Non-Canonical Translation in Vertebrates

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Chew, Guo-Liang. 2015. Non-Canonical Translation in Vertebrates. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467487">http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467487</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Non-canonical translation in vertebrates

A dissertation presented by

Guo-Liang Chew

to

The Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biochemistry

Harvard University

Cambridge, Massachusetts

April 2015
Non-canonical translation in vertebrates

Abstract

Translation is a key process during gene expression: to produce proteins, ribosomes translate the coding sequences of mRNAs. However, vertebrate genomes contain more translation potential than these annotated coding sequences: translation has been detected in many non-coding RNAs and in the non-coding regions of mRNAs. To understand the role of such translation in vertebrates, I investigated: 1) the distribution of translation in vertebrate long non-coding RNAs, and 2) the effects of translation in the 5’ leaders of vertebrate mRNAs.

To quantify and localize translation in a genome-wide manner, we produced and analyzed ribosome profiling data in zebrafish, and analyzed ribosome profiling data produced by others. The nucleotide resolution afforded by ribosome profiling allows localization of translation to individual ORFs within a transcript, while its quantitative nature enables measurement of how much translation occurs within individual ORFs.

We combined ribosome profiling with a machine-learning approach to classify lncRNAs during zebrafish development and in mouse ES cells. We found that dozens of proposed lncRNAs are protein-coding contaminants and that many lncRNAs have ribosome profiles that resemble that of the 5’ leaders of coding mRNAs. These results clarify the annotation of lncRNAs and suggest a potential role for translation in lncRNA regulation.

Because much of the translation in non-coding regions of mRNAs occurs within uORFs, we further examined the effects of their translation on the cognate gene expression. While much is known about the repression of individual genes by their uORFs, how uORF repressiveness varies within a genome and what underlies this variation had not been characterized. To address these questions, we analyzed transcript sequences and ribosome profiling data from human, mouse and zebrafish.
Linear modeling revealed that sequence features at both uORFs and coding sequences contribute similarly and substantially toward modulating uORF repressiveness and coding sequence translational efficiency. Strikingly, uORF sequence features are conserved in mammals, and mediate the conservation of uORF repressiveness in vertebrates. uORFs are depleted near coding sequences and have initiation contexts that diminish their translation. These observations suggest that the prevalence of vertebrate uORFs may be explained by their functional conservation as weak repressors of coding sequence translation.
# Table of Contents

Acknowledgements .......................................................................................................................................................... vi

**Chapter 1**  
Post-transcriptional gene regulation: of translation, by translation ................................................................. 1

**Chapter 2**  
Translation of many long non-coding RNAs resemble translation at 5’ leaders ................................................... 17

**Chapter 3**  
Sequence features mediate the conservation of upstream open reading frame repressiveness ......................... 29

**Chapter 4**  
Conclusion and future directions .......................................................................................................................... 41

**Appendix**  
Supplemental materials for Chapter 2.................................................................................................................... 51  
Supplemental materials for Chapter 3 .................................................................................................................... 64

**References** ................................................................................................................................................................. 81
ACKNOWLEDGEMENTS

My time in graduate school has been overwhelmingly positive. It has been a wonderful time of learning and growth: scientifically, professionally, as well as personally. My work would not have been possible without some people, and a lot less pleasant if not for some others. I would like to take this space to acknowledge those who have played a part.

First of all, I would not have embarked on a PhD if not for a great undergraduate research experience in Hidde Ploegh’s lab at the Whitehead Institute / MIT. I cut my teeth doing mass spectrometry, biochemistry and chemical biology with John Antos as an awesome postdoc mentor; this gave me the confidence to continue my pursuit of science in graduate school.

I have had a number of faculty serving on my qualifying exam and dissertation advisory committees over the past few years. Many thanks to Andy McMahon (served as chair; I also rotated in his lab), Erin O’Shea and Bill Gelbart as members of my qualifying exam committee; and to Sharad Ramanathan as a past member of my dissertation advisory committee. I am also deeply appreciative of Andrew Murray (served as chair; I also had a fun time teaching MCB291 with him and Cassandra Extavour) and Craig Hunter who have been on my dissertation advisory committee throughout, and who form my thesis defense committee along with Rich Losick (who graciously agreed to take the place of Sharad).

The MCB department has been a nurturing place for me to learn my way around research as a graduate student. I have had fruitful interactions with many of the faculty here, including Venki Murthy (my neutral advisor), Catherine Dulac (chair of the department when I first started, and director of graduate studies), Susan Mango (who I did a rotation with), Briana Burton, Vlad Denic, Andres Leschziner, and many others. John Calarco did a stellar job of organizing the RNA data club, bringing together people doing anything related to RNA in the Harvard Cambridge area; I have had many
engaging discussions with many data club attendees, including Brian Zid, Rasi Subramaniam, Alicia Darnell, Adam Norris and Dave Shechner.

Life at MCB is not all science and research; I would like to thank all the administrators and staff of the MCB department and the MCO program for making life here a whole lot smoother. In particular: Mike Lawrence, our caring and dedicated MCO program administrator, assisted by Patty Perez; me and my inbox thank both of you. Debbie Maddalena patiently helped me navigate the maze that is MCB finance. Michelle Cicerano organized the many scientific seminars (and post-seminar lunches) that I so diligently attended. Renate Hellmiss and Jack Conlin attended to many of my graphics and AV needs respectively.

My fellow students in the MCO program, especially my classmates, have been a key part of my community here at Harvard. Through classes, volleyball, data clubs, and just hanging out, I’m glad to have made friends with this diverse bunch of people. Special shout-outs to Nichole Collins (who I TFed MCB291 with), Sara Leiman (who I organized the MCO retreat with), Nick Weir, Ezgi Hacisuleyman, Fred Chang, Iris Odstrcil, Phil Shiu and Eddie Wang (fellow volleyballers): may our paths continue to cross in the future.

I have also benefitted from an open and friendly scientific community that exists outside of Harvard. The translation field is a welcoming one, and I have had the opportunity to consult with the pioneers of ribosome profiling (Jonathan Weissman and Nick Ingolia), as well as have in-depth discussions with fellow practitioners (Sebastian Leidel, Pasha Baranov, Huili Guo, Danny Nedialkova, Ivaylo Ivanov, Eric Mills and Joshua Dunn). I would also like to specifically acknowledge Rasi and David Koppstein for introducing me to iPython notebook and pandas (respectively), and thus revolutionizing the way I code and analyze data.

The Schier lab has been an ideal place to do science. Although our interests are diverse, there is no lack of cooperation and collaboration. The postdocs have been especially helpful in mentoring and
guiding us graduate students, while Steve Zimmerman, our fearless lab manager, ensures everything in
the lab and the fish facility runs smoothly (often going that extra mile in the snow to do so). It has also
been a lot of fun playing on the lab’s volleyball team (No Arms No Chocolate) captained successively by
Julien Dubrulle and James Gagnon, with players including Steve, Nadine Vastenhoew, Farhad Imam,
Albert Pan, Peng Huang, Patrick Muller, Andi Pauli, Owen Randlett, Katherine Rogers, Tessa Montague
and Meg Norris. Other members of the lab that I have overlapped with include Jason Rihel, Ian Woods,
Jen Li, Drew Robson, David Schoppik, Martin Haesemeyer, Florian Merkle, Constance Richter, Summer
Thyme, Jeff Farrell, Michal Rabani, Mehdi Goudarzi, Nate Lord, Laila Akhmetova, Alix Lacoste and Shristi
Pandey.

I have had the privilege of working very closely with two amazing postdocs, Eivind Valen and
Andi Pauli, for my thesis projects. Their pioneering work annotating lncRNAs in zebrafish was
foundational for our subsequent analysis of their translation: Eivind showed me the ropes during my
early days of learning bioinformatics analysis, while Andi guided me through developing proper
experimental controls and workflows. We had many good times in the tea room discussing all aspects of
science over countless cups of espresso. I was fortunate to be the first of many graduate students (and
even more rotation students) that Andi mentored in the Schier lab; Andi has been extremely patient and
generous in her time and energy, and I can only hope that her future graduate students in her own lab
will be easier to handle 😊.

In addition to cultivating a great lab to work in, Alex has shaped how I think and do science.
From our interactions, I have developed a better sense of what are important scientific questions and
key experiments / analyses. Most importantly, Alex has been a mentor who gives a shit: he has cared
about all aspects of my life in graduate school, from project and scientific aspects, to career
development, to personal life. Should I ever have the opportunity to helm my own research group, the
Schier lab will be what I would model it on.
Lastly, my family has brought me great joy during my time in graduate school. Shirleen, who went from being girlfriend to fiancé to wife (and co-parent) over this period, has been a constant companion and chief influence in our joint journey through life and science; I look forward to the road ahead together. We picked up a little (and growing) passenger along the way: Linnea, with the life constraints that a young child places on graduate student parents, is the reason why I became an after-hours bioinformatician. Raising our daughter during this time has been a source of immense shared happiness, and I hope that she will read and understand this (and Shirleen’s) thesis one day.
CHAPTER 1
Post-transcriptional gene regulation: of translation, by translation

Translation produces polypeptides from messenger RNAs (mRNAs), thus converting the nucleotide-based language of the genetic code to that of amino acids. During translation, ribosomes engage and traverse segments of mRNAs in triplet codons, using each codon as a cue for stringing together the chains of individual amino acids that comprise proteins. While much of translation is intended for the functional production of proteins from the genes that encode them, the process of translation itself can have other consequences for gene expression. Moreover, recent work surveying where translation occurs in the genome suggests that translation outside of annotated coding regions is widespread.

This chapter introduces the many facilitators of non-canonical translation in vertebrates, beginning with more general features of eukaryotic translation and vertebrate transcriptomes, highlighting known processes and mechanisms, and summarizing the recent developments in the field that have enabled my work.

Eukaryotic translation offers opportunities for regulation

The core mechanisms of translation are similar and largely conserved between prokaryotes and eukaryotes (Rodnina and Wintermeyer 2009). However, from a gene regulatory perspective, there are a number of critical features of eukaryotic translation that distinguish it from prokaryotic translation.

Firstly, eukaryotic translation is spatially and temporally separate from transcription, occurring after mRNA processing and export from the nucleus to the cytoplasm; in prokaryotes, due to the lack of a nucleus, translation may occur concurrently with transcription (Miller et al. 1970).
Figure 1.1. Increased complexity of vertebrate transcripts presents more points of regulation
From prokaryotes to “lower” eukaryotes to vertebrates, transcripts show increasing complexity, with more mRNA processing and longer non-coding regions. Thick bars represent coding sequences, diamonds represent 5’ caps, and kinked lines indicate splice junctions.

In addition, to initiate translation, the 43S pre-initiation complex (comprising the 40S ribosomal small subunit, initiation factors and the charged methionine initiator tRNA) typically scans the mRNA from the capped 5’ end until it encounters a start (AUG) codon in good context (Kozak 2002); in prokaryotes, the ribosome initiates translation directly at start codons downstream of Shine-Dalgarno sequences (Dahlberg 1989).

Moreover, eukaryotic mRNAs typically produce only one species of functional protein (monocistronic), while prokaryotic mRNAs may produce a number of different proteins (polycistronic) that often function in concert with one another as operons (Kozak 1983).

These features of eukaryotic translation allow for elaborate mechanisms of post-transcriptional modulation and regulation of gene expression. The 5’ caps and poly-adenylation (poly-A) tails of processed mRNAs are necessary for typical eukaryotic translation (Muthukrishnan et al. 1975; Furuichi et al. 1975), and are often the targets of processes that regulate mRNA stability and translation (Mendez and Richter 2001; Weill et al. 2012; Groppo and Richter 2009; Beelman and Parker 1995). The requirement to scan the 5’ leaders (also known as 5’ untranslated regions or UTRs) of eukaryotic mRNAs makes them a site of translational regulation as well (Gebauer and Hentze 2004). There is further cross-talk between these mechanisms, where elements in 5’ leaders and 3’ trailers (also known as 3’ UTRS)
may direct modifications to mRNA 5’ caps and the poly-A tails (Meijer et al. 2013; Thompson et al. 2007).

**Longer non-coding regions in vertebrate mRNAs amplify their regulatory potential**

Much of what we understand about eukaryotic translation comes from studies in simple unicellular eukaryotes such as yeast. The form of vertebrate mRNAs is similar to yeast mRNAs – they are spliced, have 5’ leaders and 3’ trailers, and are 5’ capped and 3’ poly-adenylated (Figure 1.1). However, the exaggeration of such features in vertebrates mRNAs amplify the potential for post-transcriptional and translational regulation. While the 5’ leaders and 3’ trailer regions of yeast mRNAs are relatively short (with mean lengths of 96 and 147 nucleotides respectively; Lin & Li, 2012), those of vertebrate mRNAs are substantially longer (with mean lengths of ~160 and ~450 respectively; Mignone, Gissi, Liuni, & Pesole, 2002); note that within species, the lengths of 5’ leaders and 3’ trailers vary considerably more than between species. In addition, vertebrate mRNAs are extensively spliced, with many examples of alternative splicing, often in a tissue-specific manner (Pan et al. 2008; Barbosa-Morais et al. 2012; Merkin et al. 2012).

