Exploring the prospect that upstream open reading frames (uORFs) produce functional peptides

April 29, 2014

Introduction

Upstream open reading frames (uORFs) are short sequences containing an initiation codon within the 5' untranslated region (5' UTR) of mRNA and a termination codon either before or after the start of the downstream coding gene. In many eukaryotic mRNAs, one or more uORFs precede the initiation codon of the downstream coding gene. It is well established that uORFs can regulate the translation of downstream gene products (Calvo *et al*., 2009; Sonenberg and Hinnebusch, 2009); however, it has yet to be determined whether many of these uORFs are actively translated into functional peptides.

A leading hypothesis for the role of uORFs is that their sequences serve as a "sponge" to attract ribosomes to the upstream region of mRNAs so that fewer ribosomes will be available to translate the downstream coding sequence of a gene. Evidence of this theory comes from studies of the *S. cerevisiae GNC4* gene, which has four uORFs with AUG start codons. Deletion of these uORF sequences results in increased translation of *GNC4* without increasing the levels of its mRNA. This result suggests that the uORF sequences are cis-acting regulatory elements that repress the process of translation (Hinnebusch, 1984; Mueller and Hinnebusch, 1986).

Studies in mammalian systems reveal that uORF sequences are present in about half of total mRNA transcripts and presence of a uORF sequence is correlated with reduced protein expression of the downstream gene. Additionally, mutations that alter uORF presence in human samples increase protein expression by 30-60%, which can lead to variation in human phenotype and disease (Calvo *et al*., 2009; Barbosa *et al*., 2013). The four properties of uORF sequences that correlate with increased repression of downstream gene translation include a strong AUG start codon, evolutionary conservation, increased distance from the cap, and multiple uORFs in the 5'UTR (Calvo *et al*., 2009).

The presence of uORFs can also influence the response to cellular stress by promoting increased expression of stress-related mRNAs (Spriggs *et al*., 2010). Some uORFs have also been shown to have positive regulatory effects that lead to increase translation of the downstream gene even in the absence of stress conditions (Mueller and Hinnebusch, 1986). It is suspected that these positively regulating uORFs provide a platform for the ribosome to initiate more efficiently than the start codon of the downstream gene (Grant and Hinnebusch, 1994). Clearly there are several different mechanisms of uORF function since each uORF does not regulate protein translation in the same way.

Recent studies by ribosome profiling have shown that ribosomes occupy over a thousand putative uORF sequences in *S. cerevisiae (*Ingolia *et al*., 2009; Ingolia, 2014) and around three thousand uORF sequences in human cells (Guo *et al*., 2010; Fritsch *et al*., 2012). However, it is still a large debate whether ribosome occupancy on uORFs leads to active translation, and if so, whether the peptide products are functional. Analysis using the ribosome release score (RRS) metric suggests that most 5' UTRs follow the model of an untranslated sequence rather than a translated one (Guttman *et al*., 2013). However, this estimation may not be valid for all uORFs and it does not provide experimental proof that uORFs are not translated into peptides. In fact, there is emerging evidence that uORF are

translated into peptides *in vivo* as several examples have been uncovered by proteomic analysis (Oyama *et al*., 2004; Menschaert *et al*., 2013; reviewed in Andrews & Rothnagel, 2014). These studies each found a small number of uORF peptides and did not attempt to distinguish them from alternative initiation of the downstream gene; thus, more advanced technology and detailed verification will expose many more uORF peptides if they exist.

In order to address whether uORFs have regulatory roles beyond seizing ribosomes from downstream coding sequence, I propose a directed search for uORFs undergoing active translation and producing functional peptides *in vivo*. I will first combine multiple computational approaches using previously published information to identify uORFs in the human reference genome that have the most protein coding potential. Next, I plan to use mass spectrometry analysis for material enriched in low abundance proteins and short peptides in order to reveal whether translation of uORFs produce detectable levels of protein. Finally, if preliminary analyses demonstrate the presence of uORF peptide products, I will verify that they exist as independent peptides and probe the function of the uORF peptides *in vivo* using mutational analysis and overexpression.

