extracellular binding domain binds to its compatible ligand to cause intra-membrane conformational change, generating a cascade of signaling events within the cell. Should the binding sites sequence change, the ligand that could bind to this GPCR might also change (Hilger et al., 2018). The predicted site of frameshift here was precisely on the ligand-binding domain. Using a ligand binding prediction server (Wass et al., 2010), I predicted that the ligand bound to the new sequence is altered (fig 26). To date, there is not much information on this GPCR specifically, and since this is an orphan gene, cross organism comparison is unavailable. Should this protein be implicated in some conditions, ligand binding should be tested to see the effect and understand the mechanism of action.

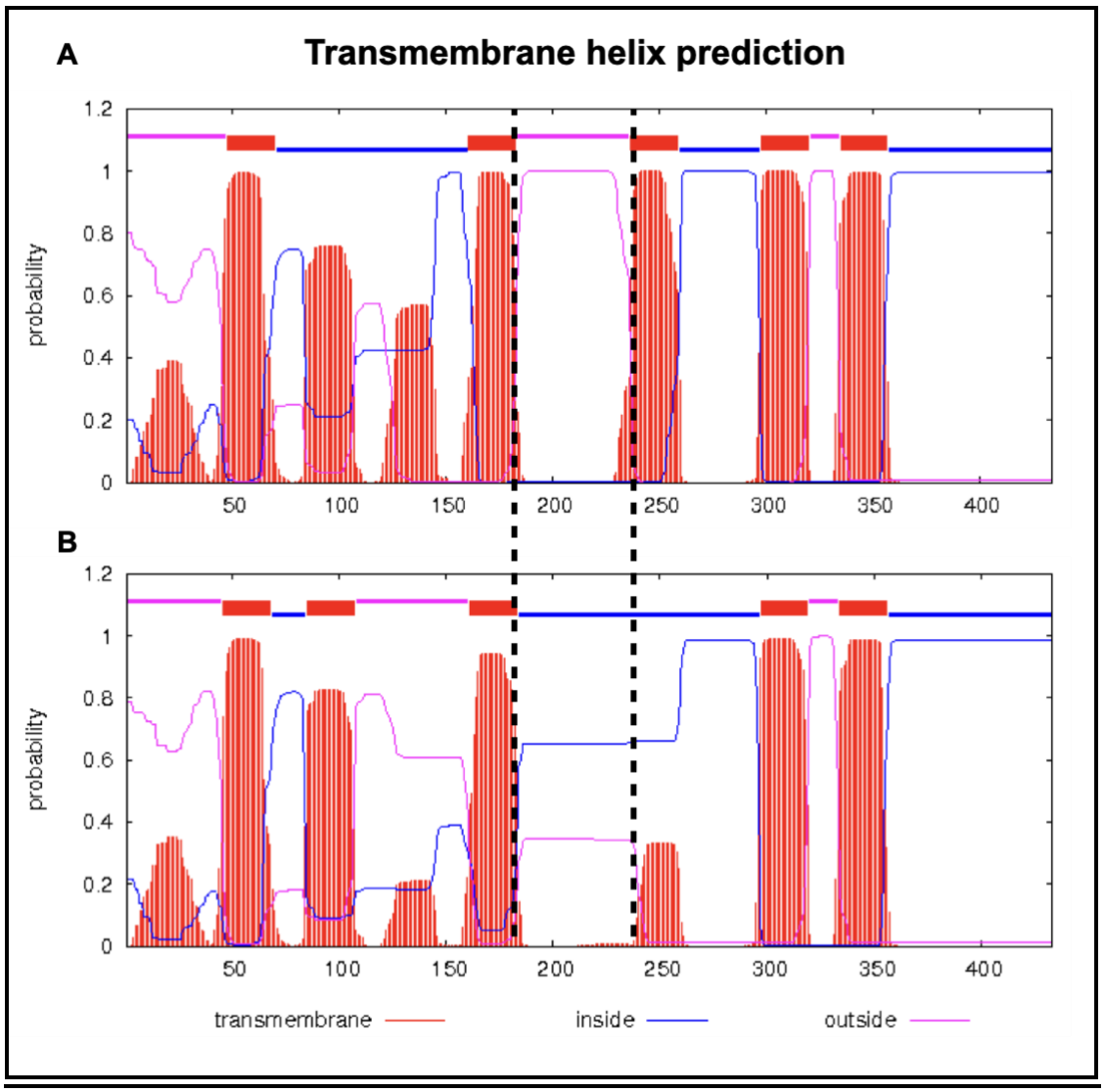


Figure 26: GPCR 150 frameshift could change binding properties. (A) Transmembrane helix predictions for the original protein shown the region predicted to be translated non canonically should be extracellular. **(B)** When using the same tool for the new protein version as predicted by my model, a alpha-helix has been lost resulting in an extracellular domain to become intracellular. Black horizontal dashed lines mark the locations of predicted frameshift. This area related to the ligand binding area as well. Using ligand binding prediction algorithm resulted in that the ligand predicted to bind to the new peptide is different than the one binding the original sequence, possibly affecting signaling.

8.3. Repetitive sequence signals

Some interesting signals arising from possible frameshifts seemed to be represented in my predictions. For one, having a frameshift may cause a sequence containing poly-Methionine (*Poly-Met*) signals. It was suggested that *poly-Met* signals might have effects on proteins' life-span (Giglione et al., 2003), and perhaps having these frameshifts are a generator of loss of function switch that can be altered using

ribosomal frameshift. The counter situation is also observed, where a poly-*Met* signal is lost due to a frameshift.

Another interesting signal represented is one where multiple STOP codons are encountered after a frameshift occurs. This might indicate that the frame-shifted sequence is very damaging and should not be translated. Perhaps this simply generates a truncated version of the protein, eliminating some functional domains in the original protein. In these genes, the alternative frame had several STOP codons that exceeded the 95 percentile of the distribution of the number of STOP codons in random sequences of the same length.

Lastly, I found cases where the translation of a protein is predicted to start in a different frame moving to the canonical one after the canonical start site is also interesting. This may indicate the presence of *upstream Open Reading Frames (uORFs)* not yet discovered and may uncover new functional domains or altered translation regulation pathways.

Methods for non-canonical protein translation detection in the human genome

1. Three-way periodicity analysis using multiple sequence alignments (building the main features for the models)

To utilize the pattern of the 3rd un-conserved position (3-way periodicity) of a nucleotide sequence across evolution, multiple-sequence alignments of orthologous genes were examined. Each set of orthologous sequences was aligned using *clustal-omega* standalone version 1.2.4 to produce a multiple sequence alignment (MSA), with the human sequence as the reference (Sievers et al., 2011). For each nucleotide position in the MSA, a variation score was produced using *rate4site* version 3, including gaps (Mayrose et al., 2004). The *rate4site* scores are calculated as the phylogenetic distance of every nucleotide position relative to the total distance between the specific gene analyzed on the species in the analysis.

Variation vectors were used to test for periodicities in un-conserved positions within codons. Each vector is analyzed in a "sliding window" manner. Each window consists of 45 codons (135 nucleotides), with a single codon (three nucleotides) gap to reach the following window. 45 codons were chosen to be the size of the window for these reasons:

1) This is the average size of a functional unit in a protein
2) To account for the minimal genetic variability in the data (~2%). This restriction will enable us to get a minimal periodic signal even under low genetic variability.
3) Using the average genetic variability in the data (~10%) and understanding that I wanted the highest resolution possible in frameshift location. This limited me to an error of at most 23 codons.

To eliminate low-frequency signals that may interrupt the algorithm, every window was normalized using a z-score to have a mean of 0 and a standard deviation of 1 in variation scores. In each window, three vectors were defined to represent the most variable sub-codon position in each possible translational frame (Fig. 27): The vectors are a summation of variability scores in each of the sub-positions of the codon:

(3)

$$frame\_0\_vec = sum(variability\_vec[3:3:end])$$

$$frame\_1\_vec = sum(variability\_vec[1:3:end])$$

$$frame\_-1\_vec = sum(variability\_vec[2:3:end])$$

Each vector had a predefined direction, and the three vectors were situated with an angle of 120 degrees between them. Negative variability scores (conserved positions) would result in the vector changing its direction with an angle of 180. The vectors were summed radially for each window to produce one resulting vector. This vector's length (*magnitude*) and direction (*phase*) would help determine the translation potential and frame of the window:

(4)

$$result\_vec\_x = frame\_0\_vec * \sin(90) + frame\_1\_vec * \sin(-120) + frame\_-1\_vec * \sin(-30)$$

$$result\_vec\_y = frame\_0\_vec * \cos(90) + frame\_1\_vec * \cos(-120) + frame\_-1\_vec * \cos(-30)$$

$$magnitude = \sqrt{result\_vec\_x^2 + result\_vec\_y^2}$$

$$phase = \text{atan}\left(\frac{result\_vec\_x}{result\_vec\_y}\right)$$

Gathering all resulting vectors for all windows along an MSA had constructed the *Gene Profile*. When the window is encoded with the 0 frame, and there is no apparent frameshift, the resulting vector would show a phase of close to 90 degrees and a high magnitude (depending on the overall variability of the gene), as shown in figure 27.

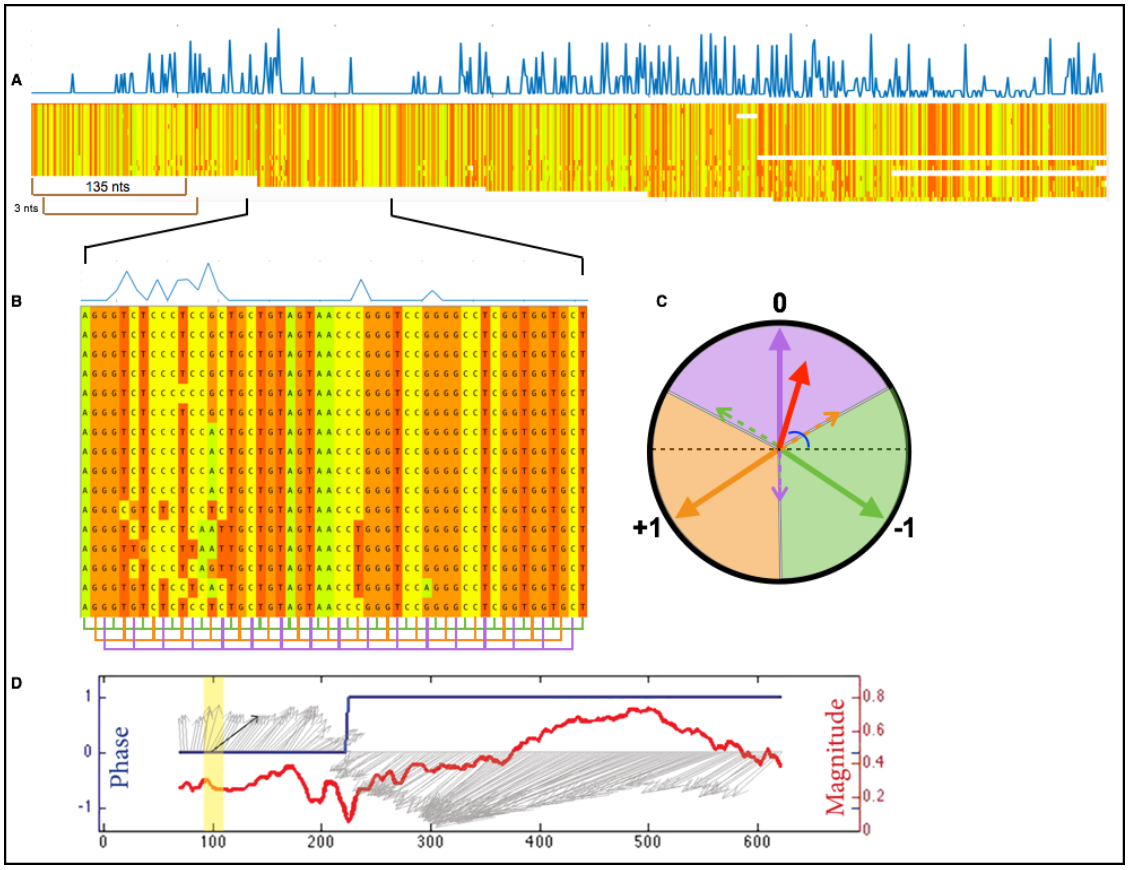


Figure 27: Periodic analysis of variability scores. The general analysis of each sliding window along the gene, to generate the gene profile for later use. **(A)** The MSA was analyzed in a single nucleotide position manner to produce the variability score. Construction of the gene profile as seen in **(C)** was done in sliding widows of 135 nucleotides with a gap of 3 nucleotides between consecutive windows. **(B)** For each window, the variability scores were summed creating three vectors representing the three different sub-codon positions. **(C)** Predetermining the direction of each sub-codons position final score, the vectors were radially summed to prduce the final windows' outcome – a single vector. This vectors length represents the magnitude (red) and it angle above (or below) the x=0 axis is the phase (blue). **(D)** Taking the rsulted vectors of all windows together constructed the gene profile having two properties: magnitude (in red) and phase (in blue).

Pseudo code for calculating the Gene Profile:

```
function CalcGeneProfile(V)
wavelet_magnitudes = []
wavelet_phases = []
for i = 1:3:len(V):
      curr_wavelet = V[i:i+135]
      S = sqrt(sum(v - mean(curr_wavelet) for v in curr_wavelet) / (len(curr_wavelet) - 1)
      standardized_wavelet = (v - mean(curr_wavelet)) / S for v in curr_wavelet
      frame_0_v = sum(standardized_wavelet(3:3:len(standardized_wavelet))
      frame_1_v = sum(standardized_wavelet(1:3:len(standardized_wavelet))
      frame_2_v = sum(standardized_wavelet(2:3:len(standardized_wavelet))
      X = frame_0_v * sin(90) + frame_1_v * sin(-120) + frame_2_v * sin(-30)
      Y = frame_0_v * cos(90) + frame_1_v * cos(-120) + frame_2_v * cos(-30)
      magnitude = sqrt(X^2 + Y^2)
      phase = atan(X/Y)
      magnitudes = [magnitudes, magnitude]
      phases = [phases, phase]
return magnitudes, phases
```

An optimization profile that shows the full translation potential in frame 0 for the specific window was constructed for each window. The vector was built by re-ordering all the variability scores in the current window in a manner that will produce the resulting vector with the highest magnitude and a phase closest to 90 (fig 28). The re-ordering was done by taking the most conserved position first, followed by the second most conserved, followed by the least conserved position, and so on. The ratio between the actual magnitude calculated and the “optimization profiles” magnitude helped find features of translation along the gene. While the optimization profile theoretically describes the conservation profile for a 0 frame translation, I could compare the observed result to find deviations. This magnitude ratio aided in discovering regions of mixed frames (transition windows), where the observed magnitude of the windows was lower than the theoretical optimization profile, as it possesses a mixture of frames creating destructive interference between the signals (Figure 28C-D).

Pseudo code for calculating the optimization profile:

```
1. function CreateOptimalWavelet(V)
2. sorted_wavelet = sort(V)
3. optimal_wavelet = []
4. while len(sorted_wavelet) >= 3:
5.        optimal_wavelet = [optimal_wavelet, sorted_wavelet[1:2], sorted_wavelet[end]]
6.        sorted_wavelet = sorted_wavelet[3:end-1]
7. return optimal_wavelet
```

```
 1. function CalcOptimalGeneProfile(V)
 2. wavelet_magnitudes = []
 3. wavelet_phases = []
 4. for i = 1:3:len(V):
 5.       curr_wavelet = V[i:i+135]
 6.       S = sqrt(sum(v – mean(curr_wavelet) for v in curr_wavelet) / (len(curr_wavelet) – 1)
 7.       standardized_wavelet = (v – mean(curr_wavelet)) / S for v in curr_wavelet
 8.       optimal_wavelet = CreateOptimalWavelet(standardized_wavelet)
 9.       frame_0_v = sum(standardized_wavelet(3:3:len(optimal_wavelet))
10.       frame_1_v = sum(standardized_wavelet(1:3:len(optimal_wavelet))
11.       frame_2_v = sum(standardized_wavelet(2:3:len(optimal_wavelet))
12.       X = frame_0_v * sin(90) + frame_1_v * sin(-120) + frame_2_v * sin(-30)
13.       Y = frame_0_v * cos(90) + frame_1_v * cos(-120) + frame_2_v * cos(-30)
14.       magnitude = sqrt(X^2 + Y^2)
15.       phase = atan(X/Y)
16.       magnitudes = [magnitudes, magnitude]
17.       phases = [phases, phase]
18. return magnitudes, phases
```

Since the goal was to explore the human genome in search of new translational options for protein-coding genes that deviate from the canonical way, before calculating the profiles, all positions in the MSA where the human gene had gaps in them were removed. This way should a frameshift occur; it would have to result in the human gene as well.

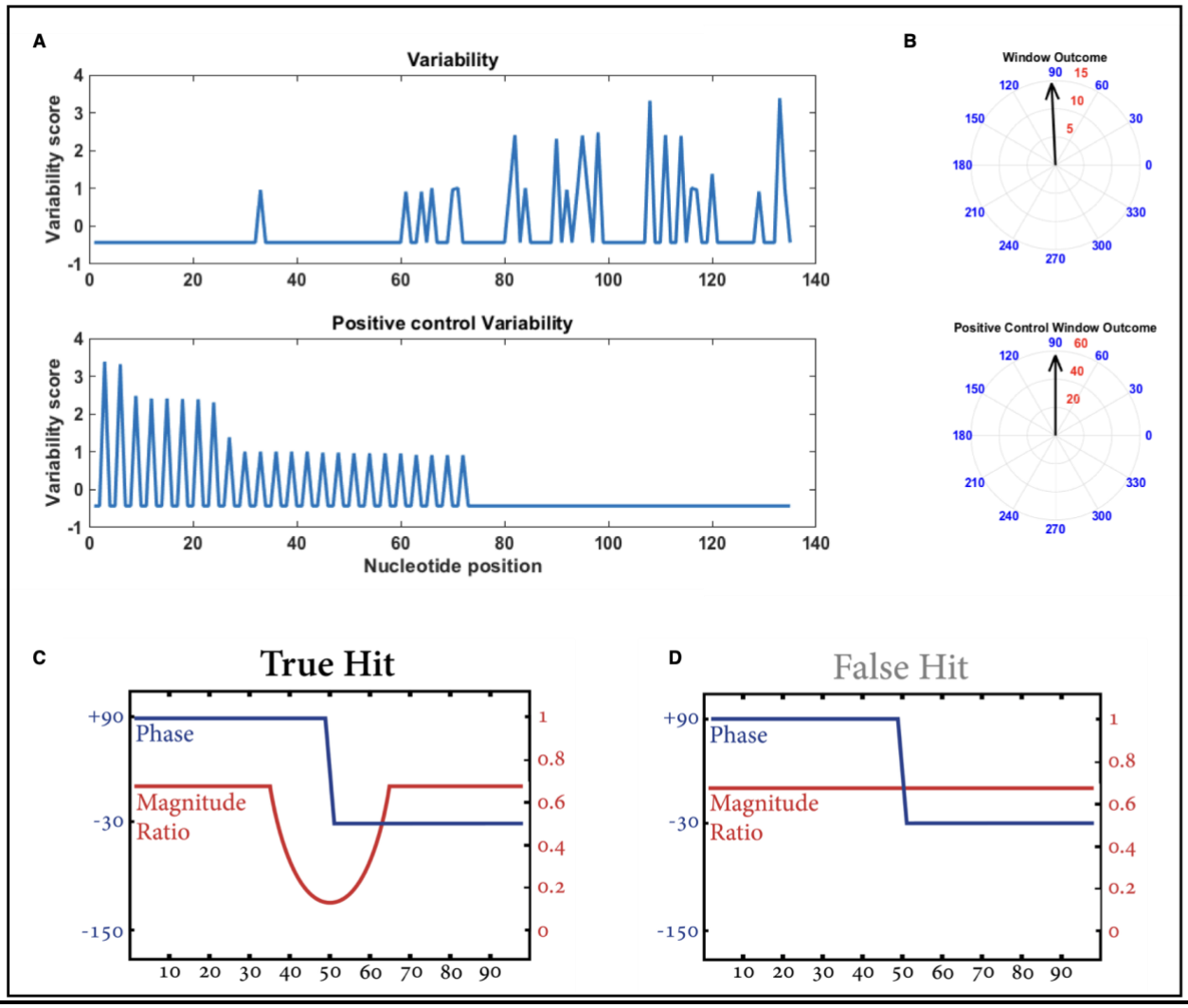


Figure 28: Eliminating false hits. (A) Constructing the optimization profile. The windows' scores were arranged in such a way that will produce the optimum 0 coding profile. **(B)** For both the raw scores and the optimal profile scores, the measure of magnitude and phase were calculated. **(C-D)** The magnitude ratio between the raw profile and the optimal profile helped in determining if an observed frameshift will be considered as a candidate **(C)** or is an artifact **(D)**.

2. Frame determination process

2.1. Rule-Based

For every window, the phase was placed within a range predefined by me (Fig 27C). The angle that the windows score presents determined this window's translational frame. This was applied to all windows independently.

Once all frames for a gene profile were calculated, their confidence was determined. This step helped eliminate frame determination when the periodic signal is not strong enough to decide. The confidence of a

window was determined by its magnitude ratio with that of the optimal profile and by its optimal profile magnitude. First, the ratio must have exceeded the 25 percentile of all ratios from all windows of the gene (thus, the period potential of the window is utilized). This threshold was selected so that there would not be stretches of uncertainty that are too long, thus affecting false-negative greatly. Should I have considered only very high confidence windows, I would be left with insufficient consecutive data to make a good enough interpolation, but using too low of a threshold would have inserted much noise into the smoothing process. Using the 25 percentile gave good results and accuracy on the known genes and also on the simulation data set. Second, the optimal profile magnitude must have exceeded the 25 percentile of all optimal profiles' magnitudes for all of the gene's windows. The two conditions require the specific window to both have a strong periodic pattern relative to the complete period pattern of the gene (first condition) and utilize the potential of periodicity within the gene (second condition). These were optimized based on simulated sequences, as will be described later.

After the non-confident frame determinations were eliminated, the gaps must be filled. This was done by simply interpolating values using the confident frame determinations. Next, the signal was smoothed and rounded to remove small stretches of frames (under ten consecutive windows) and to finally produce a frame that is one of [-1, 0, 1].

Pseudo code for rule based frame prediction:

```
1. function GetVFrame(phases)
2. frame = [];
3. frame[phases <= 150 & phases > 30] = 0;
4. frame[phases <= -90 | phases > 150] = 1;
5. frame[phases <= 30 & phases > -90] = -1;
6. return frame
```

```
1. function interpolate(X, Y, XQ)
2. YQ = []
3. for xq = XQ:
4.        i = find i such that X[i] < xq & X[i+1] > xq
5.        m = (Y[i+1] – Y[i]) / (X[i+1] – X[i])
6.        b = Y[i] – m * X[i]
7.        YQ = [YQ, m*xq + b]
8. return YQ
```

```
1. function InterpolateNonConfidentWavelets(magnitudes, phases, optimal_magnitudes,
                                            optimal_phases)
2. mag_ratio = magnitudes / optimal_magnitudes
3. high_confidence_wavelets = mag_ratio >= prctile(mag_ratio, 25) & optimal_magnitudes >=
                              prctile(optimal_magnitudes, 25)
4. frames = GetVFrame(phases)
5. frames[high_confidence_wavelets ==false] = null
6. frames = interpolate(find(high_confidence_wavelets == true),
                        frames[high_confidence_wavelets == true],
                        find(high_confidence_wavelets == false))
7. return frames
```

2.2. Gradient Boosting model

The second approach I had taken for frame determination was done using the CatBoost library (Dorogush et al., 2018) for classification; I was able to train a model for a windows' frame determination using the windows' calculated magnitude, phase, optimal profile magnitude and evolutionary distance as the features for the classification model.

Given that only a handful of samples known to be frameshifts exist, I needed to generate data for model training. I used the synthetic MSA generated from the COL1A gene, taking only those MSAs where the genetic divergence was 10%, and the number of sequences was 19, to match the properties of my actual data. The training set contained 80% of these MSAs; the validation set was the remaining 20%, including all samples with all genetic variation values and the number of sequences. The test set was all the MSAs generated using the HBA1 gene as described before. The model presented 92% accuracy in frame determination on the evaluation set and 88% accuracy for the test set.

The training was done at the single window level (not an entire gene as a sample), and I gave the label (coding frame) as I determined the exact location of the frameshift. I have also generated dual coding samples as a separate class. Overall, the classifier was trained as multiclass with 5 distinct classes (0, +1, -1, +1+0 (dual coding) and -1+0 (dual coding)). The predicted label for the next steps of frameshift detection was one of [0, +1, -1], meaning that mixture frame predictions were labeled as a pure frame for later use.

I examined the performance of the model relative to each of its features to find which feature drives mistakes and which aids significantly in confident predictions.

Figures 29-31 show the distributions of each of the four features used in this model for each label. The behavior seemed to be consistent between all frames. The phase (as expected) was a driving force for a prediction in each class (Figs 29A, 30A, 31A). It was visible that both true positives and false positives fall under the same range of values, while the negatives (true and false) were dispersed throughout the spectrum. The magnitude, however, completed the story. As can be seen in figures 29C-D, 30C-D, and 31C-D, false positive and false negative were driven from samples where both the magnitude is low, and the positive magnitude is low. This means that there isn't enough periodic information encoded within these samples, and the confidence should be low. Lastly, I could also see that false negatives and false positives were driven by low genetic variation, correlating with the fact that not enough variation information was encoded in those samples (Figs 29D, 30D, 31D).

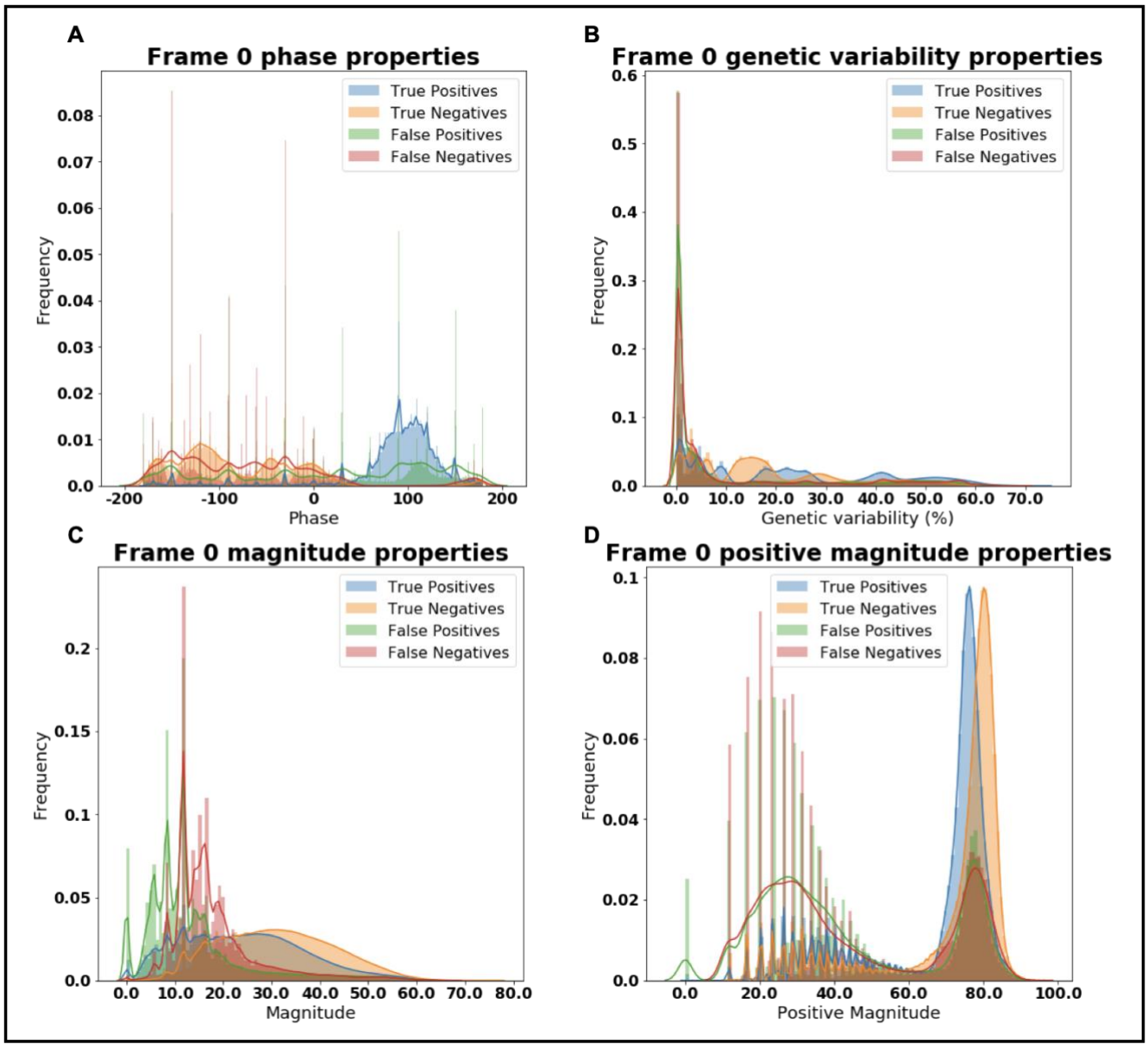


Figure 29: Features values distribution for the 0 frame label. To understand the driving forces of my gradient boosting model, I examined the value distribution of each model feature in each label (frame). The compared groups are true positive (label=prediction=frame), true negatives (label=prediction!=frame), false positives (label!=prediction, prediction=frame) and false negatives (label!=prediction, label=frame). **(A)** Phase values distribution for all groups. It was evident that the true and false positives fell under the same range, differing from that of true and false negatives. **(B)** Genetic variability values describing the level of conservation in the analyzed sequence. Low variability drove mistakes as not enough variability information was encoded to drive periodicity. **(C)** Magnitude values. Low magnitude drove mistakes. Low magnitude could be a result of low variability or destructive interference, when there is more than one frame encoded (like in transition windows). **(D)** Positive magnitude values. This was a measure of potential for detection, and was a direct measure of variability, so it came as no surprise that here also, low values drove mistakes.

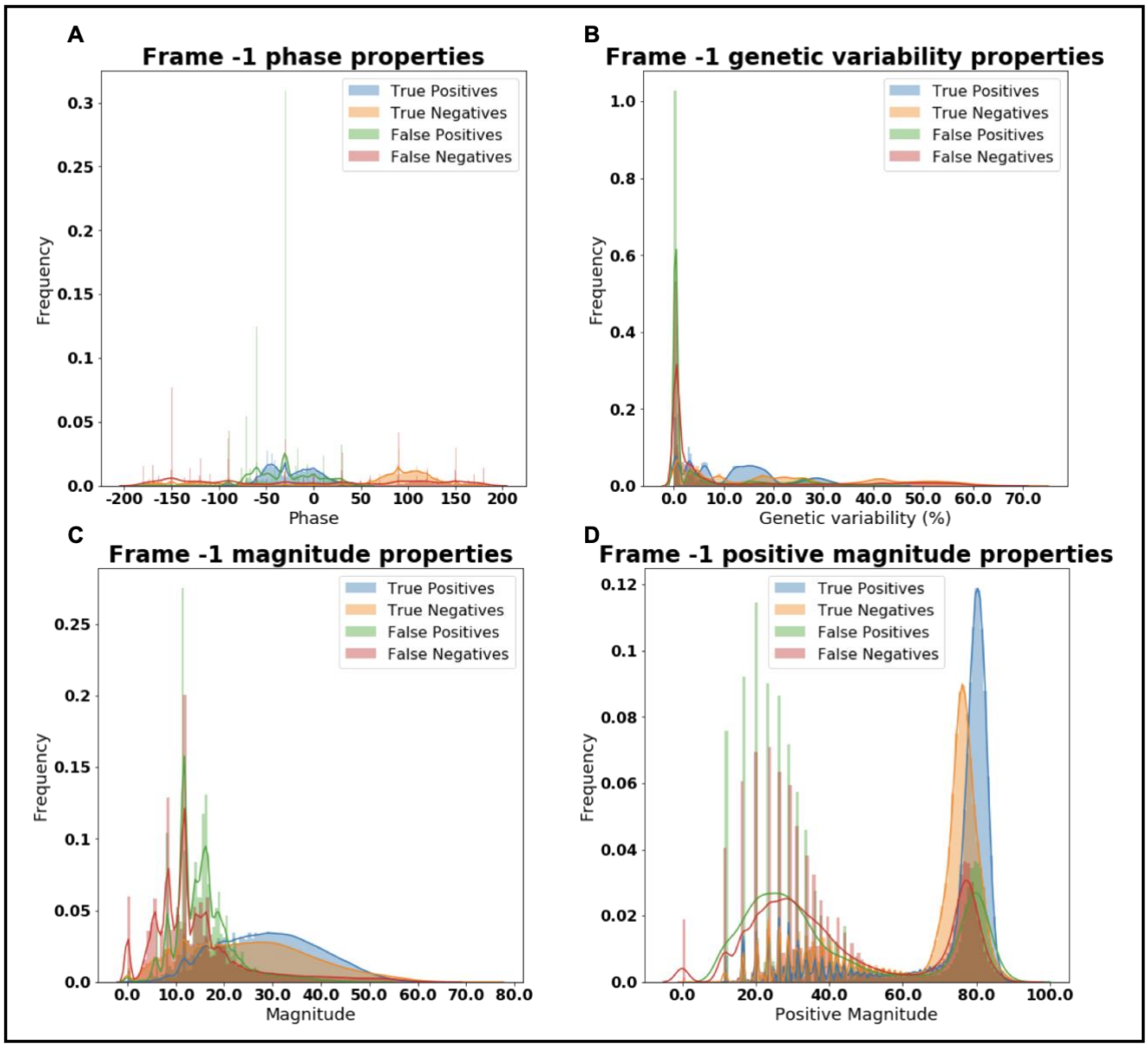


Figure 30: Features values distribution for the -1 frame label. To understand the driving forces of my gradient boosting model, I examined the value distribution of each model feature in each label (frame). The compared groups are true positive (label=prediction=frame), true negatives (label=prediction!=frame), false positives (label!=prediction, prediction=frame) and false negatives (label!=prediction, label=frame). **(A)** Phase values distribution for all groups. It was evident that the true and false positives fell under the same range, differing from that of true and false negatives. **(B)** Genetic variability values describing the level of conservation in the analyzed sequence. Low variability drove mistakes as not enough variability information was encoded to drive periodicity. **(C)** Magnitude values. Low magnitude drove mistakes. Low magnitude could be a result of low variability or destructive interference, when there is more than one frame encoded (like in transition windows). **(D)** Positive magnitude values. This was a measure of potential for detection, and was a direct measure of variability, so it came as no surprise that here also, low values drove mistakes.

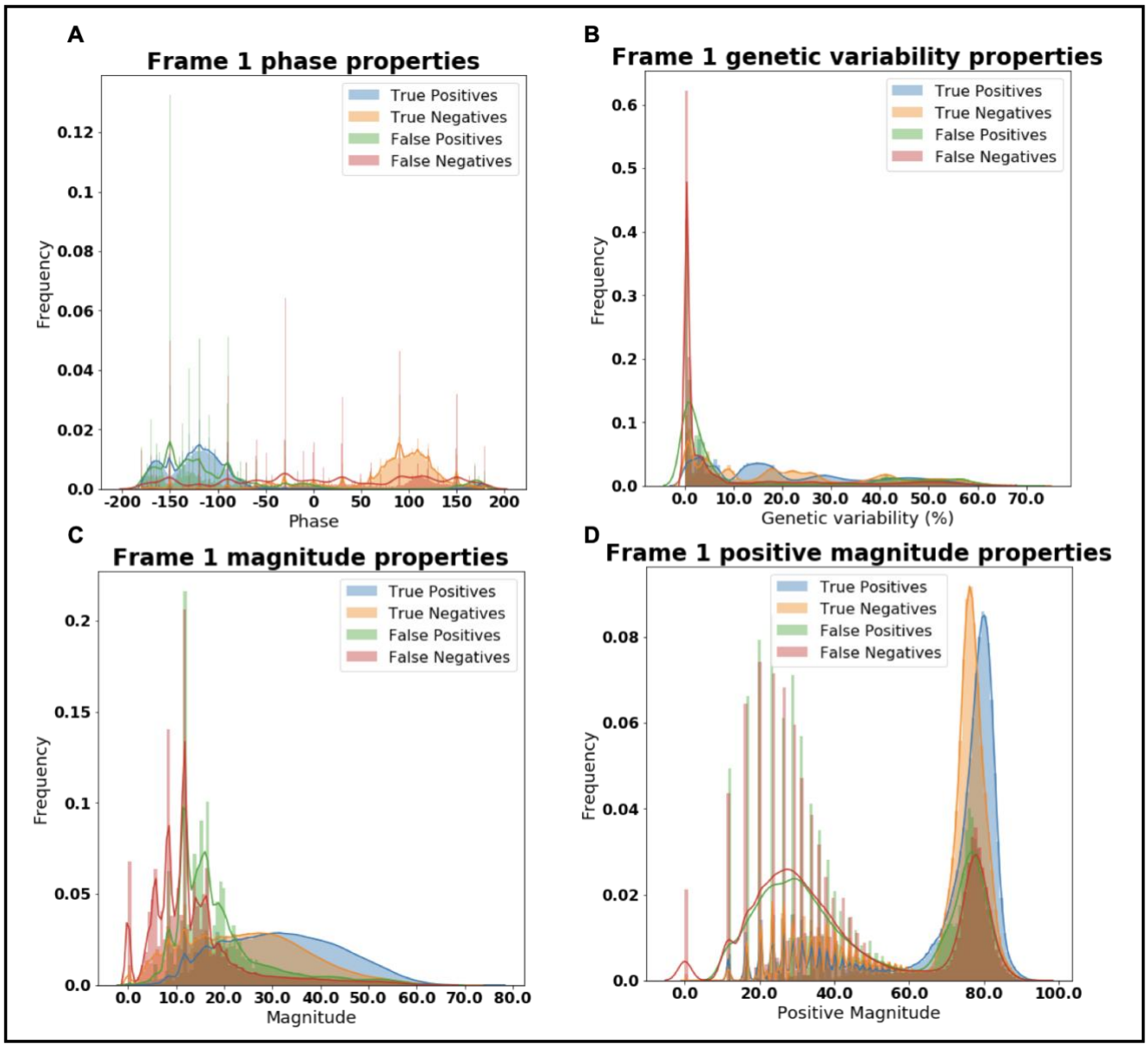


Figure 31: Features values distribution for the +1 frame label. To understand the driving forces of my gradient boosting model, I examined the value distribution of each model feature in each label (frame). The compared groups are true positive (label=prediction=frame), true negatives (label=prediction!=frame), false positives (label!=prediction, prediction=frame) and false negatives (label!=prediction, label=frame). **(A)** Phase values distribution for all groups. It was evident that the true and false positives fell under the same range, differing from that of true and false negatives. **(B)** Genetic variability values describing the level of conservation in the analyzed sequence. Low variability drove mistakes as not enough variability information was encoded to drive periodicity. **(C)** Magnitude values. Low magnitude drove mistakes. Low magnitude could be a result of low variability or destructive interference, when there is more than one frame encoded (like in transition windows). **(D)** Positive magnitude values. This was a measure of potential for detection, and was a direct measure of variability, so it came as no surprise that here also, low values drove mistakes.

To construct the full gene profile, I gathered the windows from the gene and applied post-processing of the classification results in a manner very similar to the one described for the rule-based frame determination. The main difference between this method was confidence determination. While in the rule-based method, I used the

magnitude scale as a measure of confidence, here I used the classifier's confidence in the form of prediction probability for the selected class. I used the frame inferred by the classifier only if its' probability exceeded the 25% of all prediction probabilities from all windows across the gene.

The overall analyses presented in this thesis apply to the rule-based frame inference, but the boosting method served as a means of validation for some of the results.

3. Translation frameshift calling (rule-based determination)

Once the translation frame for each position in each gene is deduced, genes were selected for further exploration using the following rules (Fig 32):

- Only two distinct frames were observed in the sequence
- No more than two frameshifts were detected
- A frame was considered only if it holds for at least 20 consecutive sliding windows
- A transition was considered successful only if it holds the characteristic magnitude ratio "valley" as shown in figure 28C. This is due to the fact that while a frame is changing, the analyzed sliding windows hold a mixture of frames, affecting the amplitude in each active frame. Once the window is re-stabilized in the new frame, the magnitude should recover to a baseline representing the new frame's amplitude.
- *BLASTp* search of the new peptide arising from frame change showed no hits (E-value < 10^{-2}) against the human genome, suggesting a novel translation option, eliminating known splice variants that appear frameshifted relative to the main transcript.