The increased target size of 5’ leaders and 3’ trailers in vertebrate mRNAs allows for greater variation and combinations in the mechanisms of regulation. Indeed, there are many well-studied RNA cis regulatory elements on 5’ leaders and 3’ trailers, including secondary structure elements, protein binding sites, internal ribosome entry sites (IRESes) and upstream open reading frames (uORFs) (Gebauer and Hentze 2004; Gebauer et al. 2012); the splice junctions themselves, often within the coding region, are also important sites for regulation of transcripts. Trans-acting factors such as RNA-binding protein complexes and small complementary RNAs interact with these cis regulatory elements to bring about regulation of transcripts through modulating translation, destabilizing the transcript, and/or localizing the transcript to a different subcellular compartment.
Figure 1.2. **uORF translation inhibits downstream coding sequence (CDS) translation**

uORFs (red) in the 5’ leader capture scanning 40S ribosomes (a) to initiate translation together with the 60S ribosome. The resultant 80S ribosome translates the uORF (b). During translation termination (c), the 80S ribosome dissociates from the transcript. Thus, uORFs can prevent scanning 40S ribosomes from reaching downstream coding sequences (CDSes), hence inhibiting downstream translation.

I draw specific attention to uORFs because through their direct role in modulating translation of transcripts, they themselves are translated. With longer 5’ leaders in vertebrates and a constant genetic code, the probability of uORFs occurring in vertebrate transcripts is significantly higher than that in simpler eukaryotes. uORFs are generally inhibitory (Figure 1.2): scanning 43S pre-initiation complexes encounter uORFs before their coding sequences (CDSes), and may translate these uORFs at the expense of downstream coding sequences (Kozak 2002; Calvo et al. 2009).

Unsurprisingly, there is generally selection against the presence of uORFs in 5’ leaders, including in vertebrates (Neafsey and Galagan 2007a; Iacono et al. 2005). Even as selection pressure against the presence of uORFs suppresses their frequency to below that expected by random chance, the longer 5’ leaders of vertebrates remain a larger mutational target for the *de novo* creation of uORFs over evolutionary time and therefore still contain more uORFs than in lower eukaryotes.

*Protein-coding translation in vertebrates is under-annotated*

The exact number of genes that encode functional proteins in vertebrates remains elusive. Even in the deeply studied vertebrate genomes of human, mouse and zebrafish, we continue to find new functional proteins (Anderson et al. 2015; Pauli et al. 2014). Table 1.1 below summarizes the number of
protein-coding in these vertebrate genomes, as well as how much of the genome they occupy (according to the gene annotations from Ensembl Release 78).

<table>
<thead>
<tr>
<th>Organism</th>
<th>No. of protein-coding genes</th>
<th>Protein-coding sequence</th>
<th>Protein-coding transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total length (nt)</td>
<td>% of genome</td>
<td>Spliced length (nt)</td>
</tr>
<tr>
<td>Human</td>
<td>21,796</td>
<td>37,049,509</td>
<td>1.18%</td>
</tr>
<tr>
<td>Mouse</td>
<td>22,154</td>
<td>35,457,944</td>
<td>1.27%</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>26,448</td>
<td>42,463,141</td>
<td>2.96%</td>
</tr>
</tbody>
</table>

Table 1.1 Coding vs. non-coding regions of vertebrate genomes

As summarized, the annotated protein-coding sequences and the transcripts that contain them make up only a small proportion (1-4%) of vertebrate genomes; significant fractions of vertebrate genomes are intronic. There may be many more functional protein-coding sequences in the genome: large-scale annotations of protein-coding genes have previously relied on the combination of homology-based methods and arbitrary protein length cut-offs (usually around 100 amino acids). As a result, there has been a bias towards discovering long and conserved proteins, while potentially missing out on short and/or lineage-specific peptides (Pauli et al. 2015).

High-throughput genome-wide methods such as ribosome profiling and mass spectrometry have found many translated regions in the genome outside of annotated coding regions (Bánfai et al. 2012; Ingolia et al. 2011; Slavoff et al. 2013a; Wilhelm et al. 2014; Kim et al. 2014; Fritsch et al. 2012a), such as within the 5′ leaders and 3′ trailers of coding transcripts, as well as within long non-coding RNAs (lncRNAs). A handful of cherry-picked translated regions have been characterized and found to encode small functional proteins (Pauli et al. 2014; Kondo et al. 2010; Petersen et al. 2011); many of these novel protein-coding genes are indeed conserved between species (this conservation the likely reason that follow-up characterization was done in the first place), but were too short to be confidently annotated as protein-coding by conservation alone. While it is likely that many more functional proteins will be found and characterized from these unannotated translated regions, much of this additional translation
is probably a combination of biological noise, and translation for surveillance and regulatory purposes; these will be outlined in the following sections.

*mRNA surveillance mechanisms involve their translation*

Ribosomes often encounter erroneous coding transcripts originating from errors in transcription, processing, as well as partially-degraded transcripts. Because translating such aberrant transcripts has the potential to produce dysfunctional proteins with deleterious consequences, various mRNA surveillance mechanisms exist in all organisms to detect and destroy such aberrant transcripts, as well as to rescue and recycle the translational machinery from these transcripts.

If the transcript is structurally intact (i.e. being capped and poly-adenylated), the cell has little means of determining whether it contains errors without translating it. Therefore, many mRNA surveillance mechanisms involve the translating ribosome (Shoemaker and Green 2012), detecting a combination of where a ribosome is on the transcript, whether a ribosome is stalled, and whether it is in proximity with other mRNA-bound complexes.

![Diagram of mRNA surveillance mechanisms](image)

**Figure 1.3. Canonical targets of mRNA surveillance mechanisms**

Various mechanisms of mRNA surveillance identify erroneous coding transcripts by stalled ribosomes in different contexts.

- **Non-stop decay (NSD)** targets are transcripts lacking a stop codon, and are identified by a ribosome stalled at the 3’ end of a transcript.
- **No-go decay (NGD)** targets are stalled ribosomes in general, and include ribosomes that translate mis-polyadenylated transcripts (stalled by poly-lysine tracts in the peptide exit tunnel).
- **Nonsense-mediated decay (NMD)** targets are ribosomes terminating at premature termination codons (PTCs) far from the 3’ end of a transcript and often upstream of a splice junction.
There are three well-studied modes of mRNA surveillance, characterized by the classes of aberrant transcripts they detect through specific locations and states of stalled ribosomes: non-stop decay (NSD), no-go decay (NGD), and nonsense-mediated decay (NMD) (Figure 1.3). They have been extensively reviewed (Shoemaker and Green 2012; Wilusz et al. 2001; Doma and Parker 2007; Behm-Ansmant and Izaurralde 2006; Chang et al. 2007), but I will summarize relevant points below.

Mechanisms of NSD exist in all kingdoms of life. Truncated coding transcripts (that end in the middle of their CDS before their original stop codon) present the same problem to prokaryotes and eukaryotes: translation may initiate on the transcript, but cannot terminate without a stop codon, resulting in a stalled 80S ribosome at the 3’ end of the transcript. Many prokaryotes handle 3’-stalled translating ribosomes (70S in this case) through the tmRNA system (Hayes and Keiler 2010; Janssen and Hayes 2012; Moore and Sauer 2007), where tmRNA binds the empty A-site of stalled ribosome as a tRNA, templates the addition of an ssrA tag to the nascent peptide, then terminates translation; the tagged peptide is now targeted for proteolysis. No homologous tmRNA system has been found in the nuclear genome of eukaryotes (they are occasionally found in mitochondria); instead, 3’-stalled eukaryotic ribosomes may recruit either Ski7 (some yeasts; Frischmeyer et al., 2002; van Hoof, Frischmeyer, Dietz, & Parker, 2002), or the Dom34/Hbs1 complex (other yeasts/eukaryotes), with either Ski7 or Dom34 binding within the ribosome’s A-site to induce ribosome recycling (Tsuboi et al. 2012).

NGD targets are also stalled ribosomes; the difference here is that the ribosome in question is stalled mid-transcript, either by impediments to ribosomal translocation on the transcript itself (e.g. stable RNA secondary structure), or impediments to the peptide in the exit channel (e.g. specific peptide sequences, antibiotics). Molecularly, NGD targets are distinguished from NSD targets by potentially having a tRNA in the A-site, making their detection slightly different. Dom34/Hbs1 were actually first implicated in NGD (Doma and Parker 2006) thus blurring the lines between NGD and NSD; in fact, the endonuclease cleavage of transcripts during NGD results in NSD targets.
Given that start and stop codons are not uncommon in sequences of RNA, the above mechanisms of mRNA surveillance would not be typically deployed on intact transcripts. However, mis-spliced or mis-transcribed transcripts may still be correctly capped and poly-adenylated, allowing their export to the cytoplasm where they could be translated. These transcripts likely contain unintended ORFs, and their detection is similarly dependent on the translation machinery. In such a situation, the NMD machinery (comprising UPF1/2/3; UPF1 being the most critical) detects terminating ribosomes that are upstream of exon junction complexes (EJCs) (Popp and Maquat 2013; Hurt et al. 2013; Kashima et al. 2006). The presence of downstream EJCs is an indirect positional cue: the vast majority of coding sequences end in the last exon of a transcript, therefore, an EJC downstream of a terminating ribosome indicates that the translating ribosome did not reach the last exon, and likely terminated at a premature stop codon, thus indicating that the transcript should be targeted for degradation.

Classifying translation by its outcome

As highlighted in the previous section, ribosomes can serve two roles: they can be the site of protein synthesis, as well as the focus of RNA surveillance mechanisms. Hence, from the ribosome’s perspective, we can more formally define what is meant by canonical vs non-canonical translation:

**Canonical translation** – where the functional outcome of a ribosome’s transit over an RNA transcript is a peptide.

**Non-canonical translation** – where the functional outcome of a ribosome’s transit over an RNA transcript is a regulatory event.

This distinction between canonical and non-canonical translation can be relatively sharp; in all three modes of mRNA surveillance where translation results in transcript degradation, the defective nascent peptides are often also targeted for degradation. In addition, these definitions allows for a third classification:
**Background translation** – where the end product of a ribosome’s transit over an RNA transcript is largely inconsequential to the cell.

All translation necessarily produces peptides as products, hence it is difficult to distinguish the three classes of translation from each other merely by the presence of detectable peptides (Bánfai et al. 2012). Instead, where the functionality of the peptide product is suspect (due to low stability, lack of conservation), the regulatory consequences for translation should be considered.

In intact, full-length transcripts, one can envision a scenario where NGD and NSD mRNA surveillance mechanisms acting within coding sequences have an endogenous role in regulating expression. Because of its constant need for charged aminoacyl-tRNAs and ATP/GTP, translation is tightly coupled to nutrient and energy availability. Hence the elongating ribosome can readily act as a sensor, with pausing along the length of the coding sequence as a signal of starvation, to be read out by NGD/NSD mechanisms. While such internal pauses and stalls have been historically difficult to study *in vivo*, new technologies and analytical methods can now detect them genome-wide, finding position-specific ribosome stalling in response to stress, codon usage and sequence context (Subramaniam et al. 2013; Liu et al. 2013; Shalgi et al. 2013; Li et al. 2012).

There are documented circumstances in which translation is both regulatory and produces functional protein; this regulation usually acts directly on the production of the translated protein. Genes regulating circadian rhythm have been found have suboptimal (for translation) codon biases necessary for their function (Zhou et al. 2013; Xu et al. 2013), while certain serine codons are especially sensitive to serine levels in a gene controlling biofilm formation, hence controlling its translation (Subramaniam et al. 2013). However, much of what we already know about non-canonical translation occurs outside of coding regions.
Non-canonical translation outside coding regions of mRNAs

Non-canonical translation has been most studied in 5’ leaders of transcripts, although thus far, the focus has been on the characterization of specific examples. The trp attenuator in bacteria (Yanofsky 1981; Zurawski et al. 1978; Lee and Yanofsky 1977) is an interesting example where co-transcriptional translation of a leader peptide is sensitive to tryptophan levels because it contains tryptophan codons; successful translation of the leader peptide promotes formation of a stem loop in the mRNA that terminates transcription. In eukaryotes, uORFs within 5’ leaders may also sense the state of the metabolic state of the cell: the translation of GCN4 in yeast (Hinnebusch 2005) and ATF4 in vertebrates (Jackson et al. 2010) are well-studied examples where their multiple uORFs sense cellular stress levels through the interaction of eIF2α phosphorylation and reinitiation. Here, reduced eIF2α phosphorylation during stress decreases the efficiency of reinitiation after translation of the first uORF, resulting in the downstream translation of GCN4 or ATF4 itself instead of another uORF.

In addition to being repressive of downstream translation, uORFs are also NMD targets: their presence are associated with a reduction in transcript levels (Calvo et al. 2009), while their translation is predictive of their regulation by NMD (Hurt et al. 2013). Given the prevalence of uORFs in vertebrate genomes, NMD could act constitutively on a large subset of transcripts, and may be an important part of regular gene expression. While I explore the genome-wide translational-repression aspect of uORFs in Chapter 3 of this thesis, NMD’s effects on genome-wide transcript levels will also warrant further study.