Specific Aim 1: Identify uORF sequences in the human reference genome with protein coding potential.

Most datasets for uORFs only contain sequences with the canonical AUG start codon and do not rank them based on their likelihood of encoding a protein using parameters such as similarity to coding genes, evolutionary conservation, and ribosome occupancy. Therefore, I plan to develop my own list of uORFs in the human reference genome with AUG or noncanonical start codons and score them based on different marks of protein coding potential. I will develop the original uORF list by identifying all sequences in 5' UTRs with a start codon (canonical or nonconononical) and the closest stop codon at least nine nucleotides away and either before or after the start codon of the downstream gene.

First, I will employ the program, sORFfinder (Hanada *et al*., 2012), which is optimized to find short open reading frames that have similar nucleotide composition of coding genes using a hidden markov model. In the sORFfinder algorithm, Bayes' theorem is utilized to identify the posterior probability that a given 30-300 nucleotide sequence appears in the coding region of a genome. I will match the scores from this analysis to my original list of uORFs. One limitation of this program is that it only analyses sequences between 30 to 300 nucleotides, so I will not uncover smaller or larger uORFs (even though they may exist within the genome) with this program.

Evolutionary conservation of uORF sequences is largely correlated with regulation of downstream protein expression (Calvo *et al*., 2009), so I plan to identify the uORFs that are highly conserved among eukaryotes. The program, uPEPeroni (Sharshewski *et al*., 2014), measures the conservation of sequences within the 5'UTR and calculates substitution frequency for the uORF as well. The uORF sequences that do not share conservation with other eukaryotes will not be ruled out as having protein coding

potential because they likely evolved more recently within the lineage; however, these uORFs will be scored lower than uORFs with high levels of eukaryotic conservation. Additionally, I plan to score the collection of uORFs with ribosome occupancy information using published ribosome profiling datasets (Guo *et al*., 2010; Lee *et al*., 2012). To perform this analysis, I will calculate average ribosome occupancy across all uORFs in my dataset compared to the average occupancy across the surrounding 5' UTR. Using a three codon window, any three sites that contain an average ribosome occupancy score that is one standard deviation unit above the determined 5'UTR average will be tagged as codons within putative a uORF. Sequences for the entire uORF will be defined as the nearest start codon (canonical or noncanonical) within or upstream of the three codons and the nearest stop codon downstream of the three codons. Neighboring three codon windows that are tagged as residing within the same potential uORF will be scored higher.

The scores from sORFfinder, conservation analysis, and ribosome profiling information will be added together so that the top scoring uORFs will be ranked highest for protein coding potential. The weights from each of these scores will be normalized so that the highest scoring values from each test are equal. In the end, scores will be increased for uORF sequences if they follow the known patterns of uORFs with strong regulatory roles, such as a distance greater than 50 nt from the 5' cap, a length greater than 25nt, and a strong AUG initiation codon (Calvo *et al*., 2009). I will also analyze putative uORF sequences for the presence of functional domains with the program Pfam (Punta *et al*., 2012). Positive scores will be given to putative coding elements in upstream regions containing functional domains since the peptide products of these uORFs are most likely to be functional.

Specific Aim 2: Determine if uORFs produce detectable peptides.

In order to detect the presence of uORF protein products, I will perform mass spectrometry with material enriched for uORFs and other short peptides by isolating actively translated short peptides with sequential enrichment by click chemistry and size selection chromatography. For both analyses, I will use HEK293 cells because they are less resistant to transfection and ribosome profiling of these cells has revealed ribosome occupancy on uORFs (Lee *et al*., 2012).