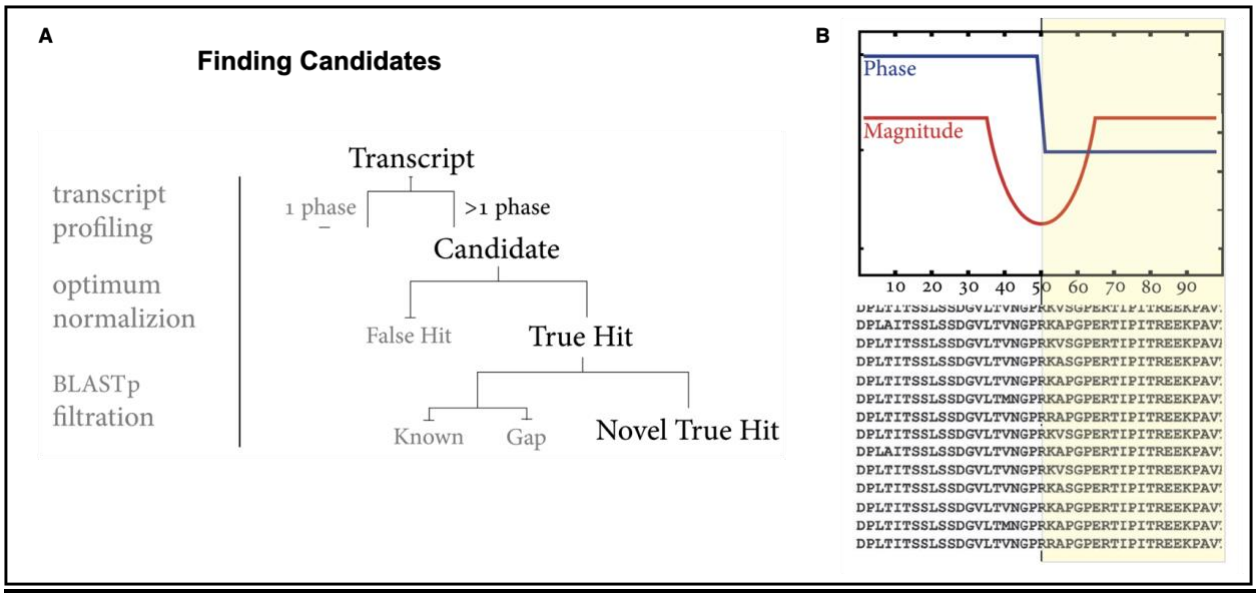


Figure 32: Determining final candidates for possible non canonical translation in an alternative frame. (A) The process of finalizing a candidate started with the gene having more than a single translational frame, but not having more than two. Next the transition windows were examined for the existence of the signature magnitude change. Lastly, blasting the conceptual translation of the frameshifted sequence **(B)**, should have produced no hits from the human proteome, eliminating known cases, and alternative transcript variants.

3.1. Determining candidates

Enforcing the first rule of having no more than 2 distinct frames gave the first list of candidate genes. The threshold of at least 20 consecutive windows maintaining the same frame was set to accept for a region that transitioned rather than signal noise. Since the probability of having a frameshift in an annotated transcript should be low, only those profiles that present a single transition to a new translation frame were chosen.

The area of frame transition was examined as a second step. If indeed translation shifted from one frame to another, when analyzing windows that include codons from the region before the frameshift as well as codons from the region after, destructive interference of the magnitude signal (as periodicity gets lost) should have appeared. Once the frameshift sequence was no longer included in the window, a single frame is re-instated and this interference should have seized. This pattern of magnitude change along the sequence was important for filtering out false frameshifts whose prediction could result from noise in the variability vector.

4. mRNA secondary structure signature analysis

To evaluate the secondary structure energy signature from mRNA, 101 nucleotides from each gene predicted to have a frameshift were extracted, surrounding the frameshift location (50 nucleotides up and downstream from the FS location). The sequences were grouped as having -1 or +1 frameshift and each group was analyzed separately. For each sequence, the *rnafold* function from the *ViennaRNA* package 2.0 (Lorenz et al., 2011) was applied. The folding energy for sliding windows of length 20 nucleotides with a step of 1 nucleotide between consecutive windows was calculated. The resulting energy vectors were clustered into 16 groups using the *k-means* clustering algorithm (Dubes and Jain, 1980). The decision regarding cluster number was done by inspecting both within-cluster sum of squares and silhouette scores, which are standard in cluster number determination when using the k-means clustering (Rousseeuw, 1987). Of these, gene groups where the average profile contained an "energy well" with a minimum of lower than -5 were chosen. Time warping (Sakoe and Chiba, 1978) of the profiles was then applied to the coordinates of the energy minimum to produce a meta profile. This was compared to the profile calculated on the human gene Ornithine decarboxylase Antizyme 1 (OAZ1), which was the only known gene in the human genome that undergoes +1 programmed ribosomal frameshift. This was compared to the same analysis on the sequences after random shuffling of the codons. The shuffled profiles did not show any structure at all and the energy level remained constant (figure 18).

5. Simulating sequences for prediction evaluation

The human collagen coding gene (COL1A – NM_000088) sequence was mutated to simulate the evolution of many sequences in different evolutionary distances in the range of 0.5%-50%, and a different number of sequences ranging from 5 up to 19. Simulated mutations were generated randomly by changing nucleotides while maintaining amino acid substitution frequency determined by BLOSUM matrices. Each set of parameters (mutation rate and the number of sequences) was used to generate 100 groups of sequences that were later aligned to produce a *Multiple Sequence Alignment (MSA)* along with the original COL1A to

calculate the variation signal. The number of times where the algorithm detected a frameshift where non was present was used to evaluate the false positive rate.

The abovementioned process was repeated but instead of using the original COL1A sequence a specific nucleotide was chosen and deleted to generate a -1 frameshift; or added a random nucleotide to generate a +1 frameshift. The number of times the algorithm was successful in detecting the frameshift, in the right location, for each evolutionary distance was counted to produce true positive rates.

Lastly, a region within the COL1A gene where there is a substantial lack of STOP codons in a frame different from the canonical was chosen, and mutations were generated that keep the amino acid substitution fidelity in both the canonical frame and the chosen second frame (to simulate dual coding). Counting the number of times where the algorithm could detect the mixture of correct frames in the correct location is used to evaluate true positive rates.

To create the test set for the synthetic MSA, a similar process was done for the human hemoglobin subunit alpha coding gene (HBA1 – NM_000558). The main difference was that the sequences were simulated to have a genetic variability of 0.1 (10%), and the number of simulated sequences per MSA was 19.

6. Sequences, orthologous groups, and alignments

All mammalian sequences used were downloaded from NCBI multiple sequence alignment of 20 mammals (16 primates) database (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way/). Sequences were stripped from gaps and then re-aligned using *clustal-omega* stand-alone version 1.2.4 (Sievers et al., 2011).

For COVID 19 sequence analysis I used 44 Sarbecovirus genomes that were selected starting from all betacoronavirus and unclassified coronavirus full genomes listed in NCBI on 5-Mar-2020. I excluded any and all sequences that differed in more than 10,000 positions in pairwise alignments, including all SARS-CoV-1 and SARS-CoV-2 genomes other than the reference sequence NC_045512.2.

7. Validation using Dual reporter essay

Using a construct with two fluorophores residing in different translation frames, and an interchangeable sequence between them, one can experimentally test whether the predicted regions from our algorithm indeed undergo translational frameshift (Mikl et al., 2018). In Mikl et al., a set of sequences showing a high probability of ribosomal frameshift have been already tested irrespectively of the work done in this thesis. Unfortunately, no overlap was found between our subset of sequences and the sequences tested.

8. Ribosome foot-printing as validation for predictions

Ribosome profiling is an experimental procedure where translating mRNAs are treated with nucleases that degrade unprotected fragments of the mRNA. mRNA-ribosome complexes were isolated and treated with different compounds to detach proteins from RNA, and the resulting fragments were then sequenced. Using computational methods and known characteristics of mammalian ribosomes, the exact codon location of the ribosome translation stage could have been extracted (Ingolia et al., 2009b). Since translating ribosomes should present a 3-way periodic pattern of density along mRNAs, as they must maintain the codon-wise translation, the same rules of periodicity analysis that I presented above for periodicity in conservation, should apply to this type of data as well. Sparseness was taken into consideration by changing the sliding window size from 135 nucleotides to 180. The same algorithm for frame prediction was used in this case, as described before.

Summary of Chapter 2:

In this chapter, I showed a novel computational approach that can detect regions in the genome that may be translated in a different coding frame than the canonical one. There have been other studies using spectral approaches that aim to predict coding regions using the periodic pattern of the conservation across a gene (Kotlar and Lavner, 2003). The novelty of my approach is translated into a few key added values: First, this approach is not aimed at finding coding regions but rather aimed to find a deviation from a canonical translation dogma with high resolution. So, in my approach, I focused more on the phase of the spectral analysis rather than the amplitude. Phase analysis was sometimes not an easy thing to do since it possessed much more noise than the amplitude, and therefore harder to give an accurate estimate. Furthermore, since the method presented in (Kotlar and Lavner, 2003) focused on short gene discovery, they emphasized a single measure, not allowing them to account for any deviations. Moreover, should there be a deviation from the canonical translation frame, as presented in this work, their signal will be impaired due to destructive interference, presenting a low signal and false results. The second, and very important, is the fact that my approach "simplified" the more complex discrete Fourier transform, allowing me to capture all the harmonies of a specific periodicity in one shot calculation. Since the 3-way periodic pattern could appear in any multiplicity of 3 within the window analyzed, using the approach presented I summarized all the harmonies at once, reducing noise and complexity. Lastly, as presented in this thesis, I could generalize my approach to account for different conservation levels, different data sources (not only sequences but also ribosome foot-printing data), and with slight modifications even more periodicities, and the adaptation was simple and approachable.

My analysis revealed hundreds of new functioning units that may have major impacts, and possibly unravel more regulatory elements and processes that have yet to be discovered. Ribosome profiling data could give even stronger indications of the predicted genes. I showed a strong proof of concept by successfully identifying the known gene that undergoes ribosomal frameshift or has frameshifted transcripts, as well as its location with high

accuracy. By finding a short list of candidate genes I laid some solid ground for further investigation of specific candidates.

The fact that I was able to modify the algorithm, slightly to apply it to different types of databases (MSA conservation scores for different data sources (mammalian, viral) vs. P-site locations) brings the high potential for further generalizations and even more databases that could be explored.

Not only could my model be used to explore different types of data to discover translational frames, but it could also help in determining translational potential by using the magnitude calculated as the strength of the periodic pattern in a specific region. This may help by exploring untranslated regions in the genome, suspected to undergo translation under some conditions.

The entire pipeline presented was able to predict a frameshift event in a SARS-Cov-2 gene, that was later given additional support in an unrelated paper (Firth, 2020). The fact that this was established using both the rule-based method and the trained model, gives additional ground to using the predictions presented in this thesis as grounds for further testing.

I showed an example of a protein, METTL27, where the predicted altered translation is found as a stand-alone transcript in many close species. This may indicate an evolutionary process that has evolved with humans to incorporate two versions of a protein under one transcript, with some regulation. It might also be that the annotation today is wrong, and the actual transcript holds a frameshift during splicing. Either way, these revelations must be further explored.

Lastly, by simulating many genomic sequences I was able to not only help in determining success and false-positive rates but also use them as a training set to build a solid classifier for better prediction and generalization. Furthermore, since I could create simulated cases for many different combinations, I could train this model to also distinguish between single-frame coding and dual-frame coding and to potentially distinguish between frameshift events and alternative ORF encoded under the same locus.

It is also possible to use these simulated sequences to train another model, on top of the frame classification model, that can distinguish between solid frame sequences, and frame transition sequences, potentially adding higher accuracy to the frame prediction pipeline. By having high confidence in areas where the sequence starts changing its frame I could both reduce the

false-positive rates and increase the accuracy of the frameshift location prediction.

As more and more such events are revealed, the model can continuously be trained to gain better accuracy and present more and more high potential candidates presenting such noncanonical behaviors, in a variety of organisms.

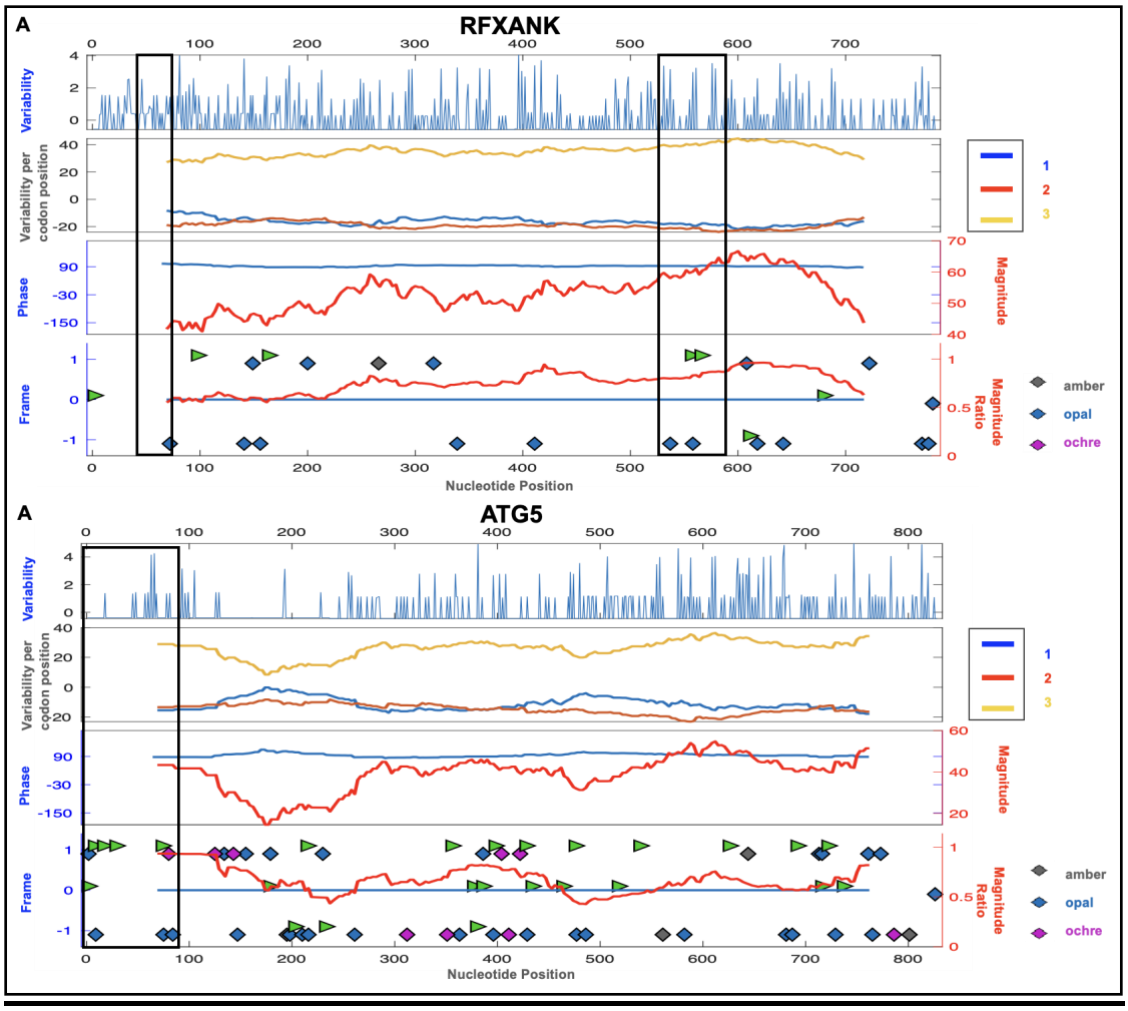
Supplemental Figures for Chapter 2:

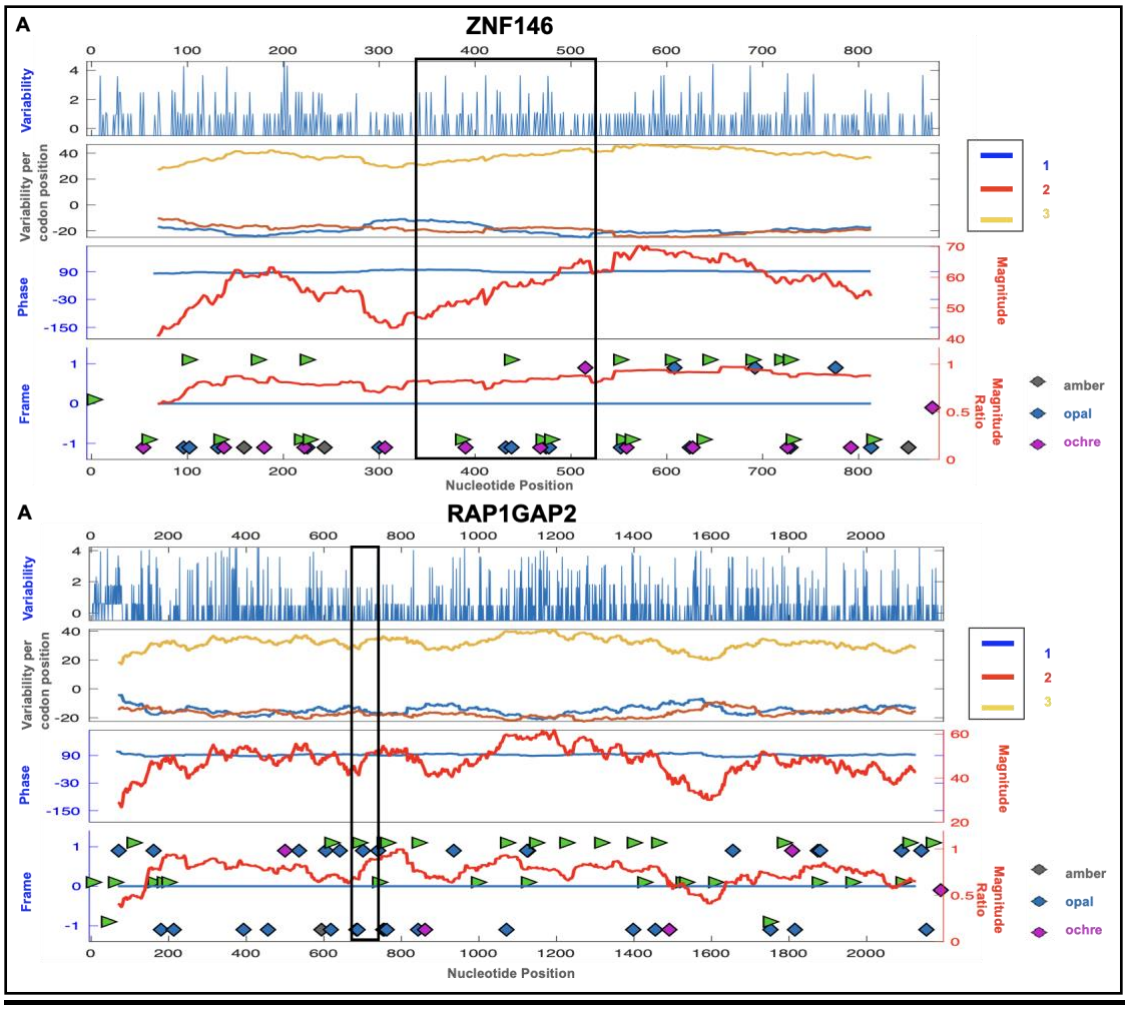
The following figures show a comparison between the results reported in (Michel et al., 2012) as novel findings of noncanonical encoding, and the profiles as calculated by the algorithm presented in this chapter.

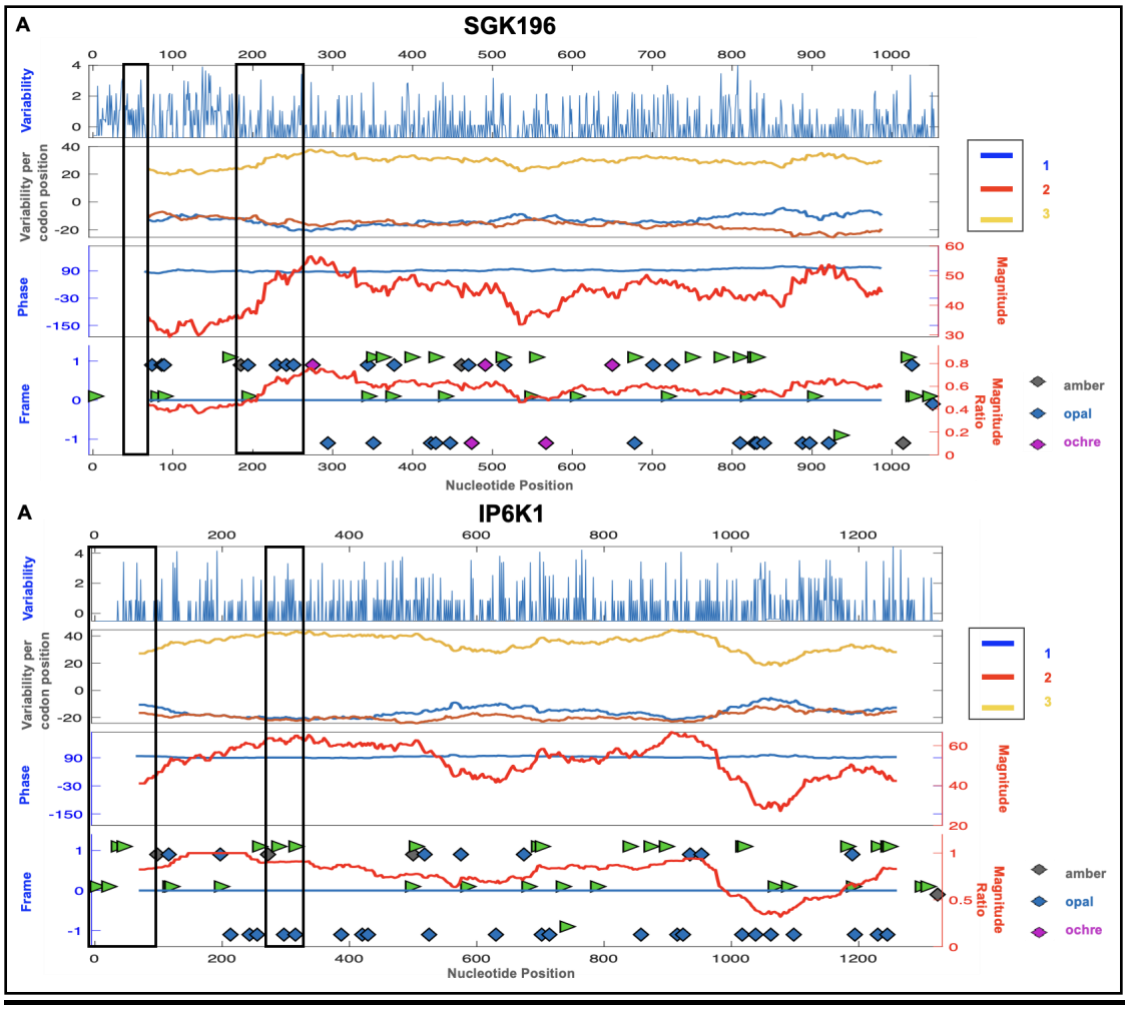
Black boxes mark areas where (Michel et al., 2012) report high RPF counts in the second sub codon position. They did not supply further insight other than graphical representations of their finding, so the comparison is qualitative.

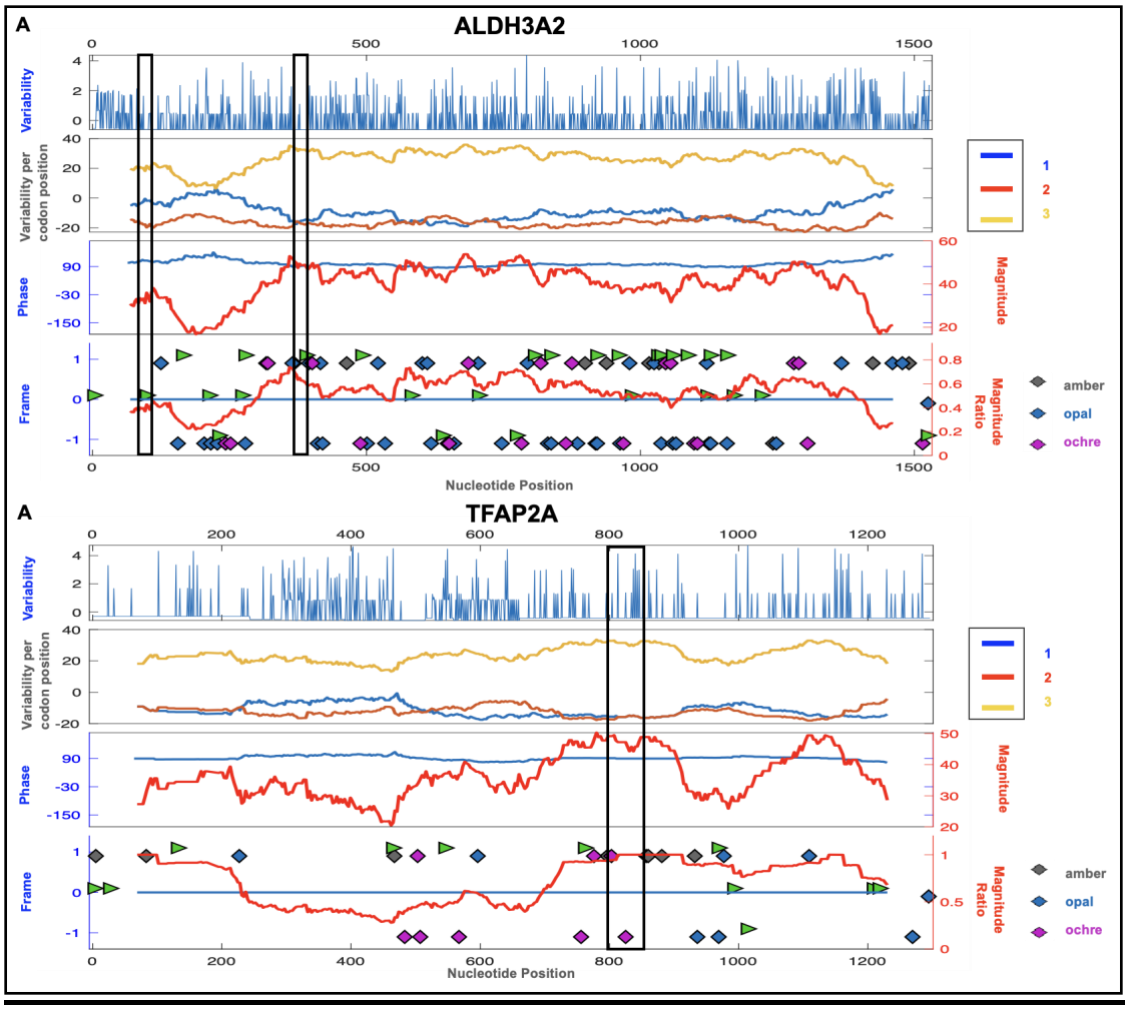
General legend for all supplemental figures:

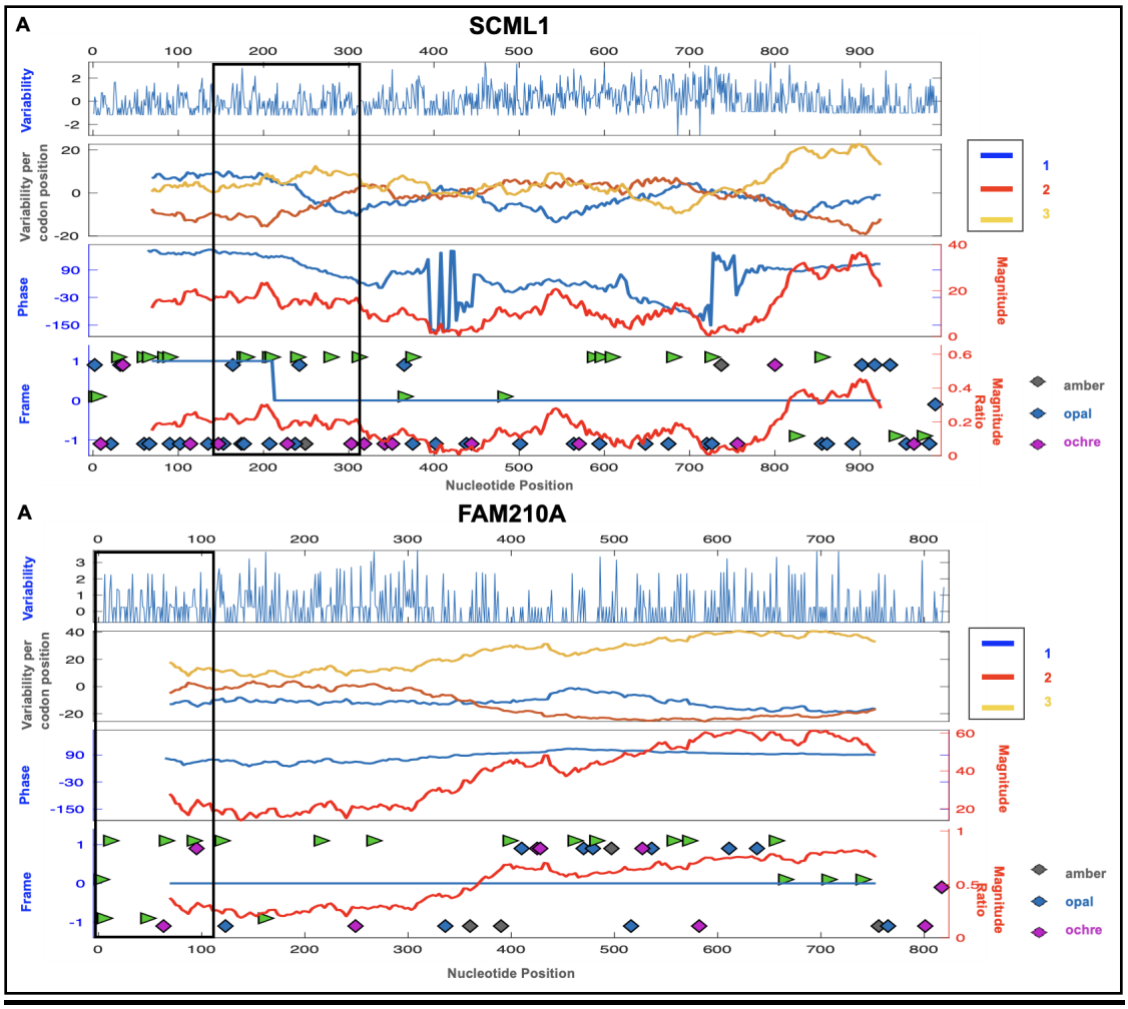
The top panel shows the variability scores for each nucleotide position. The second panel shows the variability per sub-codon position for each sliding window along the gene (summation of all sub-codon positions of that window). The third panel shows the raw output from the analysis – magnitude and phase. The bottom panel in the actual data used for frameshift detection – Inferred translational frame and magnitude ratio taking optimum normalization into account.

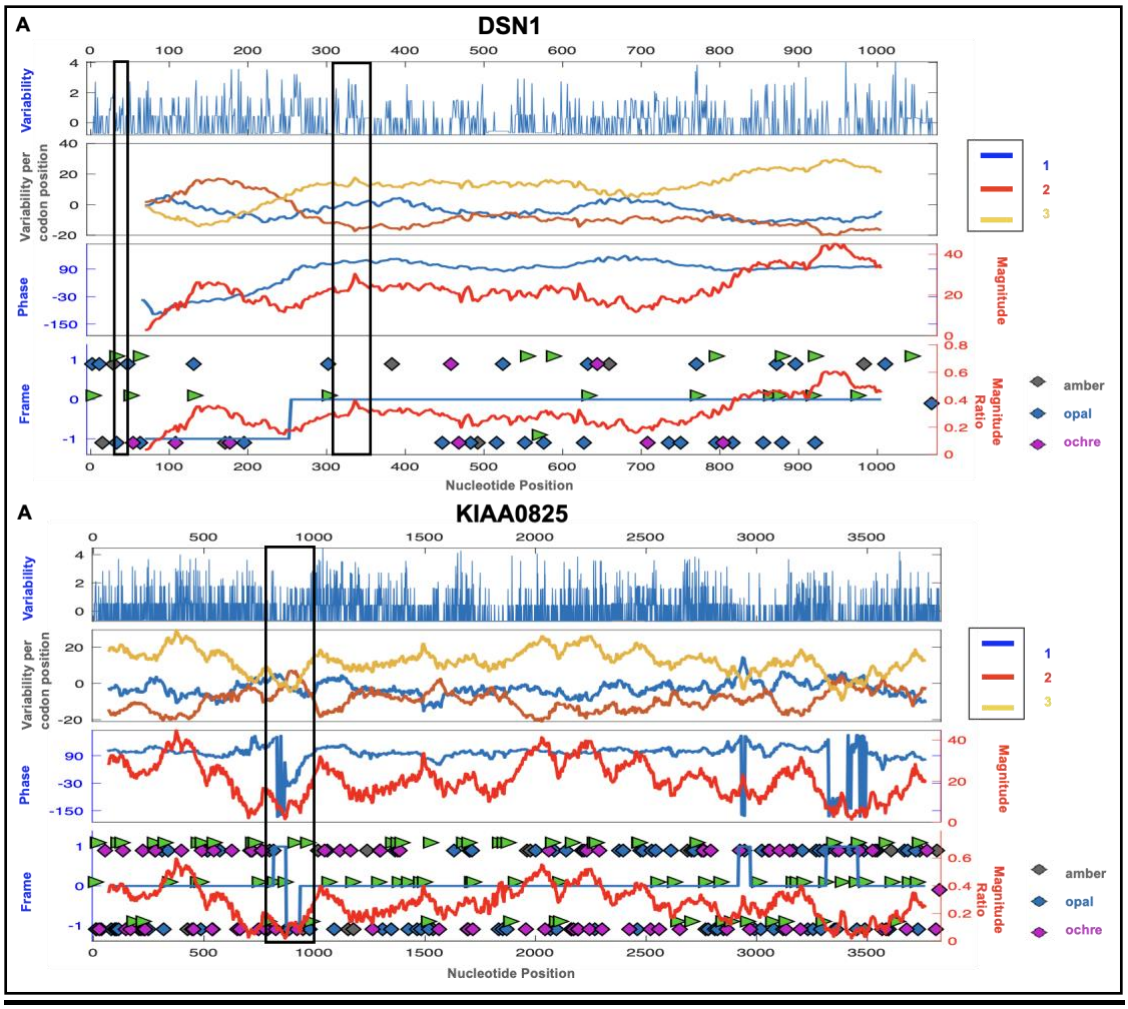


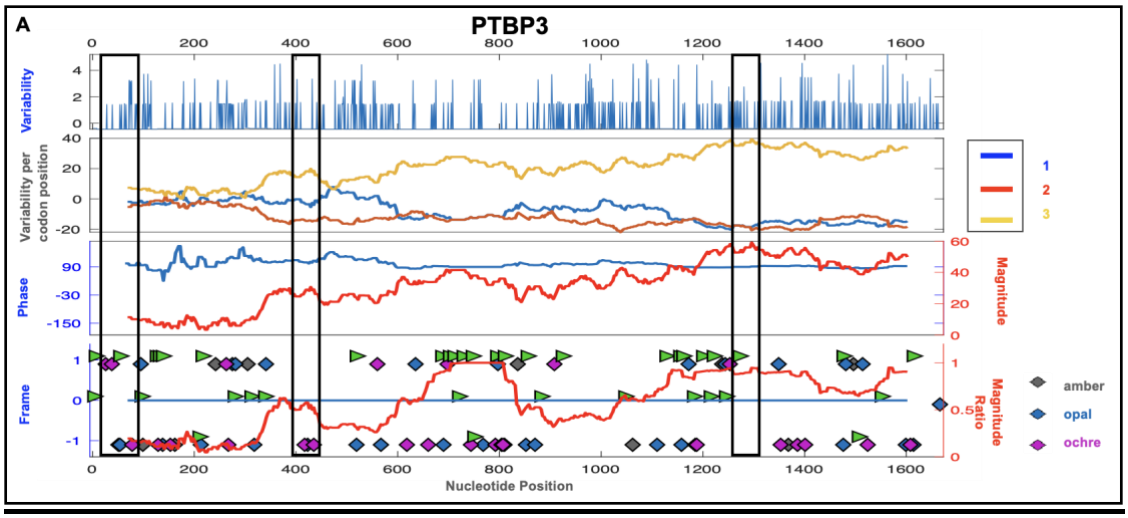

A
ZNF146
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position
A
RAP1GAP2
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position


A
SGK196
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position
A
IP6K1
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position


A
ALDH3A2
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position
A
TFAP2A
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position


A
SCML1
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position
A
FAM210A
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position


A
DSN1
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position
A
KIAA0825
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position


A
PTBP3
Variability
Variability per codon position
Phase
Frame
Magnitude
Magnitude Ratio
1
2
3
amber
opal
ochre
Nucleotide Position

Appendix 2.A: Data used to validate the model for predicting frame shifts

Gene Name	Data Source
OAZ1	https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way
PEG10	https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way
HIV gag-pol	https://www.ncbi.nlm.nih.gov/nuccore/166025821
HPV E2	https://www.ncbi.nlm.nih.gov/nuccore/X94164.1?report=fasta&from=2719&to=3873
SARS-Cov-2	https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2

Appendix 2.B: List of all genes predicted to have a frame shift

Following is the list of predicted frame shift gene names along with the predicted novel peptide sequence. The header for these fasta records is coded as:

>gene_name FS_event_# frame (1,-1)

For example:

Gene name	FS event number	Predicted new frame
NM_000328	1	1
NM_001002251	1	-1
NM_001002251	2	1

Some genes were predicted to have more than a single FS events, thus both novel peptides will be displayed

>NM_000328 1 frame 1
IAVPQEMYCRGLYQHVCGEERGRGLQILFQGEHYLQKGLLAFLLVFSPIQSFHDVLRETS
KRVSYLNRTSCSQRNQIICMKPKKQRIILQLKALEKLLISTHTSAIPMKSHNYHQFRNKR
NNKQLGNRRIQLLLKTMIVMNMKKCQKKKGKHVNNMCHKGFSRSQLRLSKHFQMRKRSQR
RRKEQRIQKEMERSKRKQMRKMRCMEEERRKQRSYQMTLQTKQRIMNFLKLRNNKMWMRK
LMLKMWKARRKLWEMMKVFLQVITVKQKEQKEPMMIAQLKLLKRKKKPTRNGPFVSTMKT
QKDTCLMMQIAV
>NM_000521 1 frame 1
WSCAGWGCPGRPCCWRCCWRHCWRRCWRCLRWRWWCRWRRRLGPRASRPSRGRRCGPCRS
WRPRTCCISPRRTSTSATAPIPRRAPPAPCWRKRFDDI
>NM_000616 1 frame 1
CGSVCVTRDRSCWNPTSRFCPHGPPRCSQWPLCWGASPASCFSLGASSSVSGAGTEGAKQ
SGCLRSRDSSVRRRPASVLTGFRRHVAPF
>NM_000893 1 frame -1
EKASRFFTFPIITNRGNKRRNNSPKVLRVQGSTPKGRGRASIEGGLL
>NM_001001325 1 frame 1
WPNLSQYSHFCPLSYIWCYLLFQALDTGGHHVELLRNVHMRKT
>NM_001002251 1 frame -1
ATLHVSAGAESWLPSGWAWGPCPRSLPRSFWPLRTRVCDLVTGWARRKDQGASVWSSQSY
GEG
>NM_001002251 2 frame 1
DQPGRRAARARVAFFPGHRGRGCCGRAPRGQPWKWAQLRQTQAGAAELEGRRPCLPLREV
GAPGTEPGASERKAAAPRFPAGQPPAQRPGRKPPRRRFLSRRRGAGCPARRAVE
>NM_001002252 1 frame -1
ATLHVSAGAESWLPSGWAWGPCPRSLPRSFWPLRTRVCDLVTGWARRKDQGASVWSSQSY
GEGRTSRGGGRRARGSPS
>NM_001002252 2 frame 1
PGHRGRGCCGRAPRGQPWKWAQLRQTQAGAAELEGRRPCLPLREVGAPGTEPGASERKAA
APRFPAGQPPAQRPGRKPPRRRFLSRRRGAGCPARRAVEP
>NM_001004303 1 frame 1
LQRNPGKSWKQKEPRLFLPRSTWWPPKYYITSLKIPLLFLNISEKAGKTHLL
>NM_001008387 1 frame 1
CCLPWPCPVCPGCCFPASFSCVRFKVKKPRRNCPLHGSAVPKAPRPMAPPAMPCFCHQ
>NM_001009613 2 frame 1