Not all RNAs are coding (in fact, the bulk of a cell’s RNA is ribosomal, which is decoding rather than coding). Conversely, not all non-coding RNAs are not translated; given the minimal requirements to be translated comprising a 5’ cap, a poly-A tail (in eukaryotes) and an ORF, we should actually expect that a large number of longer non-coding RNAs will inadvertently be translated. Indeed, we can detect some peptide products emanating from lncRNAs (Bánfai et al. 2012; Slavoff et al. 2013a; Wilhelm et al. 2014; Kim et al. 2014); however we still cannot tell if this is non-canonical translation, or background
translation. Against the backdrop of existing mRNA surveillance mechanisms, it should not be surprising if some or all of them are similarly co-opted in the post-transcriptional processing of lncRNAs for the lncRNAs to get to their final functional state. In fact, one of the most abundant lncRNAs is heavily processed and shows evidence of translation in its 5’ end (Wilusz et al. 2012), however, whether its translation is necessarily for processing and function is still unknown. Translation may additionally be involved in the biogenesis of small non-coding RNAs such as snoRNAs (Tani et al. 2013) and piRNAs.

Ribosome profiling quantifies translation genome-wide at nucleotide resolution

At around the time work began for this thesis, ribosome profiling had just been developed in yeast (Ingolia et al. 2009), and its application was being extended to other model systems (Guo et al. 2010; Brar et al. 2012; Bazzini et al. 2012). Ribosome profiling enables quantification of translation genome-wide at a resolution that can discern which ORFs were being translated. The method is well-documented (Ingolia et al. 2012) and even commercialized, and some of the biological findings emerging from its application (including that contained in Chapter 2 of this thesis) have been reviewed (Ingolia 2014; Jackson and Standart 2015). I will briefly highlight key features of the technique, as well as its application to the discovery and functional analysis of non-canonical translation.

Figure 1.4. Summary of ribosome profiling protocol
Ribosomes are stalled on mRNAs by translational inhibitors, then digested by nuclease. Protected fragments are isolated and made into Illumina-compatible sequencing libraries. Reads are mapped to the genome and reduced to single-nucleotide positions representing P- or A-site occupancy by the ribosomes. A representative ribosome profile over a highly-expressed transcript (zebrafish EF1a; Chew et al. 2013) is depicted above the transcript’s structure, showing coding sequences in green, a translated uORF in salmon, and splice junctions with kinked lines.
Essentially, ribosome profiling is the high-throughput sequencing of RNA fragments that had been protected from nuclease digestion by stalled ribosomes (Figure 1.4). By treating biological samples with various translational inhibitors, one can stall translating ribosomes on RNA – these stalled ribosomes protect the contained RNA segments from subsequent treatment by nucleases. The resultant RNA fragments can be isolated, sequenced, and mapped to the genome to determine which RNAs were being translated, and where ribosomes were translating within each RNA.

Because the ribosome moves in codon steps (i.e. 3 nucleotides at a time) when translating RNA to peptide, under certain experimental parameters, one can derive ribosome profiles that depict this triplet phasing over translated ORFs (Ingolia et al. 2009; Lareau et al. 2014). This suggests that ribosome profiling data is at least at near-nucleotide resolution, allowing us to reduce the mapped read data to single nucleotide positions corresponding to peptidyl- or aminoacyl-tRNA sites (P- or A- sites respectively), and to quantify the translation over individual ORFs.

_Ribosome profiling has high(er) sensitivity and low(er) biases than other genome-wide methods_

Other methods exist for the quantification of translation genome-wide; however, they lack ribosome profiling’s sensitivity and dynamic range in quantifying translation, and introduce other biases that can confound downstream analyses. I will highlight polysome profiling and mass spectrometry as the most frequently used methods for quantifying translation.

In polysome profiling, mRNA from the polysome fraction(s) of polysome profiles are quantified (by microarrays, or by RNA-seq) to determine the mRNAs that were polysome-associated and thus efficiently translated. Because mRNAs are either present or absent in these polysome fractions, there is limited dynamic range and resolution by which to assess individual mRNA species for how efficiently they are translated. In addition, polysome profiling neither accounts for differences in ORF lengths, nor for the number of ORFs in a transcript. Thus, an mRNA may be more frequently found in the polysome...
fractions simply because it has a longer coding sequence to contain more translating ribosomes, or has ribosomes on multiple ORFs (e.g. uORFs).

In mass spectrometry, peptides from the enzymatic digestion of proteins are detected by their mass. Length biases exist in mass spectrometry as well, but for different reasons: not all peptide fragments can be detected by mass spectrometry, and even detected peptides may not be detected equally well. Hence, our ability to quantify translation of gene products becomes increasingly stochastic the shorter these fragments are, and is practically impossible below ~20 amino acids.

These methods do have their valid uses for relative measurements of translation when translation of the same species (if it can be detected) is compared between different treatments - the biases become accounted for internally. However, for absolute measures of translation (i.e. how much each gene in a sample is translated), ribosome profiling may present fewer biases. Ribosome profiling is sensitive enough to detect translation of even the most minimal ORF (AUG-Stop; Chew et al., 2013) with coverage over entire translated ORFs, thus translation can be quantified over the relevant ORFs and sequencing biases are averaged out.

Ribosome profiling is not without its own biases and artifacts. Treatment by different translational inhibitors, as well as varying methods of treatment (e.g. pre- vs post-lysis) introduce different 5’ and 3’ biases (Ingolia et al. 2011; Gerashchenko and Gladyshev 2015); such biases are known to the community and can be corrected for, usually by truncation of ORFs during quantification of translational efficiency. In addition, because certain ribosome profiling protocols rely on an ultracentrifugation step to pellet ribosomes through a sucrose cushion, other RNA fragments (including snoRNAs and microRNAs; Chew et al., 2013) co-sediment with ribosome protected fragments, and show up as reads in ribosome profiling data; these reads usually stand out as sharp peaks out of the context of an ORF, and can be ignored. The triplet phasing of ribosome profiling data (typically denser in the first nucleotide of a codon, less dense in the second and third nucleotides) can additionally be used to
statistically distinguish ribosome protected fragments over translated ORFs from other RNA fragments (Bazzini et al. 2014; Michel et al. 2012). Some of this non-specific isolation of RNA fragments can be further mitigated by affinity purification of tagged ribosomes (Ingolia et al. 2014; Jan et al. 2014; Williams et al. 2014).

However, there remain other ribosome profiling artifacts of unknown origin to be wary of. Ribosome profiles are typically “spiky” (Figure 1.3), being decidedly non-uniform even over translated ORFs; this “spikiness” has not been properly ascribed to experimental biases (from sequencing or isolation of stalled ribosomes), or biologically meaningful pauses/stalls (Shoemaker and Green 2012). Ribosome profiles also tend to be noisier over transcript 5’ leaders compared to 3’ trailers, with a significant fraction of reads outside of possible ORFs. While some of these reads may be bona fide translation from non-canonical starts (Ingolia et al. 2011; Lee et al. 2012; Michel et al. 2014), it is also possible that ribosome profiling inadvertently captures scanning 40S-protected mRNA as well (Gerashchenko and Gladyshev 2015), although such reads are generally the minority of 5’ leader ribosome profiling reads. In addition, depending on sample preparation, the 3’ trailers may also contain many ribosome profiling reads at densities similar to 5’ leaders and coding sequences; while this was observed in a gene-specific manner in some ribosome profiling datasets (Bazzini et al. 2014), but not others (Chew et al. 2013), the source of this difference has not been investigated.

**Non-canonical translation detected by ribosome profiling**

In addition to quantifying translation over transcript coding regions, the initial ribosome profiling papers reported translation in many non-coding regions of the genome: in the 5’ leaders of transcripts, as well as on many lncRNAs. Ribosome profiling in yeast (Brar et al. 2012) and mouse embryonic stem cells (mESCs) (Ingolia et al. 2011) detected ribosome profiles in many non-coding transcripts; for mESCs, as much as half of annotated lncRNAs had regions of high translational efficiency.
similar to that of coding sequences. Ribosome profiling has also detected translation in the majority of transcript 5’ leaders of various eukaryotic organisms (Brar et al. 2012; Ingolia et al. 2011, 2009; Guo et al. 2010; Lee et al. 2012), with much of this translation occurring within uORFs (Ingolia et al. 2014). While it was not surprising that some translation would be detected outside of annotated coding regions, the extent (even after accounting for biases in ribosome profiling) was unexpected.

The observation that many lncRNAs were translated raised questions about the annotation of these transcripts as bona fide non-coding RNAs: could these translated transcripts instead be coding for novel peptides? Is translation of these RNAs an unavoidable consequence of these mRNAs being capped and polyadenylated? Or might translation be somehow required for the function of these non-coding RNA? In the first part of my thesis work (Chapter 2), I analyzed the translation over vertebrate IncRNAs more closely to distinguish patterns of translation that corresponded to canonical, non-canonical, or background translation.

uORFs have long been known to regulate translation of their cognate coding sequences, although detecting and quantifying their translation has generally been technically difficult due to limitations of previous technologies (as discussed in the previous section). While ribosome profiling has shown that the translation of uORFs (and thus, how repressive they may be) changes over time and in response to different stresses (Ingolia et al. 2011; Gerashchenko et al. 2012; Stumpf et al. 2013; Liu et al. 2013; Shalgi et al. 2013), others have questioned whether these results may be due to experimental artifacts arising from treatment with translational inhibitors (Gerashchenko and Gladyshev 2015). These experimental artifacts originate from the interaction of altered cell physiology under stress with the slow penetration of translational inhibitors, resulting in a bias in the accumulation of ribosome profiling reads at the beginning of transcripts (including at uORFs) that is not physiological, as evidenced by changes in this biases with translational inhibitor concentration; these biases are mitigated by saturating
increases in inhibitor concentration, and largely abolished if treatment with inhibitors occurs after cell lysis.

Regardless of whether their use is regulated, the widespread translation of uORFs even within a single cell type/state warrants further study. For the first time, with ribosome profiling, we are able to measure uORF translation sensitively and quantitatively genome-wide, allowing us to address how repressive uORFs generally are, how they differ between transcripts, and what mediates these differences. In the second part of my thesis work (Chapter 3), I quantified uORF repressiveness genome-wide in human, mouse and zebrafish, determining how their sequence features modulate their repressiveness, and how this repressiveness was conserved over evolution.
CHAPTER 2

Translation of many long non-coding RNAs resemble translation at 5’ leaders

This chapter was previously published in Development on July 1, 2013 (Chew et al. 2013). The project followed-up on the work that identified long non-coding RNAs during embryogenesis (Pauli et al. 2012b), and was conceived by Andrea Pauli, Alexander F. Schier and Eivind Valen. I adapted and applied ribosome profiling to zebrafish and processed the raw sequencing data. All of us analyzed and interpreted the data, with Eivind implementing the random forest classifier. We wrote the manuscript with input and discussions from John Rinn (Harvard Dept of Stem Cell and Regenerative Biology) and Aviv Regev (MIT Dept of Biology, Broad Institute of MIT and Harvard).

ABSTRACT

Large-scale genomics and computational approaches have identified thousands of putative long non-coding RNAs (IncRNAs). It has been controversial, however, as to what fraction of these RNAs is truly non-coding. Here we combine ribosome profiling with a machine-learning approach to validate IncRNAs during zebrafish development in a high throughput manner. We find that dozens of proposed IncRNAs are protein-coding contaminants and that many IncRNAs have ribosome profiles that resemble the 5’ leaders of coding RNAs. Analysis of ribosome profiling data from ES cells reveals similar properties for mammalian IncRNAs. These results clarify the annotation of developmental IncRNAs and suggest a potential role for translation in IncRNA regulation. In addition, our computational pipeline and ribosome profiling data provides a powerful resource for the identification of translated open reading frames during zebrafish development.
INTRODUCTION

Long non-coding RNAs (lncRNAs) have emerged as important regulators of gene expression during development (Rinn and Chang 2012; Pauli et al. 2011). LncRNAs were initially discovered for their essential roles in imprinting (Bartolomei et al. 1991; Barlow et al. 1991; Jinno et al. 1995; Sleutels et al. 2002) and mammalian X chromosome inactivation (Brockdorff et al. 1992; Borsani et al. 1991; Brown et al. 1992). Studies of Hox gene regulation in mammals and of flowering control in plants have identified additional lncRNAs such as HOTTIP (Wang et al. 2011) and COOLAIR (Ietswaart et al. 2012; Swiezewski et al. 2009). The past decade has seen an explosion of genome-wide studies that have identified thousands of putative lncRNAs in a range of organisms (Kapranov et al. 2002, 2007; Carninci et al. 2005; ENCODE Project Consortium et al. 2007; Fejes-Toth et al. 2009; Guttman et al. 2009; Cabili et al. 2011b; Collins et al. 2012; Derrien et al. 2012; Djebali et al. 2012; Pauli et al. 2012b; Tilgner et al. 2012; Okazaki et al. 2002; Bertone et al. 2004; Guttman et al. 2010; Ravasi et al. 2006). While the developmental roles of the vast majority of these novel transcripts are unknown, recent studies in zebrafish and ES cells have indicated roles for lncRNAs during embryogenesis, pluripotency and differentiation (Guttman et al. 2011; Ulitsky et al. 2011).

