Apr 17th, 11:00 AM - 12:00 PM

Controlling gene function with dual-coding exons: Exploring patterns of conservation and expression

Sarah Walling
University of Montana, Missoula, sarah.walling@umontana.edu

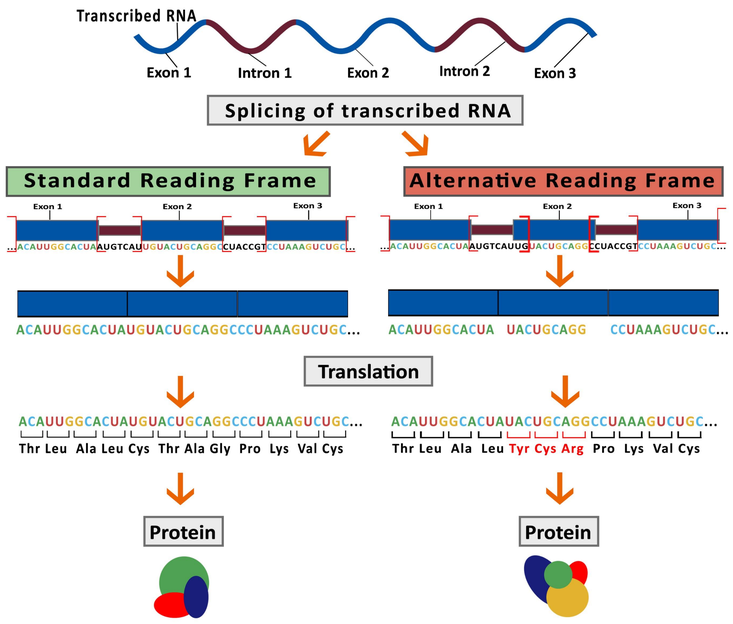
Controlling gene function with dual-coding exons: Exploring patterns of conservation and expression

Sarah Walling, Alex Nord, Kaitlin Carey, and Travis Wheeler, Dept. of Computer Science

Introduction

Gene expression is the process by which DNA is converted into RNA and then into a functional product. The RNA sequence is then read in groups of three, called codons; this division into triplets is the "reading frame" for that sequence. Genes can encode different protein products through a mechanism called alternative splicing, which typically uses different subsets of its exons. Alternative splicing impacts the proteins that are produced in different tissues or developmental timepoints; changes to an exon can cause errors in a subset of these proteins, which is the source of some genetic diseases and cancers.

We are interested in a different (and under-appreciated) form of alternative splicing, in which a single exon can encode two different protein fragments, depending on what "frame" of the exon is used. The research here investigates these surprising "dual-coding exons".



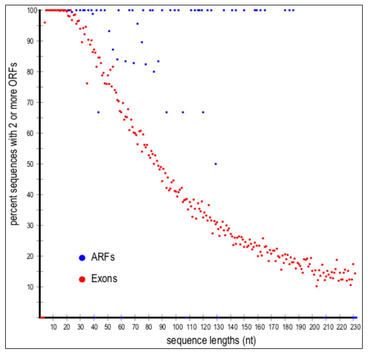
Abstract

This research seeks to further contextualize the alternative reading frame (ARF) variants of dual-coding exons by determining their tissue-specific expression level in RNA sequence data, as well as investigating the relative abundance of the variants.

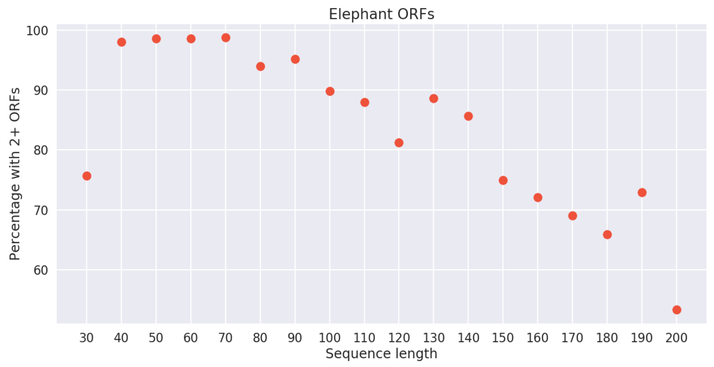
We also seek to gain insight into the role of ARF variants by evaluating their level of conservation across greater than 100 million years of evolution.