KQNIQHCFTAGRIRKEIQINWRITSLQRAPLIQSKRKETTYLQDLHRMVGRI
>NM_001010886 2 frame -1
RVALRGEGVRGQSVSNAGVLWPISVSLPAEPDLYIFKEEMAHRLWPLSENWKAEVGENVL
L
>NM_001011548 1 frame 1
CLLSRRVSTASLRKALRPKKRPWAWWVHRLLLLRSRRLLSPPPLLWSLAPWRKCLLLSQQ
VLPRVLREPLPYPL
>NM_001011549 1 frame 1
CLLSRRVSTASLRKALRPKKRPWAWWVHRLLLLRSRRLLSPPPLLWSLAPWRKCLLLSQQ
VLPRVLREPLPYPL
>NM_001011550 1 frame 1
CLLSRRVSTASLRKALRPKKRPWAWWVHRLLLLRSRRLLSPPPLLWSLAPWRKCLLLSQQ
VLPRVLREPLPYPL
>NM_001012409 1 frame 1
PIVMTAPEIYLRIYRKFLLKKLNFQDKENHFKKIRYL
>NM_001012412 1 frame 1
EWTPIVMTAPEIYLRIYRKFLLKKLNFQDKENHFKKIRYLLFLKTHWELILIQLHHLKLS
SHLILARISPMSPCILLKSEDFLFLQKRIKQAQQWLCLNVGAQPATIRSPPSLRNEEGTL
LQICVFILLFSSRKRIDVLKKEPWRYHLPKKQFLFYIMFENLFRDSQTVGNVNLKPTSAG
>NM_001012505 1 frame 1
SPLGVPAVGLRDFQALGSWKSCRSPCGEQLLRRRVDLCVSLTLPHF
>NM_001012631 1 frame 1
PCCSGCRPGGTGFWPGRRRWWPWSMQCRPSGNSSRVSAALCQSSSCPLSSPTEPHGGTRR
SHPRSALNPNPQN
>NM_001012632 1 frame 1
PCCSGCRPGGTGFWPGRRRWWPWSMQCRPSGNSSRVSAALCQSSSCPLSSPTEPHGGTRR
SHPRSALNPNPQN
>NM_001012718 1 frame 1
PCCSGCRPGGTGFWPGRRRWWPWSMQCRPSGNSSRVSAALCQSSSCPLSSPTEPHGGTRR
SHPRSALNPNPQN
>NM_001013627 1 frame 1
PTLQVPPARAIKSMKMGKILPVGILGSLTKSHLWFLRRLLPS
>NM_001018049 1 frame 1
CCASCSPWAWPWSVVSRPWTSPRPSRTWSSQSWQGPGTPWPWRPTTSPSWRHRPLGSTSP
HCCPPPRTTWRS
>NM_001024594 1 frame 1
WRPGRSGHGRVPRSAGNLPPPRRQHLSGELGSDSSSAPSALLTRETAAAPRP
>NM_001024680 1 frame 1
CIKTPRETIIEFLTQKRTLWMLRRKKMRVKTTFLNCCLQKSLLKFSVSWTFGVCAGLHHA
GAGMTQETVTLYGNLTALELCAE
>NM_001025591 1 frame -1
EGDIRHLCSSAGSDLQRPAGGCLPGYRTACECCVCVPRPPEGGA
>NM_001030288 1 frame 1
WPRFSFSLGCWWAQTLWGAQQQCRHPPPESLWSLLASPAQRCTPLQQVTLRPTALGTRPQ
PYLPQLPSMRDPLFGLPLVPALVPLYLSQQPTRKFPSRCHQCPRKPLMQPVILLFPQQTL
DPTPQVEPQRTLQKPPVGPVEPLLPRQLALWRPPEAPLDPLLPWQLSLWRLPKAPLDPLL
PWQLTLWRPPLGPLDP
>NM_001031836 1 frame 1
KGMKSSHCKSHMKLIKHHRQQRHIQTQIVLPPLIQLLRHCIHQSILTSRELTPSLFLSKH
GIRVEQTVLYHLRYLVTMQKKMKGKLQMRFMMRIPLHIQSHY
>NM_001037671 1 frame -1
SACQRVPSPGLSSAARRPLPAQPHPTTPSSGRRIAPIQVCMHLFSPVLHLL
>NM_001039481 1 frame 1
CSAAARGPAPTTGIFSESGPGSVQLLSRPGSATVPPPDVLLPLARPSQRPQP
>NM_001039664 1 frame -1
GAAAAGLRGGGGGAPPGAAGGPGPGR
>NM_001040011 1 frame 1
CSGVARGTCGVTTKAAHATAVPAHLGRGAGPAFLGCARSFLCASSHSGSEFLARCTEVAL
RLPRWTLLRHTPSGGLGPQGSG
>NM_001040437 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001040438 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001040616 1 frame -1
IHKYLWLCPLTCPRPKLQPNNTPSFDCSSSQRCVCSALLG
>NM_001042479 1 frame 1
GTYLLRFFPKALTIMRLRILSPSMTIMWPGRTTPAVLHISEDLGS
>NM_001042480 1 frame 1
GTYLLRFFPKALTIMRLRILSPSMTIMWPGRTTPAVLHISEDLGS
>NM_001077181 1 frame 1
LTFLAVQALLKEPPDLLQGKAVLKVSPFQGLKQSCV
>NM_001077203 1 frame 1
WTRESSGDGHLHPKSSQKEKGKSHLLIYRRERCMQNQRMSMFNHHCPNSEAQNAGLSLCS
GKEAGIKSSLTIKIKNISEGVLLLPSHHQKGNSKLCRMSYGRIDENSERPYLETMLIYVM
PTRCNQTHCLRHLLTARHVKNNLFAKALIYLKGAHNEVRQMTILQSRLRTIKKNEERMMA
FLFYLILSLKTLTVEVEVVIISNRKAETRMLNILIQKWNSLFPGRQREGLEIIYLILNIV
LLWISQQNRQKNKKMTQQYPLSLKSQVKTIIRIQNCLKKLQLNLQKVILLNYPHLTVRSL
VMPPKVPLPVQPLKPLRTLIPLILWGFLPWL

>NM_001077624 1 frame 1
NWREAILQRNCMTLTILEKSSMNTHFLLTLTLERKLLRIIRVEKPERTFLIVFTRKVMLR
GKCLSVLNMKKPSTSFQILLGRIKLTHKRNCVNAKTVGELFLIS
>NM_001080429 1 frame 1
WWRRRPPGPGRCGGPFCPRSSRWTSTISRGPRRRPAWRGCCGPRSGAQSTCPRSCGSPSI
PTSTRRSMSPRHGCCSQPSSTAEPGARHCHSLKHPPTPLHCWPCWRGGAQCLLCPSARCA
RPTGTSPFSWEPTLSLMPSWLPLPSSGQRPCQVPWPPAWSTSCFSGWTRPSGGCRRRPSR
KRPSEPLQQPLQTGRPRRSPRALRAGTGSWFLTQLPSVRSRGANPPSDTARTVWWRDVPP
ASRRPASRTWPVGPRWPPPSTAI
>NM_001080493 1 frame 1
WTQWPLKMWLTSHKRSGLCWVHHRRVSTEMSCRKPLGTWTVKNGRTRTLEISAKMPREIE
VIHVKLKMTVNVEKLLARFQIVLTRTLLEIHVTVVSVEKSSWVIRLLIATSELTLDTNHV
SIRNM
>NM_001080851 1 frame -1
GCPSQSSHPCQEFDHQSRSPHWEGEQHAAARDPHL
>NM_001080851 2 frame 1
QPPRRGAPKGLCCLARPQNLCLCPVLQPWSSNWFLLLGIPTGPLPCPAYAQLSSSPVPPL
TLFYCSHRSPKSLTRLWFLP
>NM_001085398 1 frame 1
WHPSNCSIPGEGGGHGQVSTLSFPKRPWGLGRGQSLLPGVTGQ
>NM_001098496 1 frame 1
RRLLSRLLLDCSVQTFSNTRSSTVERNPKD
>NM_001099439 1 frame 1
WRPAPVHTRCASSSAGCSSVSRCFWDPGGLGPPRKLSSWIPKPPRPSWAGLHCQVMGGRR
SAAWMNTTVPSARTKCAMCWSPTRTTGCRLAGAVAAGSASSWNCSSHSVTAAASLAPRVP
ARRPSTSTTWKLRPTWAVGVPAAAAGPAKSTRSRRTRASRRATWVSARSTQRCARSDRSA
GGVSTWPFRTWAHAWRLSRCASTTSSAAPPCGAWPRSQPPQPRAPSPHWWKWPERAWRTR
KGSLAAPHACTAAPTASGWCLWAAAAAARDSRSVVTSAKPVPQGFTRCPRGGPSAHRAQS
TAGPWKTPPPSACARTAMRAHPPTRPRLPAPGRRRRRG
>NM_001099686 1 frame 1
CALLRSVGHTELKLQNAMTMVALFKEERKVGVLSGIILTRGAVIMNMVGMSARLHTARRM
MEAWRGMSTRTNNDTLLIASDAKEENGIVKTKSVLPRGEIENLRREKVRTHRMDTQGTGL
R
>NM_001100595 1 frame 1
WPRRWGTIWSRRKLRAAEAGPESSSLLLHQREPP
>NM_001100595 2 frame -1
LVSKPPRFQKAESDLDYIQYRLEYEIKTNHPDSASELSPL
>NM_001100878 1 frame -1
GRPAAGQCYGGTCRPPPARPLRRLAPSASQRGRPAASHGYGLLHRAVAEPAHRTAPAGGG
HPGATPHLAGRPRTHCALVGPAGPGPPRAESQEGAA
>NM_001101376 1 frame 1
WRDTQKRRFQMRSIRTRSCGNCTSRSYEPRNSTRSITIPSARFIQSPGSPCLGMIT
>NM_001101391 1 frame 1
AASPSWAWSSSASCCCSCGAAAAGSTKTTSRWSTPSARWMGRPPRRAREARASSTRS
>NM_001102416 1 frame -1
KASRFFTFPIITNRGNKRRNNCKSTPHFHGTCTRRAGFRKRTRAYSTLGPKTKK
>NM_001105549 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001105550 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001105551 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001105552 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001105576 1 frame -1
GPARRGREPGTHGLSRADLAEPAVRPAAPPLESGHWPGQVLGQSRSRRLPG
>NM_001115152 1 frame 1
IGVGLRDLELILGSKFKPLTQAHKLQSQNGQPQQQAWLSQLQPPRRPAAPSPGPRSRAPT
SCSCSSWSCLCSACWGPSSGTDHKEGL
>NM_001126049 1 frame 1
WIARGQAPRAPAGPCTFGVTGLSGKYGTVGSCSPASGRGEETEGSKGGGRIHGPQSELLS
EGGHVCPLGNFPNSHSPVIAPE
>NM_001126049 2 frame -1
VVFFLCSGCSCLVQAAEPQPFLRRGGNRGSDQPPEGALSGLALGELVTQAPTSKHVPPPP
RLLAAADSYRTRGESPQTGATPRLLPEQAKGL
>NM_001127179 1 frame -1
EERCHTDPEALAGSQLEELRADASGLPAAAGPAPLPEAAPAVPPGPPAHHPVPGPLPRLS
GAQGLPPPPLGCAHRAGLCPGHDRPQAAPTPQGVSVAPRGENAAGGGREASEGDERQEGQ
GGGRAQASGAPGPAGSGRAGAEGEGGRSAEEGAPGADGKGPPACQSLRHGGQDVWLPGDF
RWPARPGGPGTWLGPGARAEGDGGGGPGCSPAPARGGGPLVICQVRGHLLPGDNHALLHP
AATQTATALPRRGPAGSPGGLDHHPPLHGGPPAQVPHSHEWQEDPCDDQDLDPGQEDVQE
GAAGPAGRGRGPAPRGPEEEQCEAQAGAFDSEKEVQAHRGGDQEAARGVHSAGQQHAGGP
AHLQPGEAALHHRQWHPAASTPVSAGRRGHQGL
>NM_001127393 1 frame 1

NCMELAGQTFPISSVRCFKVLLKTYGSPSPTT
>NM_001128931 1 frame 1
LVQVYLTDIVASDPDHCSQVSHTNFLVSKCIKLRLHATLVYVCN
>NM_001129828 3 frame -1
VPKTTRKGKGTHQGSSRNKRLSL
>NM_001130009 1 frame 1
LLIYVFRICLRNEYLQNYHILRIIYNQMSTKLCPYLVKNLVLLTVVLIVHHIFQRIFQEF
PCKMNPETLKFKETSCFKKTIKSILLSLILSVTQWRPAMLDHQILLIIVEKLICKPLG
>NM_001130142 1 frame -1
GALLWPTHPPPGASAAEEYLCEREHLRVCGWCVCNFELREGESSFGGLLCVPHGRLCCLQ
LGLGGWEENCSRITRQDEGPHQLESHLPGPPGLLIGGGQQLQGCLLLQCGPPTWVEGGSH
PEVCAGAASGSRWGSALCAPSCPESIPVLWVVGQLPCEDSYSPCGGPALHTQHGRHHRFP
AWHEGPIQLPLESYRVPRRGQDFCSGFPGCWTQVSGRGTPDLLQGAYPQRGFGDGDAHEA
RSFDGRSICNGEFLSKYPRRSTIKYLWRVYLSHGPLGKYAEPHEPGYISAANTGSQGNTD
FAAEEFTYRLLFQHLWIWLFLGMLSGECEVHSANNGGGSGESEAYAGRPRGHNLGTTPEH
LQGTLHPRPPPTAFCLYRWRSYRHVCNRSDQQTETQVFLIWYWRRHLHQPNKRYCPGIRG
HLRIYHRQRQDAVQGSQDSETLSAACGRGCLSELAFASWSVCNAFPRTDCHLGSEINQLC
PADREDASSRDNRRSMPQIYTPGQDFGGDISSTTQACQPHHSPPCCQVLAPDQGHGPQGD
SSKKRCIEPPVWCHKLLHSFHCYQGAQQAGSGASGSGRPKANSVGCFCPIEDKMPIRFSK
GLTLPSSFCISAQRGTYVLGQDIPDGRLQSLWVDKSQGPAQSRLWRESP
>NM_001130168 1 frame -1
GAPLSPSLHTAHHLEGAPAHSGDSQALHLQQQIKPQGGHGGCELNLSDSGRKLPVVDEWS
EPPYVSQVAVVNQQDPLSIGCHK
>NM_001130168 2 frame 1
SDPLKTGSSFYPVSREMKQDPINVKGTNMVASAVTQSPMSSMVQTSPEFTLHSPITVQEK
SSTCPVLRTLTHRHSILGQLMGSFSYQDKSSLSPKLLQSIAGSMLALFVTQPLARKAPNP
QKSLTGHYP
>NM_001131065 1 frame -1
VEVFQELSTFFLIYTSSPIWNSLPQAVGMSSSFAFWAFQLSCHLSDQFLSEHESWLCTRS
K
>NM_001131066 1 frame -1
VEVFQELSTFFLIYTSSPIWNSLPQAVGMSSSFAFWAFQLSCHLSDQFLSEHESWLCTRS
K
>NM_001134657 1 frame 1
WAAGPAAPAPALRPGGDSSQEDQALPSAADWRSPRAPNPERRPARKTRRGPRPWTRSPPW
WSWTRAVPCVCPWRTSTWCWSSRQCRSCECLLVDTPSSSPRSSAPSTNAQERRATGLPAW
KWTFSWALTGKTSSSSRKSAHLSQRSLPRKRPTRRTRTLSSRSSGWTPQPAQPLGSTPPL
EVCSAPTGRAPSEGPVLWPPTPVQR
>NM_001135098 1 frame 1
CWPQAVSWTQTKKWGQGLKMCLLPLQLPPLVSGLPAMLPVQPVELLFTREWLQLYLPPIG
PDLTGIHMLEKGLFTLKILTWIWMKTFSITWIWEQ
>NM_001135197 1 frame 1
LVWAPYSLENLMSGVKEQRMMISKKRLHCQHLGPMKEIGMRTRWCRLTARTGLLLKQVPR
TMVPASQAIRFPVTGTWFPKEPHSSHHWRTFQPALRMPAKNIKPKVCFCVTHVNIWL
>NM_001135746 1 frame -1
ARRAVGEEISPFGLPLCGSAGGQDRGQQEATARQPAVSVAPGHPGPWKEIIPISRRNGAP
WQQPNTRASTKRSAEGRGCGPKAGRGAQRAEDPWILT
>NM_001135770 1 frame 1
KRDLPVSTQACPVTPSSSWFWESWFFSCWGSGFISIGPNVPVRSFGTVICVPRVSITRAA
VI
>NM_001135947 1 frame -1
CFSGKALGLSSGTYRVAGSSAPLQPHTEAAGVSHSRGRDGLGTWEYSSIL
>NM_001136233 1 frame -1
WLCPGSSAGCPAPLCLGAAATAAAAAAATTTATADPAPDTAFESPAAASVFGNRRCSDSA
ATSRCASREPVGRSEEGRQANPSSSL
>NM_001136508 1 frame 1
RSKEKRKQHIKQLKIILKMNPNLQQKISFVIPQRPAPQQIAAVLHAYQHYHLILITAKVK
QLMTGFLMILKGTLQCLLSGNLWKKYFHTCQPFHKRVLKVYMTLY
>NM_001136533 1 frame -1
VPPRRQHRRLTRLRDKPVQQPRGAVWSRGGDRGLLRHHSDLRAEC
>NM_001136570 1 frame 1
KNILPKSTWDLPRRRLCQTQANGFMKKSHI
>NM_001142285 1 frame -1
IGKNFPGREMHRCEELEGKRKPGGCVAGRSARTVERVDMWQIAEGLWQVFTGGRYLEEDK
QRTVLPGVPFGEGTGQKRSLHVACLTCSCWFSPPCGRLGPPLARSLVSNLDSVTPHPGHQ
QVFWPFLNPPGIQLLLPTFLPLWSGPL
>NM_001142306 3 frame -1
DLHFPVEKVPEEALGGQLPSPSTPSPPRASVQTSTCEPPATWGAWESFSHQLPCVPTGSP
GSSGLHAVISPEDLRRVLGVSIIPELLQATRASVSPEAPFTPATCEHPITKPDQLHRILG
GDPTCRQVEAGVPLSSAWMCQRCWTHKTPKSQTLPWTPNIRHPRTIGNLL
>NM_001142401 1 frame 1
CRGSPAHCFGPPPAWACSACCPRTRTRPSTRTRLRPSPTPRRRRPSRWSPLRHQKPVKVE
TAAFPVLMLALLI
>NM_001142402 1 frame 1
CRGSPAHCFGPPPAWACSACCPRTRTRPSTRTRLRPSPTPRRRRPSRWSPLRHQKPVKVE
TAAFPVLMLALLI
>NM_001142579 1 frame 1
FLHWLPNHLLQLFIFKGESIAESALHLVISLCIVAWKSY

>NM_001142683 1 frame 1
WGPEGRAHTARKRGACLSFAHRERGAEPLEDRLLGLIRPRRLNSSGLDQPGLEQGAVSRV
AGQHRDPLSTERINYGSFPHLHIVPASCRRRFLMLALIPWQSPSVNYTVAVTFRMNAWIP
GLLRAPEGSFHRIIVIILSSQRS
>NM_001144875 1 frame 1
VKEREPMTYQASLMRVCCVPTQCQALGDTAVFRTALVLFSGSRHSRGREVSRLPCCVPSS
APAPGAILGASHARKRHASRALTAQPLTATPPQLHTPLAHP
>NM_001145082 1 frame 1
PQLSPTVPAPSSLLCLPNIPALPPHQGLPYLLHMQWYFLPLCLSSCCFPHLKRPTLHLSK
HIFFRISLFLGSHPFRARILCLTPC
>NM_001145083 1 frame 1
YPQLSPTVPAPSSLLCLPNIPALPPHQGLPYLLHMQWYFLPLCLSSCCFPHLKRPTLHLS
KHIFFRISLFLGSHPFRARILCLTPC
>NM_001145093 1 frame 1
YPQLSPTVPAPSSLLCLPNIPALPPHQGLPYLLHMQWYFLPLCLSSCCFPHLKRPTLHLS
KHIFFRISLFLGSHPFRARILCLTPC
>NM_001145139 1 frame 1
GASAWIPSSPLRKAPPGCLPTTLIPQMRAATYHRAFAATKEARPSPAAPRSEQTHPD
>NM_001145140 1 frame 1
GASAWIPSSPLRKAPPGCLPTTLIPQMRAATYHRAFAATKEARPSPAAPRSEQTHPD
>NM_001145176 1 frame -1
WGCRGPRGRGPRGQHSLRRVRGHQLRPGRGRRRQSGLQQGVPCAPGAPGYQRAQSSGGHS
ETSGGAWDPPGPGPKQLPDPGDRRSRQHAPQDAQEPLQHREDPQPRLASLAAAQPLLALL
HAVLLSVLPQPLASAPVLRGCREIRAEPLLFPQEAGREGSGRGLPEAPAPGIAPRHAAPA
RAAAAGHGGRLSHGGLGRPRGAGPGARAQPPAGHVPGAPAGRWWPFALRVHAAARAPGVA
GAPAAVLARGPAPESRAPAVAPAQPRAAPRPHRPGQPGRLALQPAAGARGAPNTAARAAA
TVQPARPPPPRPARLAAAGLDVRRRRGRGHTGPQAALPAPGHAGAAAVHPGPPPAPAPAH
GQQERGDCL
>NM_001145531 1 frame 1
WGPVDGPPNLTGVLLRRGRDTRDVHTQRQDTGHRDKAATCKPGRGASGKPTLPTPSWTSS
LQNSRIVKWVIMSDTVQLFLKELSPHVGLSASLVASLTIQGPLKTPRKHSPCCGA
>NM_001145640 1 frame -1
GDESVFRLRAGRRPAVQSRSGDHSPGPCIWASLC
>NM_001145873 1 frame 1
WPYQPPCSCRWPCCSTPPGRASSGCRRWIGPGTWARQWSSARCCCPTRRRAARGSSSRAA
PPPVPPSS
>NM_001159709 1 frame 1
NITGRWTEDTIKGGAWECAGMMWTGGRSTLCLPIMGTGSSDMENICISHIPVLSA
>NM_001160417 1 frame 1
PQTHNPEVTFLETLVSPSLPATRSSPPSWKLLLETGVTKLQKAATMWMKPHTRGAGGEVG
F
>NM_001160418 1 frame 1
QVVVPQTHNPEVTFLETLVSPSLPATRSSPPSWKLLLETGVTKLQKAATMWMKPHTRGAG
GEVGF
>NM_001162483 1 frame 1
RQKPARCCHLILDLPASRTASPVELPPPLWFSAGRAGSGEKCQFIHLTLSI
>NM_001163023 1 frame 1
IQKRNPPLTHQQENLDKHQTQTYRGHRWVSMLPPKNMIILSKKR
>NM_001163075 1 frame 1
GARLQTTPMSSTAPSQWCPQVVTRRWRTGRVPWRAPSLPCRSTGAPAVVKRSLGGSGGMG
RSEDAKQQPC
>NM_001164431 1 frame -1
KGTAGIPQEGGGPGTAQQDDSEECFHRDGPPLSAPRAAPQAPQRQGEAAGRRGSRGGADN
GALPGSA
>NM_001166451 1 frame -1
EKASRFFTFPIITNRGNKRRNNSPKVLRVQGSTPKGRGRASIEGGLL
>NM_001166663 1 frame -1
AGASGHPHTPPAPQGVSGQRMPGISPCGHLGSASSVTTKQHTDEGQHCMEEVAALTKWIS
SHIEVGEWLFAFQYFQIQFYSQELESSHQGSSAAGQWPLLPGGHQYIWKSSD
>NM_001168478 1 frame 1
WLTLGQKQGLEERLRLACKMESVVLPLLEMVKPRPRQWLRQNKQNQPRPKLVMEQPGHIQ
PTGRLWLQGKSRWKIQLRLESWLRLRQNPWQNAVCHKPSQRPCLCLGSVLPSLKSRLLLS
LRQILGPMPSHMIRPIL
>NM_001168479 1 frame 1
WLTLGQKQGLEERLRLACKMESVVLPLLEMVKPRPRQWLRQNKQNQPRPKLVMEQPGHIQ
PTGRLWLQGKSRWKIQLRLESWLRLRQNPWQNAVCHKPSQRPCLCLGSVLPSLKSRLLLS
LRQILGPMPSHMIRPIL
>NM_001168480 1 frame 1
WLTLGQKQGLEERLRLACKMESVVLPLLEMVKPRPRQWLRQNKQNQPRPKLVMEQPGHIQ
PTGRLWLQGKSRWKIQLRLESWLRLRQNPWQNAVCHKPSQRPCLCLGSVLPSLKSRLLLS
LRQILGPMPSHMIRPIL
>NM_001168482 1 frame 1
WLTLGQKQGLEERLRLACKMESVVLPLLEMVKPRPRQWLRQNKQNQPRPKLVMEQPGHIQ
PTGRLWLQGKSRWKIQLRLESWLRLRQNPWQNAVCHKPSQRPCLCLGSVLPSLKSRLLLS
LRQILGPMPSHMIRPIL
>NM_001168485 1 frame 1
WLTLGQKQGLEERLRLACKMESVVLPLLEMVKPRPRQWLRQNKQNQPRPKLVMEQPGHIQ
PTGRLWLQGKSRWKIQLRLESWLRLRQNPWQNAVCHKPSQRPCLCLGSVLPSLKSRLLLS

LRQILGPMPSHMIRPIL
>NM_001169574 1 frame -1
ESFLKCTTRYSEIRYKILTKSSQSTFRKCTKCKIITINARHFSSNNTEGLVISYKQNDIS
TPKT
>NM_001177548 1 frame 1
CREPRKPPPQRCYRCCCPCCGQGPWLRSGDSSWRGQSHRCRRVCASSYPADCPLPFQPRT
MVMATGSWKGLMFQWPQTTQTKKCRRRPGADSTSSGIPEGRTAPASEMPGGGTMLHTSFG
SPNGNTVIHLPSSLCVWPPTGPTSPS
>NM_001177597 1 frame -1
KQDFWIVGDKLHCLSQNYWLWASEVAAGIQSQDSWSAEPNLQVPGPNPRIPEQDTRTLEW
NSW
>NM_001177597 2 frame 1
SFLDPHAGPEPRTFPQEHQTQAPCHPTSSLDILLPQPILLLDSIRSSLFHPPCPPLWSSS
TPCFLTLLLQRPPLPALFTHPTPTPRICLRKG
>NM_001177663 1 frame 1
SKITHLVCLIKQNVCQIIRFIFLNLQWSMTIQSLKQQQWKMLYYWKLILR
>NM_001177664 1 frame 1
SKITHLVCLIKQNVCQIIRFIFLNLQWSMTIQSLKQQQWKMLYYWKLILR
>NM_001177665 1 frame 1
SKITHLVCLIKQNVCQIIRFIFLNLQWSMTIQSLKQQQWKMLYYWKLILR
>NM_001184743 1 frame 1
CLCLTQSGPRHWRLRKKQPRPLLQRSGHIVLGGTSVGTQLRRQLASVRLKKLLKIDCLKQ
SKKLLKKWNKVEKPQESP
>NM_001184771 1 frame 1
VILLKKEVIQLEMLTKILKMSLLMTVVQTSLHLLIPISIRKPTQTGSQAQQPPRKMLFLK
QQKTASSKQRSKKINEKKIKKSFFYFRPPFPENWSLTSPMIDQDPTQGKLR
>NM_001185063 1 frame -1
PENNGQGKSWYTQCTEMASSFSKEEKKETCNLTRASKTYGSSW
>NM_001190316 1 frame -1
LARRCERSWPWKALAAIPGALTPSHSVLSLPLLLLLLLALSPASHTLSFQSLV
>NM_001190317 1 frame -1
LARRCERSWPWKALAAIPGALTPSHSVLSLPLLLLLLLALSPASHTLSFQSLV
>NM_001190864 1 frame -1
GGRRAPSMAVSAVEGKDSACILHCRLLGHSGPDTPAGKNTRKKATKKKGTTSD
>NM_001190980 1 frame -1
WQLRREETGIPGTWKKKEYSTQGSWPRTEKWAFKRGL
>NM_001190986 1 frame 1
IGLIDYMSLMNKEHLVPVAQITVTMDYAPMVASTKISIVTVKVSSHPVNISWSGTVARPP
AIVQTAFI
>NM_001193374 1 frame 1
KLSVSSSSLSWGCWCLARPCAPWKKPSMRGSRRSPAPYLGQAALAWSAR
>NM_001193388 1 frame -1
PAAATPAFRAGHTGLCSLILQPPRFERCPREHLGLCFRQLLPQPLPHPECRPPILFECKS
KRSAGPGQAAAGRGQEEDPAVGGVLAAGEAGLRCLAARGAGGQGACPCGRRPAAGAAEEG
GGGGTGDLPAPRQAVCGLPGAGPRCCPAALSAPHPLAVCGHRTVPLLQGPAPAVV
>NM_001193508 1 frame -1
RSGGAATSHGACSDGGCPDTACSRACSDGGGSGGACSEGAAASRGACSDGGCPNCTCSHG
AASSHGDCSDGGCPNGACSHGTCSDGGCPGRICSHAGG
>NM_001193552 1 frame 1
HLSGEWKNVIALWAPVSEMTGNAKASFSTKIIRSDIWKKQ
>NM_001195014 1 frame 1
NPTSRFCPHGPPRCSQWPLCWGASPASCFSLGASSSVSGAGTEGAKQSGCLRSRDSSVRR
RPASVLTGFRRHVAPF
>NM_001195032 1 frame -1
LPRGGTSPLSFPQRRSPSSALPPSQPFSVTVLVLLSNAFVSAFTSKYLL
>NM_001197 1 frame 1
DPRTPGPGCPANRCCWRCCCCWRCCCRCSAGACTCCSS
>NM_001199251 1 frame 1
PIVMTAPEIYLRIYRKFLLKKLNFQDKENHFKKIRYL
>NM_001199254 1 frame 1
EWTPIVMTAPEIYLRIYRKFLLKKLNFQDKENHFKKIRYLLFLKTHWELILIQLHHLKLS
SHLILARISPMSPCILLKSEDFLFLQKRIKQAQQWLCLNVGAQPATIRSPPSLRNEEGTL
LQICVFILLFSSRKRIDVLKKEPWRYHLPKKQFLFYIMFENLFRDSQTVGNVNLKPTSAG
>NM_001199257 1 frame 1
LEWTPIVMTAPEIYLRIYRKFLLKKLNFQDKENHFKKIRYLLFLKTHWELILIQEPWRYH
LPKKQFLFYIMFENLFRDSQTVGNVNLKPTSAG
>NM_001199640 1 frame 1
ESTKDIWYSLPVNSSLLWSALPLVYQGSRNILEHFMIQVSQIRCYVTMSLNT
>NM_001200049 1 frame 1
SCWTWRGPTRARGTWRAWSTGDLLAVGKPIGETVRLLQKMSGDEAVNQDEASQTLKDKLL
LLQSSELLPTTLNLVLYGLPHQAIGCRPGPASHQLLGPQHLPLPLALPLCQPIPTSRLPS
PAPSSLSPSWASPQPLVQPQPGTLPQRLPARQPPGPRALPACVRHGVCVGAGAASLPGAR
TRAASPWRRWCVATRRLFRPACFF
>NM_001201407 1 frame -1
VGSVWKSLQLYSKCCFPASMHSGQISLQQLWGSVPSVIPSGTCGKSQRRRN
>NM_001201550 1 frame 1
ESSTSASPTMNFRVLNMHVVMETGQNHQDAYQNLVSFQKFNMDIYIMRIRVDHTFQLQDN
LTPIT

>NM_001204083 1 frame -1
HVCWRDGSRGAISAPKQEGHHRDHARRGAAVWHSRQPAFSTTPWLLAFLASEL
>NM_001204285 1 frame 1
HRAPSLLSSCCCSSQCLQLLRVLVMQALPQVEKRRLRLPREVQCPALLRRMLVPAAYSPA
TAPVQAPPPLRDRMSLWPRPRNQLQVQLPPGDRMSPRSQSPGQPWAPPPRQPTMSPQPRT
>NM_001204285 2 frame -1
ASPGLHRPPSPRCHLGPGHQAGPGLHRPPSPWCHLGPGQQARLGLHRPSSPQCHLG
>NM_001204286 1 frame 1
HRAPSLLSSCCCSSQCLQLPQPLNPQQLLRVLVMQALPQVEKRRLRLPREVQCPALLRRM
LVPAAYSPATAPVQAPPPLRDRMSLWPRPRNQLQVQLPPGDR
>NM_001205280 1 frame -1
ESKTSEPPSQEPCPSLVPTHHHHPDAREGCGWRLPS
>NM_001206631 2 frame 1
SWLPPDRPDLRTSTAQTISVCSMRRLRSSSVRLTRDCSVGPALSHQSTWLTATAQDGLLR
NAGYRNLRKWTIYGK
>NM_001206844 1 frame -1
AVRGCLGPPERRSGGSRAALGHRKRQARRSWSTAFCRKRLHPPQLEPTAWRLYQVASGAA
DSKRQPQVECGGADPRRHPQQVGCGGKDPKRQLQQV
>NM_001206844 2 frame 1
VRRQRPKKAALASRVRRRRPQKTALVSRIRSCRLQKTALANRVRRRRPQKTAPTGRVRRQ
RPKKTALASQVRRRRPQKMSLISRVRTARPQKTAPASRVRRIRPQKTSLTSRVRRSRLQK
TALTSPVQRSRAQTGPASRVRRSRPQKTALTRRKLSLQSLLR
>NM_001206845 1 frame 1
WPRWSSRRPSLASGHSYLPSSACYHSASPQHPCSATTGLWAHRRCPSPCARKVWQPSALT
CQCPWMEIPTHPPRRWYNTTGR
>NM_001207037 1 frame 1
CTAEHGFEMNGRALCILTKDDFRHRAPSSGQGNAQAFLAGTLVTSCMSCSS
>NM_001207037 2 frame -1
QDPAASPGVWALFWRDLQAEDAHPALSSPPGRGDWPLSDGHPKGPPAAATRPRAYQQLRP
P
>NM_001242312 1 frame 1
QRPPGLPSLSSPKGLPAPQRQLVLLPQLPAPPPSQTPPHSSHAIPPKSSRLMETCHHPQG
RLAPGITTWRNSTS
>NM_001242348 1 frame 1
AAGPPRSSSAQGVPPAFLGLSGGPATPLSCPATWGWDTWTAKEEDS
>NM_001242780 1 frame 1
DCALVLRRMQFSLSRELTTPGVRHCRHHCRHRDGHSSPPRAFVTADITADTEKAI
>NM_001242831 1 frame 1
SKQHIQRKKKKCFICVVLSIKKLFPRALCQLSSLKTHWPQLLHMPLP
>NM_001242885 1 frame -1
DPQHCGNPRHMLGHWAVEHTGAPCSGTRTTRQEHGHSSRGDVGASQKREKARRAARGRKV
LSFVSHQENANNQ
>NM_001242936 1 frame 1
ILPKSTWDLPRRRLCQTQANGFMKKSHIKWICSMKMVLVLVFMKMYAKQLVTSA
>NM_001243538 1 frame 1
RSPPLGSILPPPDTHTPATSRLLTLGQALWRCRRSRLFLKSQDWKADSHRLLSSTQQQTT
GRT
>NM_001245 1 frame 1
CREPRKPPPQRCYRCCCPCCGQGPWLRSGDSSWRGQSHRCRRVCASSYPADCPLPFQPRT
MVMATGSWKGLMFQWPQTTQTKKCRRRPGADSTSSGIPEGRTAPASEMPGGGTMLHTSFG
SPNGNTVIHLPSSLCVWPPTGPTSPSQG
>NM_001252030 1 frame -1
SQPQTADTWSSSWLWLNLPRYFIEKPRDISVHMAFSHEAWKLDTTV
>NM_001252619 1 frame -1
NNLEALEDFEKAAGARGLSTESILIPRQSETCSPGSDGQGTPWLLSS
>NM_001252619 2 frame 1
KDGTIQHLRHSQGHGKSSPPLQNAAGCTPSYHPPPSPCPAPPLLVLH
>NM_001256442 1 frame 1
WQPAALRSLRRGLRRVPRFQAKGLAILKLKLALPRSQGYQTSQRPRSQVQTP
>NM_001256714 1 frame -1
ASIAFLEKGGDSGQRLWSAAGPQCCTEQRGRGVPGYLEDKGVGKGDDHTCRLRQRLRLRV
LVGPVPGAGPAGKGCYGGRSANHHGAPASCHLARSSCGPRLPGRYSAQQPRARCAASGLC
GWTAPAADPVPSEVLASQEVQLLCAGESGSCGPTHADLQPSPGGTGEDGAARAKRDLPGS
VGERAAAPASGRLRACPGRPAGAPGPRARPPGGTAALAQPPRPQVPRAGCPGGRIPLLGW
RRERAPGVPHEPPLGLAPAPLPGLPLRRPARCQRLGPAHEAAPRGSSHSPPGV
>NM_001257118 1 frame 1
KHLEITLICKTLKEYFLPFQLLRQCRTTQLCPHPQAQKGMSSFAPKKLKGYGNKSRQRFI
QWTSQAAHV
>NM_001261456 1 frame 1
TSSGLVPKMLLLSHVPKKMPLWSKATWADTS
>NM_001270440 1 frame -1
GRSPRNAAEDEQRTLLGYRIPPHLGALISGGHGLQGPHATHRPDVGAATDLGMDPARS
>NM_001270497 1 frame -1
GNPRTFDTNCILQQQNTDYRKRECIWENTSGHECSLKKNALMSRRKQFENFRSNCCKEKQ
RKQKDSWIQWIWEAEKSFGNEKIQIQSNEESQQSKR
>NM_001271733 1 frame 1
WRLPQSPCWPLGQHPQCLAASPGSARHRMT
>NM_001271733 3 frame 1

CYMRWCPGLGRRMWQMLKSVLVAVGPWTAGRSTTMAAMVANCCHGLNTRPTQGCGILGAV
TSSRRKASGGTCPRSGMAWKRTSAVTLMATPEVLGATQQTLPCASRAAASNPAGWPRVSG
AMARNTAARTAPSQGASASAGIFSTRTSTPSSRASSSTKVWTTTIAGILTAPSGHGATLR
IRRSSESSVTSPAAGPRHSPAKRPQVSAASAGRVRATGAQPIPPPRRTLPALGRANPTSA
PIYARKIRVQVRWAGGRALGRAAAETFGRTSAGTSTAQRRPGASPCGPARAWAFATRSGV
VQTTCGPRTATTARGSSTAARSARPARVSSASAGPLRRRTSCRPPWGGMLCLGPEPGNGD
YPATILLPLPPRFTFTSEPHAQLEENFCQTQMGIAMGPGATRWTQGPHSTTVPCDAATRC
SLRSVARGWIGWISVVPSCAWLGAIRATHPGQSACGIAICLSRAMRYGWAPCSRTHNMES
QAYSGSQPRCCVGPQAPSLSCSSWRDLPTSVWPSACRLNGMWCLQGPSVRLQAGVRPKVR
VMTQSMWPCTSSPTRSVTSSTEDMCGRARCALRDCWPLWGPVRVTTGAHLPALPTTAGSK
ELESPTEYAQGRAGQPSSRVSLCLWTGFTRSDWV
>NM_001271748 1 frame -1
CVGAVDAAAQQSNPGTDMPKQSPSALNQASAQSGWL
>NM_001271908 1 frame 1
WRSIPRKPGRQTGSRRSLWTASTRGPGQP
>NM_001271909 1 frame 1
WRSIPRKPGRQTGSRRSLWTASTRGPGQP
>NM_001276253 1 frame 1
CSRTSVLHLFPSESARAVPDGLIGPRPWLGRPLPKRGAVGPLRQET
>NM_001276264 1 frame -1
GTASRPAMSQGGEGVQIGSPSQSLALWSRRVRRALGRTYWRLTTAPPRKRHRTPVAVGIE
RKEGMWSQGYVAQAGLELWPQTTFLPRPPKVLGLQACPGLLPSFGAYPLCYLPPHTSNL
>NM_001276265 1 frame -1
GTASRPAMSQGGEGVQIGSPSQSLALWSRRVRRALGRTYWRLTTAPPRKRHRTPVAVGIE
RKEGMWSQGYVAQAGLELWPQTTFLPRPPKVLGLQACPGLLPSFGAYPLCYLPPHTSNL
>NM_001276389 1 frame 1
KKEPARQSLRYQFQNTVNLKISGLKRSILLGKVFQIILLTG
>NM_001276495 1 frame -1
GAPLSPSLHTAHHLEGGPAHSITFKLLESAHNCPSHDSPATQSFGEGCSSTCPQFAPESC
WLHLVQRANDIPLPLHYIICSRRSKNYIWACIQWKRKSIFQCIPADPECHAGGCRILHLT
HHKATRWDW
>NM_001277307 1 frame 1
CLAVRRVSAVPVRNAARLEVKTNVSGVLKPPQQRRKSCHPPPLLHARVLPRASPMQAFLR
SPREPATPALLLQLFHSQVLMKVPRAKRGKVPTPSMARP
>NM_001277945 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001277946 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001277947 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001277948 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001277949 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001277951 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001277952 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_001278094 1 frame -1
VSPQALFPCLHPTLVAVLHAQEFCNCCSFCCCYVCRYGLISTILENIISREA
>NM_001278113 1 frame 1
TKSTVTVVLHIYLQIAQFNQIYGLVGVLTVAIFHLLYQPDSLKGATLRTFTIPNQDGHFS
YSFLSLLFSVYLLWKLISKGKNGELRTIQAMSNLKVRVNLKG
>NM_001278114 1 frame 1
NAMIEVYAITKSTVTVVLHIYLQIAQFNQIYGLVGVLTVAIFHLLYQPDSLKGATLRTFT
IPNQDGHFSYSFLSLLFSVYLLWKLISKGKNGELRTIQAMSNLKVRVNLKG
>NM_001278522 1 frame 1
GSNLFFLRRLGGAFLGLRIAAWMAHSLGVHTL
>NM_001278587 1 frame 1
CGPSGMKRKKEPLTHVSRCAHLQCIALGLDRNQHLPKLLCWKTVWNISPWPSGGCCPLHR
APSSVLSFSLMGKISHMIEPNVALLAGPAACLIREMEGLASTGRGP
>NM_001280560 1 frame -1
KRSQYSRSSDGDEDDEEEEENEAGPPEGYEEEEEEEEEEDEDEDEDEDEAGSELGEGEEE
VGLSYL
>NM_001282801 1 frame 1
WTRESSGDGHLHPKSSQKEKGKSHLLIYRRERCMQNQRMSMFNHHCPNSEAQNAGLSLCS
GKEAGIKSSLTIKIKNISEGVLLLPSHHQKGYPELYRMSWERSEENTGPHLLREVVIQTT
CNQSNFLHHLMAANLIKITLTRAVIYLKGAHNEVRQMTILQSRLRTIKKNEERMMAFLFY
LILSLKTLTVEVEVVIISNRKAETRMLNILIQKWNSLFPGRQREGLEIIYLILNIVLLWI
SQQNRQKNKKMTQQYPLSLKSQVKTIIRIQNCLKKLQLNLQKVILLNYPHLTVRSLVMPP
KVPLPVQPLKPLRTLIPLILWGFLPWL