A prerequisite for the functional analysis of lncRNAs is the high-confidence annotation of this class of genes as truly non-coding. The distinction of lncRNAs from coding mRNAs has often relied on the computational classification of expressed transcripts (Guttman and Rinn 2012; Dinger et al. 2008). These classifiers evaluate transcript features such as open reading frame (ORF) lengths, coding potential, and protein sequence conservation. Such computational approaches can distinguish between coding RNAs and lncRNAs (Carninci et al. 2005; Guttman et al. 2009; Cabili et al. 2011b; Pauli et al. 2012b; Ulitsky et al. 2011), but may also give rise to misclassifications: lncRNAs containing short conserved regions may be misclassified as protein-coding (false negatives), whereas protein-coding transcripts containing short or weakly conserved ORFs may be misclassified as non-coding (false positives). For example, two recent
zebrafish IncRNA catalogues (Pauli et al. 2012b; Ulitsky et al. 2011) share little overlap, suggesting that novel approaches are needed to distinguish coding from non-coding RNAs.

One approach to detect potential coding sequences is ribosome profiling (Ingolia et al. 2009, 2012). In this method, mRNA fragments protected from RNAsel digestion by cycloheximide (CHX)-stalled 80S ribosomes are isolated and sequenced. The resultant ribosome-protected fragments (RPFs) correspond to the sites where translating ribosomes resided on mRNA transcripts at the time of isolation, yielding a quantitative, genome-wide snapshot of translation at nucleotide (nt) resolution. Application of this method to mouse embryonic stem cells (mESCs) detected RPFs associated with many previously annotated IncRNAs (Ingolia et al. 2011). This study suggested that the majority of annotated IncRNAs contain highly translated regions comparable to protein-coding genes and may encode proteins. However, translation of a transcript was inferred by measuring localized densities of ribosome profiling reads relative to expression (translational efficiency; TE). As shown below, we find that this approach does not reliably distinguish the main ORFs (coding sequences; CDSes) from upstream ORFs (uORFs). This distinction is important because the vast majority of uORFs are unlikely to code for functional peptide products since their peptide sequences are not conserved, even though their presence in the 5’ leader may be (Hood et al. 2009). Indeed, a recent peptidomics study suggested that most annotated IncRNAs do not generate stable protein products (Bánfai et al. 2012). It has therefore remained unclear what fraction of currently annotated putative IncRNAs are truly non-coding.

Here we address the issue of IncRNA annotation by combining ribosome profiling during early zebrafish development with a new machine-learning approach. Our study suggests that dozens of previously annotated IncRNAs are protein-coding contaminants. In addition, we find that many IncRNAs in zebrafish and ES cells resemble the 5’ leaders of coding mRNAs, raising the possibility that translation is involved in IncRNA regulation. The methods and datasets provided in this study provide a broad resource for the identification of translated open reading frames during zebrafish development.
RESULTS AND DISCUSSION

Figure 2.1. Overview of IncRNA classification pipeline
High-throughput sequencing data (ribosome profiling and RNA-Seq) from eight early developmental stages (A) is used to train a classifier with RefSeq Coding Sequences (CDSes), 5' leaders and 3' trailers (B). (C) The Translated ORF Classifier (TOC) uses ribosome profiles and gene expression levels to classify putative IncRNAs as protein-coding (blue), leader-like (green) or trailer-like (red).
Ribosome profiling outlines translated regions of zebrafish transcripts

To identify ribosome-associated regions in the zebrafish transcriptome, we generated high-depth ribosome profiles over a time course of eight early developmental stages (Figures 2.1 and A.1; for details see Materials and Methods in Appendix). Of 220 million high-quality ribosome protected fragments (RPFs), 84.5 million RPFs mapped to RefSeq genes (see Figure 2.2a for examples of ribosome profiles). Approximately 81% of RefSeq genes expressed > 1 FPKM (12228 genes) had at least 10 normalized RPFs (Figure A.3a), while about 68% of genes have reads over at least 10% of their annotated coding sequence (CDS; Figure A.3b). Within exons of RefSeq transcripts, 95.7% of RPFs mapped to CDSes (mean density of 3.64 RPFs per nt), 0.54% of RPFs mapped to 3' transcript trailers (mean density of 0.054 RPFs per nt), and the rest (3.71%) mapped to 5' transcript leaders (mean density of 1.46 RPFs per nt). This distribution corresponds to a >65-fold enrichment of RPFs associated with CDSes compared to 3' trailers, and a >25-fold enrichment of RPFs associated with 5' leaders compared to 3' trailers, consistent with ribosome profiling data in other systems (Ingolia et al. 2011; Brar et al. 2012). As observed in previous studies, we found triplet phasing of ribosome profiles in the CDSes of coding genes, corresponding to the translocation of translating 80S ribosomes in steps of 3 nts (Figure 2.2b).

Consistent with the release of 80S ribosomes at in-frame stop codons, RPFs over 3' trailers tend to be sparse and randomly distributed (Figure 2.2a), and may represent background experimental noise inherent to the ribosome profiling method. As observed in ribosome profiling data in other systems (Ingolia et al. 2011; Brar et al. 2012; Fritsch et al. 2012b; Lee et al. 2012), 5' leaders of coding transcripts are widely associated with ribosomes, showing relatively high densities of RPFs at locations often corresponding, but not limited, to uORFs. The stop codons of annotated ORFs are significantly enriched for RPFs (Figure A.2c). We find widespread occurrence of uORFs (49.5% of RefSeq genes have RPF-containing uORFs), as well as many instances of translated, extremely short ORFs that are as small as an
AUG followed by a stop (minimal ORFs or minORFs) (Figure A.4). These results highlight the power of this approach in identifying translated regions of zebrafish transcripts.

Figure 2.2. Ribosome profiles outline translated ORFs of coding genes
(A) Representative examples of Ribosome Protected Fragment (RPF) densities associated with protein-coding genes. Gene structures are depicted as thick bars for the coding sequence (CDS), thin bars for 5' leaders and 3' trailers and dashed lines for introns. Note that the majority of RPFs map within the CDS and are flanked by the annotated initiation (START, green) and termination codon (STOP, red). The bottom three panels show examples of uORF-containing genes. For these genes, RPF reads map to the CDSes and to short ORFs within the 5' leaders. (B) RefSeq metagene analysis of relative phasing of ribosome P-sites (see Materials and Methods). As in previous studies (Ingolia et al. 2011), triplet phasing of ribosome profiles was observed.
**TOC distinguishes ORFs in annotated 5’ leaders, CDSes and 3’ trailers**

To use the ribosome profiling dataset for the classification of ORFs, we developed a random forest classifier (Breiman 2001). We tested whether ribosome profiles over RNA subregions might reliably distinguish CDSes from ORFs in 5’ leaders and from ORFs in 3’ trailers. To train the classifier, we used the RefSeq gene sets in zebrafish and mouse (see Materials and Methods for details). Our classifier, called TOC (Translated ORF Classifier), employs four features (Figure 2.3a): (1) Translational efficiency (TE) – the density of ribosome profiling reads over an ORF relative to its expression level; (2) Inside versus Outside (IO) - the ratio of bases covered within an ORF versus outside (upstream and downstream), capturing a distinct feature of coding transcripts where read coverage tends to be predominantly over a single ORF; (3) Fraction Length (FL) - the fraction of the transcript covered by the ORF, accounting for the observation that annotated CDSes tend to span a significant portion of the transcript; and (4) Disengagement Score (DS) - the degree to which RPFs are absent downstream of the ORF, building on prior knowledge that reinitiation after extended translation and stop-codon read-through are rare events (Jackson et al. 2007). These features effectively integrate intrinsic transcript information such as sequence and location of ORFs with external data such as ribosome profiling and expression levels derived from RNA-seq.

While individual features were able to separate one class of RefSeq ORFs from the other two, the combination of all four was necessary to reliably distinguish ORFs within annotated 5’ leaders, CDSes and 3’ trailers (Figures 2.3, A.5). Notably, TE distinguished 3’ trailers from 5’ leaders and CDSes, while DS helped separate uORFs in 5’ leaders from CDSes (Figures 2.3b for zebrafish; A.5 for mouse). The combination of IO and FL differentiated CDSes from ORFs in 5’ leaders and 3’ trailers (Figures 3b and A.5). The use of all four features in the TOC classifier was highly accurate in distinguishing CDSes from 5’ leader-like ORFs and 3’ trailer-like ORFs even at low RNA expression levels (Figure A.6; overall out-of-bag
error for zebrafish: 3.25%). These results establish TOC as a powerful classifier to distinguish ORFs in annotated 5’ leaders, CDSes and 3’ trailers.

Figure 2.3. TOC distinguishes ORFs in 5’ leaders, CDSes and 3’ trailers
(A) A training set is constructed from RefSeq genes using 1) annotated CDSes (coding ORFs, blue) in the context of the whole transcript, 2) RPF-containing ORFs in the 5’ leader sequence (green) in the context of the 5’ leader, and 3) RPF-containing ORFs in the 3’ trailer (red) in the context of the 3’ trailer (see Materials and Methods). The four metrics used to train the classifier are displayed in the grey box (Translational Efficiency, TE; Inside versus Outside, IO; Fragment Length, FL; and Disengagement Score, DS). After training, TOC uses RPF-covered ORFs to classify transcripts. (B) The combination of the four metrics separates coding ORFs, leaders and trailers of the training set. Transcripts lacking a protein-coding ORF cluster with trailers and leaders of the training set, as shown for three validated zebrafish IncRNAs (black). The density of each measure is shown along the axes.
**TOC refines classification of lncRNAs**

To refine the classification of putative lncRNAs, we applied TOC to the catalogues recently published for zebrafish embryos (Pauli et al. 2012b; Ulitsky et al. 2011) and mouse ES cells (Guttman et al. 2011). The application of TOC to these datasets is justified by the biochemical similarity between coding mRNAs and recently annotated lncRNAs (e.g. both are 5’ capped and 3’ poly-adenylated). Notably, TOC analysis revealed that dozens of putative lncRNAs have the same characteristics as protein-coding mRNAs: a main CDS engaged by ribosomes and few (if any) RPFs downstream (Figures 2.4, A.7). Depending on the dataset, we find that 8% to 45% of previously proposed lncRNAs are likely to be bona fide protein-coding mRNAs (Figure 2.4). These transcripts will be an interesting source to identify previously uncharacterized proteins. On the other hand, 18% to 44% of putative lncRNAs showed little or no association with ribosomes, akin to 3’ trailers of coding transcripts (Figure 2.4). These transcripts are bona fide lncRNAs and warrant functional characterization.

Strikingly, we found that the ribosome profiles over more than 40% of putative zebrafish and mouse lncRNAs resemble 5’ leaders rather than 3’ trailers (Figure 2.4). These lncRNAs contain ORFs with a higher TE than 3’ trailer-like lncRNAs, but have shorter and less conserved ORFs than the CDSes of protein-coding genes (Figure A.8). Similar to leaders, RPFs are often distributed over multiple ORFs, none of which stands out as a main CDS of a protein-coding gene. The leader-like class of lncRNAs represents a distinct subset of the previously described sprcRNAs (Ingolia et al. 2011). Unlike sprcRNAs, which are identified solely by TE, leader-like lncRNAs exclude misannotated protein-coding mRNAs and transcripts with spuriously associated ribosomes.
Figure 2.4. TOC refines classification of lncRNAs

(A) TOC-based classification improves previous lncRNA predictions. Shown are RNA-Seq and ribosome profiling read densities associated with three putative lncRNAs (Ulitsky et al. 2011), which had conflicting annotations in published zebrafish lncRNA sets (Pauli et al. 2012b; Ulitsky et al. 2011). Transcript structures are shown in black. Introns are indicated as dashed lines. The region scoring highest in PhyloCSF (Lin et al. 2011) is indicated in orange. While TOC reveals the protein-coding nature of linc-ca2, it confirms the non-coding nature of the two conserved lncRNAs megamind and cyrano. These two lncRNAs had been filtered out in the Pauli et al. lncRNA set due to their relatively high phylogenetic Codon Substitution Frequency scores (PhyloCSF > 20).

(B) Fraction of loci that are classified as coding (blue), leader-like (green) and trailer-like (red) in three collections of lncRNAs. ZF1 (Pauli et al. 2012b), ZF2 (Ulitsky et al. 2011) and mouse ES cells (Guttman et al. 2011).

The association of ribosomes with leader-like lncRNAs raises two important questions: Do the associated ribosomes generate proteins? Are these proteins functional? Several observations suggest that leader-associated ribosomes might generate proteins that are likely to be non-functional. Recent studies have shown that the CHX used in ribosome profiling protocols acts through the E-site of the 60S ribosomal subunit (Schneider-Poetsch et al. 2010), and should only stabilize the translating 80S...
ribosome during the ribosomal footprinting step. Moreover, the sizes of ribosome footprints isolated in ribosome profiling protocols (approximately 30 nts) correspond to RNA fragments protected by 80S ribosomes (Wolin and Walter 1988). The translation of ORFs within 5’ leaders is further supported by mass spectrometry data (Slavoff et al. 2013b) and by observed enrichment of RPFs over sites of translation initiation in ribosome profiling data from harringtonine (Ingolia et al. 2011), lactimidomycin (Lee et al. 2012) and puromycin (Fritsch et al. 2012b) treated samples. Thus, leader-associated ribosome profiles likely represent actual translation of ORFs rather than ribosomal subunits scanning the transcript.