To analyze the presence of uORF peptides with mass spectrometry, I will take two approaches to enrich for short peptide sequences that have been translated within one hour of protein extraction. The first approach utilizes click chemistry to insert a methionine analog, azidohomoalanine (AHA), into actively translated proteins that can be selected for by covalent linkage of the AHA onto an alkyne resin. This method will enrich for proteins of low abundance so that uORF peptides that turnover more rapidly will be more abundant in a set of proteins tagged with AHA than a set of proteins without. Moreover, this approach will also exclude small peptides that are a result of degradation products from proteins synthesized more than one hour before the analysis. In order to perform this selection, I will culture cells in methionine deplete medium for 30 minutes and subsequently pulse the media with 0.1 mM $_{L}$-AHA for one hour. A subset

of cells will not be pulsed with AHA in order to serve as a negative control for the click reaction. Cells will then be lysed and the proteins that are released into solution will be incubated with copper for approximately 18 hours in order to promote the click reaction. The solution will be washed vigorously to remove any non-specifically bound proteins.

The second enrichment approach will select for small peptides using chromatography. This step is important because it will allow more peptides from uORF sequences, that are typically around 10-20 amino acids in length, to be more concentrated within the solution for mass spectrometry. To perform this analysis, I will take the denatured material from the click chemistry solution and subject it to size-exclusion chromatography that is optimized to separate large proteins from the smaller (around 1-15 kDa) uORFs and other short peptides. This experiment may be less precise than running denatured peptides on an SDS-PAGE gel; however, it is higher throughput. The resulting solution of recently translated short peptides will be trypsin digested and prepared for mass spectrometry. Results from the mass spectrometry run will be adapted to a search database to cover potential peptides across the entire genome (and especially in my initial database of uORFs) rather than only known coding gene sequences (Menschaert *et al*., 2013).

One limitation of this approach is that short peptides may not contain many spaced lysines and arginines for tandem mass spectrometry analysis. If the trypsin digest of a uORF does not cleave multiple sites with lysines and arginines producing at least one short peptide for mass spectrometry, then the uORF will not be uncovered by this method. Before performing mass spectrometry, I will analyze the dataset of uORFs and identify how many peptides would be predicted to result from trypsin digestion. Since there are about eleven tryptic digestion sites per 100 amino acids, then I would suspect many uORFs, which are 48 nucleotides or 16 amino acids in median length for human cells, to have at least one peptide for mass spectrometry resulting from trypsin digestion. Additionally, some statistical parameters will need to be relaxed for identifying mass spectrometry hits from short uORFs as compared to standard length proteins because it is much less likely for multiple different hits to be uncovered from a short uORF.

A second limitation of this mass spectrometry approach with AHA tagging is that not all uORF sequences begin with the canonical AUG start codon. If a uORF does not encode methionine at the start of its sequence or throughout, it will not be uncovered through AHA tagging and enrichment methods. Before performing mass spectrometry analysis, I will also identify the percentage of uORFs that are expected to be uncovered by mass spectrometry (based on previous analysis for locations of lysines and arginines) and do not contain any methionines. If more than 25% of predicted uORFs do not contain a methionine, I will attempt to utilize an additional method of click chemistry with the phenylalanine analog, *para*-ethynylphenylalanine (PEP), that has been shown in *E. coli* (Grammel *et al*., 2012; Grammel & Hang, 2013).

Specific Aim 3: Probe the function of uORF peptides in vivo.