>NM_001282879 1 frame 1
QPLHYVSPLFSLAVPTLHSPQAMLITLPCPRPSSTTSGSAMTSGPWRSPGCTTSFCPTSR
QWRETLTRQLQPWRPGTITPRSRPPSSLWCKPSSARLVAGQLPRTPGKAGSLPAT
>NM_001284217 1 frame -1
PSIPDTTHSSRWRAVPTICQRASPERERRCCLLCGANLKEESQVCVHRGEKVLSSVRRDA
CSPV
>NM_001284242 1 frame 1
MLKSLIIWQTIKLVTCSLYPLLFRHLP
>NM_001284349 1 frame -1
CSHFHYTSGTTSISSRSEVLSQETKIKGEKQKVKRTSEGHESQNTFESPE
>NM_001284502 2 frame -1
SEKQRPGLGSRALQGQAQAASAGELLPDQPSPRMCRTSSSAWLACSAVVCVLCSCETGAA
WALFPSMLPTCAPGSRVSTQL
>NM_001284528 1 frame 1
WRKDSGSRVFCGPQNSVAPSSKAYSVTAKGEAVCLSLVSFMRKMLFQAPGPLHLLW
>NM_001284529 1 frame 1
WRKDSGSRVFCGPQNSVAPSSKAYSVTAKGEAVCLSLVSFMRKMLFQAPGPLHLLW
>NM_001285391 1 frame -1
ELYRPPAPRSFARASAGSLVGKSQKCPTEEKPGNARRARSVVPTAPRPAWQGPTT
>NM_001286514 1 frame 1
MQPLVMPLLVMCPLVMPPLVMPLLMISPLVIPPLVIPSQV
>NM_001286515 1 frame 1
LPLVMQPLVMPLLVMCPLVMPPLVMPLLMISPLVIPPLVIPSQVNRSLLN
>NM_001286688 1 frame -1
VCRRRRDVSSRFLHLASGVSTSSQNRQKQGNQNFLREHHFCPCFQHLRVSEPCDL
>NM_001287482 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287483 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287484 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287485 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287486 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287487 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287488 2 frame -1
LAMEFADTISSAWCWAYPKNLQLSCIILCEQLSPYWGKNAKL
>NM_001287682 1 frame -1
GNCRQPAPTAGKRGPEPRYSLGLRAAQGESLRPPAPASLHVG
>NM_001287682 2 frame 1
SARSEVRKELSPTPTRPEVPGPRWRKDPRGRLRALPGVRAPSGVSPASFTMATRASGAWA
ASRVGGGAAGLYPPAHEKRRSTI
>NM_001287746 1 frame 1
LRPFRWELMWWTPQYPDVAALMQKVLLGMPLRIYICLMAWGSIQVIYTKWKLVTLFAKLI
KPQTLKHKPPSML
>NM_001288792 1 frame 1
PQNSSPCFASGCVWATKMRKRMRNRPSPPSTPGPARWLKPRAMPSVRLIPRMHLCCARTT
LGTSRNRARQKTKLNSPSRTSLR
>NM_001288961 1 frame 1
WELFPTLPVENQGGLRMDGQRLHRGAPVKRKPQEGVRAEGSPPVVVSATAEEPGPLGVEL
ERNRLIWKFFCRNGKANSARCSGVHAQLLRL
>NM_001289023 1 frame -1
VSPPHRPPGPSALPGPDHPHAGGSAQTLHLGARHRDPPGEPCDFRVPGPGWGSNIPPGEG
Q
>NM_001290022 1 frame -1
KQDFWIVGDKLHCLSQNYWLWASEVAAGIQSQDSWSAEPNLQVPGPNPRIPEQDTRTLEW
NSW
>NM_001290022 2 frame 1
SFLDPHAGPEPRTFPQEHQTQAPCHPTSSLDILLPQPILLLDSIRSSLFHPPCPPLWSSS
TPCFLTLLLQRPPLPALFTHPTPTPRICLRKG
>NM_001294344 1 frame 1
CWGKESAVAAAPRPPSPPSCRTYALRKFTQTKRKQRRKLFTCCQVNSCFVKPAQYSMSRK
IPVSMGSMGGLSAQTSRLPSWVMM
>NM_001300730 1 frame 1
CLKKLLMQIFNSRTPVRWKKSQKLANLGKKHLQLPLMYGVQQPCFLFCAFCCSLDWESWQ
AC
>NM_001300901 1 frame -1
ASWRPGAAAHACNPSVLGGQGPPKFKRKRQSPHLSMRRMERSHIICGAGHNNFGNISSPG
QCAGLPRCQRNTGMAEGERVRQGERLGVGSGRVEGSWPECGQRWRSARLLFPALCEAKGG
ATCPESPQVRERQEKEGKSRHGKTLQKRNGFKPRK
>NM_001300913 1 frame 1
CGQLLCSFGFSPCPYLKAMRHPTIHATLSLTKCGRDSRGMHLWKQLIIKRLRMPWQQLLL
SHPKGLRQPTST
>NM_001301708 2 frame -1
LILGDSQALHLQQQLKPQGGHGGCALNLSDSGRKLPMVDEWSEPPCDSQVAAVQNQQDPL
SIWCHKVYCRTLMNTEPSECQSQPSHPESPPWSRPPQNLPFIHLLPFRRKPRLVLLHGIP

TGRVFLDNWEVSAIRTKALYPPNYKSRALCLLCSLSHWQGNLQIHDSQSLWSLPWRPDRV
SVM
>NM_001301709 1 frame 1
WGPSQPLPAHSASPGRGSCSQHHFTSGTRPPLPKSRLKPSHPKFLRGRMFFYLSTICPRI
FLATSGTKGKRTSTITLYRILMVKLYMGLHTVEEKQYIPTHPCSRMSPGRMQEPTPYTSS
EVMRLEKKFDISPSPYTMVQTSPEFTLHSPITVQEKTSTCPASRNLTHRQSIFGQLMGSF
SNQDK
>NM_001301726 1 frame -1
VGPRTAGAAAGGRRGRRRGGRRGRGRRGRGAGRAGAELHDHRPEPGGDDGGGVHAAAAHP
LLAHGGGGRRARDAPQLGAAGGGGPGRRRGRLRGGRS
>NM_001301851 1 frame 1
CLAARREGIGKREEERAVIPPRRRQTLLPAFQSPGPTRRQRAPPPRSRPGPFSRPSVVKQ
SLSRRLRPLPAVLTLQECAVWGRLRTLGRGTSGVGETPRKRAITTNMDTSQSLTLVSGNA
GDGCWRMPADT
>NM_001302819 1 frame -1
TSRQCHDNSRGYNRAASKDKSLDHVKTWSEEGNELLLHSRLWASSLLLRCLLVKARFQYQ
PDFFWKDPLASRWSGHLPPTELDFSLSSITQGAPLCKDLITFTAVSAEPRTVLGIQVLNK
YLPNDMHLASLRMPPGPEGARKIRGKLLDSCHTLHQHLLRVL
>NM_001302839 1 frame -1
LPRGGTSPLSFPQRRSPSSALPPSQPFSVTVLVLLSNAFVSAFTSKYLL
>NM_001302840 1 frame -1
LPRGGTSPLSFPQRRSPSSALPPSQPFSVTVLVLLSNAFVSAFTSKYLL
>NM_001302841 1 frame -1
LPRGGTSPLSFPQRRSPSSALPPSQPFSVTVLVLLSNAFVSAFTSKYLL
>NM_001453 1 frame 1
CRRATPCPAPTPWEWCPTSAASRATTARRPRRPGAATPPCRPPACTRTLRTPSSTRAAWP
APTGPTRRSRSPRTWSRPIATSRSSPWPSRTPRTRRSPTASTSSSWTASPSTGTTSRAGR
TASATTSRSTSASSRCRATTRSRARAATGRWTRTPTTCSRTAASCGGGGASRRRTRRTRR
RRTGCTSRSRPRPAASPRPRRRSRPTATRPVRSRRPCASRTSRPRTVRAPRRPSPCPRPP
PWAAAAPPRCPRSRAPTAAAA
>NM_001464 1 frame 1
LHIYLQIAQFNQIYGLVGVLTVAIFHLLYQPDSLKGATLRTFTIPNQDGHFSYSFLSLLF
SVYLLWKLISKGKNGELRTIQAMSNLKVRVNLKG
>NM_001563 1 frame 1
SSRERELHAGCQITLKIKHTKLVLKSSKINKITRSVKEILNYPNMKNLTIKIGKEI
>NM_001698 1 frame 1
WRPRWRRHLGPWDPCMLAAPAWWPLAVRGSARGGCPARWQAGERARRSGPRAGYLRPG
>NM_001768 1 frame 1
WPYQPPCSCRWPCCSTPPGRASSGCRRWIGPGTWARQWSSARCCCPTRRRAARGSSSRAA
PPPVPPSS
>NM_001782 1 frame -1
PGRQTEDEGDLAKGATEEGLGAEAEQHGEQTEALLHMRLSRHLLSVGMDNASEKLLLHLT
YFKKLAGEPKTMNSVFQAGHIQNLSTITLLLLLKFTVAKWWFREFILDWPQLQGLEVDYT
THDLCSKLKMQGTNLVMVDTGVRVMKFSSLHLDDSFQVSRL
>NM_001974 1 frame 1
QKAPRKMATSAAKGFSSNVRKM
>NM_002155 1 frame 1
SMGRTAGTSVPCAGCAQPVSAPSAPCPPAPRPPWRTPCSRAWTSTRPSLVPALRNCAQTS
SAAPWSRWRRPCGMPSWTRPRFMTSSWWGAPHASPRCRSCCRTSSTARSTRASTLMRLWP
MGLLC
>NM_002155 2 frame -1
GRVDGGQMESAGSPAAGCGSPVSGAGDSRWGDDHADPEERHYPHQADPDFHHLLGQPAWG
LHPGVGEGHDQGQQPAGAFTQWHPSCPTWSPPDRGDFHCWHPERDSHQEHRGQDHHHQQG
PAEQGGGGEDGSSRAVQGGGPEGQSGCQKLAGGPCLPCERFFARGKPGQDSRRGQAQNAR
QVSGSPCLAGAQPAGREGGVASEEGAGANLSPHLLQALWGAWCPWGQQLWHSSPPGGPQH
RPHHGGL
>NM_002287 1 frame -1
VSPPHRPPGPSALPGPDHPHAGGRSAQTLHLGARHRDPPGEPCDFRVPGPGWGSNIPPGE
GQIHIQYRCVSSS
>NM_002348 1 frame 1
TSSGLVPKMLLLSHVPKKMPLWSKATWADT
>NM_002362 1 frame 1
CLLSRRVSTASLRKALRPKKRPWAWWVHRLLLLRSRRLLSPPPLLWSLAPWRKCLLLSQQ
VLPRVLREPLPYPL
>NM_002493 1 frame 1
GGKWSMGYTKRVSLFSLMYLYLSGLFIITSIMFLKNHMALKRSPEYSLVIQFWRLEKFHQ
KNFLINII
>NM_002571 1 frame 1
CCASCSPWAWPWSVVSRPWTSPRPSRTWSSQSWQGPGTPWPWRPTTSPSWRHRPLGSTSP
HCCPPPRTTWRS
>NM_003123 1 frame 1
WPRFSFSLGCWWAQTLWGAQQQCRHPPPESLWSLLASPAQRCTPLQQVTLRPTALGTRPQ
PYLPQLPSMRDPLFGLPLVPALVPLYLSQQPTRKFPSRCHQCPRKPLMQPVILLFPQQTL
DPTPQVEPQRTLQKPPVGPVEPLLPRQLALWRPPEAPLDPLLPWQLSLWRLPKAPLDPLL
PWQLTLWRPPLGPLDP
>NM_003154 2 frame -1
IRLWVWPLSASSRTTTIPTTIPTTIPTIYLL
>NM_003185 1 frame -1

GGGLGSAGRGLLQQRGGRESGERPGGLAGVAAGGQRGPPPPPRAAHARGAGRGRRRAREP
CCERQPGRSRGRRAGRPRRGRARSGAGAAPRRSAAGGRGAAAPGPPLTAPPPCPRRARAA
RREAEAAARGQRGVLRPGARRRRRRRGARARPRRPRQARRPRRAGRPRRPRPRARPRPRP
RPWPWQARRPRRRANFEWERRAAELAPRRRTCCQPGQQRARRAAAAAQARRPRHCHPDAP
LRGRRRAPRARRALAPRRPRARRPRRRPAPATPRARHSGPAARPPRRTPDRRARRAAPRR
RPERGQRRGSPRPRPGRRGPRGGQRPTRARRGGCGAGAGGQGRVAQEGGAGGAPGGADPG
GQRPGQHG
>NM_003395 1 frame -1
QVRQGIPGQTVKQGSASPCGLPQQPRGCEGDQGWGGDHLQVPRRVRLMHGADLLAAVGAF
PGGQASEAQVDGTQGGQHHQSCRRGRCHLPTTGPCLGGRWQRPAAPHSRAGAPGLALLPG
WPLLPGHRWPVPPEELREHLLWPRPHTEPGGDKALPVPGALVLLCGVQAVHAAGGLHLQG
L
>NM_003814 1 frame 1
DIFSLKVMQQQCSMKYLTLSIWIPSIILWRLMFLELIYGLHQIHFLPVETIMFRTFLFGR
IITLIIDYNMMLHIFSKTHKASLVLPMLKEYARILLILELMFLKTTGWSFLQLLWATSLV
IIWVCNMTPSG
>NM_003943 1 frame 1
WAPSGPPCWLEGVWPEHFSFGCCGAALATPGRTGMRSRRKTPLLGELRFREAIRVAAADA
LDLPGRSWSPNQSIFKKAMDIFLRPKTLVNCKQHHGDCRILPGKSVTIQENMFLLDSFQT
QKLQLPLRPVTLGVTLKFQEMKALNLLWENGDSKKDKRYLLKQ
>NM_004221 1 frame 1
PCCSGCRPGGTGFWPGRRRWWPWSMQCRPSGNSSRVSAALCQSSSCPLSSPTEPHGGTRR
SHPRSALNPNPQN
>NM_004322 1 frame 1
CSRSQSLSRVSRKTPALQRGAWAPAPQGTGPQAPASIIARP
>NM_004323 1 frame 1
WLSAGGRGDREATGSGWVPGCAPFGQAGSRASRSPRPSVVRLPLGVHLPGVLPAGMTDPP
GAPPPALAGRGRRKPGAARPGARSPGARSPVRKRPGVKRRPRVRRRPRAKRIGARRPGTR
SRPGARRPGRKWRQLGSP
>NM_004585 1 frame 1
AWKMWWEAVAIGSTTAWTMSTNHGPWRSSVLRRRWLVRRSTVLAGTVSTLSPSDMASPAV
NRWKRPRLKSVWPRRLESWLLLDALLRLGDTKKKRQP
>NM_004590 1 frame 1
RSPRLPCLSLSSSLSLLRLLAASQKFLSGTP
>NM_004651 1 frame 1
PWIRAASQSSTGSWSPRKARSRIYVWLCPNTRASRQRGWWLMSSVTASISSISRSLAASW
TVMISSSMRCQVALRPLRAQERTSWFLSTCGSAPLPVTTTTPTTACFLDTPSWYQCPGTA
SPGRACITSCTGSHATPNPTQMMRTMGMRKKMTRRIKMTSLGPQLGAASETLSQSRLGPA
LESRTGARSSWTIALAHLSGPQGDDASSCSPCRRTPMGPATAQPPLKKSMPSRTLLSTGS
QRRSVTMTRRLRATSMTASGTRRLPCGCRSALSSSPLWRPWRRKTPGTALPASSTSWQPR
SWTCGCCRRFSSSTNAFPTPSSPERSWTPSWSFLSGTWTSLSLSSSHRMSRIRSCTNMTS
SRFPTIMGACVMDTTQHLPATRTAASGTTLMTTASPLSMRIRSSPRQPMSSSTNARTWRD
ACCPRPAHLAPQPPLPAAPHPALSSWMLI
>NM_004828 2 frame 1
RPPGLPATWSLHRPRPRAVCLPLQEPDKPLSLHLPSLSLHSHRTPRSALALQPPLPWCLC
SVDSSPRAWCCQPCSSGGGTYGGKPWSSGAWIPKKPPATFNRSRTFPGPQFPHLREKYYI
TLLQGLRAMMMMNTLC
>NM_005169 1 frame -1
GLLLPQFVRLVRGGHGGVRLRRLWRLQPARRLPIQPPAARFPRGRAALPRARLLQLRTWR
PTRPPARALLGSALQVLPRAIRPAREAQA
>NM_005354 1 frame -1
GRCRRRRRRRGALGHGHGLRAPRRAGPGGGRARSACLREPEQLRGRRRGRGGRRDGRLRC
RTCALPAA
>NM_005354 2 frame 1
HPQARWGRRAWLRSRTSHRRCPTCRASARARRCRPSTWTRRSASRRSASGCATASPPPSA
ASASWSASRAWKRKRPSRVRTRSWRPRRACCASRWRSSSRKSSATSTAAASCCPSTRCPR
T
>NM_005475 1 frame 1
TGLPCSPPRPLPRPQPPRRRPRGAGASSVSCTPRRPGSWPASTGCSPGSIRSTRRCAPSW
CRCSSPTSSSATSAARCATDGRRAATTGTQAVGPQPRPRRPRSQAPAPPPLACPRPAALR
SWPRRGRPGPAPSSTFAAASATSSAAARPGSCQRP
>NM_005612 1 frame -1
RSGGAATSHGACSDGGCPDTACSRACSDGGGSGGACSEGAAASRGACSDGGCPNCTCSHG
AASSHGDCSDGGCPNGACSHGTCSDGGCPGRICSHAGG
>NM_005623 1 frame -1
LHKNHQHPMSQGSCDLQDQTGQGGLCPQGEMGQGFHEASGPNISKSEAM
>NM_005632 1 frame -1
GHGRRVVLCALHLPEPGRPAPVLHLRGSPAQARPQPHPAAQRGGAEMALRPLHLPQLPGQ
GGLRGVRLHPGACAWGCLPASPQRGPPQATRHPGGAQGQLPGGSRSSEDCGA
>NM_005699 1 frame -1
DHETQLDTRPQPFVGPAPVCPRRHSPGQ
>NM_005867 1 frame -1
LQHRFLPEFQPASFSYGFWTCQPHNHNQRGAAILGKMSDLEIPTLPPWAPQEIRWEKGQA
DKRKPINL
>NM_006016 1 frame 1
CRGSPAHCFGPPPAWACSACCPRTRTRPSTRTRLRPSPTPRRRRPSRWSPLRHQKPVKVE
TAAFPVLMLALLI
>NM_006061 1 frame 0.5

MFANIVLLVIGLIDYMSLMNKEHLVPVAQITVTMDYAPMVASTKISIVTVKVSSHPVNIS
WSGTVARPPAIVQTAFI
>NM_006417 1 frame 1
FIVKIILLEHMQKRVTRKESMLPSSFLHFKILKFQNGNDYVHQKHCFVVMLQNITPQLIS
RMEEIEKLWT
>NM_006505 1 frame 1
KRDLPVSTQACPVTPSSSWFWESWFFSCWGSGFISIGPNVPVRSFGTVICVPRVQSMPAP
QLMGMSPIQLAERTALPRIHRQRAQG
>NM_006632 1 frame 1
SLVKPLGGPLSSISLEVLAVSAAFSGLLF
>NM_012473 1 frame -1
FQSPADPTMQSWWPDCNTQPSPDNIH
>NM_013431 1 frame -1
HCGPRNHLHCPDGHCVKNNSSYSLYWSTGAEQFFPEKNAESTSLWPLSGVDYIFQQLLLH
WGKKNLGRKSLLACASKNSDLLSI
>NM_013981 1 frame -1
VRAPAAAAAARARCVPQDVGGAPALAPLAPQRAGGAARTGGEGLAVAEQRLGRRLSLGVG
RRRGRRGRGAGGREHTFPGPAWGARRAALGLAATVPGGRQQDLLLTGQPQHAGQQQTQPR
AAPAGQAGLGATL
>NM_013983 1 frame -1
APRPGTRTRARARARRRHAAQLQLLLPRGGARTAARDLRARRQPGQPACQPLPHPRGRRV
RDHAGVRAPAAAAAARARCVPQDVGGAPALAPLAPQRAGGAARTGGEGLAVAEQRLGRRL
SLGVGRRRGRRGRGAGGREHTFPGPAWGARRAALGLAATVPGGRQQDLLLTGQPQHAGQQ
QTQPRAAPAGQAGLGATL
>NM_014117 1 frame 1
SMMSITLSYQKKWHSTWTMVELESVPQPSVAMSLQTHQPINATES
>NM_014143 1 frame 1
LNWSSQNYLWHILQMKGLTWFWEPSYYALVHHSSSVEKGEWMKNVASKIQTQRSKVIHIW
RRR
>NM_014334 1 frame 1
AHGWRGRPGSTRGSGRRCSCCYRGPPPAQPAPRTTVRARGAGDPGDARGGTRAPTRRCRA
TTPPTAPSRGS
>NM_014385 1 frame -1
LSGDLAWGRRDHEQDHPTQCVLPSSELDCDCLPRRRHSIHSSGEQLISFSPRGPVSALGL
CCQQSPCQAELDLEESDPVPLTALKPSGTGAASAPGGRGIHLSSSELSGFPARFPEPLPA
TGVHRQNEACIRSVAGGGRGSWSHSPGLPLLLCHLHCSEVLQEEIGKASSGRGRHRHEGC
KHHQGLSLSGPDVLGRPPTPWPGCPLLRGGKRDPVCTPQLSGGASGPIRTRSHQQVLRDQ
DPQV
>NM_014442 1 frame 1
CCCCCCCCPCSGGQRGWRETDNMGMVTCCKCRSWRCRRACVSMCPAPSPTPRMAGLTLTQ
FMATGSGQE
>NM_014481 1 frame -1
VGAVDAAAQQSNPGTDMPKQSPSALNQASAQSGWL
>NM_014622 1 frame -1
GALLWPTHPPPGASAAEEYLCEREHLRVCGWCVCNFELREGESSFGGLLCVPHGRLCCLQ
LGLGGWEENCSRITRQDEGPHQLESHLPGPPGLLIGGGQQLQGCLLLQCGPPTWVEGGSH
PEVCAGAASGSRWGSALCAPSCPESIPVLWVVGQLPCEDSYSPCGGPALHTQHGRHHRFP
AWHEGPIQLPLESYRVPRRGQDFCSGFPGCWTQVSGRGTPDLLQGAYPQRGFGDGDAHEA
RSFDGRSICNGEFLSKYPRRSTIKYLWRVYLSHGPLGKYAEPHEPGYISAANTGSQGNTD
FAAEEFTYRLLFQHLWIWLFLGMLSGECEVHSANNGGGSGESEAYAGRPRGHNLGTTPEH
LQGTLHPRPPPTAFCLYRWRSYRHVCNRSDQQTETQVFLIWYWRRHLHQPNKRYCPGIRG
HLRIYHRQRQDAVQGSQDSETLSAACGRGCLSELAFASWSVCNAFPRTDCHLGSEINQLC
PADREDASSRDNRRSMPQIYTPGQDFGGDISSTTQACQPHHSPPCCQVLAPDQGHGPQGD
SSKKRCIEPPVWCHKLLHSFHCYQGAQQAGSGASGSGRPKANSVGCFCPIEDKMPIRFSK
GLTLPSSFCISAQRGTYVLGQDIPDGRLQSLWVDKSQGPAQSRLWRESP
>NM_015672 1 frame -1
ATIATPAATTTVSPQETQQPAGRRRLEPALRSPDSLTPGPGPPRADKAELAAGQGVVGAG
APLL
>NM_016148 1 frame 1
WPCVLVQDPPRQAWRGVRWLSQKSHRCPCRRPPLCPGSCCPGRRARAHRHHLCPGPWPSL
RPQPWPQKPASSVNSAPSFSSLGAPRQLAALCPGPEEAVGEAETATTGEPAMSPRGPPPC
SGRDSPTTPSPHSSPSLSAACFRTGPNHLCRHSPPEQGSPLQPLRPQGPPHPQPPPPPRP
PATSRAWSSRCGPLCSAGPPAPRCCPPRSTRSALRPGPRPCPSCLPDPSTQASLTSVAPQ
LEGQEARLTPSPQSLCRHTRGYPGGSGEPCQGPRAPSHRPACSRCPRTSRLALNLWGSGP
SSTWLIGWSGWVWRSTEPSSWTTRSMAPTCPPPRRTTSIVPGWATATSTGLSNSSWRG
>NM_016382 1 frame -1
AGASGHPHTPPAPQGVSGQRMPGISPCGHLGSASSVTTKQHTDEGQHCMEEVAALTKWIS
SHIEVGEWLFAFQYFQIQFYSQELESSHQGSSAAGQWPLLPGGHQYIWKSSDSHVPGFCI
SETPPTG
>NM_016449 2 frame 1
KMMTVMMMMMMPRFYRHLSRLVLKIACFDAHDTKMKKKKMMMTSTQLGKVTRWRVVMKRF
IQG
>NM_016638 1 frame -1
ATLHVSAGAESWLPSGWAWGPCPRSLPRSFWPLRTRVCDLVTGWARRKDQGASVWSSQSY
GEGRTSRGGGRRARGSPSSQ
>NM_016638 2 frame 1
HRGRGCCGRAPRGQPWKWAQLRQTQAGAAELEGRRPCLPLREVGAPGTEPGASERKAAAP
RFPAGQPPAQRPGRKPPRRRFLSRRRGAGCPARRAVE

>NM_017810 1 frame 1
CVTTELYPSVSKNGCAQALHKGPSTGVPPRGRTVTSRWQQVCPGAMKRPRRSWLFLVVLN
FMENSRPVSRTARSTGPWRKDSGSRVFCGPQNSVAPSSKAYSVTAKGEAVCLSLVSFMRK
MLFQAPGPLHLLW
>NM_017851 1 frame 1
CSPQAGRPPGRRRAATCWPPTSGAASSPRPCRATS
>NM_017856 1 frame 1
GTYLLRFFPKALTIMRLRILSPSMTIMWPGRTTPAVLHISEDLGS
>NM_017910 1 frame 1
PKQLDHKDESVLRYEKTTMIWLRRITNTGVIHGNVMKSGQTMWILFIRTFQEQPKTNLHL
TQLWICILMLLCLFFTHILSMVVYVLYMTSHRLLNFMEFAPVNLLFHVKRARSLSEIGWF
ALQNRKMEFLKKNLKSTQMYNILKRKLELKVSCFKRMTMKNRILIFHMDHFPMLLDQYTG
NLVIQLFLSSGRSNHNLT
>NM_018102 1 frame -1
GNPRTFDTNCILQQQNTDYRKRECIWENTSGHECSLKKNALMSRRKQFENFRSNCCKEKQ
RKQKDSWIQWIWEAEKSFGNEKIQIQSNEESQQSK
>NM_018179 1 frame 1
MQPLVMPLLVMCPLVMPPLVMPLLMISPLVIPPLVIPSQV
>NM_018263 1 frame 1
SFYPVVQRLSPSLRPRPPQARHSLIVSLEHNYSKPPQCLQHLPSVEHAQVSHHQPTRNWI
M
>NM_018300 1 frame 1
CMGERMMHKSSLLKISLDTRSHIYQNCSYFKLKGKYINMITWKNLSTVVPFPHPNVFLLL
SKPTFLIHMNVILWIHYSHKKRKQILGQNTTNV
>NM_018988 1 frame 1
SLWWAQPGACWPWAPTCTGSATAPRSRSCWCRTPRRATPCFRRRPSATSPRPTCAAPSRC
RRCARPSRTRTTGARGMGGPSPWPPPSTTACMPC
>NM_019036 1 frame 1
QISLRPFRWELMWWTPQYPDVAALMQKVLLGMPLRIYICLMAWGSIQVIYTKWKLVTLFA
KLIKPQTLKHKPPSML
>NM_020415 1 frame 1
KLSVSSSSLSWGCWCLARPCAPWKKPSMRGSRRSPAPYLGQAALAWSAR
>NM_020959 1 frame 1
WPRPPPAPGARPWRASVARGPRRRASLQPRRPEFWISFSESGSCRLVATWCPTRRGRRCL
QRTATCPSQTRPMTTRCYGCTTSAWAFPSSSCKSATTATRVPTPS
>NM_021181 1 frame 1
SLPGSSVKVLLMTQIPPWSSCVSCWCPSCSVSLYWGYFFGFRERDKKSTLKRRREWTFVG
KLL
>NM_021225 1 frame -1
EINFLLGPVGSYFMFHTQESKILQKTISTWPAATTSTLQAKMGSTKSPTSLLKTFTTFSS
LCPRASSTIFFLSIPSSHSISTLSIGIYTTSTLSGLSKPTFPTKTLLCRTYDIKTPI
>NM_021602 1 frame 1
WPGWRCLLCPATGWWRCCCCSQLSQYQQPDRR
>NM_021706 1 frame -1
VSPPHRPPGPSALPGPDHPHAGGRSAQTLHLGARHRDPPGEPCDFRVPGPGWGSNIPPGE
GQ
>NM_021967 1 frame 1
CQLGHTSLRLQGRILAFLFQLLVAPGITWLMAALLSLPLSLWSSFLSSSLLFMRTFAVGF
>NM_022053 1 frame 1
CALLRSVGHTELKLQNAMTMVALFKEERKVGVLSGIILTRGAVIMNMVGMSARLHTARRM
MEAWRGMSTRTNNDTLLIASDAKEENGIVKTKSVLPRGEIENLRREKVRTHRMDTQGTGL
R
>NM_022147 1 frame -1
PKDWEFLTWNWCCVPRKPSQEPVSGKRGGEWVEIRAQSRPRSTEHLCLYFAACIYCSQML
YIRM
>NM_022479 1 frame 1
TTQQYCIRAMAGDHSLPATPRKASCTWVPWGPPHSSLTPAAWWTTPRVGCPS
>NM_022757 1 frame -1
EARHSAGSFPEAEARRAGQEGPGAHGGFFLPGFTSLGLGFSEAARDGQVWSSTGPG
>NM_022835 1 frame -1
AWKQPGSPGPRRHPTTLAMSPPRPSDSRYLTFACTWKPPGPSDPSQRPTVFVPGAPRHSG
SSYHTFAPATSPHRHLGPSPPNFTQAGKPPRHPGSSGCTSTSGAKPYRYTGPKTHTFVGA
EEP
>NM_022838 1 frame 1
WLTLGQKQGLEERLRLACKMESVVLPLLEMVKPRPRQWLRQNKQNQPRPKLVMEQPGHIQ
PTGRLWLQGKSRWKIQLRLESWLRLRQNPWQNAVCHKPSQRPCLCLGSVLPSLKSRLLLS
LRQILGPMPSHMIRPIL
>NM_022978 1 frame 1
CQLGHTSLRLQGRILAFLFQLLVAPGITWLMAALLSLPLSLWSSFLSSSLLFMRTFAVGF
>NM_023080 1 frame 1
WRPWDILLGRQRRPQARVLPARPAEPGFPAQFPAPGILPLSVSAQSNLRAVTLTP
>NM_024114 1 frame -1
VSKNHCGNPSKNPAKHEFWNLASLPEGTHLPHLHELLHRPGHHRLWAQLLQALFLPQLAR
HPNSYSVLMHKDNTAEKPQNHSIEEDGFPCQKSQSLAIPELGANVWHSQGDKEDVLSGQE
PALFAVLQLSGAPVSQTLSRVGCGTLGEAFKENAVFMGKSLKSEKPECG
>NM_024616 1 frame 1
WTWPNPHSQTSWSSRCSSGSQKRTRLSMVPTRSSLLWTRRSPRPRLTRKAPGAPVIRMWL
EDANCGSSPPFSVSLQSSA

>NM_024650 1 frame 1
ELCALPGPRARQRPPRAARKPPRPPSLLEEGSTARLTAGPWRRAGRASEAAWRTCCGCRR
SPTCQSGVPALGP
>NM_024786 1 frame -1
EFKTSSSEERSIRANGQRSSPARSWRPGLICTGSQSQELPADSQALMSLLHFSKPGWGFD
GTGRGRPVSICTWSQGQELPADLQAPVSVLHSCTPRRGLDGTGSRCPEYIYTWAATRNNR
AHENQCKRL
>NM_024840 1 frame 1
CWRTIATSCQWGIKPANQMHSSSWNKESHGQKMKSTAKSVQKSRKLTIIYRCTHKSKDVR
EWNNAINIMHLETSFIRGKVIFLGKIMIHLTYMGKYNQIVSTRTKGMKSRILWGLMEMGN
PSFMPSMNN
>NM_025099 1 frame 1
WRLAGPRSLPPNKPGLRMLRSSSKRPCVQLSRSLMSSLHWLIVRLSGCPREGTKVLHCPS
AIASSQYRTSRLTSVSHAAATCRGAVVHTRPGPKRLDQMGTPCPESSCYFGHQTYRQTWN
KSAGTEASMEITLASAVSSTWTFLGWAIFFCSPVGVTSLLPGGIPQGKGTWSCGMPLCQC
FLPSVLAPSRLSLSSTQRVLPACSGSETSSEVCSETWLGVFDVLWKVNRKLTSSCLLVDH
TQLSPTCPSSCRSLPSWCGTEPFGLVQPMCQNCECPRSVVSASMFGPVSPPVCCCNQNVC
RSWNWSWKDPSRLTPSHSPCPATRRTRRIQKVLSGILDSYPIREQSLACMSPLASMSWMG
SWGSALPTSSSVALGGCDLECVCSSRMFTCSSQWEGGQEGQCSPPASVAPFCFKASLVRS
LGLTHPVKPTGPPCTSSWCGNVSDFPSTCGLPRPWRSWPASCVPMCDTTSSCNIPLLGAP
AWDCNSWLLPWIFLRQAALFGMHTMRSLKSHITVPSRNTLGCRLPPPSPLWPPKKKDSVR
PGPPLTLRPFCPSRRPPTCPAANSIAAWLGPGSVCC
>NM_030642 1 frame -1
AMWQTRKFASSRFQGVTWLGRRLRNVASKGNLRRGLGEVPRTVPLTREPVPELENQF
>NM_030642 2 frame -1
TESTCEAGAGAGPAHPAPPAPAAEGEPDLFQLPGQGCSRIPCGTRRVSLTSPLACCGAPA
AGPWRGTEDTKEDSLCPKDAWPPASPTSTSKKGETGPGKTPTM
>NM_030763 1 frame 1
CPKERLQVKVIGRSQREDLPGCLLCLCQLHQRSLKEHQVQGKRQKVIWKKTIQVPKQLLK
PSKKQLLKKTTMKMLKMEKPKLQRHQLLKKKLWKKKKILKMPQKREEKRKKQWQQKKMKK
KIRKKMKKIKTKRKGKLEKKTKMK
>NM_030764 1 frame 1
CCCGHCWSSLMQSLNRQIRPLWRPLLSSKETASFNARENRTGKFRRWLT
>NM_030764 2 frame -1
DSWSSLGTVWCPWFHWCCFAVVCLVPQDIRRKFCHTQRGFQAKSSRVHLFKPNPRHGGAA
ASVCQCGLCRCGCGLFSGLEHAAARKLSKHQDTSGEQGLPSHLLFCEEII
>NM_030776 1 frame 1
PQTHNPEVTFLETLVSPSLPATRSSPPSWKLLLETGVTKLQKAATMWMKPHTRGAGGEVG
F
>NM_031289 1 frame 1
WPRWSSRRPSLASGHSYLPSSACYHSASPQHPCSATTGLWAHRRCPSPCARKVWQPSALT
CQCPWMEIPTHPPRRWYNTTGRLGM
>NM_031289 2 frame 1
RGAEVSLNLHHQPREVRKDYWNLPRCKAHVTPLSDLEGSGWRRLPSPPLPWGFVAKILWL
SLGTQITYIGLQFISFLLLLTDLLLTGNPACGLKLSAFAAVSSVLSGLLGMVAHMMYSQV
FQATVNLGPEDWRPHVWNYGWAFYMAWLSFTCCMASAVTTFNTYTRMVLEFKCKH
>NM_031440 1 frame 1
LGPYRLQASPHLRPQEYTPFTRWRRLSPGPQERMSIPTHAKTTSVGTAFSAVVSFSLLSW
LLKLLY
>NM_032343 1 frame -1
QLESPSQRIHTAQVGEQWWPAAL
>NM_032507 1 frame 1
CLCLTQSGPRHWRLRKKQPRPLLQRSGHIVLGGTSVGTQLRRQLASVRLKKLLKIDCLKQ
SKKLLKKWNKVEKPQESP
>NM_032989 1 frame 1
CSRSQSLSRVSRKTPALQRGAWAPAPQGTGPQAPASIIARP
>NM_033048 1 frame 1
VILLKKEVIQLEMLTKILKMSLLMTVVQTSLHLLIPISIRKPTQTGSQAQQPPRKMLFLK
QQKTASSKQRSKKINEKKIKKSFFYFRPPFPENWSLTSPMIDQDPTQGKLR
>NM_033163 1 frame 1
WAAPAPRAACCCTCWSSASKPRKARAGALRWAGSSLPCSGLAGSPRVSPNRL
>NM_033188 1 frame 1
AASPAAAAPIAVNPAAAAPAAARPPAAGPPAAGPPAAAPAA
>NM_033292 1 frame 1
KHLEITLICKTLKEYFLPFQLLRQCRTTQLCPHPQAQKGMSSFAPKKLKGYGNKSRQRFI
QWTSQAAHV
>NM_033293 1 frame 1
WPTRSRRRESCLSVPWVKLLRQCRTTQLCPHPQAQKGMSSFAPKKLKGYGNKSRQRFIQW
TSQAAHVL
>NM_033294 1 frame 1
WPTRSRRRESCLSVPWVKLLRQCRTTQLCPHPQAQKGMSSFAPKKLKGYGNKSRQRFIQW
T
>NM_080387 1 frame 1
RMNPTTLREKTVLFLFITKINGPGMMFLVTLKQVGFVKYLEQHT
>NM_080659 1 frame -1
GKPGLLRRKLELPINFPEEKENRKPNKTDTEAAATTAAAESPKGPNNR
>NM_080759 1 frame 1
PQSPLPAAAAAAAAAEAAAAAAATEAAVAAAAVAATATPTWRPRATAAAAAAAASALAAA