The lack of conservation of most uORFs suggests that the protein product might not be functional (Neafsey and Galagan 2007b; Calvo et al. 2009; Hood et al. 2009). Instead, ribosomal engagement with leader-like IncRNAs may be regulatory. Given the regulatory role of uORFs in some coding transcripts (Arribere and Gilbert 2013; Calvo et al. 2009; Hood et al. 2009; Johansson and Jacobson 2010; Hinnebusch 2005), 5’ leader-like translation may affect IncRNA stability and/or subcellular localization. Translating ORFs within IncRNAs might target the transcript for nonsense-mediated decay (Tani et al. 2013), degrading it in the cytoplasm and/or retaining it in the nucleus (de Turris et al. 2011), resulting in the predominantly nuclear localization of most IncRNAs (Derrien et al. 2012). Prime candidates for such regulation are the minORF-containing IncRNAs where the single amino acid product of their translation could not conceivably be functional. Alternatively, association of ribosomes with leader-like IncRNAs might be translational noise caused by the cytoplasmic location of 5’-capped and poly-adenylated transcripts. Such spurious translation may only be functional on evolutionary time-scales as the source of novel coding genes (Carvunis et al. 2012).

In summary, our ribosome profiling data and Translated ORF Classifier allow the high-confidence annotation of coding and non-coding RNAs, and complements and extends previous computational approaches such as PhyloCSF. As demonstrated by our previously published pipeline (Pauli et al. 2012b),
these more traditional computational approaches can exclude the large majority of potential false-positives but misannotate some conserved IncRNAs as coding RNAs (e.g. cyrano and megamind (Ulitsky et al., 2011; Figure 2.4). The use of additional approaches such as mass-spectrometry will further improve the annotation of coding- and non-coding RNAs in zebrafish (Slavoff et al. 2013b).

While our study has focused on the classification of IncRNAs, the accompanying ribosome profiling data will be a rich resource for the discovery of novel protein-coding genes that act during development. Our dataset increases the depth of previous ribosome profiling datasets in zebrafish by an order of magnitude (Bazzini et al. 2012) and expands the temporal coverage to five days of development. The nucleotide resolution of the data allows annotation of translated subregions of transcripts and the identification of potential protein isoforms, furthering ongoing efforts to refine the zebrafish genome annotation (Kettleborough et al. 2013). Finally, the quantitative nature of ribosome profiling combined with existing RNA-seq data will enable studies of post-transcriptional and translational regulation during zebrafish development.
CHAPTER 3

Repressiveness and sequence features of uORFs are conserved in vertebrates

This chapter was a manuscript in preparation for submission. The project continued analysis of ribosome profiling data generated in zebrafish (Chew et al. 2013), while analyzing published data from human (Stumpf et al. 2013) and mouse (Ingolia et al. 2011). I performed all analysis and wrote the manuscript with input from Andrea Pauli and Alexander F. Schier.

ABSTRACT

Upstream open reading frames (uORFs) are ubiquitous repressive genetic elements in vertebrate mRNAs. While much is known about the regulation of individual genes by their uORFs, the range of uORF-mediated translational repression in vertebrate genomes has been unexplored. Moreover, it is unknown whether the repressive effects of uORFs are conserved in orthologous transcripts. To address these questions, we analyzed transcript sequences and ribosome profiling data from human, mouse and zebrafish. We found that uORFs are depleted near coding sequences (CDSes) and have initiation contexts that diminish their translation. Linear modeling reveals that sequence features at both uORFs and CDSes modulate the translation of CDSes. Strikingly, uORF sequence features are conserved between human and mouse, and correlate with the relative levels of translation over 5’ leaders and CDSes. These observations suggest that the prevalence of vertebrate uORFs may be explained by their functional conservation as weak repressors of CDS translation.
INTRODUCTION

Ribosomal pre-initiation complexes typically scan across the 5' leaders (also known as 5' untranslated regions or 5' UTRs) of eukaryotic mRNAs before initiating translation of at the start codon of coding sequences (CDSes) (Jackson et al. 2010; Hinnebusch 2011). Open reading frames (ORFs), as defined by a start codon and a downstream in-frame stop codon, can occur upstream of CDSes in the 5' leader; these upstream open reading frames (uORFs) are often repressive, because translation of uORFs occurs at the expense of translation of downstream CDSes (Kozak 2002; Sonenberg and Hinnebusch 2009; Hood et al. 2009).

Indeed, ribosome profiling (Ingolia et al. 2011; Stumpf et al. 2013; Lee et al. 2012; Fritsch et al. 2012b; Chew et al. 2013) and mass spectrometry (Oyama et al. 2004; Slavoff et al. 2013b; Menschaert et al. 2013; Bazzini et al. 2014; Kim et al. 2014; Wilhelm et al. 2014) has detected the widespread translation of uORFs, which are prevalent in vertebrates (present in ~50% of human and mouse mRNAs; ~65% of zebrafish mRNAs) (Iacono et al. 2005; Matsui et al. 2007; Calvo et al. 2009), and many uORFs are translated, as evidenced by ribosome profiling. While this suggests that uORFs are broadly repressive, how this repressiveness varies in vertebrates and whether their repressive effects are conserved has not been explored.

Here, we address these questions by comparing uORF repressiveness among orthologous vertebrate transcripts in human, mouse and zebrafish. By taking advantage of the nucleotide resolution and quantitative nature of ribosome profiling data (Ingolia et al. 2012), we quantified the range of uORF-mediated translational repression and determined how various uORF sequence features modulate repressiveness in three independently generated vertebrate ribosome profiling datasets (Stumpf et al. 2013; Ingolia et al. 2011; Chew et al. 2013). Our analyses indicate that the repressiveness and sequence features of uORFs are conserved in vertebrates.
RESULTS

*uORFs vary in their sequence features, trend toward weak repressiveness*

Previous studies have identified sequence features that modulate the repressive effects of uORFs on the translation of CDSes: uORF initiation context sequence and secondary structure affects how efficiently translation may initiate at uORFs (Hinnebusch 2011), while the distance between uORFs and CDSes affects the efficiency of reinitiation following translation of a uORF (Kozak 1987, 2002). We used these well-established sequence features to analyze the repressive potential of human, mouse and zebrafish uORFs. Unless otherwise stated, results discussed in main figures and text are for mouse ES cell ribosome profiling data (Ingolia et al. 2011); similar results observed in the analyses of zebrafish and human data are further discussed in supplemental materials.

*Initiation context sequence:* To identify the sequence motifs that promote translational initiation, we constructed weighted position-specific scoring matrices (PSSMs) from the initiation contexts of CDSes (±10 nucleotides around AUG start codon); we used initiation contexts in mRNAs that lacked uORFs and weighted their contribution using translational efficiency values (density of ribosome profiling reads over CDS normalized by transcript expression levels) calculated from ribosome profiling data (Figure 3.1a). While these PSSMs resemble the Kozak consensus sequence (which is typically derived from unweighted PSSMS of whole transcriptomes), weighting for translational efficiency accounts for transcript-specific variation in translational efficiency and sequence motifs (Figures A.9, A.10). This Weighted Relative ENTropy (WRENT) score (see methods) allowed us to compare the initiation contexts of uORFs, as well as other ORFs (including the CDS) in transcripts (Figure 3.1b; see methods). We found that although initiation context sequences at uORFs vary widely, they generally scored unfavorably for translation initiation (Figure 3.1b): only ~15% of uORFs had more favorable initiation contexts than the median CDS WRENT score.
Figure 3.1: Genome-wide distribution of uORF sequence features

a. Sequence motif representing efficient initiation at CDSes is used to score initiation contexts at all ORFs. Motif was constructed from the CDS initiation context (+/- 10 nucleotides around the annotated start, indicated with the blue bar over transcript schematic) of coding transcripts lacking uORFs, weighted for their TE. Height on vertical axis represents weighted relative entropy (WRENT). Little information is gained from performing this analyses on uORFs (insert), indicating limited selection for uORF TE.

b. Cumulative distribution of WRENT scores around AUGs at various positions in coding transcripts. Sequences around AUGs at the start of CDSes are more favorable for translation initiation that sequences around AUGs everywhere else on the transcript, including at uORFs.

c. Meta-profile of predicted secondary structure (in sliding 35 nucleotide window) around uORF and CDS starts; uORF profiles are subdivided by the number of uORFs on each transcript (dashed lines). Red line at -10 nucleotides indicates position where secondary structure was considered for further analyses. Secondary structure alone can define the starts of coding regions, less average secondary structure (less negative ensemble free energy) is observed with increasing number of uORFs.

d. Cumulative distribution of initiation context secondary structure of uORFs in transcripts with varying number of uORFs. Less average secondary structure (less negative ensemble free energy) is observed with increasing number of uORFs. Average secondary structure at CDS initiation contexts is indicated by green dashed line.

e. AUG (solid lines) and stop codon (dashed lines) trinucleotide biases in the 5' leader as a function of distance with respect to CDS start, for all 3 vertebrates. AUGs, but not stop codon trinucleotides are specifically depleted near the CDS start, in all 3 vertebrates examined; the effect is most drastic in zebrafish transcripts.

f. uORF ends depletion (observed minus expected from shuffled sequences) with respect to CDS start. uORF ends are specifically depleted in the 5' leader near the CDS, most so in zebrafish transcripts.
Initiation context secondary structure: Secondary structure at the beginning of ORFs need to be unwound prior to translation initiation (Pestova and Kolupaeva 2002; Hinnebusch 2011). We therefore estimated the effect of secondary structures at initiation contexts by calculating the ensemble free energies (EFEs) of predicted secondary structures at initiation contexts using the Vienna RNAfold package (Lorenz et al. 2011). This allowed us to compare the energies required to unwind secondary structure prior to translational initiation among uORFs and other ORFs.

We found that uORFs also vary in their initiation context secondary structure EFEs. The average secondary structure at uORF starts is slightly less stable than at CDS starts, which would favor translational initiation at uORFs. However, mammalian transcripts with fewer uORFs tend to have more stable secondary structures around their uORF starts (Figures 3.1c-d). For example, on single-uORF transcripts, uORFs starts are on average ~2.1 kcal/mol more stable in their secondary structure than CDS starts. These opposing trends suggest dual modes of selection happening at 5’ leaders: for less stable secondary structure in the total length of 5’ leaders, and for more stable secondary structures specifically at uORF initiation contexts.

uORF position with respect to CDS: Following uORF translation, post-termination 40S ribosomal subunits may remain attached and continue scanning to reinitiate at downstream CDSes (Kozak 2002; Jackson et al. 2007). The efficiency of reinitiation has been observed to decrease as the distance between uORFs and CDSes decreases (Kozak 1987). To characterize the potential of uORFs to allow reinitiation at downstream CDSes, we examined the positional distribution of uORFs in vertebrate 5’ leaders. While uORFs are broadly distributed in 5’ leaders, uORF starts and uORF ends are depleted near the CDSes (Figures 3.1e-f). The position-specific depletion of AUGs is observed in all three frames, whereas stop codon trinucleotides (which often occur outside of ORFs) are not specifically depleted near CDSes (Figure 3.1e). These observations indicate that uORFs are depleted in a position-specific manner
within vertebrate 5’ leaders from approximately 100 nucleotides upstream of the CDS to the CDS, which coincides with the region where uORF position diminishes the efficiency of reinitiation (Kozak 1987).

Taken together, our analysis reveals that vertebrate uORFs tend to have sequence features associated with weak repressiveness: they have unfavorable initiation sequences and secondary structures, and they are depleted from regions closest to the CDS.

**uORFs are modestly repressive on average**

The sequence features of uORFs and previous proteomics data (Calvo et al. 2009) suggest that uORFs are only modestly (~15 to 30%) repressive for downstream CDS translation. To directly quantify the translational efficiency of CDSes (as opposed to inferring it from protein and RNA levels), we calculated the density of ribosome profiling reads over individual CDSes and normalized it by transcript abundance. This approach allowed us to compare the translational efficiency of CDSes in mRNAs with or without uORFs. We observed that the presence of uORFs in 5’ leaders was associated with a modest decrease in downstream translation (averaging 17-30% reduction; Figure 3.2a). uORFs were associated with a reduction in CDS translational efficiency in a “dose-dependent” manner: the more uORFs in a transcript, the greater the reduction in translational efficiency (Figure 3.2a).

Although uORFs can be repressive, studies during yeast meiosis have suggested that they might not be the dominant factor in regulating CDS translation (Brar et al. 2012): instead of a negative correlation between uORF and CDS translational efficiencies, uORF and CDS translation had been found to be weakly correlated. To determine if a similar trend holds true in vertebrates, we compared uORF and CDS translational efficiencies in transcripts with only one non-overlapping uORF. Indeed, we observed a significant and positive correlation between uORF and CDS translational efficiencies in all three vertebrate ribosome profiling datasets (Figure 3.2b). This observation suggests that other forms of
translational regulation are dominant in regulating CDS translation and drive both uORF and CDS translation.