Exploring Patterns of Conservation

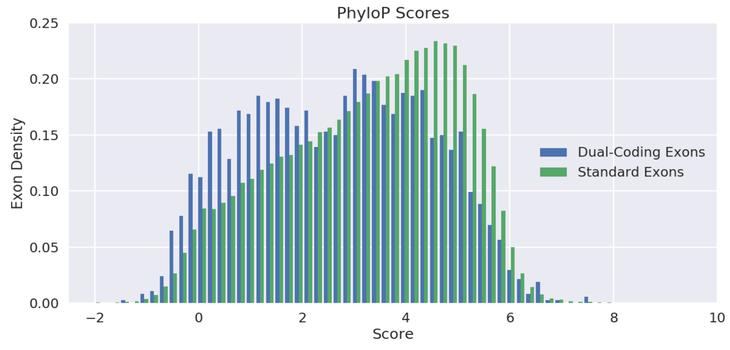
Previous research in our lab identified over 2,500 human proteins that conditionally encode alternative protein products using dual-coding exons. To confirm they are not merely the result of noisy transcription, dual-coding exons were mapped to homologous mouse exons; 97.8% were found to have two open reading frames in mouse, a shocking level of conservation suggesting an unknown functional role.



As an additional measure of conservation, we lifted the dual-coding exons over to homologous exons in elephant, megabat, and opossum which share commons ancestors with humans between 100 and 150mya. While not as conserved as mouse, these exons had a much higher proportion with at least two open reading frames (ORFs) than exons that do not show evidence of dual-coding function.

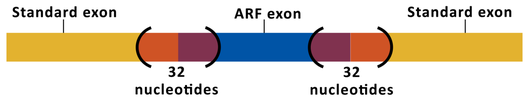


We also used two existing metrics for conservation (PhastCons and PhyloP), both based on the alignment of human to 99 different species ranging from ape to fish. We also created a new metric by manually calculating the percent identity between all homologous human and mouse sequences. All 3 metrics showed that dual-coding exons are less conserved than exons with no alternative product.

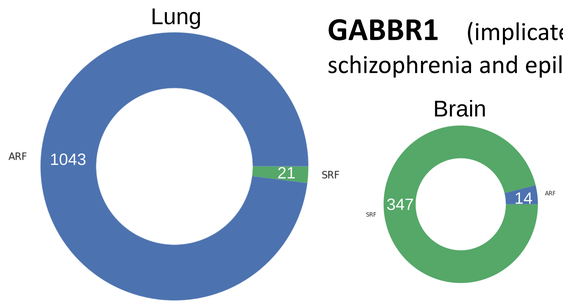


Tissue-Specific Expression Level

For each dual-coding exon, we constructed sequences to determine which reading frame of a dual-coding exon is present in a sequenced sample. These sequences bridge the junction of exons with 16 nucleotides from each side, concatenated together to make length 32 sequences. Preliminary searches in RNA sequence and mass spectrometry data confirmed that dual-coding exons do exist in humans, with both frames occurring frequently in tissue samples.



We have now used these distinguishing sequences to search against a specialized dataset with ~500GB of RNA sequences consisting of 13 distinct tissue types sampled from 714 individuals (from GTEx). We sought evidence of changes in relative expression (the ratio of expression level between variants) between tissues. With existing tools (SRA and Bowtie), we identified 13 genes which have a high number of reads and in which this ratio is greatly reversed between different tissues (one frame is >80% of all product in one tissue, and <20% in another). This suggests differential function of the alternative frames.



Future Research

There is still much to be learned about where and why dual-coding exons occur. We will build a new pipeline specific to this type of exact-match searching instead of using the very slow existing tools, which will allow analysis of many more RNA sample datasets in the future. We will continue to seek patterns in these observations as well as look at protein activity via mass spectrometry analysis, test for overrepresented sequence motifs, and identify common structural features.

Acknowledgments

Research funded by NIH, P20GM103546 (courtesy of a CBSD Junior Investigator grant)