SLPAPPSTPAPAAAAAAVAAA
>NM_130759 1 frame 1
RCMSWRRCCAGQALRSGSGGWRSAWQPGCRGGHGAPGCRPGCGSGSPPGAGGWAWPCCWG
ARSCSGCCSTGGGRRPLRRSGLT
>NM_133273 1 frame 1
TKITRRRTSAWPWQDWSSWLSWPYWLKIGTAIRHTRKPRQMWLNRAGANRCVSQDPLHEH
QVSAS
>NM_138770 1 frame 1
TFPGPLLPWALAGSASTECGSREDPALPGRRAPHGHTGQEASVGVLRAVAPTL
>NM_139161 1 frame 1
WRTPGWGCFWRWACRSCWPAGAEPGGKYRPLLQMRIALFC
>NM_144505 1 frame -1
GTPPTSCGQDVDVPALAGGSLGSVWKPGPPHVVCGELAVCPFEPTVAFPALALPQRVAIQ
PS
>NM_144565 1 frame -1
GTASRPAMSQGGEGVQIGSPSQSLALWSRRVRRALGRTYWRLTTAPPRKRHRTPVAVGIE
RKEGMWSQGYVAQAGLELWPQTTFLPRPPKVLGLQACPGLLPSFGAYPLCYLPPHTSNL
>NM_144644 1 frame -1
GCRRPGKRVFDTDCGSPRQVTVTFATASSS
>NM_144706 1 frame 1
CPQLWGSVIKLKKHFHYLKRYLQITFTFVQLSWDFHLVNLLLRYLLYIWIQKWMITYE
>NM_144956 1 frame 1
PCHPSGACRPTTPVTSYRISIALATWGIHPMTLPWSCLHLSPTLNTSSPSVSRPPHLSLR
TGQT
>NM_145043 1 frame -1
ARRAVGEEISPFGLPLCGSAGGQDRGQQEATARQPAVSVAPGHPGPWKEIIPISRRNGAP
WQQPNTRASTKRSAEGRGCGPKAGRGAQRAEDPWILT
>NM_145301 2 frame -1
FFHEDHGSSSLLETCAPPFSTSSCGSRQLPTDGARPPISLLAALGLEGRWGESQQRNWAR
AKISSGIRNAL
>NM_145650 1 frame -1
VGLSQNSVDFNVVLFTSIGEPWKRKSRHKHNTEHCRSFNNGKTYFFGKSKLKL
>NM_152356 1 frame 1
WERDSLRVQKVVSVEKLSPRFRKTCTRKLFLENHVKVVHVEKSSWDIHPLIGTSELTLDT
NHISVR
>NM_152394 1 frame -1
GPLALARLRRPGEEGPEGVRGGVRGGGGGGRGGGGGGGRGGGGGGRGGGGGGGGGGGVGG
GRAGVGGPDVQRRVP
>NM_152399 1 frame -1
GFRSYQDNLGCCTNFQTWHSCRCRTDAIWCDSTEERKFTQSVPCTGFSGNGKVPGFVFGS
LAGVETWN
>NM_152400 1 frame -1
VLSRGAAAGRRRRGTRRGRGRAGGAGGGRDRVG
>NM_152404 1 frame 1
KNHTKLSGGFHLKIIKKELRSILIATKQHWMAEKNLKP
>NM_152475 1 frame 1
RSHLSSASLGRTFCPVQDCAKKQLLRRQTVKLCMAHPFRREKLITVVENAQKPSAPNTQL
FH
>NM_152499 1 frame -1
SPLRGCTSPVLPACTSSGALPSTSRPVGYPPLGTAASVQPQGRASFHTHVQCSTPSPSL
>NM_152544 1 frame -1
TRPGFPRRGPGSQVAIRTRAGHGMLRTGGPGPVPARGGTEGSRLSAPE
>NM_152554 1 frame 1
SIHLTSSGTFPGSRGSALLPCLPREKGHLMAWSHSASQEPGPLGGYSMPQIATGSSPLPG
PALSAPTPQFLPHLPTAVAVRPGAFP
>NM_152554 2 frame -1
HFHGSHTLQDQQPGHSLPSEKPRNHLPSEGDQSFL
>NM_152559 1 frame -1
PPVDRWELATSELEVVSGISAKDGFISGIVYLYRKWKATQVEEV
>NM_152569 1 frame -1
STAAASGTGRPRSQARPAADANGLVRPRPLQRSSASHQRWAVRPQPARRGGRLTCSARTR
GALRSAEWPWPAGGSVASRVAGCGKRGPGAPQSSRGTCPAKRKPRLSVGAGNFRISTFAE
ITPGEGPSSEQRRPAPPPGLARRQRRARRAWAQRGTGLPILGQTARSARGRRPPAEGSAG
GRQATTARRPEAPSARSCEAARECRQVGGAPQSWRGRESPQGFV
>NM_152577 2 frame 1
VLFSSAREHTCLRKAYRLISNAAIRELENKLPALRLKKFVLILLRHKLKSLTSLKDCKTG
TIAIFHQNSTPWCHYRLCNICYKSNIISHIRISRLLPQNYLVCLFPSSGKPLVFLRENMA
IQLIIFRITQILVLRSQHLPTIILSVKVNQRYRPLIILYLRNSIMRHLHLHHHQTIKCHI
LLRMLLLTRFVAKCQL
>NM_152577 3 frame -1
DLRSIIWIYYCKGATYEFSSTTCSPVSKWSICKICFSPLPHFTSVHRKSRNLEPSVYTNR
CNGSKVACMETTVTLSTNAESTSFNPTWSITSFTPEKTTVL
>NM_152769 1 frame 1
GVDLWALAPLRPPLGGPGSALQAPWTRARLPPAARPLPEADSLGPPLQHGPEEGSRERLG
>NM_153038 2 frame 1
SRRYPTPPGRLLRAHSPQTPGGLPALGHVLSRPLTLSKLGKRRRRKRRKREREEKRERRK
EILKN
>NM_153714 1 frame -1

DGLGPGSQGSLCHEHSYMGSLLFLLLEGDLWHTLGGHESQGHRAAGLLRAEARSSGIQAP
ANARVHE
>NM_153823 1 frame 1
WPRWSSRRPSLASGHSYLPSSACYHSASPQHPCSATTGLWAHRRCPSPCARKVWQPSALT
CQCPWMEIPTHPPRRWYNTTGR
>NM_170664 1 frame -1
VQAGAAGLFLQPGIAGCHCGSSPALPDDQVQPPEGLPGWRPEAQSRTPGHCHGEPDPQDP
GFFVGC
>NM_170685 1 frame 1
CCLASPCFSWSCPCALWQVMVERNRHSALKQRPGLWPWRKALAPAFSSSCRR
>NM_172131 1 frame -1
GVLKEGKQKCRTNFALGFWLHLQTIPQKEEQHNPFSQSSCIVHRNQGLEATQHR
>NM_172314 1 frame 1
CTRWLHSWQWSWEPTPTATGPAAAPAKGRTPLRSCGGALCLCLPSLLGPTATQS
>NM_173499 1 frame -1
PQGPGVACKGREQLFSRKDPKGSKAKGAICFPTDSEDKQEKCPVSPILLVL
>NM_173625 1 frame 1
GSCANASGCGDGKRREASHLGQLNPSLTLSPRRLDKMLPIHQTQRKLQRSLLSTRHTS
>NM_174881 1 frame 1
WRTPGWGCFWRWACRSCWPAGAEPGGKYRPLLQMRIA
>NM_175739 1 frame 1
CKDRAGEEEPAKTYFVPKWHLTFMEYSLLLASVLQSTVCPRPMPPVHTPALPPQRAPLPH
R
>NM_175872 1 frame 1
KALSSVQTFPGNRCSRTYTTLSERRRARLPLRPAETTHQISFPPAGRVGRTLWPQQGFCS
VRSLPAMGSRTR
>NM_176786 1 frame 1
CLRLVMGLGRCIWEVTAARTDRHCCRELATVFHAVAGGSRLHAPFERGPQPSPLQAGPWR
VRPGETWAPGSWPASASAPVSAWESLSQGKDKG
>NM_176870 1 frame 1
WTPTAPAPLVSPAPAPAPARAKSANAPPAR
>NM_178173 1 frame 1
LVWAPYSLENLMSGVKEQRMMISKKRLHCQHLGPMKEIGMRTRWCRLTARTGLLLKQVPR
TMVPASQAIRFPVTGTWFPKEPHSSHHWRTFQPALRMPAKNIKPKVCFCVTHVNIWL
>NM_178553 1 frame -1
KAEGSRASGQGYSGLWASIASISASSSSGPGRPSPRHTLPDQELQVCRPSITKLPTMFLR
ASTSGTKTGHYLPEAQALPWPKEASFSGAHSPKVISL
>NM_181352 1 frame 1
MQPLVMPLLVMCPLVMPPLVMPLLMISPLVIPPLVIPSQ
>NM_181684 1 frame 1
CVIPAAPRAASQPAARPAPASQPVVCPAPARHPAVCLWA
>NM_182506 1 frame -1
ASRSEETPCQGKTPSGPRRAGRFDRCSGHFRRGGRISPLCLCLFEGCFPEFTWGIQQSPW
TSGSPIHQHICYSCFTHKTSRRSQRPNGRK
>NM_182625 1 frame 1
LLIYVFRICLRNEYLQNYHILRIIYNQMSTKLCPYLVKNLVLLTVVLIVHHIFQRIFQEF
PCKMNPETLKFKETSCFKKTIKSILLSLILSVTQWRPAMLDHQILLIIVEKLICKPLG
>NM_182646 1 frame 1
AALARPQPRARAAPPSPPMRAWPPRRRTPRRAPWSPPTTLCSTVPVTSCPEVRLARAPSA
A
>NM_182704 1 frame 1
RPSWGPSRRPSPYRIVRKPSPPPARTSRWTRGSFEPTVAS
>NM_182739 1 frame 1
LGGKWSMGYTKRVSLFSLMYLYLSGLFIITSIMFLVIQFWRLEKFHQKNFLINII
>NM_198180 1 frame -1
GKALPPDLLPLPAAGRLLPSTGQKRAHRRHGWPRSWRTLGRPGHGAPTPLRVGFLSVAES
F
>NM_198448 1 frame 1
CCLPWPCPVCPGCCFPASFSCVRFKVKKPRRNCPLHGSAVPKAPRPMAPPAMPCFCHQ
>NM_198451 1 frame -1
HFSFPSAPTKRRRVLLRGQSGRVSAIFLQAVSFTEAGYPFPQLAH
>NM_198480 1 frame 1
ITCKIKVFRRVNSAMNRICLEILLIRTKVISCSKIVIRLTYMKNLNQIVLKTRKGALART
LLSLIEMGNPFFMLTINNFILKSFLQL
>NM_198546 1 frame 1
LGQNLPPWEPRSPPPCLALSCWALPWTWAGEGWNSCTRAAREPATPRRGRSRKSRASKSF
I
>NM_198845 1 frame 1
CREPRKPPPQRCYRCCCPCCGQGPWLRSGDSSWRGQSHRCRRVCASSYPADCPLPFQPRT
MVMATGSWKGLMFQWPQTTQTKKCRRRPGADSTSSGIPEGRTAPASEMPGGGTMLHTSFG
SPNGNTVIHLPSSLCVWPPTGPTSPSQGPWSLAIPAIPALCPGS
>NM_198846 1 frame 1
CREPRKPPPQRCYRCCCPCCGQGPWLRSGDSSWRGQSHRCRRVCASSYPADCPLPFQPRT
MVMATGSWKGLMFQWPQTTQTKKCRRRPGADSTSSGIPEGRTAPASEMPGGGTMLHTSFG
SPNGNTVIHLPSSLCVWPPTGPTSPSQGPW
>NM_198928 1 frame 1
REAMTPGAWHYPRNRVSIQNLQRPAQPYLKAIS
>NM_198928 2 frame -1

SSDATNNPIQQFLHQEGNPPRSTVIACKLNHLTNCGPSSFFHFHQPPSSSSL
>NM_198930 2 frame 1
IQTREAMTPGAWHYPRNRVSIQNLQRPAQPYLKAISRLLRCHQQPHPAVPSPRSPRTRIS
N
>NM_199243 1 frame -1
RRAPGRPRLAGGASLPRDLRAPAALAPAGLRVLRGRRGLR
>NM_199483 1 frame -1
IIFYTAIHYDWCTAPKCSLSSPKDPLDYSTGTSLGNPSPLELGRPVFPGPRISVLGISVF
CKGCDL
>NM_203304 1 frame -1
AQLARPARRRRGRGRRRRRRGGGGGGPRTRTCAPARGRPGGRARAPAAARTRRRGRRAPP
GAGPAVGARAGGRWRHGRGGGGRGRRSGGGGRGRRGGSGACAPRRTGRRAPDPGPRRGPR
VAAA
>NM_203373 1 frame -1
AGAHGQKEPPSRSSLRFCPHARLPAHLSWGGCVPAV
>NM_213633 1 frame 1
WGPSQPLPAHSASPGRGSCSQHHFTSGIRPQLPKSRLKPSHPKFLRGRMFFYLSTICPRI
LLATFGTKGKHTSTITLHHMTVKELYMGLHTVEEKEYIPMHPCSRMSRRRMQDPTPYTSS
DAMGLEELDI
>NM_213633 2 frame -1
LHLTPGDSQALHLQQQLKSQGGHGGCDLNLSCDSSRKLPVVDEWSEPPYDSQVAAVQNQQ
DPLYIWCHKVYCRTLMNTEPSECQPQPSHPESPPWSRPPQHLPFIHLLPFRRKPLLVLLR
RVPTGTIFLDNWEVSAIRTKALYPPNNYKAWALCLLCSLSHWQGKLQIHHSQSLLDITL
>NM_214711 1 frame 1
SFSFGPALYVLLLQGRDGSPSLVRMTMTMVTHFIHLIFLMAYGIYHLLFIIAQIQSPVTL
GILTLTQGYLRIPGFLLLDSPMSITSVVFPLLSMFLLSLLGVSRLSLLQGFFQQLQHPLP
HLLQLSLLQLHLLQPHLQLS
>NM_214711 3 frame 1
LHQLSLLQPSLLPQNLTLLPLLNRQIS

Chapter 3: Predicting prion proteins in *C. elegans*

Overview of Chapter 3:

Prion proteins are misfolded proteins where their abnormal three-dimensional structure is suspected to have infectious properties (Prusiner, 1991). Once interaction with a prion protein starts, a cascade of “infection” of the prionic state starts to spread across different cells. In mammals, the well-known prion protein, *PRnP*, spreads the prionic infection from the gut to the brain causing severe neurodegeneration and death (Kupfer et al., 2009). Finding such proteins experimentally is a difficult task depending on many observations and scans. In the last few years more and more algorithms that predict the prionic potential of a protein began to rise, with the leading one being *Prion-Like Amino Acid Composition (PLAAC)*. These algorithms depend on known examples of prion proteins and characterize their sequence properties to find more proteins alike. These algorithms evolve as more examples are validated. The core of them depends on stretches of Q/N amino acids in the sequence that implies propagation.

I applied *PLAAC* on the nematode Caenorhabditis *Elegans,* to find prion proteins and further investigate their implications on the animal under different stress conditions.

Scientific background for prediction of prion proteins

1. Prions

Prions are proteins that switch between structurally and functionally distinct states, one or more of which is transmissible. In mammals, the prion *PRNP (Major Prion Protein)* had already been well characterized in its two metastable states (normal functioning and miss folded disease-causing) (Prusiner, 1991). It has been shown that certain point mutations can cause an initial conformational change in the protein that can act as a seed for propagation. In recent years the pursuit of finding and characterizing new prion proteins was focused on the yeast *Saccharomyces Cerevisiae* (*S. cerevisiae*).

The main computational tool used to discover new prion proteins is a bioinformatics approach testing protein sequence properties. One of the most commonly used method was termed *PLAAC (Prion-Like Amino Acid Composition)* and is focused on finding sequence elements enriched with Q/N amino acids (Alberti et al., 2009). In their algorithm, if a protein is found to have a prion-forming domain above a certain threshold (according to a specific test set of proteins), a *PrLD (Prion-Like Domain)* was defined and scored. The *PrLD* included the amino acid sequence that will hypothetically be responsible for prion behavior under an appropriate condition (Alberti et al., 2009). Proteins presenting a *PrLD* that is longer than 60 amino acids were called core prion proteins, and would usually score high. Only those proteins that were considered “cores” could act as a prion. To date, about a dozen new prion proteins have been shown to act as prions in the yeast *S. cerevisiae*, and further screens are being conducted rapidly in other species as well (Alberti et al., 2009, 2010). An additional approach reported termed *PAPA (Prior Aggregation Prediction Algorithm)* computed the propensity score for each AA in a sequence to be part of the prion forming domain. It used the most investigated prion in yeast arose from the translation termination factor *SUP-35*, which was documented as having a strong *PrLD*. The PAPA method used a scrambled version of SUP-35 to identify those sequences that could still form prions experimentally. Under normal conditions, this protein acts as a release factor of the ribosome. When in its prion state [*PSI+*], translation termination of the target STOP codon is

suppressed, and read-through of the STOP codon often occurs. This read-through can and will affect other proteins and functions in the cell. It has also been shown that the [PSI+] state can transfer from cell to cell thus confirming a transmissible nature and possibly an epigenetic inheritance machinery (Nussbaum-Krammer et al., 2013). The mechanism behind SUP-35's transmissibility and aggregation is yet to be described and is of great interest. Should the mechanism be explained, it might help in understanding more aggregation mechanisms, and how to prevent and treat them.

2. The *C. elegans* protein ABU-13

ABU-13 (*Activated in Blocked Unfolded protein response*) is a protein found in Caenorhabditis *Elegans* (*C. Elegans*). It is named so after its activation in animals with blocked unfolded protein response under ER stress. It is also named Prion-like (Q/N) Domain baring protein 46 (PQN-46) due to having a Q/N rich protein region categorizing it as prion-like. It has an important role when nematodes are exposed to the pathogenic bacteria *Pseudomonas Aeruginosa*, by up-regulation of the protein (Sun et al., 2011). This up-regulation due to exposure to the pathogen indicates that the non-canonical *Unfolded Protein Response* (*UPR*) pathway may be required for *C. Elegan's* immune response against it.

A mechanism regarding ABU-13's behavior suggested it becomes a prion seed that can help the nematode survive exposure to the stress conditions described (Sun et al., 2011).

3. The *C. elegans* protein MUT-16

MUT-16 (MUTator) is a Q/N-rich protein that is essential for mutator complex formation (mutator class genes cause activation of Tc1 (DNA transposons) in the germline). Generally, mutator genes are essential factors in RNA silencing, yet their specific roles in small RNA pathways are poorly understood. Studies showed that MUT-16 has a crucial role in endogenous siRNA production, as well as exogenous RNAi (Phillips et al., 2012; Zhang et al., 2011). MUT-16 is uniquely required for the formation of mutator foci and is essential for mutator complex formation.

Combining these facts, I hypothesized that MUT-16 could be a prion protein induced by RNAi stress. I was also interested in the role MUT-16 might have in siRNA inheritance (similar to those of hrde-1).

4. Nematode conditioning to pathogen exposure

C. elegans has been shown to have abrasive behavior when exposed to certain pathogenic bacteria such as *EPEC* (*Entero-Pathogenic E. Coli*), as a means of survival. When studying the pathway causing this behavior, it was shown that the nematodes could be “taught” to survive longer on a lawn of *EPEC* with less abrasive behavior (Anyanful et al., 2009). This was accomplished by exposing the nematodes to the pathogen for a short period, before administering lethal exposure time (conditioning).

Research goal: identifying potential prion proteins in *C. elegans*

My goal was to test and characterize top-scoring prion candidates in *C. elegans.* I sought to find the environmental conditions causing the proteins to act as prions, test the transference mechanisms, and find the implications on the entire animal. I used genetic engineering tools to create knockout strains causing them to lose their prion-forming domain and hoped to see a lack of prionic behavior under the tested conditions.

By finding prion proteins and characterizing their effects I hoped to further investigate interactions with other proteins, and hopefully find underlying mechanisms for prionic behavior in multicellular organisms.

Results for prediction of prion proteins candidates

1. Prion protein prediction algorithm

Previous work (Alberti et al., 2009) Showed the use of a prediction algorithm (*PLAAC*) to identify proteins displaying sequence elements that may act as prions. By implementing the algorithm for all *C. Elegans* proteins, I gathered the repertoire of *PrLDs* scores, location, and sequence. *C. Elegans* had some genes annotated as PQN- (Prion-like (Q/N) Domain baring protein), which contain Q/N rich regions. I expected that their core score would be significantly higher than that of the entire proteome. Indeed, both for the PRD and core scores were relatively higher, and the hypothesis that the scores are sampled from a population with similar averages is rejected under a t-test with a p-value << 0.01. Moreover, while only 628 protein-coding genes (out of the 20,222 in the *C. elegans* proteome) were identified as having any PRD, all the PQN proteins were identified as having one. These comprise 10% of the group of proteins having a PRD. With that, not all PQN proteins score on the higher spectrum of core scores. As expected, both linear (Pearson's) and rank (Spearman's) correlation for PRD and core scores were positive (figure 33), suggesting that the higher the PRD score the more likely it is the protein could be prion forming.

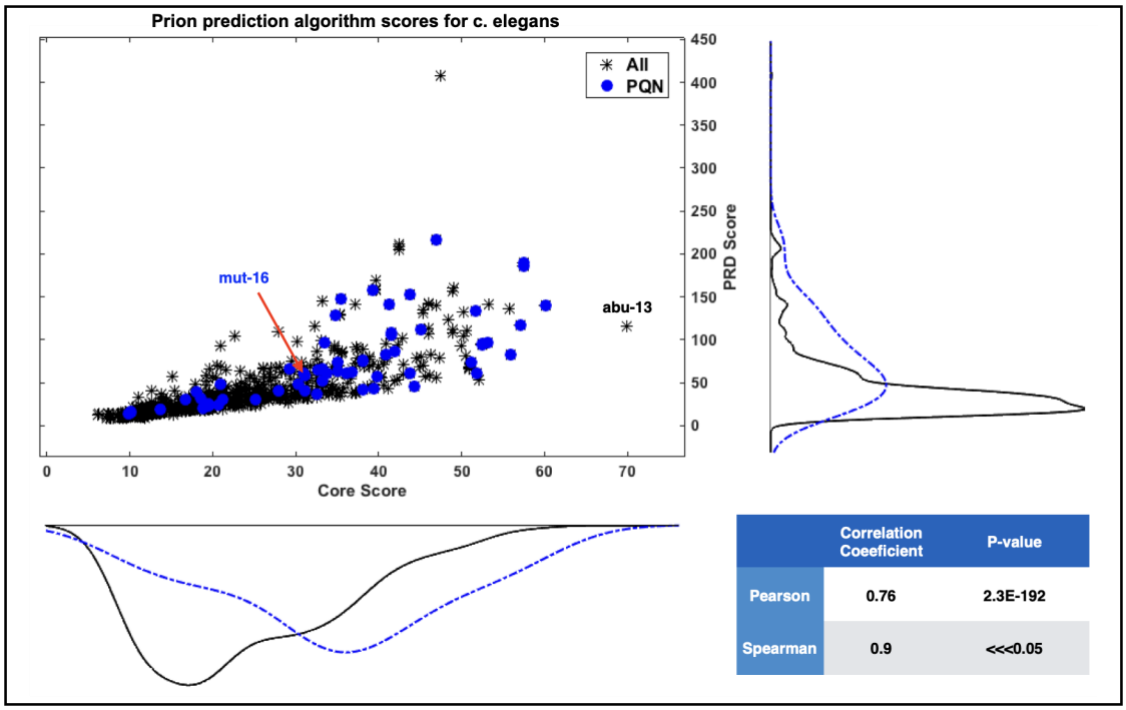


Figure 33: PLAAC algorithm prion domain scores for *C. Elegans* genes. The algorithm outputed two main scores for consideration Prion Domain (PRD) which included the entire sequence predicted to be the prion forming domain and a Core score depicting the core of pricing behavior that would elicit aggregation and more. Poly Q/N rich region genes (PQN) seemed to generally have higher Core cores. ABU-13 exhibited the highest Core score of all *C. Elegant* genes. MUT-16 (PQN-3) was located well within the peak of PQN scores which is slightly higher than that of all other genes.

To test the validity of the methods on genomes other than *S. cerevisiae*, I tested the correlation of scores between both species, for all orthologs found. Understanding the level of correlation could provide insights into the validity of these methods on sequences that are different from those they were trained for, i.e. the generalization potential for the models. As can be seen from figure 34, PLAAC scores correlation between *C. elegans* and *S. cerevisiae* were much lower than those calculated using PAPA. Still, the rank correlation showed significant positive results which suggest that these algorithms could serve as good indicators when accounting for relative scores within the species, and the highest-scoring genes can most likely act as prion proteins.

I found that ABU-13 was the top-scoring protein in having a strong prion-like core domain, suggesting prionic behavior that may imply stress response (Fig 33).

Other than prion prediction algorithms (based on specific sequence features), I examined the protein's secondary structure prediction. I searched for proteins with high *PLAAC* scores that also fail to have a well-

defined single secondary structure. I found that overall protein secondary structure predictions were ambiguous and most of the proteins do not show strong predefined structures. This led me to the conclusion that prion secondary structure predictions are not reliable enough to provide more information about the prionicity of a sequence. The proteins' potential to have prion-like behavior then lies within its *PrLD* score and some biological evidence.

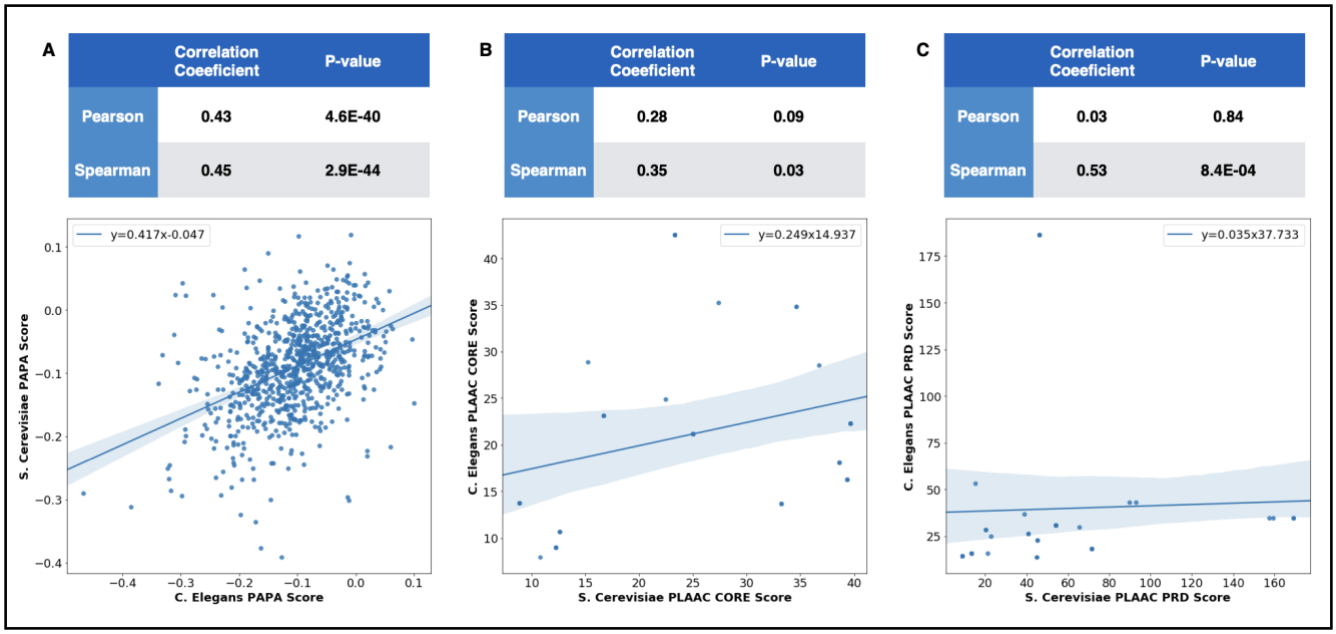


Figure 34: PAPA and PLAAC algorithms scores correlation between *C. Elegans* and *S. cerevisiae*. To test the level of validity of these algorithms on other species, the level of correlation between orthologous genes was tested. **(A)** PAPA scores seemed to have significant positive correlation. It is important to note that the correlation was calculated for all orthologous proteins, even those that did not pass the 0.05 threshold mentioned in (Toombs et al., 2012). **(B-C)** Correlation of PLAAC Core (B) and PRD (C) scores between orthologous proteins. While the correlation still maintained a positive sign, it appears that linear correlation was not significant using Pearson's correlation, but under Spearman's rank correlation with p-value of 0.05, both scores were correlative between the two species.

2. ABU-13 conditioning essay using *P. Aeruginosa* and cross-reactivity conditioning

ABU-13, also known as PQN-46 (prion-like Q/N rich protein) is believed to have a role in pathogen exposure resistance (Sun et al., 2011). I hypothesized this resistance is due to a prionic process that creates ABU-13 prion seeds as aggregates, which will somehow allow the nematode to resist a pathogenic process better. I wished to prove this using pathogen exposure conditioning. In this case, short exposure to a pathogen (exposure that is not lethal) should induce the prionic state of ABU-13. The conditioning period should be followed by a resting period where the

nematodes will not be exposed to a pathogen. After a short rest, the lethal exposure should begin. I expected the nematodes to be more resistant following conditioning.

ABU genes (Activated in Blocked Unfolded protein response) were shown to over-express when nematodes are exposed to ER stress conditions, such as *Tunicamycin* treatment (Urano et al., 2002). Perhaps ER stress exposure also activates the prionic ABU-13 behavior, which can help the nematode react better to Tunicamycin exposure, and even to pathogen exposure.

Once I have established that *pseudomonas* and *Tunicamycin* treatments show a difference in their effectiveness on nematode lifespan shortening under different genetic backgrounds, I wished to test the hypothesis of resistance due to prion priming. I tried to cause ABU-13 priming in two different ways:

1) Short exposure to the propagating stress followed by a resting period before the phenotypic causing exposure
2) For *Pseudomonas* stress, growing the nematodes on heat-killed bacteria from hatching.

I used nematodes with different genetic backgrounds that cause hypersensitivity and hyper resistance to test the changes in durability to the stress at hand. I also used a strain knocked out for ABU-13 to prove that the pathway originates from ABU-13 propagation for resistance.

The next stage would have been to connect the pathways. I wished to cause conditioning using one stress and test for a phenotype in the other stress. This would have helped me in finding an underlying pathway for ABU-13 prion protection against stresses activating similar defense mechanisms.

I first started by conducting control experiments to find lethal exposure periods on the pathogen. I found that when following documented protocols for pathogen killing as described in (Sun et al., 2011), I was not able to recreate fatality rates. I have repeated the experiment numerous times and was not able to see high mortality rates even after 72 hours. Since *P. Aeruginosa* I should be highly pathogenic, experiments were conducted off-site, contributing to many temperatures and other technical factors affecting

the experiment. Due to this, I was unable to generate a proper setting for testing and the project was terminated.

3. MUT-16 as a prion and hereditary component of RNAi

MUT-16 was shown to have a role in RNAi silencing (Zhang et al., 2011). It also scores very high as a prion candidate using *PLAAC* and *PAPA* (Alberti et al., 2009) (Fig 29). A recent study has shown that MUT-16 tends to be expressed as foci in the germ-line (Phillips et al., 2012). Using immunostaining I wanted to show that MUT-16 creates these foci in the soma as well, after exposure to RNAi stress, which I hypothesize is the propagating stress. Unfortunately, I could not obtain consistent immunostaining of adult nematode, obtained only a weak signal, and terminated these analyses.

Methods for prediction of prion protein candidates

1. Prion protein prediction algorithm

Previous work (Alberti et al., 2009) showed the use of a prediction algorithm (*PLAAC*) to identify proteins displaying sequence elements that may act as prions. The algorithm is based on a *Hidden Markov Model (HMM)* with a training set of previously established yeast prion proteins. The algorithm was based on identifying sequence properties that imply a high probability to form a prion domain. The training set included sequences of proteins from the yeast *S. cerevisiae* and only those sequences with high Q/N content, therefore this approach was biased toward a particular class of prions.

The second algorithm mentioned and used was the PAPA algorithm. This algorithm could give the prion propensity of each amino acid. For each amino acid, the prion propensity is defined as the log-odds ratio of the frequency of occurrence of the amino acid among the prion-forming clones relative to the starting library of SUP-35 scrambled sequences. PAPA employed a sliding window approach to calculate the average log-odds ratio of each AA in the window to give its final score. The final score would be a measure of the maximal calculated average odd-logs ratio of the windows along the protein.

Both algorithms could be implemented on any organism with simple alterations to increase precision. I applied both methods to the model organism *C. elegans* proteome. (Toombs et al., 2012) reported that using a cut-off of ~0.05 was efficient in differentiating proteins with clear prion activity from those proteins lacking prion activity, as reported by (Alberti et al., 2009). Thus, I combined the results from the *PAPA* algorithm passing the 0.05 threshold, with the *PrLD* scores calculated by the *PLAAC* algorithm to suggest possible prion-forming proteins in C. elegans. By slight modification to the algorithm, it was possible to also find proteins that have a domain that is very non-prion-like.

Other than prion prediction algorithms (based on specific sequence features, the nematode's protein secondary structure was examined (using predictions). The RaptorX (Källberg et al., 2012) tool was used to generate

the secondary structure of the proteome. Its' output described the probability of each amino acid (AA) residue, forming one of three types of secondary structures: alpha-helix, beta-sheet of coiled-coil. Additionally, it could calculate intrinsically disordered domains based on the aa sequence.

2. ABU-13 CRISPR

Plasmids for CRISPR/Cas9 genomic editing of *sec-5* were constructed as described previously (Dickinson et al., 2013). The homologous guide RNA sequence (5′-GATATCAGTCTGTTTC-GTAA-3′) from plasmid pDD122 was replaced with the sequence (5′-TTGTGGAACAAATGCGTAT-3′), which was designed to direct cleavage near the ABU-13 N-terminus. The plasmid was ligated after PCR amplification and injected into young-adult N2 animals. Since the ABU-13 knockout phenotype is unknown, micro-injection was performed using a co-marker CRISPR plasmid to create a strong ROL-6 phenotype (Arribere et al., 2014). Animals presenting ROL-6 phenotype were isolated, as they are positive for CRISPR reaction, and sent to DNA sequencing after laying eggs. The mutation that arose from sequencing was a 17-bp insertion near the cleavage site of ABU-13, close to the translation start site, which caused an early stop codon and thus, loss of function (figure 35A).

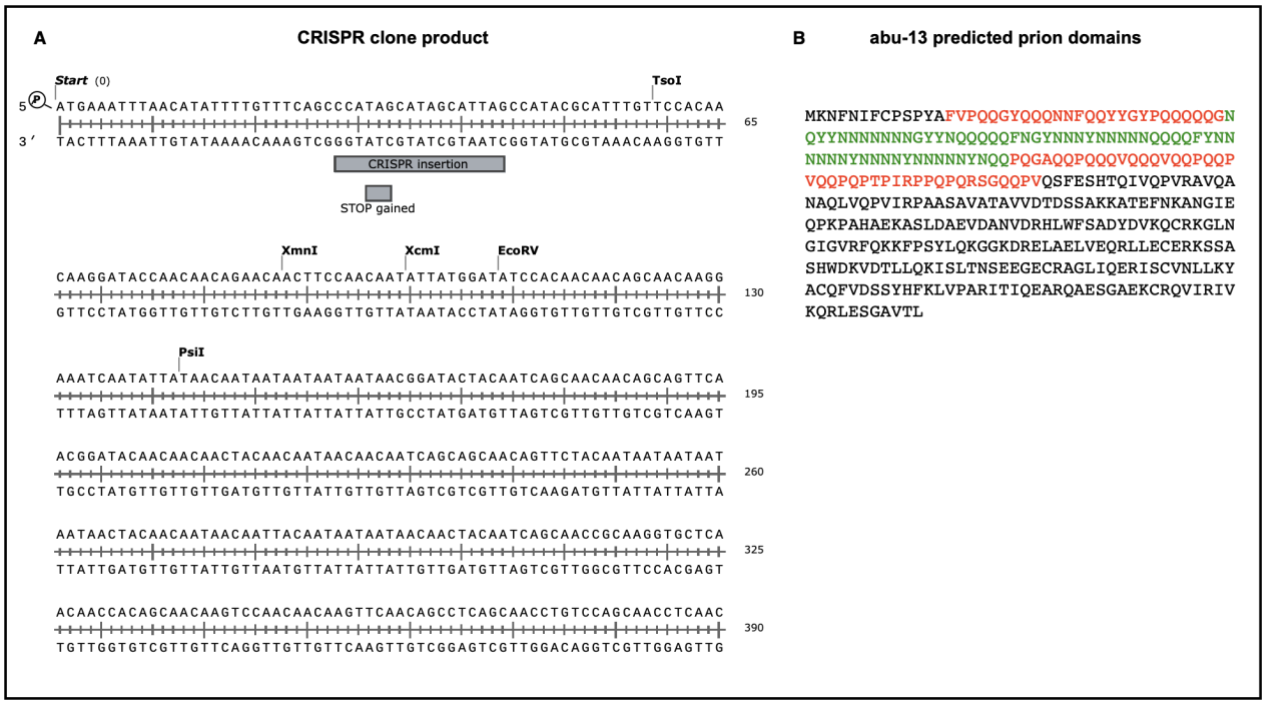


Figure 35: ABU-13 knock out strain using CRISPR-Cas9. (A) sg-RNA was targeted to be close to the starting site of translation in order to eliminate translation. Since the prion core is located early in the gene, I aimed to generate a version without the prion domain if a translation product still exists. CRISPR process resulted in a 17 bp insertion at the cleavage site, which generated an early STOP codon, resulting in a knockout version. **(B)** mapping of the prion domains as predicted by PLAAC. Red letters mark the prion forming domain (PRD). Green letters mark the prion core.

3. *Pseudomonas Aeruginosa* conditioning essay

Conditioning using live bacteria was performed by exposing nematodes for a short period (1-8 hours) on plates seeded with *P. Aeruginosa* to allow prion priming. The nematodes were then placed on NGM plates seeded with normal OP50 for different time courses ranging from 5 to 24 hours. The nematodes were then transferred to NGM plated seeded with *P. Aeruginosa* for a period that was found to be lethal (in control experiments).

Plates were seeded with overnight grown culture at 37°C of *P. Aeruginosa*, incubated at 37°C for 12~16 hours, and then cooled down at room temperature for at least 1 hour before placing synchronized young adults on to them. Nematodes were then kept at 25°C and transferred daily onto freshly seeded plates.

Conditioning using heat-killed bacteria was performed by seeding nematodes as un-hatched eggs on plated seeded with heat-killed *P. Aeruginosa*. When they reached young adult age, they were transferred to

plates seeded with live *P. Aeruginosa* (as explained before). Nematodes were then kept at 25°C and transferred daily onto freshly seeded plates.

Summary of Chapter 3:

Prion proteins are usually associated with having negative effects. Whether it is PRnP in mammals, causing neurodegenerative disease, or SUP35 in yeast creating massive STOP codon readthrough.

Using computational tools that can identify sequence properties of discovered prion proteins (in specific species), we are now able to predict more and more proteins that may hold prionic behaviors.

In this chapter not only did I predict a handful of proteins that may have prion-like behavior in *C. elegans*, but I suggested that at least two of these, may have positive effects such as pathogen immunity. By examining their sequence, and some documented evidence, I was quite convinced that both ABU-13 and MUT-16 could result in proteins that may have roles in stress survival by prion seeding. I searched for evidence showing that minor exposure to the activating stress related to these proteins, would activate a prion response generating immunity later in the nematode's life.

While I was unable to produce the experimental evidence myself, I have gathered enough convincing evidence for future work and research in hopes of elucidating the roles of the proteins discussed in this chapter, and hopefully much more.