**Figure 3.2: uORFs are weakly repressive on average**

a. Cumulative distribution of TEs of CDSes in transcripts with increasing number of uORFs. The presence of uORFs has a modest effect on TE (between 19-49% reduction with increasing number of uORFs, averaging 30%).

b. Scatter plot of TEs of cognate uORFs and CDSes, for transcripts with one non-overlapping uORF. Red line indicates locally-weighted regression (LOWESS) fit. Figure insert schematic describes how ORF TE is calculated. uORF TEs correlates weakly, but significantly and positively (Spearman’s rho = 0.33, p < 10^{-23}) with cognate CDS TE, suggesting that repressive effects of uORFs are weak relative to other upstream forms of translational regulation.

**uORF sequence features correlate with uORF translation and repressiveness**

To integrate the above analyses, we asked if there is a relationship between uORF sequence features and uORF translation and repressiveness. We found that more favorable initiation context sequences and less stable secondary structures lead to increased uORF translational efficiency (Figures 3.3a-b). In contrast and as expected, uORF translational efficiency is unaffected by its distance from the downstream CDS (Figure 3.3c).

To estimate the repressive effects of uORFs on CDS translation, we calculated the ratio between uORF and CDS translational efficiencies and correlated it with uORF sequence features, reasoning that a highly repressive uORF would be well-translated at the expense of downstream CDS translation, resulting in a high ratio. When scored this way, individual uORFs varied in their repressiveness.
approximately 100-1000 fold (range of 95% confidence interval). Importantly, each individual uORF sequence feature correlated significantly with uORF repressiveness: increased uORF repressiveness correlated with more favorable initiation context sequences, less stable initiation context secondary structure, and reduced distance from the CDS (Figures 3.3a-c).

From the above analyses, we found that individual sequence features correlate with uORF repressiveness. To determine if the repressiveness associated with individual uORF sequence features is additive, we constructed linear models that combined them. The combination of all three uORF sequence features correlated substantially better with uORF repressiveness than individual sequence features \( (p < 10^{-11} \text{ versus } p < 10^{-5}, p < 10^{-5} \text{ and } p < 0.05 \text{ for uORF initiation context sequence, secondary structure, and uORF-CDS distance, respectively) (Figures 3.3d). Moreover, while the individual sequence features were associated with 1.8- (uORF-CDS distance), 5.5- (uORF initiation context secondary structure) and 7.1-fold (uORF initiation context sequence) ranges in uORF repressiveness, their combination was associated with a 14-fold range. Thus, the repressiveness associated with individual uORF sequence features is at least partially additive.}

\textit{uORF sequence features contribute to CDS translational efficiency}

To determine the extent that uORF and CDS sequence features contribute to CDS translational efficiency, we extended the linear model to include CDS initiation context sequence and secondary structure. We found that the combination of uORF and CDS sequence features correlated substantially better with CDS translational efficiency than uORF or CDS sequence features \( (p < 10^{-23} \text{ versus } p < 10^{-12} \text{ for both uORF and CDS sequence features alone) (Figure 3.3e). Examining the coefficients of the linear modelling revealed that uORF and CDS sequence features contribute similarly in magnitude to CDS translational efficiency \((-0.57 \text{ vs } 0.41 \text{ per unit WRENT score for initiation context sequence, } 0.60 \text{ vs } 0.83\).}
per kJ/mol for initiation context secondary structure). These analyses suggest that uORF sequence features can be as important as CDS sequence features in specifying CDS translational efficiency.

Figure 3.3: Quantifying contributions of uORF sequence features to uORF repressiveness a, b, c. LOWESS fits of relationship between uORF initiation context sequence (a), secondary structure (b), and uORF-end position w.r.t. CDS (c) against uORF TE, CDS TE and uORF/CDS TE. uORF repressiveness (estimated from uORF/CDS TE) increases with more favorable sequence, decreased secondary structure at uORF initiation context, and reduced uORF-CDS distance. All lines represent significant rank and linear correlations except for uORF TE against uORF-end position w.r.t. CDS (c). d. Scatter plot of linear modelling of uORF sequence features against uORF/CDS TE; red line is LOWESS fit line. Good correlation between a score that combines all measured uORF sequence features with uORF repressiveness ($r = 0.23; p < 10^{-11}$) over a 14.2-fold range in uORF-repressiveness. e. Scatter plot of linear modelling of uORF and CDS sequence features against CDS TE; red line is LOWESS fit line. Good correlation between a score that combines all measured sequence features with CDS TE ($r = 0.33, p < 10^{-23}$) over a 5.2-fold range in CDS TE. Numbers in figure schematic represent normalized coefficients of the linear model, showing that uORF and CDS sequence features contribute approximately equally toward specifying CDS TE.
uORFs are conserved via their sequence features in mammals

To determine whether the repressiveness of uORFs in orthologous genes is conserved amongst human, mouse and zebrafish, we initially sought to compare the ratio between uORF and CDS translational efficiencies amongst orthologous genes. However, due to the scarcity of orthologous transcripts with single non-overlapping uORFs in all species, we instead compared the ratio of ribosome profiling read densities between entire 5’ leaders and CDSes (Figure 3.4a) to estimate the repressiveness of 5’ leaders. This approach revealed significant conservation of 5’ leader repressiveness between mouse and human orthologs and weaker conservation between mammals and zebrafish orthologs (Figures A.19d-e).

To investigate whether uORF sequence features are conserved between orthologous transcripts, we compared the orthologous uORFs with the highest repressiveness as predicted by their individual sequence features (e.g. single uORFs with the most favorable initiation context in their 5’ leaders). We found that human and mouse orthologs share similar uORF sequence features (Figures 3.4b-d) (Pearson correlations for uORF initiation context sequence: r = 0.45; secondary structure: r = 0.34; uORF-CDS distance: r = 0.49). In contrast, no such similarities were found between mammals and zebrafish.

Despite the general conservation of 5’ leader repressiveness amongst orthologous genes, there is also a substantial divergence in 5’ leader repressiveness between some orthologs. We therefore investigated whether differences in uORF sequence features might account for this divergence. Strikingly, differences in the uORF sequence features of orthologous transcripts correlate with differences in uORF repressiveness (Figure 3.4e). This observation held true even when comparing mammalian and zebrafish data (Figures A.19g-h). These findings underscore the strong influence of uORF sequence features in specifying of uORF repressiveness.
Figure 3.4: uORF sequence features are functionally conserved between mouse and human

a. Scatter plot ratio of 5' leader to CDS ribosome profiling read density, between human and mouse orthologous transcripts. Good correlation ($r = 0.60$, $p \sim 0$) observed over large range (>100-fold) of ratios.

b, c, d. Scatter plots of uORF sequence features between human and mouse orthologous transcripts for initiation context sequence (b), secondary structure (c), and uORF-end position w.r.t. CDS (log magnitude transformed; d. Multiple uORFs were considered in this analysis, but only the uORFs with the sequence features most expected to repress downstream translation per transcript were considered. Good correlation observed between orthologous transcripts for all uORF sequence features ($r = 0.45$; $p < 10^{-104}$ for initiation context sequence, $r = 0.34$; $p < 10^{-55}$ for initiation context secondary structure, $r = 0.49$, $p < 10^{-122}$ for uORF-CDS distance). Individual points are colored for their divergence: red for higher in human, blue for higher in mouse.

e. Scatter plot of linear modelling of combined divergence of uORF sequence characteristics against divergence of ratio of 5' leader to CDS ribosome profiling read density, between human and mouse orthologous transcripts. Very good correlation ($r = 0.47$, $p < 10^{-112}$) observed.
DISCUSSION

Our study uncovers the wide range of uORF-mediated translational repression in vertebrates and shows that some of the repressive effects and sequence features of uORFs are conserved. Our genome-wide analyses confirm and extend previous studies that analyzed the roles of uORFs in translational regulation (Calvo et al. 2009; Pelechano et al. 2013; Arribere and Gilbert 2013; Brar et al. 2012; Wethmar et al. 2014) and reach five major conclusions. First, uORFs can vary 100-1000 fold in their repressiveness. Second, sequence features at uORFs (uORF initiation context sequence and secondary structure, and uORF-CDS distance) contribute significantly to their repressiveness. Third, these uORF features are as important as CDS sequence features for determining how efficiently CDSes are translated. Fourth, 5’ leaders in orthologous mammalian genes are broadly similar in their repressiveness and sequence features. Fifth, differences in uORF repressiveness among vertebrates correlate with differences in their sequence features.

While we have described sequence features that correlate with uORF repressiveness, these features alone are not sufficient to predict the repressive effects of individual uORFs on cognate CDS translation (Figure 3.3d). Additional features in mRNAs that likely modulate both uORF and CDS translation include secondary structures in 5’ leaders outside of uORFs, binding sites for proteins and microRNAs, and mRNA localization elements. Such elements might also account for the observed positive correlation between uORF and CDS translational efficiencies (Figure 3.2b) and the relatively modest average repression by uORFs.

The analytical strategies we used to examine uORF sequence features can be applied to determine how these features can have varying effects in different contexts. For example, the optimal translation initiation context in one condition (e.g. signal A) might differ from another condition (e.g. signal B), resulting in different uORF translational efficiencies and repressiveness. Such analyses could reveal broad changes in gene expression brought about by translational regulators.
Our discovery that uORF repressiveness is conserved between mammalian orthlogs supports the idea that uORFs contribute to the precise tuning of protein levels. It is important for cells to express proteins at the right levels and in the right ratios (Li et al. 2014a) over a dynamic range spanning six orders of magnitude (Ghaemmaghami et al. 2003). While transcriptional control accounts for the majority of variation in gene expression (Li et al. 2014b; Jovanovic et al. 2015), translational tuning via uORFs may act to refine expression levels to their optimum.

Finally, the finding that differences in uORF sequence features correlate with differences in uORF repressiveness in orthologous genes between species raises the possibility that sequence variation at individual genes within a species contribute to expression level and phenotypic diversity. In humans, sequence variations in 5’ leaders have been statistically associated with variation in gene expression (Battle et al. 2014), while mutations at uORFs have been shown to contribute to disease (Barbosa et al. 2013). The analytical strategies developed here lay the foundation for a molecular framework to understand the effects of uORF sequence variation on gene expression and physiology.
CHAPTER 4

Conclusion and future directions

The work that I present in this thesis has explored two major aspects of non-canonical translation in vertebrates. I have classified the translation of vertebrate lncRNAs (Chapter 2), finding that while approximately half of them are translated, their translation resembles the translation of 5′ leaders rather than translation of coding sequences. I subsequently analyzed the translation of vertebrate 5′ leaders (Chapter 3), finding that uORF translation and repressiveness are modulated by their sequence features, which in turn mediate the conservation/divergence of uORF repressiveness in vertebrates.

These genome-scale analyses of non-canonical translation were conducted on only a limited number of datasets: HeLa cells, mES cells and the developing zebrafish embryo. However, the computational methods developed for these analyses are broadly applicable, and can be used (in conjunction relevant ribosome profiling data) in the classification of other lncRNA gene sets by how they are translated, as well as in determining how sequence features affect uORF usage in various other combinations of developmental stage, tissue, organism and treatment.

In this chapter, I will discuss a number of logical extensions of these analyses of non-canonical translation: relating translation of lncRNAs to their function, determining if such translation is tissue-specific, exploring the role of uORFs in dampening gene expression noise and mediating gene-specific translational regulation, and predicting phenotypic variation and disease in humans from sequence variation at uORFs.
At around the time I started this thesis work, there was much excitement in the IncRNA field. Many IncRNAs had just been identified in various organisms (Guttman et al. 2010; Ulitsky et al. 2011; Pauli et al. 2012a) and in various human tissues (Cabili et al. 2011a). Much of what drove the excitement in the field was the huge gap between what we already knew about the identity and expression of these IncRNAs, and what we did not know about their biological functions. Given that these IncRNAs presumably do not code for proteins, we could not easily assess what constituted loss-of-function mutations in them (Sauvageau et al. 2013), making their characterization more difficult.

Arguably, the gap between identity and function of IncRNAs remains wide. In the past few years, although we have identified even more IncRNAs from various genome-wide analyses of RNA-seq data (Wang et al. 2013; Mattick and Rinn 2015; Pauli et al. 2015), the functions and mechanisms of only a dozen or so more IncRNAs genes have been thoroughly investigated. The field is beginning to ask what proportion of these genes is indeed non-coding; indeed, with ribosome profiling becoming more accessible, various groups are beginning to identify functional translated peptides from these IncRNAs (as well as other non-coding regions of the genome) in a high throughput manner (Bazzini et al. 2014; Menschaert et al. 2013; Aspden et al. 2014).

We initially intended to use ribosome profiling data to validate our IncRNA gene sets so that we could begin functional characterization of a set of bona fide non-coding RNAs for their function as RNAs; instead, we found that about half of them were translated, albeit in a manner akin to 5’ leaders. This distinction may yet turn out to be a useful sub-classification of IncRNAs, in much the same way that the sub-classification of proteins by localization (e.g. membrane-bound, extracellular, cytoplasmic, nuclear) is informative of possible function.

We had postulated that the translation of IncRNAs might be related to their subcellular localization: the 5’ leader-like translation of IncRNAs makes them optimal NMD targets, degrading them...
in the cytoplasm and restricting them to the nucleus. Since then, single-molecule in situ techniques have enabled the visualization of the subcellular localization of many of these lncRNAs (Cabili et al. 2015), and a logical follow-up would be to compare the translation status of these lncRNAs (translated like 5’ leaders or not) to their localization to determine if there is a correlation.