If the mass spectrometry analysis uncovers peptides from uORF sequences, I will first confirm that uORF peptides are isolated from the protein in the downstream gene and not

a result of an alternative initiation site within the 5' UTR. For this analysis, I will design plasmids with a HIS-tag at the C-terminus of the downstream gene because it is furthest from the uORF (unless this site is predicted to alter function of this protein). Additionally, I will design plasmids with a HIS-tag at the C-terminus of the downstream gene and a myc-tag at the C or N-terminus of the uORF (I will likely test both combinations as it is very plausible that these tags will alter the function of uORFs). I will transfect these plasmids, in addition to HIS-vector and myc-vector controls, into HEK293 cells and extract proteins from the cell lysate to be run on SDS-PAGE gels. An antibody targeting HIS will confirm that the protein is stable with the addition of the HIS-tag. This western blot will also reveal the size of the downstream gene and whether it is the expected size or possibly contains the addition of its uORF. An antibody targeting myc will be utilized to display the size of the protein containing the uORF sequence. If the anti-myc band is very small, then it suggests that uORF peptides are separate from the protein product of the downstream gene; however, if the anti-myc band is the same size as the anti-HIS band, then it is assumed that this gene has an alternative initiate site and a larger protein isoform containing a uORF. For each of the analyses with myc-tagged uORFs, I will confirm that the HIS-tagged downstream gene has similar expression to the samples without myc-tagged uORFs because this will reveal whether the tag on the uORF is altering its function.

Next, if the uORF peptides prove to be isolated from the downstream gene, I will assess their function by mutating the uORF peptide sequence on the plasmid with myc-tagged uORF and transfecting into HEK293 cells to observe changes in protein expression of the downstream gene by western blot against the HIS-tag. Most previous studies have attempted to assess uORF peptide function by altering the start codon or truncating the peptide by inserting a premature stop codon; however, these methods do not distinguish altered peptide function from altered regulation by ribosome occupancy on the uORF. Therefore, I intend to mutate nucleotides in the middle of the uORF to create nonsynonymous changes that will alter the peptide product. If the changes in protein sequence affect the translation levels of the downstream gene, then it suggests that the peptide product of the uORF is likely regulating the protein expression of the downstream gene.

Finally, if the mutagenesis experiments reveal that the uORF peptide sequence is important for regulating the downstream gene, I will make additional plasmids with only the uORF-myc sequence. By overexpressing the uORF peptide using this system, I will be able to show whether increased levels of the uORF peptide affect translation of the target gene (the gene that is downstream of the original uORF). Since most uORFs lead to repression of downstream genes, I expect that overexpression of the uORF peptide will lead to decreased translation in the target gene.

References

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

Barbosa, C., Peixeiro, I. & Romão, L. Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* **9,** e1003529 (2013).

Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are. **106,** (2009).

Fritsch, C. *et al*. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22,** 2208–18 (2012).

Grammel, M. & Hang, H. C. Chemical reporters for biological discovery. *Nat. Chem. Biol.* **9,** 475–84 (2013).

Grammel, M., Dossa, P. D., Taylor-Salmon, E. & Hang, H. C. Cell-selective labeling of bacterial proteomes with an orthogonal phenylalanine amino acid reporter. *Chem. Commun. (Camb).* **48,** 1473–4 (2012).

Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466,** 835–40 (2010).

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–51 (2013).

Hanada, K. *et al*. sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26,** 399–400 (2010).

Hinnebusch, A. G. Evidence for translational regulation of the activator of general amino, acid control in yeast. **81,** 6442–6446 (1984).

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–23 (2009).

Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15,** 205–13 (2014).

Lee, S. *et al*. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2424–32 (2012).

Menschaert, G. *et al*. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of

alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics* **12,** 1780–90 (2013).

Mueller, P. P. & Hinnebusch, a G. Multiple upstream AUG codons mediate translational control of GCN4. *Cell* **45,** 201–7 (1986).

Oyama, M. *et al*. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res*. **14,** 2048–52 (2004).

Punta, M. *et al*. The Pfam protein families database. *Nucleic Acids Res*. **40,** D290–301 (2012).

Skarshewski, A. *et al*. uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics* **15,** 36 (2014).

Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136,** 731–45 (2009).

Spriggs, K. a, Bushell, M. & Willis, A. E. Translational regulation of gene expression during conditions of cell stress. *Mol. Cell* **40,** 228–37 (2010).