Appendix 3.A: Prion prediction results for C. elegans

SEQid	COREscore	COREstart	COREend	CORElen	PRDscore	PRDstart	PRDend	PRDlen
F57B9.9	69.857	38	97	60	116.107	14	142	129
F29C12.1a	60.215	333	392	60	139.35	192	432	241
F29C12.1b	60.215	331	390	60	139.35	190	430	241
F21C10.8a	57.517	278	337	60	189.204	16	381	366
F21C10.8b	57.517	264	323	60	185.393	16	367	352
F40F4.8	57.11	196	255	60	117.256	165	428	264
T23F1.6	55.875	155	214	60	82.313	127	324	198
Y75B8A.3	55.857	536	595	60	136.085	433	640	208
T10A3.1a	53.281	209	268	60	141.219	170	415	246
T10A3.1b	53.281	209	268	60	141.219	170	415	246
Y71G12B.21	53.13	107	166	60	96.422	39	188	150
E01G4.4a	52.539	35	94	60	95.473	0	187	188
Y41C4A.5	52.477	223	282	60	93.738	194	338	145
Y39E4B.3a.1	52.06	334	393	60	53.042	328	394	67
Y39E4B.3a.2	52.06	334	393	60	53.042	328	394	67
F53G2.4a	51.892	197	256	60	60.457	185	268	84
F53G2.4b	51.892	197	256	60	60.457	185	268	84
C05B5.3	51.745	220	279	60	133.771	69	350	282
C24A8.3	51.131	943	1002	60	73.562	903	1012	110
C07A9.3a	50.858	162	221	60	68.305	76	226	151
C07A9.3b	50.858	162	221	60	68.305	76	226	151
C07A9.3c	50.858	95	154	60	65.641	0	159	160
C07A9.3d	50.858	81	140	60	67.401	0	145	146
C07G1.5.1	50.6	598	657	60	110.016	529	718	190
C07G1.5.2	50.6	598	657	60	110.016	529	718	190
F13H8.5a	50.48	44	103	60	81.215	25	150	126
F13H8.5b	50.48	44	103	60	81.215	25	150	126
M88.5a	50.258	136	195	60	132.765	0	299	300
M88.5b	50.258	12	71	60	104.852	8	175	168
M88.5c	50.258	136	195	60	132.765	0	299	300
M88.5d	50.258	12	71	60	104.852	8	175	168
Y79H2A.3a	50.04	1815	1874	60	93.086	1807	1950	144
Y79H2A.3d	50.04	1808	1867	60	93.086	1800	1943	144
Y79H2A.3e	50.04	1803	1862	60	93.086	1795	1938	144
Y79H2A.3g	50.04	1173	1232	60	93.086	1165	1308	144
ZC21.3b	49.002	102	161	60	160.921	26	423	398
ZC21.3a	48.87	198	257	60	157.011	0	320	321
ZC21.3d	48.87	148	207	60	107.629	26	270	245
ZC21.3e	48.87	229	288	60	111.4	26	351	326
H20J18.1a.1	48.374	152	211	60	123.483	0	270	271

H20J18.1a.2	48.374	152	211	60	123.483	0	270	271
H20J18.1b.1	48.374	152	211	60	123.483	0	270	271
H20J18.1b.2	48.374	152	211	60	123.483	0	270	271
F47A4.2	47.463	2746	2805	60	408.153	2721	3497	777
F52E4.6	47.404	297	356	60	78.365	237	377	141
F39D8.1a	46.999	229	288	60	216.664	73	512	440
F39D8.1b	46.999	206	265	60	216.664	50	489	440
F39D8.1c	46.999	211	270	60	216.664	55	494	440
T04C10.1	46.992	737	796	60	55.956	737	820	84
Y75B8A.8	46.947	176	235	60	140.116	0	334	335
ZC116.1a	46.348	44	103	60	82.401	14	177	164
ZC116.1b	46.348	99	158	60	82.401	69	232	164
F55A12.6	46.177	131	190	60	58.471	113	194	82
C18E9.3a	46.043	486	545	60	142.087	270	622	353
C18E9.3b	46.043	483	542	60	142.087	267	619	353
C18E9.3c.1	46.043	423	482	60	142.087	207	559	353
C18E9.3c.2	46.043	423	482	60	142.087	207	559	353
C18E9.3d	46.043	426	485	60	142.087	210	562	353
C18E9.3e.1	46.043	320	379	60	113.386	206	456	251
C18E9.3e.2	46.043	320	379	60	113.386	206	456	251
C18E9.3f.1	46.043	421	480	60	141.676	218	557	340
C18E9.3f.2	46.043	421	480	60	141.676	218	557	340
C18E9.3g	46.043	323	382	60	113.386	209	459	251
F52D1.3	45.983	465	524	60	106.92	410	602	193
C24H11.7a	45.577	1831	1890	60	57.651	1803	1911	109
C24H11.7b	45.577	341	400	60	57.651	313	421	109
Y67H2A.10a	45.426	418	477	60	131.161	317	607	291
Y67H2A.10b	45.426	415	474	60	135.119	317	604	288
T22H6.7	45.423	102	161	60	70.086	79	189	111
W01C9.3a	45.161	493	552	60	111.955	426	736	311
W01C9.3b	45.161	526	585	60	111.955	459	769	311
Y43H11AL.3	45.113	74	133	60	79.779	0	173	174
T04D1.4	44.981	484	543	60	67.957	480	620	141
T21B6.3	44.564	291	350	60	79.364	204	351	148
K07D4.8	44.288	22	81	60	45.169	18	105	88
Y73B6BR.1a	43.778	20	79	60	60.456	20	141	122
Y73B6BR.1b	43.778	20	79	60	60.456	20	141	122
DY3.5	43.72	299	358	60	153.121	31	446	416
R10E11.1a	43.039	1946	2005	60	71.463	1884	2005	122
R10E11.1b	43.039	1957	2016	60	71.463	1895	2016	122
R10E11.1c	43.039	1917	1976	60	59.603	1881	1976	96
T13H2.5a	42.858	2194	2253	60	102.034	2066	2285	220
T13H2.5b	42.858	810	869	60	102.034	682	901	220

T05A10.1a	42.401	195	254	60	205.278	0	420	421
T05A10.1b	42.401	218	277	60	208.506	0	434	435
T05A10.1d	42.401	195	254	60	205.039	0	411	412
T05A10.1e	42.401	195	254	60	205.039	0	411	412
T05A10.1f	42.401	195	254	60	205.039	0	411	412
T05A10.1g.1	42.401	195	254	60	205.039	0	411	412
T05A10.1g.2	42.401	195	254	60	205.039	0	411	412
T05A10.1i	42.401	195	254	60	205.039	0	411	412
T05A10.1j	42.401	195	254	60	205.039	0	411	412
T05A10.1k	42.401	218	277	60	208.506	0	434	435
T05A10.1l	42.401	234	293	60	211.017	0	450	451
T05A10.1m	42.401	218	277	60	208.506	0	434	435
Y40B1A.4	42.326	69	128	60	94.648	0	216	217
K10G6.3a	42.17	1049	1108	60	53.107	1009	1114	106
K10G6.3b	42.17	904	963	60	53.107	864	969	106
K10G6.3c	42.17	1049	1108	60	53.107	1009	1114	106
K10G6.3d	42.17	807	866	60	53.107	767	872	106
K10G6.3e	42.17	278	337	60	53.107	238	343	106
C37A2.2	41.962	676	735	60	86.035	676	887	212
F13E9.15	41.946	47	106	60	52.472	17	109	93
W03D2.1a	41.601	169	228	60	107.96	23	340	318
W03D2.1b	41.601	163	222	60	107.96	17	334	318
W03D2.1c	41.601	184	243	60	105.273	40	355	316
F10F2.9	41.276	438	497	60	140.884	160	504	345
C01G10.15	41.1	27	86	60	43.12	17	99	83
K08E3.8	41.015	2	61	60	78.183	0	280	281
ZK1236.6	40.934	146	205	60	82.751	0	205	206
K08F8.6	40.839	1382	1441	60	95.315	1215	1456	242
C32A3.1a	40.774	389	448	60	78.554	319	489	171
C32A3.1b	40.774	367	426	60	78.554	297	467	171
F39H11.2a	40.672	132	191	60	71.068	0	213	214
F39H11.2b	40.672	135	194	60	71.518	0	216	217
Y46G5A.38	40.606	23	82	60	46.431	21	96	76
M110.4a	40.348	50	109	60	55.821	0	132	133
M110.4b	40.348	50	109	60	55.821	0	132	133
R10E4.2q	40.29	43	102	60	56.459	0	137	138
R10E4.2r	40.29	43	102	60	56.459	0	137	138
C26C6.1a	40.002	1702	1761	60	68.126	1632	1764	133
C26C6.1b.1	40.002	332	391	60	68.126	262	394	133
C26C6.1b.2	40.002	332	391	60	68.126	262	394	133
E01G4.4b	39.781	0	59	60	57.151	0	124	125
D2045.1a	39.639	664	723	60	169.215	514	958	445
D2045.1b	39.639	339	398	60	169.215	189	633	445

D2045.1c	39.639	664	723	60	157.443	514	925	412
D2045.1d	39.639	751	810	60	159.489	601	1025	425
F20D1.3	39.494	379	438	60	47.832	371	453	83
ZC15.8	39.455	16	75	60	43.451	16	94	79
F13B9.1a	39.338	47	106	60	71.375	0	181	182
F13B9.1b	39.338	47	106	60	71.375	0	181	182
F13B9.1c	39.338	47	106	60	71.375	0	181	182
T06E4.11	39.254	259	318	60	157.513	88	385	298
B0041.2a.1	38.836	152	211	60	77.138	0	245	246
B0041.2a.2	38.836	152	211	60	77.138	0	245	246
B0041.2b	38.836	70	129	60	68.927	0	163	164
B0041.2c.1	38.836	152	211	60	77.138	0	245	246
B0041.2c.2	38.836	152	211	60	77.138	0	245	246
B0041.2d	38.836	194	253	60	80.669	0	287	288
F57A8.2a	38.626	33	92	60	45.173	14	92	79
F57A8.2b	38.626	33	92	60	45.173	14	92	79
ZC518.2	38.237	187	246	60	69.094	0	248	249
D1007.14	38.114	111	170	60	42.156	95	173	79
Y57A10A.18a	38.062	1305	1364	60	75.55	1275	1455	181
Y57A10A.18b	38.062	1305	1364	60	76.288	1275	1458	184
Y57A10A.18c	38.062	1334	1393	60	76.125	1301	1484	184
Y57A10A.18d	38.062	1328	1387	60	75.55	1298	1478	181
Y57A10A.18e	38.062	1332	1391	60	75.55	1302	1482	181
F35B12.3	38.059	121	180	60	88.834	17	206	190
ZC308.1a	37.766	965	1024	60	41.332	963	1047	85
ZC308.1b	37.766	723	782	60	41.332	721	805	85
ZC308.1c	37.766	888	947	60	41.332	886	970	85
ZC308.1d	37.766	659	718	60	41.332	657	741	85
C54G7.3a	37.705	180	239	60	72.712	145	280	136
C54G7.3b	37.705	180	239	60	72.712	145	280	136
T07C4.9a	37.445	116	175	60	89.44	22	176	155
T07C4.9b.1	37.445	74	133	60	78.524	0	134	135
T07C4.9b.2	37.445	74	133	60	78.524	0	134	135
C49H3.5a	37.316	616	675	60	91.789	530	749	220
C49H3.5b	37.316	435	494	60	91.789	349	568	220
Y46G5A.36	37.313	23	82	60	37.736	23	83	61
R10E4.2a.1	37.213	1	60	60	38.378	0	70	71
R10E4.2a.2	37.213	1	60	60	38.378	0	70	71
R10E4.2a.3	37.213	1	60	60	38.378	0	70	71
R10E4.2a.4	37.213	1	60	60	38.378	0	70	71
R10E4.2b.1	37.213	1	60	60	38.378	0	70	71
R10E4.2b.2	37.213	1	60	60	38.378	0	70	71
R10E4.2f.1	37.213	1	60	60	38.378	0	70	71

R10E4.2f.2	37.213	1	60	60	38.378	0	70	71
R10E4.2f.3	37.213	1	60	60	38.378	0	70	71
R10E4.2f.4	37.213	1	60	60	38.378	0	70	71
R10E4.2g.1	37.213	1	60	60	38.378	0	70	71
R10E4.2g.2	37.213	1	60	60	38.378	0	70	71
R10E4.2i.1	37.213	1	60	60	38.378	0	70	71
R10E4.2i.2	37.213	1	60	60	38.378	0	70	71
R10E4.2l.1	37.213	1	60	60	38.378	0	70	71
R10E4.2l.2	37.213	1	60	60	38.378	0	70	71
R10E4.2m.1	37.213	1	60	60	38.378	0	70	71
R10E4.2m.2	37.213	1	60	60	38.378	0	70	71
R10E4.2n.1	37.213	1	60	60	38.378	0	70	71
R10E4.2n.2	37.213	1	60	60	38.378	0	70	71
R10E4.2n.3	37.213	1	60	60	38.378	0	70	71
R10E4.2n.4	37.213	1	60	60	38.378	0	70	71
R10E4.2p.1	37.213	1	60	60	38.378	0	70	71
R10E4.2p.2	37.213	1	60	60	38.378	0	70	71
R10E4.2c	37.174	6	65	60	71.821	0	149	150
F13E9.4	37.132	62	121	60	140.896	21	387	367
C01G10.6	36.826	22	81	60	38.465	15	88	74
W06B11.2	36.744	236	295	60	65.571	122	323	202
T13H2.4a	36.727	1224	1283	60	62.499	1165	1298	134
Y41G9A.10	36.627	67	126	60	54.326	27	145	119
R11A8.7a	36.271	2339	2398	60	49.051	2314	2418	105
R11A8.7b	36.271	2317	2376	60	49.051	2292	2396	105
R11A8.7c	36.271	1124	1183	60	49.051	1099	1203	105
R11A8.7d	36.271	1146	1205	60	49.051	1121	1225	105
R11A8.7e	36.271	1202	1261	60	49.051	1177	1281	105
R11A8.7f	36.271	2341	2400	60	49.051	2316	2420	105
R11A8.7g	36.271	2319	2378	60	49.051	2294	2398	105
R11A8.7h	36.271	1204	1263	60	49.051	1179	1283	105
R119.4	36.111	479	538	60	61.471	438	711	274
C10C5.6a	35.756	634	693	60	43.834	587	708	122
C10C5.6b	35.756	634	693	60	43.834	587	708	122
C10C5.6c	35.756	658	717	60	43.834	611	732	122
C10C5.6d	35.756	637	696	60	43.834	590	711	122
C10C5.6e	35.756	637	696	60	43.834	590	711	122
F10E7.2	35.666	61	120	60	39.054	53	124	72
Y113G7B.23a	35.606	641	700	60	86.016	591	788	198
Y113G7B.23b.1	35.606	537	596	60	86.016	487	684	198
Y113G7B.23b.2	35.606	537	596	60	86.016	487	684	198
Y113G7B.23c	35.606	641	700	60	84.628	591	791	201
F26G5.9	35.481	526	585	60	83.452	369	616	248

C34E7.1a.1	35.43	161	220	60	147.005	0	429	430
C34E7.1a.2	35.43	161	220	60	147.005	0	429	430
Y67D8C.7	35.397	98	157	60	37.976	82	157	76
F15E6.3	35.355	35	94	60	34.49	35	117	83
T13F2.3a	35.279	171	230	60	129.179	28	511	484
T13F2.3b	35.279	173	232	60	128.105	28	513	486
F35B3.5a	35.257	540	599	60	69.594	493	693	201
F35B3.5c	35.257	542	601	60	69.594	495	695	201
C34E7.1b	35.049	172	231	60	73.344	0	248	249
C34E7.1c	35.049	143	202	60	63.051	0	219	220
R10E12.1a	35.014	755	814	60	59.93	755	860	106
R10E12.1b	35.014	776	835	60	59.93	776	881	106
R10E12.1d	35.014	392	451	60	59.93	392	497	106
F32D8.14a	34.917	11	70	60	35.488	0	79	80
Y73B6BL.6a.1	34.777	244	303	60	41.697	192	304	113
Y73B6BL.6a.2	34.777	244	303	60	41.697	192	304	113
Y73B6BL.6a.3	34.777	244	303	60	41.697	192	304	113
Y73B6BL.6a.4	34.777	244	303	60	41.697	192	304	113
Y73B6BL.6a.5	34.777	244	303	60	41.697	192	304	113
Y73B6BL.6a.6	34.777	244	303	60	41.697	192	304	113
Y73B6BL.6a.7	34.777	244	303	60	41.697	192	304	113
F53A3.4a	34.745	1854	1913	60	128.205	1615	2013	399
F53A3.4b	34.745	1889	1948	60	128.205	1650	2048	399
F53A3.4d	34.745	267	326	60	128.205	28	426	399
F39H11.3	34.689	383	442	60	88.001	373	587	215
Y56A3A.4a	34.645	73	132	60	92.842	18	217	200
Y56A3A.4b.1	34.645	40	99	60	89.944	0	184	185
Y56A3A.4b.2	34.645	40	99	60	89.944	0	184	185
Y39E4B.3b.1	34.545	58	117	60	43.005	24	131	108
Y39E4B.3b.2	34.545	58	117	60	43.005	24	131	108
Y39E4B.3c.1	34.545	58	117	60	43.005	24	131	108
Y39E4B.3c.2	34.545	58	117	60	43.005	24	131	108
Y54E10A.9a.1	34.477	0	59	60	39.929	0	76	77
Y54E10A.9a.2	34.477	0	59	60	39.929	0	76	77
Y54E10A.9b.1	34.477	0	59	60	37.333	0	69	70
Y54E10A.9b.2	34.477	0	59	60	37.333	0	69	70
C18A3.5a	34.421	348	407	60	48.188	318	407	90
C18A3.5b	34.421	316	375	60	48.188	286	375	90
C18A3.5e	34.421	245	304	60	48.188	215	304	90
C18A3.5f	34.421	235	294	60	48.188	205	294	90
F52G3.1	34.355	1006	1065	60	68.173	843	1171	329
Y73B6BL.6b.1	34.32	241	300	60	42.22	192	307	116
Y73B6BL.6b.2	34.32	241	300	60	42.22	192	307	116

Y73B6BL.6b.3	34.32	241	300	60	42.22	192	307	116
Y73B6BL.6b.4	34.32	241	300	60	42.22	192	307	116
Y73B6BL.6b.5	34.32	241	300	60	42.22	192	307	116
Y73B6BL.6b.6	34.32	241	300	60	42.22	192	307	116
Y73B6BL.6b.7	34.32	241	300	60	42.22	192	307	116
C27B7.4	34.012	655	714	60	35.777	652	716	65
C09E7.2	33.763	144	203	60	61.132	27	211	185
F39B2.4a	33.683	1524	1583	60	60.888	1419	1586	168
F39B2.4b	33.683	1526	1585	60	60.888	1421	1588	168
Y46G5A.13	33.663	344	403	60	39.498	336	433	98
W07B3.2a.1	33.646	472	531	60	46.648	439	544	106
W07B3.2a.2	33.646	472	531	60	46.648	439	544	106
W07B3.2b.1	33.646	495	554	60	46.648	462	567	106
W07B3.2b.2	33.646	495	554	60	46.648	462	567	106
W07B3.2d.1	33.646	471	530	60	46.648	438	543	106
W07B3.2d.2	33.646	471	530	60	46.648	438	543	106
W07B3.2f.1	33.646	491	550	60	46.648	458	563	106
W07B3.2f.2	33.646	491	550	60	46.648	458	563	106
Y111B2A.14a.1	33.527	1345	1404	60	96.207	1123	1415	293
Y111B2A.14a.2	33.527	1345	1404	60	96.207	1123	1415	293
Y111B2A.14b.1	33.527	1375	1434	60	96.207	1153	1445	293
Y111B2A.14b.2	33.527	1375	1434	60	96.207	1153	1445	293
ZK596.1	33.329	169	228	60	38.144	159	240	82
F56D1.7	33.211	274	333	60	67.57	268	417	150
C34G6.7a	33.199	384	443	60	44.933	373	456	84
C46G7.4a	33.193	544	603	60	52.492	541	728	188
C46G7.4c	33.193	413	472	60	66.068	410	654	245
C17G1.4a	33.188	114	173	60	144.972	0	458	459
C17G1.4b	33.188	114	173	60	144.972	0	458	459
Y53C12B.3a	33.068	403	462	60	99.002	360	592	233
B0379.3a	32.829	788	847	60	71.453	709	866	158
B0379.3b	32.829	784	843	60	71.453	705	862	158
T16G1.1	32.76	527	586	60	65.005	483	627	145
R10E12.1c	32.665	776	835	60	37.044	776	845	70
Y116A8C.32	32.651	574	633	60	36.317	574	683	110
Y53C12B.3b	32.625	252	311	60	86.726	90	314	225
F58A3.1a.1	32.567	513	572	60	76.992	404	591	188
F58A3.1a.2	32.567	513	572	60	76.992	404	591	188
F58A3.1b	32.567	499	558	60	76.992	390	577	188
F58A3.1c	32.567	550	609	60	76.992	441	628	188
F44A6.1a	32.559	373	432	60	32.559	373	432	60
F44A6.1b	32.559	387	446	60	32.559	387	446	60
W07B3.2c.1	32.543	480	539	60	46.165	433	541	109

W07B3.2c.2	32.543	480	539	60	46.165	433	541	109
C34E7.1d	32.53	13	72	60	36.806	0	72	73
T28C6.1a	32.5	184	243	60	69.977	43	256	214
T28C6.1b	32.5	184	243	60	69.977	43	256	214
T28C6.1c	32.5	184	243	60	67.634	43	261	219
T28C6.1d	32.5	196	255	60	67.634	55	273	219
T28C6.1e	32.5	112	171	60	58.44	17	189	173
T28C6.1f	32.5	196	255	60	69.977	55	268	214
R08B4.1a.1	32.391	788	847	60	52.231	779	920	142
R08B4.1a.2	32.391	788	847	60	52.231	779	920	142
R08B4.1b.1	32.391	811	870	60	62.312	779	943	165
R08B4.1b.2	32.391	811	870	60	62.312	779	943	165
K01A6.4	32.284	29	88	60	115.275	23	283	261
F14B8.5b	32.093	18	77	60	39.519	15	96	82
T04F8.8a	32.088	18	77	60	39.077	18	105	88
C06G1.4	31.778	282	341	60	69.282	151	343	193
F42A6.7b.1	31.526	206	265	60	58.019	175	308	134
F42A6.7b.2	31.526	206	265	60	58.019	175	308	134
F42A6.7b.3	31.526	206	265	60	58.019	175	308	134
F42A6.7b.4	31.526	206	265	60	58.019	175	308	134
F42A6.7d.1	31.526	244	303	60	58.019	213	346	134
F42A6.7d.2	31.526	244	303	60	58.019	213	346	134
Y75B8A.6	31.504	221	280	60	31.705	220	280	61
R07E3.2	31.5	99	158	60	55.657	20	164	145
B0302.1a.1	31.382	1034	1093	60	31.583	1033	1093	61
B0302.1a.2	31.382	1034	1093	60	31.583	1033	1093	61
B0302.1b.1	31.382	927	986	60	31.583	926	986	61
B0302.1b.2	31.382	927	986	60	31.583	926	986	61
W10D5.3a.1	31.375	616	675	60	37.37	608	678	71
W10D5.3a.2	31.375	616	675	60	37.37	608	678	71
W10D5.3a.3	31.375	616	675	60	37.37	608	678	71
W10D5.3c.1	31.375	717	776	60	46.545	682	779	98
W10D5.3c.2	31.375	717	776	60	46.545	682	779	98
W10D5.3c.3	31.375	717	776	60	46.545	682	779	98
W10D5.3d	31.375	614	673	60	37.37	606	676	71
W10D5.3e.1	31.375	600	659	60	37.37	592	662	71
W10D5.3e.2	31.375	600	659	60	37.37	592	662	71
W10D5.3e.3	31.375	600	659	60	37.37	592	662	71
K01A6.8	31.329	23	82	60	49.716	23	111	89
T19B10.4a	31.059	21	80	60	58.024	0	141	142
T19B10.4b	31.059	21	80	60	40.448	0	105	106
K04G2.8a	30.984	524	583	60	34.565	490	583	94
F42A6.7a.1	30.934	274	333	60	54.884	213	345	133

F42A6.7a.2	30.934	274	333	60	54.884	213	345	133
F42A6.7c.1	30.934	236	295	60	54.884	175	307	133
F42A6.7c.2	30.934	236	295	60	54.884	175	307	133
F42A6.7c.3	30.934	236	295	60	54.884	175	307	133
F42A6.7c.4	30.934	236	295	60	54.884	175	307	133
R09B5.5	30.927	196	255	60	61.085	31	382	352
T12F5.5a	30.919	478	537	60	67.84	360	561	202
T12F5.5b	30.919	380	439	60	67.84	262	463	202
Y41G9A.5a	30.868	25	84	60	32.478	21	104	84
Y41G9A.5b	30.868	34	93	60	32.478	30	113	84
Y106G6H.2a.1	30.864	479	538	60	63.144	403	559	157
Y106G6H.2a.2	30.864	479	538	60	63.144	403	559	157
Y106G6H.2a.3	30.864	479	538	60	63.144	403	559	157
Y106G6H.2a.4	30.864	479	538	60	63.144	403	559	157
Y106G6H.2a.5	30.864	479	538	60	63.144	403	559	157
Y106G6H.2b.1	30.864	416	475	60	63.144	340	496	157
Y106G6H.2b.2	30.864	416	475	60	63.144	340	496	157
Y106G6H.2c.1	30.864	419	478	60	63.144	343	499	157
Y106G6H.2c.2	30.864	419	478	60	63.144	343	499	157
Y106G6H.2c.3	30.864	419	478	60	63.144	343	499	157
Y106G6H.2c.4	30.864	419	478	60	63.144	343	499	157
Y106G6H.2c.5	30.864	419	478	60	63.144	343	499	157
C03A7.14	30.699	238	297	60	74.116	31	439	409
R10E4.2e	30.643	41	100	60	39.242	0	100	101
K04G2.8b	30.553	525	584	60	35.356	490	585	96
C35A5.10	30.525	117	176	60	59.998	28	185	158
F58A4.11	30.516	316	375	60	36.52	288	388	101
ZC101.1	30.368	636	695	60	34.827	636	704	69
C46G7.4b	30.363	1	60	60	47.624	0	178	179
C45H4.13	30.272	263	322	60	32.15	235	348	114
H14N18.1a	30.263	11	70	60	46.391	11	174	164
H14N18.1c	30.263	10	69	60	42.827	0	173	174
F12F6.6	30.252	1	60	60	97.865	0	350	351
R10E4.2d	30.15	29	88	60	37.939	0	97	98
R10E4.2j	30.15	247	306	60	39.411	212	315	104
Y77E11A.11a.1	30.012	24	83	60	31.822	24	96	73
Y77E11A.11a.2	30.012	24	83	60	31.822	24	96	73
F52C9.8a	30.005	489	548	60	41.396	453	564	112
F52C9.8b	30.005	489	548	60	41.396	453	564	112
Y47D7A.13	29.781	170	229	60	62.172	41	259	219
R12B2.5a.1	29.757	193	252	60	62.534	173	364	192
R12B2.5a.2	29.757	193	252	60	62.534	173	364	192
R12B2.5a.3	29.757	193	252	60	62.534	173	364	192

R12B2.5b.1	29.757	190	249	60	62.534	170	361	192
R12B2.5b.2	29.757	190	249	60	62.534	170	361	192
R12B2.5b.3	29.757	190	249	60	62.534	170	361	192
Y111B2A.22a	29.57	2081	2140	60	74.827	1981	2166	186
Y111B2A.22c	29.57	560	619	60	74.827	460	645	186
Y111B2A.22d	29.57	1935	1994	60	74.827	1835	2020	186
W10D5.3f.1	29.554	646	705	60	38.466	627	714	88
W10D5.3f.2	29.554	646	705	60	38.466	627	714	88
W10D5.3f.3	29.554	646	705	60	38.466	627	714	88
W10D5.3g.1	29.554	578	637	60	38.466	559	646	88
W10D5.3g.2	29.554	578	637	60	38.466	559	646	88
C14B9.6a	29.534	783	842	60	33.443	783	846	64
C14B9.6c	29.534	783	842	60	33.443	783	846	64
F54E2.3a	29.463	159	218	60	46.799	110	288	179
F54E2.3c	29.463	159	218	60	46.799	110	288	179
F54E2.3d	29.463	159	218	60	46.799	110	288	179
ZC416.2	29.434	1	60	60	29.997	0	61	62
Y20F4.2	29.389	211	270	60	55.099	117	361	245
F48F7.4	29.21	438	497	60	66.167	367	540	174
T23D8.9a	29.116	46	105	60	26.479	0	105	106
T23D8.9b	29.116	46	105	60	26.479	0	105	106
Y75B8A.27	28.826	15	74	60	34.312	8	78	71
ZK418.9a	28.815	420	479	60	34.216	413	556	144
ZK418.9b.1	28.815	373	432	60	34.216	366	509	144
ZK418.9b.2	28.815	373	432	60	34.216	366	509	144
F40F9.1a.1	28.781	2	61	60	30.088	0	65	66
F40F9.1a.2	28.781	2	61	60	30.088	0	65	66
F40F9.1a.3	28.781	2	61	60	30.088	0	65	66
F40F9.1b.1	28.781	2	61	60	30.088	0	65	66
F40F9.1b.2	28.781	2	61	60	30.088	0	65	66
F40F9.1b.3	28.781	2	61	60	30.088	0	65	66
F53A9.9	28.712	14	73	60	41.936	0	106	107
T01D1.6	28.693	210	269	60	63.807	47	387	341
D1046.1c.1	28.547	268	327	60	40.689	202	369	168
D1046.1c.2	28.547	268	327	60	40.689	202	369	168
D1046.1c.3	28.547	268	327	60	40.689	202	369	168
D1046.1c.4	28.547	268	327	60	40.689	202	369	168
F42A10.2a	28.529	494	553	60	31.15	463	553	91
F42A10.2b	28.529	494	553	60	31.15	463	553	91
F42A10.2c	28.529	494	553	60	31.15	463	553	91
F59B10.1	28.507	109	168	60	41.751	44	177	134
T11G6.5a	28.5	1163	1222	60	34.435	1157	1233	77
T11G6.5b.1	28.5	1106	1165	60	34.435	1100	1176	77

T11G6.5b.2	28.5	1106	1165	60	34.435	1100	1176	77
Y50D4C.3	28.358	454	513	60	62.76	364	553	190
W05F2.4a	28.263	1733	1792	60	33.925	1709	1792	84
W05F2.4b	28.263	628	687	60	33.925	604	687	84
W05F2.4d	28.263	656	715	60	33.925	632	715	84
T02E9.2a	28.187	56	115	60	33.413	44	139	96
T02E9.2b	28.187	56	115	60	33.413	44	139	96
H12D21.6	28.143	97	156	60	41.496	62	162	101
C30F2.3	28.129	268	327	60	68.296	19	330	312
R09A1.1	28.101	76	135	60	30.277	76	146	71
H15N14.1c	28.064	54	113	60	46.196	0	161	162
F35D11.2a	27.997	332	391	60	40.719	293	419	127
F35D11.2b	27.997	340	399	60	40.719	301	427	127
Y47D7A.15	27.997	296	355	60	108.858	43	355	313
T03G11.1	27.856	640	699	60	43.17	529	711	183
H15N14.1g	27.842	53	112	60	31.767	0	114	115
C05C9.3	27.737	1279	1338	60	67.255	1140	1435	296
D1046.1a.1	27.728	268	327	60	40.288	202	367	166
D1046.1a.2	27.728	268	327	60	40.288	202	367	166
D1046.1a.3	27.728	268	327	60	40.288	202	367	166
D1046.1a.4	27.728	268	327	60	40.288	202	367	166
D1046.1b.1	27.728	270	329	60	40.288	204	369	166
D1046.1b.2	27.728	270	329	60	40.288	204	369	166
D1046.1b.3	27.728	270	329	60	40.288	204	369	166
D1046.1b.4	27.728	270	329	60	40.288	204	369	166
D1046.1e.1	27.728	73	132	60	37.712	0	171	172
D1046.1e.2	27.728	73	132	60	37.712	0	171	172
D1046.1e.3	27.728	73	132	60	37.712	0	171	172
D1046.1e.4	27.728	73	132	60	37.712	0	171	172
D1046.1e.5	27.728	73	132	60	37.712	0	171	172
D1046.1e.6	27.728	73	132	60	37.712	0	171	172
D1046.1e.7	27.728	73	132	60	37.712	0	171	172
F25H8.5a	27.563	125	184	60	46.678	84	244	161
F25H8.5b	27.563	125	184	60	46.678	84	244	161
F25H8.5c	27.563	125	184	60	46.678	84	244	161
F25H8.5g	27.563	148	207	60	46.678	107	267	161
F25H8.5k	27.563	148	207	60	46.678	107	267	161
ZK973.9	27.511	154	213	60	55.599	76	232	157
F13D12.3	27.471	31	90	60	32.451	25	127	103
Y48B6A.3	27.394	900	959	60	38.818	821	974	154
F32B4.4a	27.015	1012	1071	60	52.179	848	1072	225
F32B4.4b.1	27.015	602	661	60	52.179	438	662	225
F32B4.4b.2	27.015	602	661	60	52.179	438	662	225

F32B4.4c	27.015	941	1000	60	52.179	777	1001	225
F27D4.2a.1	26.988	140	199	60	38.693	93	212	120
F27D4.2a.2	26.988	140	199	60	38.693	93	212	120
F27D4.2b.1	26.988	140	199	60	37.902	93	210	118
F27D4.2b.2	26.988	140	199	60	37.902	93	210	118
R74.5a	26.988	253	312	60	33.518	193	403	211
R74.5b.1	26.988	135	194	60	33.518	75	285	211
R74.5b.2	26.988	135	194	60	33.518	75	285	211
R74.5b.3	26.988	135	194	60	33.518	75	285	211
R74.5c	26.988	228	287	60	33.518	168	378	211
C03A7.8	26.978	231	290	60	67.473	31	432	402
C03A7.4	26.881	231	290	60	60.259	31	382	352
C03A7.7	26.83	231	290	60	59.899	31	382	352
C01G8.9a	26.801	747	806	60	52.037	668	888	221
C01G8.9c	26.801	828	887	60	52.037	749	969	221
Y59A8B.10a	26.791	244	303	60	36.254	224	313	90
Y59A8B.10b	26.791	143	202	60	36.254	123	212	90
Y71H2AM.19a.1	26.783	579	638	60	30.596	565	638	74
Y71H2AM.19a.2	26.783	579	638	60	30.596	565	638	74
Y71H2AM.19b	26.783	644	703	60	30.596	630	703	74
R06C1.6	26.731	186	245	60	28.813	164	246	83
F41F3.4	26.547	224	283	60	35.152	183	284	102
Y42H9AR.1.1	26.346	364	423	60	41.641	348	505	158
Y42H9AR.1.2	26.346	364	423	60	41.641	348	505	158
F56A8.6.1	26.321	233	292	60	33.291	206	292	87
F56A8.6.2	26.321	233	292	60	33.291	206	292	87
R06A4.9a	26.266	605	664	60	32.102	602	682	81
R06A4.9b	26.266	389	448	60	32.102	386	466	81
F38B7.3	26.171	300	359	60	56.527	138	366	229
F16B4.4	26.156	31	90	60	31.323	15	90	76
Y63D3A.5.1	26.155	282	341	60	67.344	253	485	233
Y63D3A.5.2	26.155	282	341	60	67.344	253	485	233
K02E11.10	26.108	174	233	60	65.58	77	359	283
C36E6.1b	26.107	398	457	60	48.718	397	558	162
C16B8.3	25.782	70	129	60	34.969	35	163	129
W04B5.3a	25.781	302	361	60	36.456	231	405	175
W04B5.3b	25.781	302	361	60	36.456	231	405	175
K09B3.1a	25.773	17	76	60	25.773	17	76	60
Y106G6D.7	25.742	770	829	60	40.177	715	857	143
F26F12.5b	25.74	53	112	60	26.364	53	114	62
F46F2.3	25.729	45	104	60	29.748	43	136	94
C26C6.5a	25.67	585	644	60	39.488	583	700	118
C26C6.5b	25.67	588	647	60	39.488	586	703	118

ZC116.5	25.666	49	108	60	26.238	19	110	92
W02A2.7	25.557	57	116	60	46.483	0	168	169
T04F8.8b	25.498	18	77	60	27.668	18	87	70
Y69H2.14	25.393	86	145	60	44.048	86	323	238
AH6.5	25.262	73	132	60	51.209	0	167	168
Y102A11A.1	25.212	181	240	60	26.601	181	248	68
C27H5.3.1	25.176	16	75	60	33.253	0	134	135
C27H5.3.2	25.176	16	75	60	33.253	0	134	135
K11D12.2.1	25.161	124	183	60	30.787	108	193	86
K11D12.2.2	25.161	124	183	60	30.787	108	193	86
K11D12.2.3	25.161	124	183	60	30.787	108	193	86
K11D12.2.4	25.161	124	183	60	30.787	108	193	86
Y43F8C.2	25.148	21	80	60	22.902	15	89	75
F56F3.1	25.124	628	687	60	38.798	603	702	100
F01F1.1a	25.024	85	144	60	38.684	31	154	124
F01F1.1c	25.024	80	139	60	38.684	26	149	124
F18H3.3a.1	25.013	452	511	60	54.204	440	594	155
F18H3.3a.2	25.013	452	511	60	54.204	440	594	155
F18H3.3a.3	25.013	452	511	60	54.204	440	594	155
F18H3.3a.4	25.013	452	511	60	54.204	440	594	155
F18H3.3b.1	25.013	335	394	60	54.204	323	477	155
F18H3.3b.2	25.013	335	394	60	54.204	323	477	155
T28F2.5	24.74	430	489	60	36.248	411	501	91
Y95B8A.8	24.728	4	63	60	38.298	0	127	128
C04E6.6	24.723	182	241	60	34.642	178	266	89
F33A8.10	24.609	26	85	60	25.666	26	93	68
H15N14.1d	24.603	26	85	60	32.421	0	95	96
H15N14.1e	24.603	26	85	60	32.421	0	95	96
H15N14.1f	24.603	26	85	60	31.14	0	92	93
C18B2.4	24.565	221	280	60	36.002	141	295	155
C04A2.3a	24.498	932	991	60	29.201	930	998	69
C04A2.3b	24.498	788	847	60	29.201	786	854	69
C04A2.3c.1	24.498	415	474	60	29.201	413	481	69
C04A2.3c.2	24.498	415	474	60	29.201	413	481	69
C04A2.3d	24.498	691	750	60	29.201	689	757	69
R07B5.9a.1	24.495	1142	1201	60	56.383	953	1245	293
R07B5.9a.2	24.495	1142	1201	60	56.383	953	1245	293
R07B5.9c	24.495	1433	1492	60	56.383	1244	1536	293
R07B5.9d	24.495	1149	1208	60	56.383	960	1252	293
R07B5.9f	24.495	1478	1537	60	56.383	1289	1581	293
R07B5.9g	24.495	1436	1495	60	58.29	1244	1539	296
R07B5.9h	24.495	1481	1540	60	58.29	1289	1584	296
R07B5.9i	24.495	1152	1211	60	58.29	960	1255	296

R07B5.9j.1	24.495	1145	1204	60	58.29	953	1248	296
R07B5.9j.2	24.495	1145	1204	60	58.29	953	1248	296
F13E6.4	24.465	319	378	60	42.704	248	409	162
R07H5.10a	24.457	6	65	60	23.747	0	65	66
R07H5.10b	24.457	26	85	60	24.35	0	85	86
ZK180.5a	24.412	99	158	60	41.518	29	201	173
ZK180.5b	24.412	99	158	60	41.518	29	201	173
ZK180.5c	24.412	99	158	60	41.518	29	201	173
Y37A1B.1a	24.264	41	100	60	25.915	0	100	101
Y37A1B.1b.1	24.264	41	100	60	25.915	0	100	101
Y37A1B.1b.2	24.264	41	100	60	25.915	0	100	101
Y37A1B.1c	24.264	41	100	60	25.915	0	100	101
ZK858.8	24.185	12	71	60	23.794	0	80	81
F44B9.7	23.982	0	59	60	48.246	0	259	260
F07C4.7	23.97	45	104	60	44.351	20	162	143
C26F1.1b	23.797	84	143	60	27.164	78	154	77
F25D7.3a	23.567	710	769	60	33.241	710	795	86
F25D7.3b	23.567	694	753	60	33.241	694	779	86
ZK678.5	23.501	274	333	60	24.093	274	341	68
W04B5.3c	23.496	290	349	60	26.805	231	352	122
C09G5.4	23.457	189	248	60	40.743	137	257	121
F57B9.2	23.44	779	838	60	49.482	769	1027	259
Y73F8A.21a	23.408	605	664	60	34.43	579	671	93
T21D12.11	23.401	120	179	60	42.973	12	182	171
D2005.6	23.374	73	132	60	26.688	41	133	93
Y94H6A.11a	23.363	1	60	60	32.635	0	114	115
T20B6.3	23.323	188	247	60	63.451	61	252	192
ZK328.5a	23.309	298	357	60	46.158	212	492	281
ZK328.5b	23.309	298	357	60	46.158	212	492	281
ZK328.5c	23.309	300	359	60	46.158	214	494	281
C17D12.2	23.288	292	351	60	39.093	208	409	202
Y66A7A.8	23.236	1	60	60	30.193	0	79	80
C34D4.11	23.163	72	131	60	23.593	48	135	88
Y18D10A.17	23.16	275	334	60	22.012	275	339	65
M04B2.1	23.121	565	624	60	42.868	544	685	142
C30F12.4	23.102	49	108	60	33.774	48	149	102
Y61A9LA.3a	23.059	588	647	60	26.466	585	650	66
Y47D3A.6a	23.01	451	510	60	30.362	415	510	96
Y47D3A.12	22.999	235	294	60	28.72	233	312	80
C33G3.6	22.987	383	442	60	36.49	382	529	148
R13.4a	22.825	249	308	60	47.021	145	322	178
R13.4b.1	22.825	80	139	60	39.536	0	153	154
R13.4b.2	22.825	80	139	60	39.536	0	153	154