In light of recent work by others, is not clear whether we would expect a correlation or anti-correlation between nuclear localization and translation of lncRNAs: recent evidence for nuclear translation (Apcher et al. 2013; David et al. 2012) suggests that NMD could act within the nucleus could exclude its targets from it (anti-correlation expected); however, other recent work has shown that the majority of degradation of an NMD target occurs after nuclear export (Halstead et al. 2015), suggesting that nuclear translation (if it exists) is decoupled from NMD and that cytoplasmic NMD might exclude lncRNAs from the cytoplasm (correlation expected). Comparing lncRNA localization with translation status possibly distinguishes between these possibilities. Given that we know the exact ORFs that are translated on these leader-like transcripts, we could specifically inhibit their translation with anti-sense modified oligonucleotides (such as morpholinos; complementary to the ORF’s initiation context) to determine if their translation is indeed necessary for their localization.

lncRNAs have been shown to be both tissue-specific (Cabili et al. 2011a) and developmental stage-specific (Pauli et al. 2012a) in their expression. Given that many lncRNAs are also ideal NMD targets, this raises the question of how much this tissue- and stage-specificity might be due to differential degradation by the NMD machinery in various tissues. As translation is required for NMD, differential degradation could be the result of differential translation, and would be evident in ribosome profiling data by comparing the ribosome profiles of similarly-expressed lncRNAs in various tissues and developmental stages. A similar anti-sense oligonucleotide approach to specifically inhibit the translation of these lncRNAs could be used to compare their translation-mediated post-transcriptional regulation in various tissue.
Why uORFs?

From our analysis of uORFs and their effect on downstream translation, we have found that uORFs and their repressiveness are conserved. However, this does not completely address why uORFs exist in such abundance in vertebrate genomes. I will discuss a number of possible reasons from the simple to the most complex; given the large number of uORFs in the genome, any or all of these reasons could explain the origin of each of these uORFs.

The semi-random and serendipitous existence of uORFs

Firstly, the null scenario assumes no function of the existing uORFs, just that they were not deleterious enough to be selected against. In short to medium evolutionary timescales, given a reasonable rate of mutation and the small amount of information needed to encode an ORF, we simply cannot avoid having uORFs appear in 5' leaders (Neafsey and Galagan 2007a; Lynch et al. 2005). On the other hand, uORFs that were sufficiently repressive to be deleterious to the organism are depleted (Churbanov et al. 2005). On longer evolutionary time scales, the overall deleterious effects of spontaneously appearing uORFs may also be mitigated by the evolution of mechanisms that allow them to be bypassed, thus reducing their overall repressiveness. Such a scenario might explain the current existence of mechanisms for bypassing uORFs (leaky scanning and reinitiation).

While some uORFs are not conserved and hence may have originated from the random processes described above, we do indeed detect evolutionary conservation of many uORFs, potentially as coding ORFs (Churbanov et al. 2005), or for their repressiveness (Chapter 3). For the uORFs conserved for their repressiveness, the simplest situation where uORFs affect gene expression functionally might be to fine-tune it.

Even though the vast majority of genes exist as two copies in diploid genomes, steady-state transcript levels of these genes vary over several orders of magnitude largely due to variations in
transcription rates and regulation (Li et al. 2014b; Battle et al. 2014; Jovanovic et al. 2015). In contrast, post-transcriptional regulation (in the form of transcript stability and translational regulation) does not explain a large fraction of the absolute variation in gene expression: while previous estimates had post-transcriptional regulation explaining >60% of variation in protein abundance, (Schwanhäusser et al. 2011), more recent reanalysis of the data corrects this to <20% (Li et al. 2014b; Li and Biggin 2015). These numbers are consistent with post-transcriptional regulation playing a role in fine-tuning gene expression: transcription gets gene expression levels in the right ballpark (thus explaining the majority of variation), while post-transcriptional mechanisms nudge expression levels to their optimal. Indeed, we see that widespread post-transcriptional mechanisms such as miRNAs (Selbach et al. 2008) and uORFs (Chapter 3; Calvo et al. 2009) have relatively mild effects on gene expression levels.

All roads lead to Rome, but some may dump you onto the outskirts

uORFs may do more than just serendipitously adjust for levels of gene expression. Suppose a cell needs to achieve a specific level of gene expression (with protein as the functional gene product; Figure 4.1). A cell could achieve this with many combinations of transcriptional and translational levels. Drawing attention to the boundary cases, the cell could transcribe the gene at a low level, then translate it without restraint (red path in Figure 4.1), or the cell could transcribe the gene at a high level, then translate it with low efficiency (yellow path). While both paths result in the same mean levels of gene expression, when we account for biological and stochastic noise, the red path results in greater variance in final gene (protein) expression then the yellow path (Raser and O’Shea 2005). Given that uORFs act repressively during the translation of RNA to protein, they may act to reduce the variation in gene expression; other means of dialing down translational efficiency such as microRNAs, RNA secondary structure, and codon usage could serve the same function. Thus, genes with substantial and constitutive levels of post-transcriptional repression may have been selected for precise gene expression. Indeed,
many transcription factors and genes involved in signal transduction have been observed to contain uORFs (Churbanov et al. 2005), where it would make sense that their expression were precise.

Figure 4.1. Gene expression variation depends on transcriptional and translational efficiency
From two genomic loci (in a diploid organism), a gene can be transcribed and translated with varying efficiencies to achieve the same level of protein expression. The red path illustrates a transcribe low, translate high approach, which results in greater variation in final protein expression levels compared to the yellow path, which is a transcribe high, translate low approach.

However, while we can postulate mechanisms and rationale for achieving either precise or varying levels of gene expression, whether this precision in gene expression is biologically relevant remains to be tested. To specifically test the biological importance of precise gene expression, we would have to manipulate both the transcriptional and translational efficiency of a gene. While it is trivial to relief a transcript of translational repression by deleting a uORF or microRNA binding site, it is nearly impossible to effect the complementary and precise reduction in transcriptional efficiency to restore protein expression levels to their original level. I suspect that understanding this question will require a genome-wide study of the endogenous relationships between transcriptional and translational efficiency for given levels of protein expression.

Dynamic, regulated uORF repression

The differential ribosomal occupancy of uORFs (and 5' leaders in general) under various conditions (growing vs starved yeast) was documented in the very first ribosome profiling papers, with subsequent papers reporting the differential use of uORFs over development (Ingolia et al. 2011), under various stresses (Ingolia et al. 2009; Gerashchenko et al. 2012), or through the cell cycle (Stumpf et al.
However, a recent study has shown that many of the differential ribosomal occupancies observed in the studies are likely due to artifacts that originate from translational inhibitor treatment (Gerashchenko and Gladyshev 2015). This is a particular problem when biological samples are pre-treated with inhibitors like cycloheximide to stall ribosomes while cells are still alive; to mitigate this, I had analyzed mouse and zebrafish ribosome profiling data that used translational inhibitors post-lysis, where most of these artifacts are abolished. In addition, these artifacts from translational inhibitor treatment would not affect my analysis of uORFs from single samples, where my comparisons were made amongst different transcripts in the same sample, rather than between the same transcripts in different samples. While that does not rule out the possibility that uORF use is dynamic and regulated, it does suggest that one has to pay close attention to experimental conditions when analyzing ribosome profiling data, and that many of the previous experiments may need to be redone to rule out such artifacts.

There are a number of ways that uORF repressiveness could be dynamic and regulated. uORF repressiveness is determined by its sequence features through multiple mechanisms (Chapter 3), including primary sequence, secondary structure, and the distance from a downstream ORF. While we do not know the exact mechanisms of how sequence features can mediate uORF repressiveness, we do have a few leads.

In terms of primary sequence, it is known that certain initiation factors mediate initiation site selectivity (Hinnebusch 2011; Barth-Baus et al. 2013). Specifically, the stoichiometry of eIF1, eif1a and eIF5 have been shown to affect the efficiency of initiation based on sequence context; in structural data (Hashem et al. 2013), eIF1 and eIF1a do indeed interact with the flanking sequences of the start codon during initiation. Differences in the stoichiometry of the above initiation factors (as well as other components of the translation machinery) could affect initiation site selectivity in way that depends on cell-type, tissue and/or developmental stage. This would consequently affect how efficiently translation
initiates at uORFs in a gene-specific manner, and thus regulate the uORF repressiveness of whole sets of genes.

Indeed, I do observe that over developmental time, as well as under different treatment conditions (Figure 4.2), the optimal initiation contexts can differ quantitatively. It remains to be determined whether this quantitative change is due to changes in the stoichiometry of translation initiation factors, whether this indeed results in changes in uORF repressiveness (in a manner predicted by uORF sequence features), as well as whether such changes have a functional consequence on gene expression.

**Figure 4.2 Optimal initiation context varies under different treatments**
The motifs representing optimal initiation contexts in zebrafish shield stage embryos with or without the presence of a signal. Note that the "A" peak at -3 is higher without signal, while the "GC" peaks at +3 and +4 (0-based indexing) is higher with signal.

How uORF secondary structure affects uORF repressiveness could also be context dependent. There are many genes that encode various RNA helicases, and many have been demonstrated to be involved in translation initiation (Parsyan et al. 2011; Bleichert and Baserga 2007; Jaramillo et al. 1990; Hinnebusch 2011). eIF4a itself (a core translation initiation factor) has RNA helicase activity: it exists as two different paralogs, for which one of them (eIF4a2) is required for microRNA-mediated translational repression (Meijer et al. 2013).

A quick survey of a number of RNA helicases finds that some of them are indeed expressed in a developmental stage-specific manner in zebrafish. In addition, ribosome profiling data from a cell line different than that analyzed in Chapter 3 (lymphoblastoid cells vs HeLa cells) shows a consistently different effect of secondary structure on uORF translational efficiency and repressiveness. Further work
would characterize the potential relationship between tissue-specific expression of RNA helicases with the effect of genetic elements within 5’ leaders such as uORFs.

The distance between uORFs and CDSes are thought to affect uORF repressiveness through the process of reinitiation. The relationship between uORF-CDS distance and uORF repressiveness has historically been challenging to study. As measured in Chapter 3, the effects of this distance are more subtle compared to the effects of uORF initiation context sequence and secondary structure, making it more difficult to detect. Genetically, while it is possible to delete portions of the sequence between a uORF and CDS to reduce the intervening distance, it is difficult to ascribe the resultant effect to either the reduction of that distance, or the removal of a yet-uncharacterized regulatory element. Molecularly, the phosphorylation levels of eIF2α are thought to indirectly facilitate reinitiation through availability of the eIF2-GTP-Met-tRNA ternary complex; however, as eIF2 is also required for translation initiation, it is difficult to make targeted perturbations to eIF2α phosphorylation in the context of reinitiation. eIF3 (a large multi-protein complex) is also thought to mediate the process of reinitiation (Roy et al. 2010), but how it does so is still unknown.

I have been able to quantifying uORF repressiveness as a function of uORF-CDS distance by analyzing ribosome profiling data over the transcriptome. While it would be possible to measure differences in this effect in different cells (where this distance could have a larger or smaller effect on uORF repressiveness), without molecular tools for specifically manipulating the efficiency of reinitiation, it will be difficult to probe the biological relevance of reinitiation to the organism.

Towards predicting the effect of non-coding variation

Many factors contribute towards determining translational efficiency of transcripts. Our study of uORFs examines just one of the non-coding genetic elements that can affect gene expression: while we have shown that they have significant and substantial impact, they certainly do not explain all the
endogenous variation in translational efficiency that exists in the transcriptome. Therefore, while
knowing the sequence features of uORFs alone will not allow us to predict the translational efficiency of
transcripts, we might still predict the effect of genetic variation at uORFs on gene expression.

This approach might be especially useful to the field of human genetics and genomics, where
genetic (and often, phenotypic) variation has been extensively characterized. Particularly in non-coding
areas of the genome and transcriptome, we lack the biological framework for predicting the effect of
sequence variation on gene expression; by contrast, mutations in the coding regions can easily be scored
by how they would change the translated protein product.

