Apr 27th, 3:00 PM - 4:00 PM

Quantifying the Presence of Alternative Reading Frames in the Human Genome

Sarah Walling
University of Montana, Missoula, Aurezalia@gmail.com

Let us know how access to this document benefits you.
Follow this and additional works at: https://scholarworks.umt.edu/umcur

https://scholarworks.umt.edu/umcur/2018/pmposters/11

This Poster is brought to you for free and open access by ScholarWorks at University of Montana. It has been accepted for inclusion in University of Montana Conference on Undergraduate Research (UMCUR) by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.
INTRODUCTION

Gene expression is the process by which DNA is converted into a functional product (generally proteins). Transcription, the first phase of gene expression, makes an RNA copy from DNA sequence. After this stage, non-coding introns are spliced out and only the coding parts of the genetic sequence, called exons, are retained. The RNA sequence is then read in groups of three, called codons, which specify either an end to the polypeptide or an amino acid; this division into triplets is the “reading frame” for that sequence. With such base-three grouping, shifting the reading frame over by one or two nucleotides causes dramatic changes in the final protein. Until recently, it was believed that human DNA almost always has only one biologically relevant (protein-coding) reading frame. Yet recently, we discovered that Alternative Reading Frame (ARF) genes appear in the human genome far more commonly than was thought possible. ARFs can encode different proteins contingent on where the reading frame is established.

PREVIOUS WORK

During previous research in our lab on multiple sequence alignment tools, we identified over 2,000 human proteins that conditionally encode ARFs. To confirm the validity of these suspected ARFs, the exons were mapped to homologous mouse exons. For nearly every ARF exon (97.8%), the mouse homolog was found to have two open reading frames, a shocking level of conservation suggesting an unknown functional role.

UNDERSTANDING ARF SEQUENCES

We seek to understand the role of these highly conserved ARFs. This will be aided by understanding which tissues, conditions, and developmental stages present ARF-variant activity. We can achieve this by searching in RNA-Seq datasets for sequences that belong to either the Alternative Reading Frame (ARFs) or the Standard Reading Frame (SRFs). These distinguishing sequences can be built by concatenating the 16 nucleotides ending the exon preceding the ARF/SRF, with the 16 at the beginning of the ARF/SRF (and similar flanking sequence near the end of the ARF/SRF).

We developed a pilot implementation of this test. We first created a database of distinguishing sequences as described above. We then downloaded RNA-Seq file SRA832889 from NCBI’s Sequence Read Archive. The Archive was converted to FASTA format, then indexed and searched using Bowtie with no mismatches allowed. 30% of ARF distinguishing sequences are identified in this single RNA-Seq analysis. This demonstrates that the ARF sequences are present, though it does not address abundance.

ARF DISTRIBUTIONS

We found that the majority of ARFs are 1 exon long, though the frameshift occasionally lasts through multiple exons. We analyzed the distribution of ARF exons within proteins, and found that a surprisingly large fraction are the final exon in the protein. Finally, we determined the distribution of ARFs across chromosomes, and found that they appear to have non-random distribution; chromosome 19 is particularly dense with ARFs.

FUTURE RESEARCH

There is still much to be learned about where and what ARFs are. We will do this search against an SRA file representing a more substantial RNA sequence dataset representing many tissues under many conditions. We will compare ARF-distinctive read counts to standard read counts, then seek to understand patterns in these observations. We will also look at protein activity via mass spectrometry analysis, test for overrepresented sequence motifs, and identify common structural features.

Acknowledgements

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