C26F1.1a	22.688	6	65	60	24.963	6	74	69
ZK377.1	22.659	319	378	60	26.37	303	392	90
Y53C10A.10	22.657	982	1041	60	104.208	916	1434	519
C06G4.2b.1	22.618	56	115	60	36.711	47	161	115
C06G4.2b.2	22.618	56	115	60	36.711	47	161	115
Y62F5A.1a	22.585	241	300	60	31.682	210	324	115
T01D1.2a.1	22.52	351	410	60	31.06	348	418	71
T01D1.2a.2	22.52	351	410	60	31.06	348	418	71
T01D1.2a.3	22.52	351	410	60	31.06	348	418	71
T01D1.2a.4	22.52	351	410	60	31.06	348	418	71
T01D1.2a.5	22.52	351	410	60	31.06	348	418	71
T01D1.2b.1	22.52	119	178	60	31.06	116	186	71
T01D1.2b.2	22.52	119	178	60	31.06	116	186	71
T01D1.2b.3	22.52	119	178	60	31.06	116	186	71
T01D1.2b.4	22.52	119	178	60	31.06	116	186	71
T01D1.2b.5	22.52	119	178	60	31.06	116	186	71
F29D10.4	22.513	954	1013	60	27.934	954	1030	77
W07B3.2e.1	22.479	0	59	60	24.459	0	64	65
W07B3.2e.2	22.479	0	59	60	24.459	0	64	65
F32A5.6	22.447	18	77	60	22.857	0	77	78
F36A2.1a	22.405	526	585	60	26.79	419	657	239
F36A2.1b	22.405	524	583	60	26.79	417	655	239
F36A2.1c	22.405	530	589	60	26.79	423	661	239
F36A2.1d	22.405	532	591	60	26.79	425	663	239
F13H6.1a	22.382	351	410	60	25.48	322	417	96
F13H6.1b.1	22.382	355	414	60	25.48	326	421	96
F13H6.1b.2	22.382	355	414	60	25.48	326	421	96
Y54G2A.26a	22.342	74	133	60	37.677	0	173	174
Y54G2A.26b	22.342	12	71	60	31.594	0	111	112
K09F5.6	22.166	550	609	60	22.59	550	610	61
B0513.1a	22.124	10	69	60	22.782	0	80	81
M03E7.2	22.072	42	101	60	27.006	39	113	75
F41E7.9	21.889	173	232	60	23.951	143	232	90
M02G9.1b	21.859	707	766	60	32.31	694	834	141
Y62F5A.1b	21.604	228	287	60	23.537	210	296	87
Y39A3CL.2	21.601	1005	1064	60	31.679	1004	1103	100
ZK1321.4a	21.601	123	182	60	45.517	117	313	197
ZK1321.4b	21.601	129	188	60	45.517	123	319	197
C55B7.1	21.575	32	91	60	64.268	17	269	253
T17H7.1.1	21.485	187	246	60	42.593	187	681	495
T17H7.1.2	21.485	187	246	60	42.593	187	681	495
T07D4.3	21.484	1193	1252	60	28.553	1191	1300	110
K12H6.1	21.466	632	691	60	26.339	618	707	90

C09G5.5	21.422	138	197	60	39.392	134	256	123
Y41C4A.16	21.384	220	279	60	39.742	158	285	128
M18.1	21.279	210	269	60	38.991	143	269	127
C38C10.5a	21.254	1282	1341	60	30.719	1275	1381	107
C38C10.5b	21.254	1288	1347	60	30.719	1281	1387	107
C38C10.5c	21.254	1416	1475	60	30.719	1409	1515	107
W05G11.3.1	21.239	168	227	60	34.687	113	248	136
W05G11.3.2	21.239	168	227	60	34.687	113	248	136
T20F7.5	21.211	21	80	60	35.392	19	177	159
R09F10.7	21.186	197	256	60	29.729	186	372	187
R09F10.2	21.186	197	256	60	29.729	186	372	187
F16B4.7	21.159	20	79	60	19.123	19	84	66
C53B7.3a	21.137	63	122	60	38.735	44	212	169
C53B7.3b	21.137	63	122	60	26.757	44	125	82
C53B7.3c	21.137	63	122	60	28.162	44	176	133
C53B7.3d.1	21.137	14	73	60	37.08	0	163	164
C53B7.3d.2	21.137	14	73	60	37.08	0	163	164
C32E8.10a	21.053	499	558	60	33.036	404	585	182
C32E8.10b	21.053	459	518	60	26.075	404	545	142
C32E8.10c	21.053	570	629	60	24.516	552	656	105
C32E8.10d	21.053	447	506	60	24.516	429	533	105
C32E8.10f	21.053	111	170	60	32.584	0	197	198
C32E8.10h	21.053	461	520	60	26.013	404	547	144
Y39A3CR.7	21.007	26	85	60	48.104	0	204	205
Y49E10.29	20.987	23	82	60	93.056	15	459	445
Y53H1C.2a	20.978	945	1004	60	68.032	830	1255	426
Y53H1C.2c	20.978	945	1004	60	68.032	830	1255	426
K02B9.1	20.936	294	353	60	39.044	180	353	174
F52G2.2a	20.92	1169	1228	60	33.676	1142	1261	120
F52G2.2b	20.92	979	1038	60	33.676	952	1071	120
F52G2.2c	20.92	267	326	60	33.676	240	359	120
F52G2.2d	20.92	721	780	60	33.676	694	813	120
R10E4.2k	20.914	228	287	60	28.264	212	294	83
R10E4.2o	20.914	228	287	60	28.264	212	294	83
F58H7.1	20.905	176	235	60	32.447	171	288	118
C06G4.2d	20.862	64	123	60	40.584	47	186	140
F58E10.3a.1	20.842	11	70	60	22.313	0	74	75
F58E10.3a.2	20.842	11	70	60	22.313	0	74	75
F58E10.3a.3	20.842	11	70	60	22.313	0	74	75
F58E10.3a.4	20.842	11	70	60	22.313	0	74	75
F58E10.3a.5	20.842	11	70	60	22.313	0	74	75
T11F9.9	20.833	148	207	60	37.979	148	273	126
R11G11.7	20.732	26	85	60	23.47	25	124	100

B0344.2	20.694	384	443	60	26.888	367	552	186
Y40C5A.3a	20.587	2042	2101	60	29.557	2042	2188	147
Y40C5A.3b	20.587	2089	2148	60	29.557	2089	2235	147
C02B10.5.1	20.516	581	640	60	69.635	377	697	321
C02B10.5.2	20.516	581	640	60	69.635	377	697	321
Y79H2A.1a	20.447	282	341	60	33.308	250	342	93
Y79H2A.1b	20.447	196	255	60	33.308	164	256	93
Y79H2A.1c	20.447	245	304	60	33.308	213	305	93
F58H1.2	20.447	49	108	60	22.346	48	116	69
R11E3.6a	20.444	841	900	60	33.704	722	908	187
R11E3.6b	20.444	223	282	60	33.704	104	290	187
F19B2.6	20.412	425	484	60	55.983	35	507	473
C24G6.7	20.404	18	77	60	23.175	13	79	67
C05C8.4	20.144	926	985	60	25.026	907	985	79
F38A6.1b	20.077	30	89	60	19.095	0	89	90
F55C10.2	20.054	214	273	60	36.705	148	273	126
F55C10.3	20.054	214	273	60	36.705	148	273	126
F52C9.8c	20.046	2	61	60	20.297	0	64	65
F52C9.8g	20.046	2	61	60	20.297	0	64	65
C36B1.8a	20.004	988	1047	60	26.587	947	1048	102
C36B1.8b	20.004	991	1050	60	26.587	950	1051	102
C36B1.8c	20.004	947	1006	60	26.587	906	1007	102
Y32G9A.5	19.987	64	123	60	55.636	35	247	213
Y79H2A.1e	19.962	264	323	60	24.876	250	324	75
C24H11.3	19.953	234	293	60	22.365	230	293	64
F52B5.3	19.928	1310	1369	60	27.509	1306	1406	101
C34F6.10	19.805	235	294	60	25.627	206	321	116
F58E10.5	19.766	41	100	60	21.736	0	100	101
Y43F8C.20	19.753	104	163	60	33.939	40	179	140
F53H2.3b	19.736	176	235	60	20.34	175	239	65
F53H2.3c	19.736	86	145	60	20.34	85	149	65
F08F8.9a.1	19.718	360	419	60	21.309	336	419	84
F08F8.9a.2	19.718	360	419	60	21.309	336	419	84
F08F8.9b.1	19.718	360	419	60	21.309	336	419	84
F08F8.9b.2	19.718	360	419	60	21.309	336	419	84
F08F8.9c.1	19.718	360	419	60	21.309	336	419	84
F08F8.9c.2	19.718	360	419	60	21.309	336	419	84
ZK643.8a	19.687	235	294	60	65.443	83	451	369
T06E4.6	19.676	190	249	60	30.065	133	253	121
F14B8.5a.1	19.646	16	75	60	38.67	15	132	118
F14B8.5a.2	19.646	16	75	60	38.67	15	132	118
F41B4.1	19.644	197	256	60	30.657	153	262	110
Y56A3A.6.1	19.594	251	310	60	19.594	251	310	60

Y56A3A.6.2	19.594	251	310	60	19.594	251	310	60
C15A11.6	19.593	136	195	60	38.214	136	258	123
F12A10.9	19.55	46	105	60	18.976	46	106	61
M04F3.5	19.527	345	404	60	21.701	324	404	81
ZK488.7	19.503	40	99	60	22.784	40	158	119
Y41C4A.19	19.458	199	258	60	33.914	137	263	127
C15A11.5	19.453	199	258	60	37.257	136	258	123
ZK488.10	19.44	37	96	60	25.493	37	135	99
F58D5.1a.1	19.398	469	528	60	29.635	462	575	114
F58D5.1a.2	19.398	469	528	60	29.635	462	575	114
F52B11.4	19.354	218	277	60	31.438	153	279	127
F12A10.7	19.354	53	112	60	19.354	53	112	60
C12D8.1a	19.315	409	468	60	26.079	402	482	81
C12D8.1b	19.315	431	490	60	26.079	424	504	81
C12D8.1c.1	19.315	368	427	60	26.079	361	441	81
C12D8.1c.2	19.315	368	427	60	26.079	361	441	81
C12D8.1c.3	19.315	368	427	60	26.079	361	441	81
F58G11.2a	19.299	102	161	60	47.403	95	313	219
F58G11.2b	19.299	96	155	60	51.724	41	307	267
F58G11.2c	19.299	83	142	60	47.403	76	294	219
C09H6.2a	19.297	326	385	60	23.53	326	409	84
C09H6.2b	19.297	298	357	60	23.53	298	381	84
C09H6.2c	19.297	275	334	60	23.53	275	358	84
T28F2.8	19.266	232	291	60	58.303	83	421	339
F53H10.2a	19.262	138	197	60	19.751	134	197	64
VK10D6R.1	19.144	16	75	60	17.217	16	88	73
F53F4.2	19.13	16	75	60	22.987	15	88	74
T23C6.4	19.087	264	323	60	33.69	161	323	163
T12D8.1	18.991	1194	1253	60	22.371	1194	1260	67
F57B7.3	18.988	183	242	60	30.965	137	247	111
Y59A8B.1a	18.914	112	171	60	27.511	90	176	87
M110.5a.1	18.867	252	311	60	21.194	247	343	97
M110.5a.2	18.867	252	311	60	21.194	247	343	97
M110.5b.1	18.867	252	311	60	21.194	247	343	97
M110.5b.2	18.867	252	311	60	21.194	247	343	97
M110.5c	18.867	236	295	60	21.194	231	327	97
M110.5d.1	18.867	254	313	60	21.194	249	345	97
M110.5d.2	18.867	254	313	60	21.194	249	345	97
M01E11.4c	18.812	300	359	60	19.924	297	359	63
B0285.3	18.805	32	91	60	29.231	0	110	111
C08B11.3	18.738	484	543	60	23.087	481	572	92
B0222.6	18.671	167	226	60	30.254	158	275	118
K02D7.3a	18.671	253	312	60	32.578	138	312	175

K02D7.3b	18.671	240	299	60	32.578	125	299	175
B0222.8	18.671	167	226	60	30.254	158	275	118
R07B7.3a	18.557	250	309	60	31.337	205	311	107
F02D8.2	18.544	42	101	60	21.512	28	102	75
H12D21.3	18.499	20	79	60	18.716	18	82	65
ZC412.8	18.499	20	79	60	18.716	18	82	65
Y73B6BL.34	18.486	212	271	60	28.795	146	271	126
M01E10.2a	18.463	83	142	60	26.268	55	158	104
F15B9.10a	18.442	301	360	60	37.183	249	360	112
T04C10.4.1	18.428	6	65	60	18.123	0	69	70
T04C10.4.2	18.428	6	65	60	18.123	0	69	70
T15B7.5	18.423	210	269	60	30.042	143	269	127
F25B5.7a	18.291	438	497	60	34.451	378	499	122
F25B5.7c	18.291	158	217	60	34.451	98	219	122
F11A1.3a	18.286	211	270	60	20.096	211	280	70
F11A1.3b	18.286	152	211	60	20.096	152	221	70
F11A1.3d	18.286	177	236	60	20.096	177	246	70
T18H9.1	18.275	41	100	60	25.092	18	100	83
T06E4.4	18.254	184	243	60	31.81	133	259	127
Y54E10A.9c.1	18.23	3	62	60	24.275	0	95	96
Y54E10A.9c.2	18.23	3	62	60	24.275	0	95	96
D1007.7	18.219	369	428	60	21.3	369	463	95
F49E11.1b	18.217	190	249	60	21.678	166	257	92
F49E11.1c	18.217	190	249	60	21.678	166	257	92
F49E11.1g	18.217	173	232	60	32.152	57	240	184
W05B2.6	18.211	218	277	60	33.139	153	279	127
W05B2.5	18.211	218	277	60	33.139	153	279	127
W05B2.1	18.211	218	277	60	33.139	153	279	127
F41E7.5	18.202	31	90	60	19.505	31	99	69
F31D5.3a	18.18	263	322	60	19.377	246	322	77
F31D5.3b	18.18	263	322	60	19.377	246	322	77
F31D5.3c	18.18	263	322	60	19.377	246	322	77
F31D5.3d	18.18	263	322	60	19.377	246	322	77
D1044.3	18.107	424	483	60	38.648	424	611	188
F22A3.1a	18.095	176	235	60	20.696	166	263	98
F22A3.1b	18.095	237	296	60	20.696	227	324	98
Y56A3A.30	18.081	693	752	60	20.546	693	766	74
T21C9.9	18.032	32	91	60	22.471	31	123	93
C44C10.1	18.018	159	218	60	32.922	158	276	119
F56B3.1	17.967	248	307	60	47.772	69	307	239
C06G4.2a	17.827	117	176	60	28.978	117	207	91
H14N18.1b.1	17.815	56	115	60	22.407	33	115	83
H14N18.1b.2	17.815	56	115	60	22.407	33	115	83

C16A3.7	17.771	118	177	60	22.899	117	195	79
F15B9.10b	17.765	294	353	60	36.337	249	358	110
F53F8.1	17.747	10	69	60	21.887	0	102	103
Y38E10A.17	17.718	440	499	60	20.155	422	499	78
T21B4.2	17.645	169	228	60	30.752	169	298	130
ZK1236.3a	17.622	638	697	60	22.782	637	705	69
ZK1236.3b	17.622	448	507	60	22.782	447	515	69
K08E5.3a	17.574	3610	3669	60	18.498	3608	3671	64
K08E5.3b	17.574	3026	3085	60	18.498	3024	3087	64
C05D11.4	17.556	320	379	60	27.278	320	424	105
C14C11.8b	17.529	249	308	60	22.401	204	308	105
T11B7.4d	17.498	370	429	60	18.243	365	429	65
T15B7.3	17.481	188	247	60	31.546	136	262	127
Y43C5A.7	17.433	18	77	60	17.856	18	78	61
H05C05.1a	17.386	210	269	60	22.79	161	271	111
H05C05.1c	17.386	279	338	60	22.79	230	340	111
T01D1.2d.1	17.345	228	287	60	28.109	226	331	106
T01D1.2d.2	17.345	228	287	60	28.109	226	331	106
T01D1.2d.3	17.345	228	287	60	28.109	226	331	106
T01D1.2d.4	17.345	228	287	60	28.109	226	331	106
T01D1.2d.5	17.345	228	287	60	28.109	226	331	106
T01D1.2e.1	17.345	279	338	60	29.933	250	382	133
T01D1.2e.2	17.345	279	338	60	29.933	250	382	133
T01D1.2e.3	17.345	279	338	60	29.933	250	382	133
T01D1.2e.4	17.345	279	338	60	29.933	250	382	133
T01D1.2e.5	17.345	279	338	60	29.933	250	382	133
T01D1.2f.1	17.345	279	338	60	29.933	250	382	133
T01D1.2f.2	17.345	279	338	60	29.933	250	382	133
T01D1.2f.3	17.345	279	338	60	29.933	250	382	133
T01D1.2f.4	17.345	279	338	60	29.933	250	382	133
T01D1.2f.5	17.345	279	338	60	29.933	250	382	133
T01D1.2g.1	17.345	279	338	60	27.487	250	347	98
T01D1.2g.2	17.345	279	338	60	27.487	250	347	98
T01D1.2g.3	17.345	279	338	60	27.487	250	347	98
T01D1.2g.4	17.345	279	338	60	27.487	250	347	98
T01D1.2g.5	17.345	279	338	60	27.487	250	347	98
Y39G10AR.17	17.333	327	386	60	19.542	323	390	68
F57B1.3	17.328	211	270	60	29.163	172	290	119
T15B7.4	17.299	188	247	60	31.997	136	262	127
ZK1127.9a	17.294	89	148	60	21.638	33	150	118
ZK1127.9b	17.294	80	139	60	19.413	40	141	102
ZK1127.9d	17.294	80	139	60	19.413	40	141	102
F46C3.3a	17.288	1216	1275	60	17.705	1203	1275	73

F46C3.3d	17.288	10	69	60	17.467	0	69	70
F46C3.3f	17.288	1216	1275	60	17.705	1203	1275	73
F46C3.3g	17.288	10	69	60	17.467	0	69	70
Y48G8AL.6	17.265	991	1050	60	17.466	990	1050	61
T21G5.3	17.242	18	77	60	25.111	18	166	149
E02D9.1b	17.231	2	61	60	17.978	0	64	65
E02D9.1c	17.231	2	61	60	17.978	0	64	65
C36C9.1	17.174	119	178	60	26.936	55	214	160
T10E10.1	17.136	160	219	60	30.523	159	277	119
T10E10.2.1	17.136	160	219	60	30.523	159	277	119
T10E10.2.2	17.136	160	219	60	30.523	159	277	119
T10E10.5	17.136	159	218	60	30.523	158	276	119
T07H6.3a	17.136	222	281	60	30.523	221	339	119
T07H6.3b.1	17.136	160	219	60	30.523	159	277	119
T07H6.3b.2	17.136	160	219	60	30.523	159	277	119
T10E10.6	17.136	159	218	60	30.523	158	276	119
Y47D7A.5	17.069	42	101	60	17.069	42	101	60
W08E3.2	17.067	435	494	60	28.252	414	544	131
F25E2.5a	17.06	432	491	60	34.2	386	542	157
F25E2.5b	17.06	404	463	60	34.2	358	514	157
F25E2.5c	17.06	336	395	60	34.2	290	446	157
C30F12.1	17.06	478	537	60	17.66	476	539	64
Y73F8A.16	16.805	320	379	60	18.562	317	395	79
T28C6.6	16.787	150	209	60	29.197	150	275	126
T28C6.4	16.787	150	209	60	29.197	150	275	126
F57B1.4.1	16.783	215	274	60	30.164	176	294	119
F57B1.4.2	16.783	215	274	60	30.164	176	294	119
F57B1.4.3	16.783	215	274	60	30.164	176	294	119
F34D10.4	16.732	203	262	60	16.732	203	262	60
C29F4.1	16.694	150	209	60	29.967	150	275	126
F02E9.4a	16.667	64	123	60	20.14	64	136	73
F02E9.4b	16.667	64	123	60	20.14	64	136	73
AC3.4	16.665	213	272	60	29.935	209	322	114
AC3.3	16.665	213	272	60	29.935	209	322	114
F11G11.12	16.57	142	201	60	24.764	141	258	118
F40F11.2	16.536	141	200	60	22.962	78	201	124
F15H10.1.1	16.503	214	273	60	28.094	175	293	119
F15H10.1.2	16.503	214	273	60	28.094	175	293	119
F15H10.2	16.503	214	273	60	28.094	175	293	119
Y40B1A.3a	16.463	64	123	60	25.215	59	235	177
Y40B1A.3b	16.463	64	123	60	25.215	59	235	177
Y40B1A.3c	16.463	33	92	60	25.215	28	204	177
F46F3.4a	16.335	74	133	60	16.335	74	133	60

F46F3.4b	16.335	10	69	60	16.335	10	69	60
T05E7.2	16.308	118	177	60	21.45	96	177	82
F57A8.8	16.288	16	75	60	16.226	16	77	62
R119.6	16.268	190	249	60	27.195	190	297	108
K02B12.7	16.206	140	199	60	18.194	124	201	78
F31A3.1	16.192	66	125	60	29.8	23	236	214
F09G8.6	16.072	190	249	60	21.325	139	250	112
F54B11.1	16.003	185	244	60	27.996	141	310	170
E02H4.2	16	117	176	60	22.146	98	176	79
Y105E8A.26a	15.976	901	960	60	16.957	879	960	82
Y105E8A.26b	15.976	862	921	60	17.02	838	921	84
Y105E8A.26c	15.976	860	919	60	16.957	838	919	82
Y105E8A.26d	15.976	768	827	60	16.957	746	827	82
Y105E8A.26e	15.976	892	951	60	16.957	870	951	82
F21A10.2a.1	15.923	95	154	60	21.767	11	154	144
F21A10.2a.2	15.923	95	154	60	21.767	11	154	144
F21A10.2a.3	15.923	95	154	60	21.767	11	154	144
F21A10.2b	15.923	144	203	60	21.767	60	203	144
F21A10.2c	15.923	146	205	60	21.767	62	205	144
F21A10.2d	15.923	141	200	60	21.767	57	200	144
F21A10.2e	15.923	206	265	60	21.767	122	265	144
H17B01.2	15.854	101	160	60	16.704	92	160	69
F46E10.2	15.836	60	119	60	17.406	28	138	111
F28C6.1	15.817	71	130	60	20.716	53	135	83
K01D12.5	15.813	68	127	60	21.025	46	127	82
Y65B4BR.6a	15.804	89	148	60	34.741	42	211	170
Y65B4BR.6b	15.804	89	148	60	34.741	42	211	170
R119.7	15.709	106	165	60	20.746	101	191	91
Y49E10.14a.1	15.705	240	299	60	18.743	210	304	95
Y49E10.14a.2	15.705	240	299	60	18.743	210	304	95
Y49E10.14b	15.705	29	88	60	17.399	0	93	94
F53C11.7.1	15.59	3	62	60	16.805	0	97	98
F53C11.7.2	15.59	3	62	60	16.805	0	97	98
F53F4.1	15.502	20	79	60	16.296	18	85	68
F08G5.4	15.439	193	252	60	19.224	150	252	103
F13A7.1	15.425	238	297	60	40.883	119	339	221
T21G5.5b	15.379	290	349	60	24.355	290	378	89
T21G5.5c	15.379	332	391	60	24.355	332	420	89
T21G5.5d	15.379	362	421	60	24.789	362	471	110
C45E5.6a.1	15.353	11	70	60	18.054	11	92	82
C45E5.6a.2	15.353	11	70	60	18.054	11	92	82
C15C8.2a	15.308	18	77	60	12.775	0	77	78
C15C8.2b	15.308	18	77	60	12.775	0	77	78

R11A8.1	15.261	497	556	60	15.893	497	582	86
F01G4.1	15.223	1324	1383	60	15.223	1324	1383	60
C50F7.2	15.203	365	424	60	56.898	140	495	356
B0024.1	15.167	179	238	60	20.498	172	278	107
F19G12.7	15.163	191	250	60	17.546	182	250	69
T08H4.3	15.147	42	101	60	20.759	42	139	98
C03D6.4	15.122	1212	1271	60	23.007	1197	1380	184
C17H12.12	15.068	531	590	60	16.309	531	600	70
ZK1010.7.1	15.038	214	273	60	23.219	151	274	124
ZK1010.7.2	15.038	214	273	60	23.219	151	274	124
BE10.4	14.975	34	93	60	20.465	14	95	82
T07D4.4a	14.902	490	549	60	19.027	488	555	68
T07D4.4b	14.902	106	165	60	19.027	104	171	68
T07D4.4c	14.902	81	140	60	19.027	79	146	68
T07D4.4d	14.902	75	134	60	19.027	73	140	68
C09G12.9	14.888	174	233	60	20.893	147	239	93
ZK1151.1i	14.811	196	255	60	17.199	192	259	68
R04E5.8a	14.77	914	973	60	17.037	908	973	66
R04E5.8b	14.77	70	129	60	17.037	64	129	66
F48C1.11	14.728	28	87	60	18.255	23	122	100
C02B8.6	14.727	94	153	60	24.569	93	201	109
Y48C3A.8a.1	14.544	371	430	60	16.405	362	430	69
Y48C3A.8a.2	14.544	371	430	60	16.405	362	430	69
Y48C3A.8b	14.544	167	226	60	16.405	158	226	69
F59E12.9	14.53	1507	1566	60	29.712	1422	1613	192
T16G1.2a	14.449	268	327	60	26.766	190	327	138
F26B1.2a	14.344	252	311	60	27.193	209	314	106
F26B1.2c.1	14.344	235	294	60	27.193	192	297	106
F26B1.2c.2	14.344	235	294	60	27.193	192	297	106
Y51H4A.9	14.195	210	269	60	24.164	210	324	115
F54D1.2	14.195	216	275	60	28.351	149	275	127
F54D1.3	14.195	216	275	60	28.351	149	275	127
Y48G8AR.1	14.163	67	126	60	20.147	44	126	83
B0222.7	14.156	167	226	60	26.61	147	275	129
T20B12.2.1	14.151	35	94	60	14.859	0	120	121
T20B12.2.2	14.151	35	94	60	14.859	0	120	121
F46H5.6	14.025	461	520	60	32.838	361	522	162
F56C9.8	14.001	58	117	60	21.859	53	173	121
C32F10.6	13.96	8	67	60	15.684	0	114	115
F53C11.5a	13.935	51	110	60	18.474	42	114	73
F53C11.5b	13.935	87	146	60	17.581	0	150	151
F53C11.5c	13.935	26	85	60	18.474	17	89	73
K02B9.2	13.866	522	581	60	36.64	356	639	284

C50B6.4	13.841	153	212	60	21.283	144	261	118
C55B7.12a	13.826	4	63	60	15.656	0	69	70
C55B7.12b	13.826	4	63	60	15.548	0	72	73
Y119D3B.17	13.726	329	388	60	31.51	253	390	138
T28H11.1	13.717	304	363	60	27.379	0	372	373
H06I04.6b	13.695	58	117	60	18.426	24	120	97
F58D5.1b.1	13.678	13	72	60	14.545	0	72	73
F58D5.1b.2	13.678	13	72	60	14.545	0	72	73
C37A2.5a	13.662	575	634	60	19.391	575	658	84
C37A2.5b	13.662	593	652	60	19.391	593	676	84
C37A2.5c	13.662	436	495	60	19.391	436	519	84
C37A2.5d	13.662	287	346	60	19.391	287	370	84
C09G5.6	13.649	168	227	60	20.489	153	313	161
C28A5.4	13.629	13	72	60	13.988	13	74	62
R10E9.1	13.555	254	313	60	13.973	249	319	71
F54D8.6	13.455	702	761	60	17.069	677	764	88
R07B7.2a	13.445	366	425	60	25.003	280	425	146
R07B7.2b	13.445	366	425	60	25.003	280	425	146
Y73F8A.17	13.334	311	370	60	13.882	308	379	72
W09D10.1.1	13.305	342	401	60	20.766	323	466	144
W09D10.1.2	13.305	342	401	60	20.766	323	466	144
Y116A8C.35	13.291	212	271	60	15.596	205	272	68
C34E11.1	13.283	389	448	60	15.365	389	482	94
C26G2.1	13.272	1193	1252	60	14.536	1193	1263	71
Y39B6A.18	13.266	732	791	60	14.048	732	794	63
M195.1	13.24	162	221	60	17.054	162	262	101
F02D10.1	13.185	182	241	60	24.278	181	289	109
Y41G9A.1	13.169	26	85	60	13.169	26	85	60
T20B12.6a	13.15	327	386	60	17.406	327	471	145
T20B12.6b	13.15	327	386	60	17.406	327	471	145
Y119C1A.1	13.097	50	109	60	13.455	48	109	62
B0001.8a	13.09	191	250	60	21.866	142	255	114
B0001.8b	13.09	191	250	60	21.866	142	255	114
Y49E10.17a	13.011	102	161	60	17.914	92	191	100
Y49E10.17b	13.011	39	98	60	17.914	29	128	100
F41E6.11	12.952	224	283	60	13.109	224	284	61
H39E23.1a	12.94	870	929	60	17.471	851	935	85
H39E23.1b	12.94	774	833	60	17.471	755	839	85
H39E23.1c	12.94	740	799	60	17.471	721	805	85
H39E23.1d	12.94	806	865	60	17.471	787	871	85
H39E23.1f	12.94	781	840	60	17.471	762	846	85
H39E23.1g	12.94	719	778	60	17.471	700	784	85
H39E23.1h	12.94	383	442	60	17.471	364	448	85

H39E23.1j	12.94	878	937	60	17.471	859	943	85
H39E23.1k	12.94	272	331	60	17.471	253	337	85
H39E23.1l	12.94	784	843	60	17.471	765	849	85
T12F5.4	12.933	22	81	60	14.146	0	81	82
Y47G6A.15	12.909	210	269	60	14.566	196	277	82
C02F12.8	12.863	125	184	60	23.565	82	242	161
Y80D3A.8	12.79	762	821	60	14.122	758	826	69
B0207.1	12.784	4	63	60	14.37	0	68	69
C46A5.3a	12.735	272	331	60	13.042	267	331	65
C46A5.3b	12.735	253	312	60	13.042	248	312	65
F45E4.3a.1	12.65	667	726	60	28.178	584	798	215
F45E4.3a.2	12.65	667	726	60	28.178	584	798	215
F45E4.3b.1	12.65	667	726	60	28.178	584	798	215
F45E4.3b.2	12.65	667	726	60	28.178	584	798	215
F36A4.7	12.647	1744	1803	60	40.899	1568	1855	288
F35A5.3	12.64	33	92	60	13.135	33	112	80
F28B4.3.1	12.631	1962	2021	60	14.647	1955	2021	67
F28B4.3.2	12.631	1962	2021	60	14.647	1955	2021	67
F17C8.2	12.607	216	275	60	12.607	216	275	60
Y51H4A.12	12.554	664	723	60	14.742	651	723	73
W02A2.3	12.507	211	270	60	13.981	206	270	65
K07F5.11	12.439	226	285	60	16.864	0	289	290
M142.6a	12.439	471	530	60	15.865	454	543	90
M142.6c	12.439	471	530	60	15.865	454	543	90
Y37D8A.21	12.333	154	213	60	19.05	126	213	88
Y2H9A.3	12.316	153	212	60	20.91	144	261	118
ZC477.1	12.316	304	363	60	25.664	0	370	371
B0302.5	12.254	94	153	60	14.508	93	158	66
C50D2.4	12.247	255	314	60	16.796	188	315	128
T01G1.3	12.237	784	843	60	13.107	784	847	64
B0035.1a	12.203	106	165	60	18.904	106	205	100
B0035.1b	12.203	106	165	60	18.904	106	205	100
T06D8.9.1	12.202	189	248	60	16.417	170	315	146
T06D8.9.2	12.202	189	248	60	16.417	170	315	146
F54D5.15a	12.175	353	412	60	12.859	353	424	72
F54D5.15b.1	12.175	278	337	60	12.859	278	349	72
F54D5.15b.2	12.175	278	337	60	12.859	278	349	72
F54D5.15c	12.175	351	410	60	12.859	351	422	72
F38B7.1a	12.141	267	326	60	12.835	263	328	66
F38B7.1b	12.141	226	285	60	12.835	222	287	66
Y80D3A.3	12.088	36	95	60	17.86	0	106	107
ZK121.2	12.062	426	485	60	16.141	404	488	85
ZC373.7	12.045	184	243	60	20.153	171	259	89

F56D3.1.1	12.04	184	243	60	17.481	184	272	89
F56D3.1.2	12.04	184	243	60	17.481	184	272	89
C08B11.5	12.038	307	366	60	23.17	240	387	148
W01C8.3	12.021	856	915	60	17.307	851	922	72
ZK822.4	12.008	44	103	60	15.583	21	103	83
F23B12.4a.1	11.988	270	329	60	17.401	212	329	118
F23B12.4a.2	11.988	270	329	60	17.401	212	329	118
C49A1.4a	11.964	87	146	60	17.454	84	165	82
C49A1.4b	11.964	53	112	60	17.454	50	131	82
Y113G7A.6b	11.938	17	76	60	15.138	17	116	100
F14F7.1	11.932	221	280	60	21.819	152	280	129
F38A3.1	11.93	215	274	60	24.52	167	279	113
T02E1.3a	11.93	186	245	60	14.021	181	250	70
T02E1.3b	11.93	219	278	60	14.021	214	283	70
T02E1.3c	11.93	184	243	60	14.021	179	248	70
M02E1.1a	11.902	991	1050	60	12.494	986	1050	65
K10C8.3c.1	11.862	204	263	60	16.346	204	288	85
K10C8.3c.2	11.862	204	263	60	16.346	204	288	85
T19B4.5	11.833	58	117	60	13.67	51	117	67
F46F3.1a	11.819	35	94	60	9.199	0	94	95
T28H11.5	11.806	341	400	60	13.281	308	404	97
R05F9.10	11.795	242	301	60	14.131	238	307	70
C01B12.1	11.794	159	218	60	14.289	155	227	73
F58F6.1	11.771	204	263	60	12.194	204	264	61
T23G11.7b	11.766	176	235	60	17.241	171	254	84
C45G9.6b	11.748	91	150	60	9.525	91	154	64
T05C12.10	11.705	522	581	60	14.156	522	624	103
F32G8.3	11.641	56	115	60	15.717	32	128	97
Y92H12BR.7	11.597	22	81	60	9.566	0	81	82
F15E6.1	11.545	201	260	60	21.941	201	373	173
T19B4.2.1	11.532	1120	1179	60	19.61	1040	1200	161
T19B4.2.2	11.532	1120	1179	60	19.61	1040	1200	161
T13B5.4	11.517	160	219	60	14.362	159	278	120
C56A3.1	11.47	124	183	60	21.743	75	206	132
F21C10.18	11.325	21	80	60	12.805	18	89	72
Y73F4A.3	11.321	85	144	60	15.22	34	144	111
T10E10.7	11.28	225	284	60	19.472	177	284	108
W08E3.1	11.25	98	157	60	10.855	98	159	62
Y53C10A.12	11.201	436	495	60	15.62	436	516	81
Y51H4A.4	11.134	23	82	60	12.784	23	88	66
H43E16.1	11.099	698	757	60	32.724	629	968	340
Y105C5A.13a	11.048	28	87	60	14.355	19	93	75
F58E10.2	11.018	49	108	60	15.103	35	109	75

R01E6.5	10.981	185	244	60	11.405	184	244	61
B0024.2	10.974	178	237	60	15.074	160	259	100
C24B5.5	10.963	166	225	60	14.775	163	236	74
Y41E3.8	10.959	43	102	60	9.575	43	109	67
Y43F8B.1d	10.915	215	274	60	18.054	215	285	71
Y43F8B.1e	10.915	191	250	60	18.054	191	261	71
Y60A3A.25	10.914	52	111	60	17.082	47	141	95
Y54E10BL.2	10.913	173	232	60	17.117	147	264	118
F43D9.1	10.878	1164	1223	60	11.569	1161	1225	65
T02C5.5b	10.855	1895	1954	60	12.853	1893	1962	70
T02C5.5c	10.855	1794	1853	60	12.853	1792	1861	70
T02C5.5d.1	10.855	154	213	60	12.853	152	221	70
T02C5.5d.2	10.855	154	213	60	12.853	152	221	70
T02C5.5e	10.855	1955	2014	60	12.853	1953	2022	70
Y105C5A.4	10.852	159	218	60	15.402	159	235	77
C33G3.1a	10.816	191	250	60	19.783	188	353	166
C33G3.1b.1	10.816	285	344	60	19.783	282	447	166
C33G3.1b.2	10.816	285	344	60	19.783	282	447	166
Y46E12A.4	10.804	237	296	60	13.993	236	304	69
C33D3.3	10.799	245	304	60	11.311	245	305	61
F08B4.7	10.782	79	138	60	10.728	79	141	63
ZK632.7	10.77	6	65	60	21.061	0	131	132
F14F3.1a	10.753	291	350	60	16.227	291	404	114
F14F3.1b	10.753	105	164	60	16.227	105	218	114
F14F3.1c	10.753	132	191	60	16.227	132	245	114
ZK377.2a	10.716	928	987	60	16.832	928	1003	76
ZK377.2b	10.716	928	987	60	16.832	928	1003	76
T19D12.1	10.687	918	977	60	15.318	913	1011	99
F41F3.3	10.607	87	146	60	9.948	87	155	69
W01B6.7	10.553	150	209	60	16.847	150	267	118
Y41G9A.6	10.491	129	188	60	15.105	129	205	77
C44H4.7a	10.458	40	99	60	13.238	0	102	103
C44H4.7b	10.458	40	99	60	13.238	0	102	103
W09C2.1a	10.456	145	204	60	16.846	107	280	174
W09C2.1b	10.456	73	132	60	16.846	35	208	174
W09C2.1c	10.456	144	203	60	16.846	106	279	174
W09C2.1d	10.456	145	204	60	16.846	107	280	174
H05C05.1b	10.455	100	159	60	14.144	100	182	83
C18A3.8	10.429	158	217	60	13.166	158	245	88
F09C8.2	10.405	90	149	60	21.268	0	162	163
C40H1.1	10.381	114	173	60	13.751	82	173	92
F23B2.11.1	10.358	809	868	60	13.957	793	868	76
F23B2.11.2	10.358	809	868	60	13.957	793	868	76

C46C2.1a	10.263	1767	1826	60	11.306	1767	1837	71
C46C2.1b	10.263	1606	1665	60	11.306	1606	1676	71
C46C2.1c	10.263	1714	1773	60	11.306	1714	1784	71
C46C2.1d	10.263	1769	1828	60	11.306	1769	1839	71
C46C2.1e	10.263	1608	1667	60	11.306	1608	1678	71
C46C2.1f	10.263	1716	1775	60	11.306	1716	1786	71
C46C2.1g	10.263	1770	1829	60	11.306	1770	1840	71
C46C2.1h	10.263	1609	1668	60	11.306	1609	1679	71
C46C2.1i	10.263	1717	1776	60	11.306	1717	1787	71
C46C2.1j	10.263	1772	1831	60	11.306	1772	1842	71
C46C2.1k	10.263	1611	1670	60	11.306	1611	1681	71
C46C2.1l	10.263	1719	1778	60	11.306	1719	1789	71
C14B1.4	10.22	12	71	60	12.671	12	76	65
F16D3.2.1	10.19	102	161	60	13.209	101	193	93
F16D3.2.2	10.19	102	161	60	13.209	101	193	93
C48D5.2a	10.169	514	573	60	12.944	500	603	104
C48D5.2b	10.169	112	171	60	12.944	98	201	104
C48D5.2c	10.169	77	136	60	12.944	63	166	104
C16D9.8	10.161	58	117	60	15.311	34	126	93
Y73F8A.9	10.124	244	303	60	16.825	163	319	157
Y73F8A.8	10.124	245	304	60	16.825	164	320	157
Y105C5A.5	10.124	159	218	60	14.674	159	235	77
F01G12.5a.1	10.038	348	407	60	12.056	348	410	63
F01G12.5a.2	10.038	348	407	60	12.056	348	410	63
F01G12.5b.1	10.038	349	408	60	12.056	349	411	63
F01G12.5b.2	10.038	349	408	60	12.056	349	411	63
AC3.6	9.999	185	244	60	13.092	184	262	79
ZK863.2	9.984	164	223	60	15.161	103	223	121
Y81G3A.5b	9.867	59	118	60	10.213	0	118	119
Y105C5A.6	9.814	159	218	60	14.364	159	235	77
Y105C5A.3	9.814	159	218	60	14.364	159	235	77
C53B4.5	9.729	167	226	60	14.827	140	259	120
C12D8.8	9.618	192	251	60	14.463	157	253	97
Y77E11A.15	9.614	164	223	60	13.601	164	248	85
T01B7.7	9.606	80	139	60	14.032	71	139	69
F31E3.1	9.603	270	329	60	8.87	270	337	68
Y71G12B.9a	9.499	434	493	60	16.033	419	493	75
Y71G12B.9b	9.499	434	493	60	16.033	419	493	75
D1044.7	9.441	296	355	60	13.886	290	374	85
F27E5.2	9.408	244	303	60	10.168	244	307	64
ZK1067.7	9.401	122	181	60	13.81	66	198	133
Y113G7A.6a	9.343	331	390	60	10.316	327	390	64
Y113G7A.6c	9.343	294	353	60	10.316	290	353	64

B0205.8	9.301	43	102	60	11.801	43	119	77
C35E7.2a	9.232	508	567	60	15.308	507	578	72
C35E7.2b	9.232	204	263	60	15.308	203	274	72
F52C9.8f	9.157	16	75	60	9.151	0	91	92
F38A6.1a	9.148	444	503	60	10.519	438	505	68
F38A6.1c	9.148	349	408	60	10.519	343	410	68
F48F7.1b	9.114	5	64	60	10.27	0	87	88
R04F11.4b	9.047	6	65	60	10.599	0	65	66
R02F2.5	9.025	129	188	60	12.995	129	222	94
T22B2.4a	8.995	182	241	60	13.639	160	241	82
Y104H12A.1a	8.944	115	174	60	13.118	88	183	96
Y104H12A.1b.1	8.944	5	64	60	9.632	0	73	74
Y104H12A.1b.2	8.944	5	64	60	9.632	0	73	74
B0285.1a	8.843	670	729	60	8.843	670	729	60
B0285.1b	8.843	674	733	60	8.843	674	733	60
B0285.1c	8.843	671	730	60	8.843	671	730	60
Y59A8B.10c	8.838	0	59	60	8.264	0	60	61
T08D10.1	8.686	98	157	60	12.965	64	157	94
F48F7.1a	8.573	0	59	60	9.596	0	66	67
K04H4.1a	8.557	1193	1252	60	13.452	1193	1315	123
K04H4.1b	8.557	936	995	60	13.452	936	1058	123
D1007.1	7.979	23	82	60	8.945	0	102	103
Y42H9B.1	7.965	141	200	60	13.675	141	222	82
R12B2.1a.1	7.815	1	60	60	8.178	0	60	61
R12B2.1a.2	7.815	1	60	60	8.178	0	60	61
R12B2.1a.3	7.815	1	60	60	8.178	0	60	61
C27D6.4d	7.621	1	60	60	8.463	0	64	65
ZK381.4a.1	7.618	669	728	60	8.284	661	729	69
ZK381.4a.2	7.618	669	728	60	8.284	661	729	69
ZK381.4b	7.618	710	769	60	8.284	702	770	69
T23F6.3	7.492	68	127	60	13.372	50	127	78
F25D7.2	7.328	131	190	60	13.968	114	190	77
C15C8.1	7.262	373	432	60	9.161	373	440	68
F15B9.2	7.171	144	203	60	10.739	124	204	81
Y5H2A.4	6.266	97	156	60	13.859	48	157	110
R13H8.1a	6.23	53	112	60	12.689	39	113	75

Discussion

In my thesis, I attempted to find convincing evidence for non-canonical protein translation, as a mechanism for genetic diversity. I went through exploring the 3' UTRome to find evidence of functional STOP codon read-through, developed a computational model for detecting non-canonical translation within protein-coding sequences, and finally, explored prion proteins, where the same protein can have more than a single conformation. There are many indications that the mRNA of Eukaryotes can code for more than a single protein (Calvo et al., 2009; Dunn et al., 2013; Jungreis et al., 2011; Mouilleron et al., 2016), but it is not always clear what the underlying general mechanism is. Understanding the underlying mechanism for non canonical translation, or finding a generic systematic approach of revealing such events across species seems would be able to contribute highly to the scientific society.

