8-2018

# Alternative Reading Frames in Protein-coding DNA, a functional and regulatory mystery

Travis Wheeler
*University of Montana, Missoula*

## University Grant Program 2017-2018 – Final Report

Name: Travis Wheeler
Department: Computer Science
Project title: Alternative Reading Frames in Protein-coding DNA, a functional and regulatory mystery

### Objective

We have found that ~13% of all human genes include at least one exon that can (and apparently does) encode alternate peptides in more than one reading frame. These surprisingly prevalent dual-coding exons are nearly all conserved in the mouse genome, strongly suggesting that they play an important biological role. The objective of this seed grant was to lay the groundwork for a grant proposal to characterize these dual-coding exons, understand when and where they are functional, and learn what sort of sequence and structural traits correlate with their presence.

### Summary of Results

Funds from the UGP grant were used to pay wages for two students materially involved in gathering preliminary data and results to be used in an NIH R01 proposal to be submitted in October of 2018. Below, I present some highlights of that pilot work:

**Identification of dual-coding Exons.** We identified the aforementioned dual-coding exons using a software tool developed in our lab prior to the start of this project. By running Mirage on all human isoforms in UniProtKB, we identified dual-coding exons within alternative splicing products. We found that 2,799 of the 21,980 UniProtKB families contain at least one dual-coding exon, and 68 contain a triple-coding exon (including an exon in the human MLL5 gene that encodes three open reading frames each encoding a 131 amino acid). Most dual-coding regions are short (only 18% are longer than 20 amino acids) and consist of a single exon (82% are 1-exon long, 14% involve two consecutive exons, 3% have three exons, while the remaining 1% are longer). Of dual-coding regions, 45% are 5' terminal (most are nearly full length, suggesting that nonsense-mediated decay is unlikely), while the remaining 55% revert to the standard frame in 5' flanking exons. Genes with dual-coding exons are distributed across the genome, though not uniformly: they are 10x denser on chromosome 19 than on chromosome 8, and extremely rare on Y.

**Dual-coding nature of exons is conserved.** To confirm that our observations are not simply the result of noisy splicing, we inspected mouse orthologs to human dual-coding exons, with orthology defined by UCSC whole genome alignments. We found that 98% of all these mouse exons encode at least two open reading frames as well (Fig 1, blue dots). While most of the dual-coding exons are short, and thus may have two open reading frames by chance, this feature is found in nearly all longer exons as well. This stands in stark contrast to the low frequency with which other mouse exons encode multiple open reading frames (Fig 1, red dots). The conservation of genomic sequence encoding multiple overlapping open reading frames over ~90 million years strongly suggests that many of these dual-coding exons play an important role in some unidentified biological process(es).

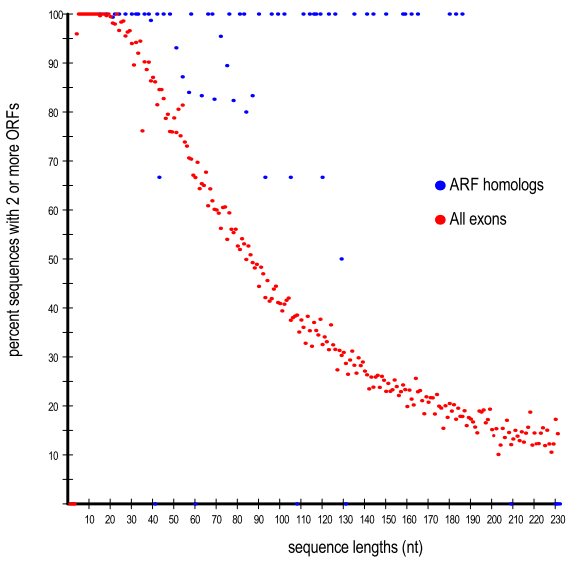


**Figure 1.** Frequency of Exons Containing Multiple Open Reading Frames, by Exon Length. Exons in mouse show a decreasing probability of containing multiple open reading frames as length increases (red dots), while nearly all mouse exons that are orthologous to human ARF exons have two open reading frames (regardless of exon length).

**Expression at the RNA level (pilot).** We confirmed the existence, in ten single-tissue RNA-Seq read sets from the NCBI Sequence Read Archive, of over 400 of the alternative variant of our dual-coding exons (we call these "Alternative Reading Frame [ARF] exons"). These were identified by seeking reads that matched a 32-nucleotide sequence that spans the splice junction on either end of the dual-coding exon (Fig 2). The left junction-spanning sequence is the concatenation of the final 16 nucleotides of the preceding exon with the first 16 nucleotides of the ARF exon; the right junction-spanning sequence follows a similar strategy. These results do not demonstrate activity relative to the standard reading frame of the dual-coding exon, merely that the ARF exon is observed as a spliced product. We have begun deeper analysis of large-scale RNA-Seq data found in the tissue-specific expression archive GTEx.

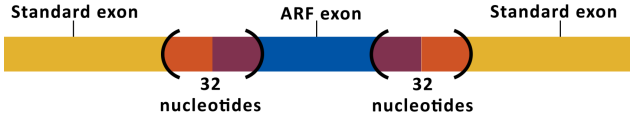


**Figure 2.** Sequence unique to the Alternative (or Standard) reading frame of a dual-coding exon can be constructed based on the sequence surrounding the splice junction.

**Presence at the protein level (pilot).** With collaborator Josh Adkins at the Pacific Northwest National Labs, we developed a pilot analysis on mass spectrometry samples from 6 tissues. The study found evidence for 296 distinct peptides from our set of dual-coding exons, of which 115 were the ARF. This surprisingly high ratio of ARF variants suggests that further MS analysis will be fruitful (note: we developed the peptide libraries, Dr. Adkins ran the MS analysis pipeline).

**Functional annotation.** We tested for gene ontology (GO) term overrepresentation among all proteins containing dual-coding exons, and found modest enrichment in genes associated with phosphorylation and DNA damage response. Significance of enrichment was only on the order of 1e-5; we anticipate that deeper analysis and clustering based on e.g. tissue and developmental state will lead to greater resolution of GO enrichment.

**Conservation.** We tested the level of conservation of exons that are dual coding and exons that are not dual coding, using UCSC Genome Browser PhyloP scores based on 100 vertebrate genomes, averaged over nucleotide positions across each exon. As Figure 3 demonstrates, dual-coding exons tend to be less conserved than single-coding exons. This agrees with the notion that exons with lower functional constraint are more likely to be able to acquire dual-coding status. Further analysis of constraint after the common ancestor with mouse (in which nearly all human dual-coding exons are also dual-coding) are forthcoming.