I was first aiming to understand if I can find evidence for an evolutionary process that can create genetic variability with a species by simply using programmed STOP codon read-through. We already know that in the yeast *S. cerevisiae*, under specific conditions, massive events of STOP codon readthrough occur (True and Lindquist, 2000), generating many new functioning units with high potential of effects on the cell. While exploring the sequences of the yeasts 3' UTR, I noticed that it does not hold the expected distribution of STOP codon encounters expected from random sequences and that sometimes the next STOP codon could reside very far from the canonical one (the *eORF*). If translated, this could have major effects on the original protein, as it may have many new functional units. After further exploration I found that for some genes, the 3' UTR holds very high conservation across evolution, reaching the levels of the actual CDSs. Since the UTRs are in general much less conserved, this was of great interest and another indication of possible unexplored protein translation options, arising from non-canonical translation events. I found that some genes even revealed trans-membrane domains in their 3' UTR, and under certain conditions changed their cellular location, an indication of transport that might originate from a new domain being translated. Out of 6 such genes, 5 were annotated as having an

“unknown function”. Perhaps if the *eORF* would be taken into consideration as part of the protein, their function could be explained. I also looked for evolutionary evidence of *eORFs* translation by finding genes with an ortholog where the *S. cerevisiae eORF* is encoded as part of the ORF. This would suggest that either annotation could be wrong, or that evolution had developed another form of regulation by allowing two versions to be translated from one mRNA molecule. Out of 7 genes exhibiting this behavior, 5 are annotated as having an “unknown function”. For these as well, it might also be that the *eORF* holds the key to finding the protein's function. When considering experimental evidence (from ribosome-profiling), I found that the *eORFs* length probably isn’t an indicator of 3’ UTR translation, as most experimental evidence of STOP codon readthrough were actually directed to relatively short *eORFs*.

Not only did I try to predict if a 3’ UTR could be translated to benefit the cell, but I also sought to find if some genes developed defense mechanisms to prevent this from happening by having a series of consecutive STOP codons (*SMS*). I saw that some genes had multiple consecutive STOP codons, creating a signal that would probably diminish the probability of a STOP codon read-through event. Examining this across evolution and finding high conservation strengthened my hypothesis that this is a defense mechanism against the phenomena. When observing the implication of translation after the sequence of consecutive STOP codons, and finding no striking or compelling evidence, combined with the fact the indeed the early nucleotides of the 3’ UTR are generally more conserved, I realized that this observation is probably due to pure chance and has no real meaning in terms of mechanism.

This phenomena is of course not specific to single cell organisms. Evidence of abundant stop codon read through had also been in *Drosophila* (Dunn et al., 2013). This gave me the desire to explore higher organisms for non canonical translation, as it is harder to come across such events empirically when handling complex organisms, so a systematic approach must come in handy.

When expending the analysis to the human genome, I was able to cross possible STOP codon read-through events with mutations and pathogenic behavior reported. When Observing the same sequence composition and conservation properties as was done for *S. cerevisiae*, I did not find more

compelling evidence that could indicate a higher probability of STOP codon read-through or, translation of the 3' UTR in general. This had an even less dramatic effect when examining genes with a sequence of STOP codons right after the canonical one. However, out of the genes having consecutive STOP codons that are conserved in evolution, 15 genes were previously reported as having frameshift mutations (*indels*) right on the canonical STOP codon and are implicated in some pathology. This alone does not give a direct connection between the mutation and the disease, but further exploration of the implication of translation beyond the consecutive sequence of STOP codons could be explored further.

After exploring sequence properties of non-translated regions and getting a feel and an understanding of what translated sequences properties have in common, I moved on to a more elaborate task of identifying events creating non-canonical translation within known coding sequences. Discovering frame shifting events is usually done empirically, focusing on specific genes, performing many experiments until reaching a discovery (Loughran et al., 2018; Michel et al., 2012). I describe a unique algorithm I developed that allows me to systematically detect alternative translation options within genes, based solely on DNA sequences. I utilized my algorithm to specifically detect frame changes, or the potential for translation in more than one active frame, as classical translation dictates. The method is based on the fact that the genetic code holds redundancies that are captured by the wobble position of codons being less conserved (Trotta, 2011). This fact generates a periodic conservation signal across protein-coding genes that can be analyzed using harmonic methods, such as Fourier transform. Not only can it capture the "strength" of this periodicity to deduce the potential to be protein-coding, but it can also determine the frame of translation by examining the phase of the periodic pattern. I developed two computational models for translational frame determination using the periodic pattern and I applied them to almost 40,000 human protein-coding genes. I then developed and applied another algorithm to predict which genes present frameshifts, and found that there could be thousands of genes with more than one active frame. This is not surprising, as the algorithm cannot distinguish between ribosomal frameshift and frameshifts that are present due to alternative splicing. When eliminating all known transcripts that are frameshifted relative to their major transcript, I was left with

~400 genes still presenting possible translation in the non-canonical frame. Both methods have been validated by observing the results of the only 3 known ribosomal frameshifts in mammals: OAZ1 and PEG10 (Atkins et al., 2016; Michel et al., 2012). I have also validated my methods by observing known viral ribosomal frameshifts in HIV and in SARS-Cov-2 genes (Baranov et al., 2005; Nikolaitchik and Hu, 2014), and a dual coding region within the HPV genome (Graham and Faizo, 2017). When examining the unknown cases I observed that for some cases, translation in the new frame is not known to produce an annotated protein from the human proteome, but there was a version in other species (that may be very close to humans) that does present this version of the protein. This might mean that either annotation of the human proteome is still lacking, or that evolution discovered a way to diversify its repertoire of proteins by applying ribosomal frameshift under varying conditions.

I have compared the results from my approach to another work aiming at systematically detecting frameshifting events (Michel et al., 2012). In their work they had tried to detect frameshift events from ribosome profiling experiments. This approach is limited as the method depends on minimal counts of P-sites to make a determination into the translation. This alone eliminates the possibility to analyze most of the human genome. While comparing (Michel et al., 2012) with mine on the novel genes, there were some overlaps. Since they report high false positive rates, it’s also hard to conclude which of the predictions is indeed expected to overlap, and be actually considered a frameshifting event. Methods like the one presented in this thesis, allows researchers to generate much more focused efforts for finding experimental evidence into some of the events presented.

Finding evidence of ribosomal frameshift event is of high importance, as it is a complex mechanism requiring much regulation, and is not well described in mammals. Viruses use this mechanism to enrich their proteome with a very compact genome (Jacobs et al., 2007; Theis et al., 2008). I tried to impose known conditions such as sequence motifs or mRNA secondary structures to explain some of the predicted events but in general, failed to do so. While it makes sense that mechanisms should be shared, as the translation mechanism for a virus is the same as its host, if these phenomena were

evolved to also happen in the host, perhaps the rules of regulations are different, and more stringent.

I could not find any common denominators between the predicted genes, even after subgrouping them under various constraints. It points out that probably every case discovered is a world on its own and deserves its unique treatment. This comes inline with the fact that other works trying to find high abundance of dual coding events in humans, have not presented common denominators (Michel et al., 2012). This strengthens the importance of being able to systematically discover new translation events, as they do not hold many commonalities between them. Having a method that only examines the sequence can find unrelated proteins, presenting similar mechanisms of translation otherwise unknown.

When observing the predictions on their own I could find some interesting cases as were documented in the main text. It was interesting to see that for some cases the frame-shifted version that I predict is annotated as its own transcript, most times the major one, in other organisms. This supports the fact that with a high probability this version of the protein should exist in human cells as well, but it makes more sense than it would result in a new transcript rather than a ribosomal frameshift.

Most known ribosomal frameshifts documented (in viruses mostly) are ones where the ribosome shifts to the -1 frame relative to the canonical one. Since many such cases were experimentally proven, motifs describing the transition have been suggested (Jacobs et al., 2007). Since most of these motifs were found and discovered in viral genomes, it was no surprise that I could not find them motifs in my predictions. Furthermore, I sought to characterize my motifs as well. I could not find any common sequence motif. Still, I did find a subset of +1 predicted frameshift cases where an mRNA secondary structure “motif” was visible, in the form of a tight secondary structure (low free energy) right at the frameshift site. It might hint into the mechanism that causes the ribosome to move to that frame, much like the slippery site described for -1 ribosomal frameshifts in viruses (Wills et al., 2006).

I was eager to find experimental support for the predictions presented in this work. I started by exploring evidence of function changes for some of the most promising candidates one by one. I quickly realized that I could also use

a more systematic way to gain some experimental insight into some of the predictions. I could apply the same rule-based algorithm I had developed (by adapting it to another data source) on extracted elongating ribosome P-sites from ribosome profiling experimental data to support the findings further. I showed that when using this data, I can still detect the known ribosomal frameshifts in the human genome. Still, due to the lack of sufficient data on our attractive candidates described in the text, I could not extract the same results to support them. The same problem holds when working with mass spectrometry data. For both cases, I would have to find data from the specific tissue where the gene is expressed and find the conditions allowing for the frameshift to take place. Since I have no prior knowledge of the environmental conditions and partial expression information, this becomes tedious and requires much computational power, data, and time.

Non the less, this technique lays the ground for ample future research, as some predictions show much potential as novel protein versions that have special functions not yet discussed. I believe that meticulous study of some may lead to attractive new prospects in many pathways and processes.

By combining the insights gained from observing known frameshift sequences and using simulated sequences as a training set for a classifier to predict translational frames, the algorithm presented a somewhat generalized method for predicting non-canonical frames in many conservational datasets and not only mammalian only. As shown, I was also able to predict a newly discovered ORF in the SARS-Cov-2 genome (Baranov et al., 2005), using the optimized model to predict such events for a mammalian dataset. Adding more simulated and evidence-based sequences and expanding the models to have more features as inputs (such as actual evolutionary distance and perhaps even organism-specific features) can produce even more accurate results in the future. I also believe that this can be further used with more databases, including human-only single nucleotide variants, using the allele frequency as the variability score. This will create a human-specific analysis based on actual reported cases, giving a higher probability of them being active. It could also be applied for less studied and annotated genomes to aid in annotation by examining the periodic pattern and suggesting translating and non-translating regions.

Finally, the algorithm is quite generic and could be expanded to find other frequencies of periodic signals. It could be used to find periodic motifs, analyze sequences with modified genetic tables, and so on.

The last part of my thesis describes my exploration of prions in *S. cerevisiae* as environment-dependent (SUP35), which serve as regulators of STOP codon read-through in response to stress conditions that might require this mechanism as a means of survival (True and Lindquist, 2000). I hoped to find similar cases in higher organisms and describe a mechanism of action that may lead to more elaborate studies, including the inheritance of prion proteins. I explored two leading candidates in *C. elegans* that seemed to have high potential as “contributing” prions in the sense that they are serving as a defense memory mechanism, reacting to specific stress conditions. ABU-13 was a strong candidate for this as it is known to respond to ER stress with high expression and is known to have immune-like responses to pathogens (Sun et al., 2011). MUT-16, on the other hand, seemed to have potential implications for RNAi inheritance in an epigenetic manner (Phillips et al., 2012). Both proteins score very high in the prediction algorithm to have a prion forming domain and appear to check all the boxes for prion behavior, but I have yet to prove this. It would be of great interest to continue with the proposed experiments and prove that ABU-13 will serve as a memory seed for pathogen resistance. Should this be demonstrated experimentally, orthologs in even higher organisms could be explored, and perhaps new mechanisms of immune systems could be discovered.

References

Alberti, S., Halfmann, R., King, O., Kapila, A., and Lindquist, S. (2009). A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. Cell *137*, 146–158.

Alberti, S., Halfmann, R., and Lindquist, S. (2010). Biochemical, cell biological, and genetic assays to analyze amyloid and prion aggregation in yeast. (Elsevier Inc).

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. *15*, 193–204.

Anyanful, A., Easley, K. a, Benian, G.M., and Kalman, D. (2009). Conditioning protects C. elegans from lethal effects of enteropathogenic E. coli by activating genes that regulate lifespan and innate immunity. Cell Host Microbe *5*, 450–462.

Arribere, J.A., Bell, R.T., Fu, B.X.H., Artiles, K.L., Hartman, P.S., and Fire, A.Z. (2014). Efficient marker-free recovery of custom genetic modifications with CRISPR/Cas9 in Caenorhabditis elegans. Genetics *198*, 837–846.

Atkins, J.F., Loughran, G., Bhatt, P.R., Firth, A.E., and Baranov, P. V. (2016). Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. *44*, 7007–7078.

Baranov, P. V., Henderson, C.M., Anderson, C.B., Gesteland, R.F., Atkins, J.F., and Howard, M.T. (2005). Programmed ribosomal frameshifting in decoding the SARS-CoV genome. Virology *332*, 498–510.

Baudin-Baillieu, A., Legendre, R., Kuchly, C., Hatin, I., Demais, S., Mestdagh, C., Gautheret, D., and Namy, O. (2014). Genome-wide Translational Changes Induced by the Prion [PSI+]. Cell Rep. *8*, 439–448.

Bowie, J.U. (1997). Helix Packing in Membrane Proteins. 780–789.

Brachat, S., Dietrich, F.S., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T., and Philippsen, P. (2003). Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. Genome Biol. *4*, R45.

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat. Rev. Mol. Cell Biol. *16*, 651–664.

Breker, M., Gymrek, M., Moldavski, O., and Schuldiner, M. (2014). LoQAtE - Localization and Quantitation ATlas of the yeast proteomE. A new tool for multiparametric dissection of single-protein behavior in response to biological perturbations in yeast. Nucleic Acids Res. *42*, 726–730.

Brenner, C., Pace, H.C., and Brenner, C. (2016). The nitrilase superfamily : Classification , structure and function The nitrilase superfamily : classification , structure and function. 1–9.

Brent, M.R. (2005). Genome annotation past, present, and future: How to define an ORF at each locus. Genome Res. *15*, 1777–1786.

Brent, M.R. (2007). How does eukaryotic gene prediction work? Nat. Biotechnol. *25*, 883–885.

Byrne, K.P., and Wolfe, K.H. (2005). The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. *15*, 1456–1461.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 165–170.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. *106*, 7507–7512.

Chan, C.S., Jungreis, I., and Kellis, M. (2013). Heterologous stop codon readthrough of metazoan readthrough candidates in yeast. PLoS One *8*, e59450.

Chng, S.C., Ho, L., Tian, J., and Reversade, B. (2013). ELABELA: A hormone essential for heart development signals via the apelin receptor. Dev. Cell *27*, 672–680.

Chung, W.Y., Wadhawan, S., Szklarczyk, R., Pond, S.K., and Nekrutenko, A. (2007). A first look at ARFome: Dual-coding genes in mammalian genomes. PLoS Comput. Biol. *3*, 0855–0861.

Clare, J.J., Belcourt, M., and Farabaugh, P.J. (1988). Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. Proc. Natl. Acad. Sci. U. S. A. *85*, 6816–6820.

Dickinson, D.J., Ward, J.D., Reiner, D.J., and Goldstein, B. (2013). Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. Nat. Methods *10*, 1028.

Dinman, J.D. (2006). Programmed Ribosomal Frameshifting Goes Beyond Viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift. Microbe Wash. DC. *1*, 521–527.

Dorogush, A.V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. 1–7.

Dubes, R., and Jain, A.K. (1980). Clustering Methodologies in Exploratory Data Analysis. M.C. Yovits, ed. (Elsevier), pp. 113–228.

Dulude, D., Baril, M., and Brakier-Gingras, L. (2002). Characterization of the frameshift stimulatory signal controlling a programmed -1 ribosomal frameshift in the human immunodeficiency virus type 1. Nucleic Acids Res. *30*, 5094–5102.

Dunham, I., Shimizu, N., Roe, B.A., and Chissoe, S. (2000). Erratum: correction: The DNA sequence of human chromosome 22. Nature *404*, 904–904.

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *3*, 1–32.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 1–7.

Exploration, G., Genomic, R., City, Y., and Hayashizaki, Y. (2001). Sequences With Frameshift Errors. 81–87.

Farabaugh, P.J., Kramer, E., Vallabhaneni, H., and Raman, A. (2006). Evolution of +1 programmed frameshifting signals and frameshift-regulating tRNAs in the order saccharomycetales. J. Mol. Evol. *63*, 545–561.

Fickett, J.W. (1996). Finding genes by computer: the state of the art. Trends Genet. *12*, 316–320.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Stein, D., Israeli, O., et al. (2021). The coding capacity of SARS-CoV-2. Nature *589*, 125–130.

Firth, A.E. (2020). A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a. J. Gen. Virol. *101*, 1085–1089.

Frith, M.C., Bailey, T.L., Kasukawa, T., Mignone, F., Kummerfeld, S.K., Madera, M., Sunkara, S., Furuno, M., Bult, C.J., Quackenbush, J., et al. (2006). Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol. *3*, 40–48.

Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., Hayashizaki, Y., and Okazaki, Y. (2003). CDS annotation in full-length cDNA sequence. Genome Res. *13*, 1478–1487.

Fusco, C., Micale, L., Augello, B., Teresa Pellico, M., Menghini, D., Alfieri, P., Cristina Digilio, M., Mandriani, B., Carella, M., Palumbo, O., et al. (2014). Smaller and larger deletions of the Williams Beuren syndrome region implicate genes involved in mild facial phenotype, epilepsy and autistic traits. Eur. J. Hum. Genet. *22*, 64–70.

Giglione, C., Vallon, O., and Meinnel, T. (2003). Control of protein life-span by N-terminal methionine excision. EMBO J. *22*, 13–23.

Godard, P., Urrestarazu, A., Vissers, S., Kontos, K., Bontempi, G., van Helden, J., and Andre, B. (2007). Effect of 21 Different Nitrogen Sources on Global Gene Expression in the Yeast Saccharomyces cerevisiae. Mol. Cell. Biol. *27*, 3065–3086.

Graham, S. V, and Faizo, A.A.A. (2017). Control of human papillomavirus gene expression by alternative splicing. Virus Res. *231*, 83–95.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Harris, K., Lamson, R.E., Nelson, B., Hughes, T.R., Marton, M.J., Roberts, C.J., Boone, C., and Pryciak, P.M. (2001). Role of scaffolds in MAP kinase pathway specificity revealed by custom design of pathway-dedicated signaling proteins Kendra Harris*, Rachel E. Lamson*, Bryce Nelson. Curr. Biol. *11*, 1–10.

Harte, R.A., Karolchik, D., Kuhn, R.M., Kent, W.J., and Haussler, D. (2010). Databases and Genome Browsers. In Vogel and Motulsky's Human Genetics, M.R. Speicher, A.G. Motulsky, and S.E. Antonarakis, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 905–921.

Hatzigeorgiou, A.G., Fiziev, P., and Reczko, M. (2001). DIANA-EST: A statistical analysis. Bioinformatics *17*, 913–919.

Hilger, D., Masureel, M., and Kobilka, B.K. (2018). Structure and dynamics of GPCR signaling complexes. Nat. Struct. Mol. Biol. *25*, 4–12.

Hunt, R.C., Simhadri, V.L., Iandoli, M., Sauna, Z.E., and Kimchi-Sarfaty, C. (2014). Exposing synonymous mutations. Trends Genet. *30*, 308–321.

Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. *15*, 205–213.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009a). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science (80-.). *12*, 7536–7545.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009b). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Jackson, R.J., Hellen, C.U.T., and Pestova, T. V. (2012). Termination and post-termination events in eukaryotic translation. Adv. Protein Chem. Struct. Biol. *86*, 45–93.

Jacobs, J.L., Belew, A.T., Rakauskaite, R., and Dinman, J.D. (2007). Identification of functional, endogenous programmed - 1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae. Nucleic Acids Res. *35*, 165–174.

Jonathan, D.D. (2012). Mechanisms and implications of Programmed Translational Frameshifting. Changes *29*, 997–1003.

Jungreis, I., Lin, M.F., Spokony, R., Chan, C.S., Negre, N., Victorsen, A., White, K.P., and Kellis, M. (2011). Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res. *21*, 2096–2113.

Jungreis, I., Sealfon, R., and Kellis, M. (2020). Sarbecovirus comparative genomics elucidates gene content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations. BioRxiv Prepr. Serv. Biol. 1–39.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. *7*, 1511–1522.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241.

Ketteler, R. (2012). On programmed ribosomal frameshifting: The alternative proteomes. Front. Genet. *3*, 1–10.

Klemke, M., Kehlenbach, R.H., and Huttner, W.B. (2001). Cde358. *20*, 3849–3860.

Kotlar, D., and Lavner, Y. (2003). Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions. Genome Res. *13*, 1930–1937.

Kovacs, E., Tompa, P., Liliom, K., and Kalmar, L. (2010). Dual coding in alternative reading frames correlates with intrinsic protein disorder. Proc. Natl. Acad. Sci. U. S. A. *107*, 5429–5434.

Kozak, M. (1987). An analysis of 5’-noncoding sequences from 699 vertebrate messenger rNAS. Nucleic Acids Res. *15*, 8125–8148.

Kozak, M. (1991). An analysis of vertebrate mRNA sequences: Intimations of translational control. J. Cell Biol. *115*, 887–903.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. Gene *234*, 187–208.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kupfer, L., Hinrichs, W., and Groschup, M. (2009). Prion Protein Misfolding. Curr. Mol. Med. *9*, 826–835.

Lek, M., Karczewski, K.J., Minikel, E. V, Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, J., and Zhang, Y. (2015). Translation with frameshifting of ribosome along mRNA transcript. 1–18.

Li, L., Wang, A.L., and Wang, C.C. (2002). Structural Analysis of the -1 Ribosomal Frameshift Elements in Giardiavirus mRNA. J. Virol. *75*, 10612–10622.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, 275–282.

Lorenz, R., Bernhart, S.H., zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Loughran, G., Jungreis, I., Tzani, I., Power, M., Dmitriev, R.I., Ivanov, I.P., Kellis, M., and Atkins, J.F. (2018). Stop codon readthrough generates a C-terminally extended variant of the human Vitamin D receptor with reduced calcitriol response. J. Biol. Chem. *293*, 4434–4444.

Lowe, T.M., and Eddy, S.R. (1996). TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M., and Saghatelian, A. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J. Proteome Res. *13*, 1757–1765.

Magrane, M., and Consortium, U.P. (2011). UniProt Knowledgebase: A hub of integrated protein data. Database *2011*, 1–13.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics *20*, 2878–2879.

Mathé, C., Sagot, M.F., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. *30*, 4103–4117.

Mauger, D.M., Joseph Cabral, B., Presnyak, V., Su, S. V., Reid, D.W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M.J., et al. (2019). mRNA structure regulates protein expression through changes in functional

half-life. Proc. Natl. Acad. Sci. U. S. A. *116*, 24075–24083.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. Mol. Biol. Evol. *21*, 1781–1791.

McGillivray, P., Ault, R., Pawashe, M., Kitchen, R., Balasubramanian, S., and Gerstein, M. (2018). A comprehensive catalog of predicted functional upstream open reading frames in humans. Nucleic Acids Res. *46*, 3326–3338.

Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F., and Baranov, P. V. (2012). Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. *22*, 2219–2229.

Mikl, M., Alon, A., Mordret, E., Pilpel, Y., and Segal, E. (2018). Extensive programmed ribosomal frameshifting in human as revealed by a massively parallel reporter assay. BioRxiv 469692.

Min, X.J., Butler, G., Storms, R., and Tsang, A. (2005). OrfPredictor: Predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. *33*, 677–680.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat. Genet. *30*, 13–19.

Molina-García, L., and Giraldo, R. (2017). Enabling stop codon read-through translation in bacteria as a probe for amyloid aggregation. Sci. Rep. *7*, 1–9.

Moon, S., Byun, Y., and Han, K. (2007). FSDB: a frameshift signal database. Comput. Biol. Chem. *31*, 298–302.

Mouilleron, H., Delcourt, V., and Roucou, X. (2016). Death of a dogma: Eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res. *44*, 14–23.

Namy, O., Hatin, I., and Rousset, J.P. (2001). Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep. *2*, 787–793.

Namy, O., Duchateau-Nguyen, G., Hatin, I., Hermann-Le Denmat, S., Termier, M., and Rousset, J.P. (2003). Identification of stop codon readthrough genes in Saccharomyces cerevisiae. Nucleic Acids Res. *31*, 2289–2296.

Namy, O., Rousset, J.P., Napthine, S., and Brierley, I. (2004). Reprogrammed Genetic Decoding in Cellular Gene Expression. Mol. Cell *13*, 157–168.

Nikolaitchik, O.A., and Hu, W.-S. (2014). Deciphering the Role of the Gag-Pol Ribosomal Frameshift Signal in HIV-1 RNA Genome Packaging. J. Virol. *88*, 4040–4046.

Novoselova, T. V., Zahira, K., Rose, R.S., and Sullivan, J.A. (2012). Bul proteins, a nonredundant, antagonistic family of ubiquitin ligase regulatory proteins. Eukaryot. Cell *11*, 463–470.

Nussbaum-Krammer, C.I., Park, K.-W., Li, L., Melki, R., and Morimoto, R.I. (2013). Spreading of a prion domain from cell-to-cell by vesicular transport in Caenorhabditis elegans. PLoS Genet. *9*, e1003351.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., et al. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat. Genet. *36*, 40–45.

Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T., and Sugano, S. (2004). Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res. *14*, 2048–2052.

Pelletier, J., and Sonenberg, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature *334*, 320–325.

Phillips, C.M., Montgomery, T. a, Breen, P.C., and Ruvkun, G. (2012). MUT-16 promotes formation of perinuclear mutator foci required for RNA silencing in the C. elegans germline. Genes Dev. *26*, 1433–1444.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: An update on mammalian reference sequences. Nucleic Acids Res. *42*, 756–763.

Prusiner, S.B. (1991). Molecular biology of prion diseases. Science (80-.). *252*, 1515–1522.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Ribrioux, S., Brüngger, A., Baumgarten, B., Seuwen, K., and John, M.R. (2008). Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC Genomics *9*, 1–16.

Rogic, S., Mackworth, A.K., and Ouellette, F.B.F. (2001). Evaluation of gene-finding programs on mammalian sequences. Genome Res. *11*, 817–832.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Sakoe, H., and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. *26*, 43–49.

Salamov, A.A., and Solovyev, V. V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res. *10*, 516–522.

Schueren, F., Lingner, T., George, R., Hofhuis, J., Dickel, C., Gärtner, J., and Thoms, S. (2014). Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. Elife *3*, e03640.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*.

Skuzeski, J.M., Nichols, L.M., Gesteland, R.F., and Atkins, J.F. (1991). The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. J. Mol. Biol. *218*, 365–373.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64.

Soemedi, R., Cygan, K.J., Rhine, C., Glidden, D.T., Allison, J., Lin, C., Fredericks, A.M., Fairbrother, W.G., and Biology, C. (2018). HHS Public

Access. *1*, 36–44.

Sun, J., Singh, V., Kajino-Sakamoto, R., and Aballay, A. (2011). Neuronal GPCR controls innate immunity by regulating noncanonical unfolded protein response genes. Science *332*, 729–732.

Sychrova, H., Braun, V., Potier, S., and Souciet, J.L. (2000). Organization of specific genomic regions of Zygosaccharomyces rouxii and Pichia sorbitophila: Comparison with Saccharomyces cerevisiae. Yeast *16*, 1377–1385.

Tan, C.L., Gunaratne, J., Lai, D., Carthagena, L., Wang, Q., Xue, Y.Z., Quek, L.S., Doorbar, J., Bachelerie, F., Thierry, F., et al. (2012). HPV-18 E2E4 chimera: 2 new spliced transcripts and proteins induced by keratinocyte differentiation. Virology *429*, 47–56.

Theis, C., Reeder, J., and Giegerich, R. (2008). KnotInFrame: Prediction of -1 ribosomal frameshift events. Nucleic Acids Res. *36*, 6013–6020.

Toombs, J. a, Petri, M., Paul, K.R., Kan, G.Y., Ben-Hur, A., and Ross, E.D. (2012). De novo design of synthetic prion domains. Proc. Natl. Acad. Sci. U. S. A. *109*, 6519–6524.

Torabi, N., and Kruglyak, L. (2011). Variants in SUP45 and TRM10 underlie natural variation in translation termination efficiency in Saccharomyces cerevisiae. PLoS Genet. *7*, e1002211.

Trotta, E. (2011). The 3-base periodicity and codon usage of coding sequences are correlated with gene expression at the level of transcription elongation. PLoS One *6*.

True, H.L., and Lindquist, S.L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. Nature *407*, 477–483.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Urano, F., Calfon, M., Yoneda, T., Yun, C., Kiraly, M., Clark, S.G., and Ron, D. (2002). A survival pathway for Caenorhabditis elegans with a blocked unfolded protein response. J. Cell Biol. *158*, 639–646.

Vanderperre, B., Lucier, J.F., and Roucou, X. (2012). HAltORF: Adatabase of predicted out-of-frame alternative open reading frames in human. Database *2012*, 1–5.

Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.M., and Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. PLoS One *8*.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. Nature *449*, 54–61.

Wass, M.N., Kelley, L.A., and Sternberg, M.J.E. (2010). 3DLigandSite: Predicting ligand-binding sites using similar structures. Nucleic Acids Res. *38*, 469–473.

Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdiscip. Rev. RNA *5*, 765–778.

Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M.A., and Leutz, A. (2014).

UORFdb - A comprehensive literature database on eukaryotic uORF biology. Nucleic Acids Res. *42*, 60–67.

Wills, N.M., Moore, B., Hammer, A., Gesteland, R.F., and Atkins, J.F. (2006). A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. J. Biol. Chem. *281*, 7082–7088.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-, K., Lander, E.S., and Kellis, M. (2010). NIH Public Access. *434*, 338–345.

Xu, H., Wang, P., Fu, Y., Zheng, Y., Tang, Q., Si, L., You, J., Zhang, Z., Zhu, Y., Zhou, L., et al. (2010). Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Res. *20*, 445–457.

Yada, T., Takagi, T., Totoki, Y., Sakaki, Y., and Takaeda, Y. (2003). DIGIT: a novel gene finding program by combining gene-finders. Pac. Symp. Biocomput. *387*, 375–387.

Yandell, M., and Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342.

Yang, X.S., Tian, Z.M., Yuan, J.J., Zhang, Y.T., and Shao, W. (2013). Numerical study on indoor wideband channel characteristics with different internal wall. Radioengineering *22*, 1169–1175.

Zhang, M.Q. (2002). Computational prediction of eukaryotic protein-coding genes. Nat. Rev. Genet. *3*, 698–709.

Zhang, C., Montgomery, T. a, Gabel, H.W., Fischer, S.E.J., Phillips, C.M., Fahlgren, N., Sullivan, C.M., Carrington, J.C., and Ruvkun, G. (2011). mut-16 and other mutator class genes modulate 22G and 26G siRNA pathways in Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A. *108*, 1201–1208.

הכלי נוסף בו השתמשתי הוא אלגוריתם שפותח כבר בעבר אשר חוזה את הפוטנציאל של חלבון מסוים להוות חלבון בעל תכונות פריוניות. הכלי פותח על בסיס רצפים שידוע שמכילים תכונות שכאלה והתבסס בעיקר על רצפים מהשמר *Saccharomyces cerevisiae*. אחת שהנחתי בסיס חישובי עבור תחזיות של חלבונים פריונים, המשכתי בחיפושים למציאת עדות ביולוגית וחשיבות לממצאים החישוביים: (1) מצאתי כי החלבונים אשר נמצאו בראשית הרשימה של התחזיות כבעלי תכונות פריוניות ב C. elegans הם חלבונים הקשורים בעקה אשר ייתכן והם קשורים בהגנה מפתוגנים ע"י חשיפה ראשונית. (2) מצאתי כי חלק מהגנים שחזיתי שעלולים לעבור קריאה מעבר לקודון העצירה ב S. cerevisiae "יקבלו" אזור טרנס-ממברנלי עקב תרגום של ה 3' UTR, וגם מראים עדות לכך שהם משנים את מיקומם התת תאי עקב שינוי סביבתי. (3) קיבלתי עדות אבולוציונית וניסיונית (דרך ריצוף ריבוזומים) לתרגום אלטרנטיבי של כמה מהחלבונים שחזיתי כמתרגמים במסגרת קריאה שונה מזו הקאנונית, מה שמחזק את הסיכוי לכך שתחזיות אלו באמת יכולות להיות יחידות מתפקדות בתא. באופן ספציפי, מצאתי כי חלבון מסוים בחומר הגנטי המרכיב את וירוס ה COVID-19 אשר חזיתי כי בעל תכונות תרגום אלטרנטיביות, הוכח נסיונית דרך ריצוף ריבוזום וספקטרומטריה, כי הוא אכן בעל תכונות אלו.

בתיזה אני מתמקדת בעיקר בממצאים שעלו מגילוי מסגרת קריאה שאינה קאנונית בבני אדם וביונקים באופן כללי, שכן אלו התוצאות המקיפות והמעניינות ביותר שמצאתי. מהממצאים אני מסיקה ששינוי מסגרת קריאה עלול לשנות באופן דרמטי את התוצר המתורגם, מה שייצור שינויים גדולים בחלבון כגון קטיעות, אובדן של אזורים תפקודיים, שינוי בתכונות קשירה ואינטראקציה ועוד. ההשלכות של שינויים אלו יכולות להיות משמעותיות, וזה נותן עוד רמה של גיוון לפרוטאום שייתכן ותוכנן כך מראש (כלומר לא נובע מטעויות תרגום).

מכניזמים כאלו ידועים באורגניזמים שלהם גנומים מצומצמים, אך לא הראו עדיין כי קיים כזה מנגנון סיסטמטי ביונקים.

אני מעלה את ההשערה כי חלק מאירועי שינוי מסגרת הקריאה שאני מוצאת בבני אדם עלולים להיות אימוץ של תכונות מוירוסים שנגרמו עקב זיהום שקרה, אשר משתמש במנגנונים שמאפשרים יצירת חלבונים חדשים כאשר יש בכך צורך.

תקציר

כיום ברור כי הרעיון המקורי של "גן אחד מקודד לחלבון בודד" היה נאיבי, כיון שתהליכים רבים כגון ספלייסינג אלטרנטיבי, מאפשרים גרסאות רבות של חלבון הנובעות מלוקוס בודד בגנום. בעבודת התיזה שלי פיתחתי גישות חישוביות חדשות לגילוי של חלבונים ופפטידים לא קאנונים. כלים אלו הופעלו על מנת לחזות תרגום לא קאנוני של חלבונים שיכול לקרות עקב קריאה מעבר לקודון העצירה ושינוי מסגרת קריאה. יותר מכך, אני מראה כי כלים אלו מאפשרים לחזות פוטנציאל פריוני (דוגמא נוספת לפעילות לא קאנונית של חלבונים). ההנחה הבסיסית היא שבכל המקרים האלו, מולקולת mRNA בודדת יכולה להיות מתורגמת כדי לייצר כמה תוצרים חלבוניים בזכות השינוי המתרחש במהלך התרגום עצמו.

הכלים עוצבו על מנת לגלות חלבונים חדשים עם פוטנציאל ליצור אזורים פריונים, קריאה מאסיבית מעבר לקודון העצירה ומקרים רבים של שינוי מסגרת קריאה. באופן ספציפי אני מוצאת יותר מ200 גרסאות חדשות של חלבונים בבני אדם שאני מציעה כי נובעים כתוצאה משינוי מסגרת הקריאה תוך כדי תרגום על ידי הריבוזום. אני מציגה ראיות שבמקרים מסוימים קריאה מעבר לקודון העצירה עלול לשנות באופן קיצוני את פעילותו המקורית של החלבון. כדוגמא מייצגת, אני גם מגלה שני חלבונים בנמטודה *Caenoharbditis elegans* בעלי פוטנציאל גבוה להיות פריונים.

תהליך הגילוי במלואו נח על ניתוחים חישוביים שבוצעו על רצפי DNA או חלבון, על מנת לקבל תחזיות לתרגום לא קאנוני או לפעילות לא קאנונית של חלבונים:

הכלי המרכזי הפותח עבור תיזה זו התבסס על אנאליזה של קבוצת מאפיינים של רצפי DNA שמקודדים לחלבונים. התכונה הכי מאפיינת הינה תבנית מחזורית בשונות של נוקלאוטידים, המבוססת על העובדה כי טבלת הקידוד היא בעלת יתירות, ובכך מאפשרת גמישות ביצירת רצפים המקודדים לחלבונים, בעוד שהם שומרים על הרצף הרצוי של חומצות אמינו. ניתן להשתמש בתכונה זו על מנת לבחון רצף DNA ספציפי כדי לקבל את הפוטנציאל שלו לקודד לחלבון, בהתבסס על "כמות" התבנית המחזורית שחבויה בו. השתמשתי בתכונה זו לא רק על מנת לקבל פוטנציאל של רצף להיות מקודד לחלבון אלא גם על מנת למצוא מסגרות קריאה אלטרנטיביות בגנים שידוע כבר שהם מקודדים לחלבון, ובכך מאפשרים לגרסאות חדשות של חלבונים להיחשף.

חיזוי תרגום לא קנוני בגנים המקודדים לחלבונים

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

מאת:

עמית אלון

הוגש לסנאט אוניברסיטת תל אביב:

27/10/2019

עבודה זו נעשתה בהנחיית פרופסור עודד רכבי מאוניברסיטת תל אביב, בשיתוף פעולה עם פרופסור צחי פלפל ממכון ויצמן

פרופ. עודד רכבי

חיזוי תרגום לא קנוני בגנים המקודדים לחלבונים

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

מאת:

עמית אלון

הוגש לסנאט אוניברסיטת תל אביב:

27/10/2019