With our analytical framework for quantifying the effect of uORF sequence features on their
repressiveness, we could compare our predictions of the effects of uORF sequence variation to
measured gene expression differences, as well as to phenotypic variation. Indeed, large-scale datasets
that compare genetic variation to gene expression differences at the transcript (RNA-seq), translation
(ribosome profiling), and proteome (quantitative mass spectrometry) levels exist (Battle et al. 2014);
however, their analysis will require additional computational tools that can handle the precise mapping
of short (ribosome profiling) reads to variant (and diploid) genomes.
Supplemental materials for Chapter 2: Translation of many long non-coding RNAs resemble translation at 5’ leaders

MATERIALS AND METHODS

Ribosome Profiling

Ribosome profiling was adapted from Ingolia et al., 2011 and applied to a zebrafish developmental time course. 400-600 embryos per stage from 2-4 cells, 256 cells, 1000 cells, dome, shield, bud, 28 hours post fertilization and 5 days post fertilization (Kimmel et al. 1995) were washed with cold PBS, flash-frozen and stored at -80°C. Embryos were lysed by repeated micropipetting in 1.5 ml of cold polysome buffer (20 mM Tris-HCl pH 7.4, 250 mM NaCl, 15 mM MgCl2, 1mM dithiothreitol, 100 μg/ml CHX) with added 0.5% Triton X-100, 500 μg/ml GMP-PNP, 24 U/ml TurboDNase (Ambion AM2238), incubated with agitation for 10 min at 4°C, and clarified by centrifugation at 1300 rcf for 10 min at 4°C. For ribosome footprinting, 20 μl RNAseI (Ambion AM2294) was added to the 1.5 ml of supernatant and incubated for 30 min at 37°C, then stopped by chilling on ice and addition of 40 μl of SuperaseIn (Ambion AM2694). Footprinted samples were pelleted through a sucrose cushion (1 M sucrose in polysome buffer with added 100 U/ml SuperaseIn) by centrifugation at 260,000 rcf for 4.5 hours at 4°C, and resuspended in 800 μl 10mM Tris pH 7.4 with 1% SDS. RNA was purified by hot acid phenol/chloroform extraction and precipitated by standard ethanol precipitation. From this point, ribosome profiling Illumina-compatible sequencing libraries were prepared as previously described (Ingolia et al. 2011). Table S1 lists the primers and subtractive hybridization oligos corresponding to the most abundant rRNA contaminants that were determined in a pilot ribosome profiling experiment.
Sequencing and Mapping of Ribosome Protected Fragments (RPFs)

Ribosome profiling libraries were sequenced on an Illumina HiSeq 2000 (1 stage per lane, 44 bp reads), resulting in a total of 880 million reads (for an overview, see Figure A1). Following adapter sequence trimming, RPFs were compared to zebrafish rRNAs from the SILVA rRNA database (Quast et al. 2013) using Bowtie2 (Langmead and Salzberg, 2012; parameters: -N 1 –L 20 –k 20). Reads matching rRNA (~50%) were discarded. The remaining RPFs were mapped by Tophat2 (Trapnell et al., 2009; parameters: no indels, no novel junctions, -M -g 10) to a zebrafish developmental transcriptome (Pauli et al., 2012) and the Zv9 genome assembly, resulting in 317 million mapped reads. To obtain near-nucleotide resolution from ribosome profiling (Figure A2), RPFs aligning at annotated start and stop codons of RefSeq genes were subdivided by read length (Figure A2b). Approximate P-site position for each read-length was determined by inspection of coverage and phasing of the read’s left-most position relative to annotated start and stop codons. Offsets were determined to be +12 for 27-28 nt RPFs, +13 for 29-31 nt RPFs, and +14 for 32 nt RPFs (Figure A2a). Based on observable phasing over the coding sequences, RPFs between 27 and 32 nts (totaling 220 million) were deemed to be high-quality and used in subsequent analysis. The remaining RPFs were likely over or under digested, and were discarded. Library sizes between stages were normalized by the number of RPFs in each stage that mapped to annotated coding regions of RefSeq genes. Mouse embryonic stem cell (mESC) ribosome profiling data was obtained from Ingolia et al., 2011.

Construction of training and lncRNA data sets

The zebrafish training set was constructed from RefSeq genes in the Zv9/danRer7 zebrafish genome assembly. Only genes expressed at FPKM >1 (summed over the developmental transcriptome; Pauli et al., 2012) were used. Similarly, the mouse training set was based on RefSeq genes in the mm9 mouse genome assembly expressed at FPKM >1 in mouse ES cells (Guttman et al. 2010). ORFs were
defined as regions starting with either an ATG or CTG and ending with an in-frame stop codon. Three classes of ORFs were defined: i) the annotated coding sequences (CDSes) in the context of their respective transcripts, ii) all RPF-containing ORFs in transcript leaders in the context of the detached 5’ leaders, and iii) all RPF-containing ORFs in the transcript trailers in the context of the detached 3’ trailers (Figure 2.1). CDSes with trailers shorter than 100nt were not included. Due to the high number of truncated transcripts annotated in zebrafish, all ORFs in the zebrafish set were required to be at least 20 nts from the transcript edge. ORFs in leaders and trailers were filtered to ensure lack of any overlap with annotated RefSeq, Ensembl or XenoRefSeq coding regions.

For classification, IncRNAs were required to be expressed at >1 FPKM over the developmental time course (for zebrafish) and in ES cells (for mouse). As a few transcripts had a clear RPF-covered coding ORF, but lacked start/stop codons likely due to truncations in transcript assembly, ORFs were allowed to extend beyond the ends of transcripts. To account for possible transcript truncations, it was assumed that the start/stop of the ORF was at the edge and a pseudo-trailer of 10 nt was added to all transcripts when calculating IO scores (see below).

Classification

For each ORF, we used four metrics designed to distinguish between the three classes and capture the features of protein coding genes:

**TE** (Translational Efficiency) - (Density of RPFs within ORF)/(RNA expression). Density is the average sum of normalized RPFs over the embryonic time course within the ORF divided by the length of the ORF. RNA expression is the average FPKM of the locus over the embryonic time course containing this transcript.
IO (Inside vs Outside) - (Coverage inside ORF)/(Coverage outside ORF). Coverage refers to the number of nt positions having any RPF divided by the total number of nts inside or outside the ORF. A pseudo-count of 1 is added to both the inside and outside sums.

FL (Fraction Length) - (Length of ORF)/(Length of transcript). The fraction of the transcript covered by the ORF.

DS (Disengagement Score) - (RPFs over ORF)/(RPFs downstream). Number of RPFs inside ORF divided by number of RPFs downstream. A pseudo-count of one was added to both the ORF and downstream sums.

A random forest classifier (Breiman, 2001; implemented in the R package randomForest) was trained using these four metrics on the respective training sets. The three classes were weighted according to size, and standard options were used (500 trees, 2 variables per split). Classes were assigned to loci in order to minimize cross-mapping between coding and non-coding isoforms. If any ORF was classified as coding, the locus was considered coding. If not, the locus was considered leader-like if at least one ORF was classified as leader-like. Finally, if all ORFs were classified as trailer-like or no ORF had RPFs, the locus was classified as trailer-like.

Public database accession numbers

The ribosome profiling data is accessible at Gene Expression Omnibus (GEO) with accession no. GSE46512. The RNA-seq data was published previously (Pauli et al., 2012) and is available at Gene Expression Omnibus (GEO) (accession no. GSE32900, Subseries GSE32898).

Acknowledgements

We thank Jonathan Weissman and Nick Ingolia for helpful advice on ribosome profiling and for sharing mouse ESC ribosome profiling data, Mitchell Guttman for sharing results before publication, and Rahul Satija and Schraga Schwartz for discussions and comments on the manuscript. G-LC is supported
by a Howard Hughes Medical Institute International Student Fellowship. AP and EV are supported by Human Frontier Science Program (HFSP) postdoctoral fellowships. This work was funded by the NIH (JLR, AR and AFS).
Figure A.1. Schematic for sequencing and mapping of ribosome protected fragments. For details, see Materials and Methods.
Figure A.2. Determination of P-site positions of mapped reads from ribosome profiling. Single-nucleotide mapping of RPF reads. (A) Schematic for the offset from the leftmost position of reads of various lengths. Position of red nucleotide corresponds to the first position of the P-site in the 80s ribosome. (B) Leftmost position of RPF reads as a function of read length. Leftmost positions of the mapped RPFs are stratified over the start (left) and stop (right) codons of a RefSeq metagene. (C) P-site position of mapped reads following offset correction, stratified over the start (left) and stop (right) codon of a RefSeq metagene. A strong peak is observed at the codon position just prior to the stop codon.
Figure A.3. Ribosome profiling statistics of expressed RefSeq genes. (A) Sum of RPF reads associated with the CDSes of expressed RefSeq genes relative to minimum transcript expression levels. The plot shows the fraction of RefSeq CDSes (y-axis) associated with at least 1, 10 or 50 summed normalized RPFs. (B) RPF-coverage of the CDSes of expressed RefSeq genes relative to minimum transcript expression levels. The plot shows the fraction of RefSeq genes (y-axis) having at least 1, 10, 50, or 100% (color scale) of all nucleotides within their coding regions covered by one or more RPFs. RPFs are evaluated at single nucleotide level, and a single RPF can therefore only cover one nucleotide. For example, 68.19% of RefSeq genes expressed at >1 FPKM have 10% of their nucleotides covered by at least one RPF. Note that the fraction of genes above the thresholds (normalized sum (A) or coverage (B)) increase with increasing RNA expression levels (FPKM).
Figure A.4. Examples of minORFs in the leader sequence of Jarid2 in zebrafish (A) and mouse (B). minORFs were identified as distinct peaks of RPF enrichment over Start-Stop sequences in 5’ leaders. For the majority of minORFs, initiation at an upstream alternative Start codon is unlikely due to the absence of upstream in frame initiation codons. Additional support for the existence of minORFs comes from ribosome profiling experiments in mouse ESCs in the presence of harringtonine (Ingolia et al., 2011). While our RPF data does not show a distinctive enrichment of RPF reads at initiation codons (no drug pre-treatment), harringtonine leads to an accumulation of ribosomes over the initiating codon (Ingolia et al., 2011). RPF densities of harringtonine-treated mESCs resemble those of cycloheximide-treated mESCs (B), suggesting that initiation does indeed occur at the AUG immediately preceding the STOP codon.
Figure A.5. TOC distinguishes ORFs in 5’ leaders, CDSes and 3’ trailers of mouse genes. A training set was constructed from mouse RefSeq genes using 1) annotated CDSes (coding ORFs) in the context of the whole transcript, 2) RPF-containing ORFs in 5’ leader sequences in the context of the 5’ leaders, and 3) RPF-containing ORFs in 3’ trailers in the context of the 3’ trailers. Training was performed as described for zebrafish (see main text and Figure 3A), using the metrics Translational Efficiency (TE), Inside versus Outside (IO), Fragment Length (FL) and Disengagement Score (DS). The combination of the four metrics separates coding ORFs (blue), leaders (green) and trailers (red) of the training set. The density of each measure is shown along the axes. ORFs of well-annotated IncRNAs (black) are plotted on top of ORFs of the RefSeq training set and cluster with trailers and leaders. The ORF with the most bases covered by RPFs was used to represent each IncRNA. Shown are 18 mouse IncRNAs: Adapt33, Gas5, H19, Malat1, Meg3, Neat1, Rian, SNHG1, SNHG3, SNHG4, SNHG5, SNHG6, SNHG7, SNHG8, SNHG10, SNHG12, Tsix and Tug1.
Figure A.6. Accuracy of TOC as a function of expression when classifying CDSes versus ORFs within trailers and leaders (purple), and when classifying ORFs within trailers versus ORFs within leaders (black). Accuracy was calculated as the fraction of correct predictions divided by all out-of-bag (OOB) predictions at a given expression level.
Figure A.7. TOC-based classification of IncRNAs in zebrafish and mouse. TOC was used to classify ORFs in previously proposed zebrafish and mouse IncRNA sets (ZF1: Pauli et al., 2012; ZF2: Ulitsky et al., 2011; mouse ESCs: Guttman et al., 2011). ORFs of IncRNAs (black) are plotted on top of ORFs of the respective zebrafish or mouse RefSeq training sets (blue: CDS; green: ORFs of 5' leaders; red: ORFs of 3' trailers). The ORF with the most bases covered by RPFs was used to represent each IncRNA. LncRNAs without any RPFs are not shown in this figure. See Figure 3 for explanation of the four metrics.
Figure A.8. Length of ORFs subdivided by TOC-based classification across three sets of IncRNAs. Coding ORFs are significantly longer than trailer-like and leader-like ORFs.
Supplemental materials for Chapter 3: Sequence features mediate the conservation of upstream open reading frame repressiveness

MATERIALS AND METHODS

All data was manipulated and analyzed via a combination of existing software (Tuxedo suite tools (Trapnell et al. 2012) for short-read alignment and quantification; ViennaRNAfold (Lorenz et al. 2011) for secondary structure predictions), custom shell and Python scripts, and existing Python libraries. These analyses are fully documented in the included iPython Notebooks (Pérez and Granger 2007); a brief description of key data sources and methods follow:

Ribosome Profiling Data

Ribosome profiling data analyzed in this study had been previously published: human data from mitotic HeLa cells (Stumpf et al. 2013), mouse data from mouse embryonic stem cells (mESCs) (Ingolia et al. 2011), and zebrafish data from shield stage embryos (Chew et al. 2013).

Gene Annotations and Mapping

Ribosome profiling and RNA-Seq data were mapped (protocol adapted from (Ingolia et al. 2012)) to GRCh37/hg19, GRCm38/mm10 and Zv9 assemblies of the human, mouse and zebrafish genomes respectively, using gene annotations based on Ensembl Release 70, as compiled in Illumina’s iGenomes collection. Only one transcript per gene (as collated by UCSC’s gene-transcript-protein tables) was analyzed: if there were multiple annotated transcripts per gene, only the transcript with the longest CDS, and the longest 5’ UTR was used. Orthologous transcripts were determined from the list of high-confidence one-to-one orthologous genes in Ensembl Release 75.
*Upstream Open Reading Frames (uORFs)*

uORFs are defined without a minimum length, beginning with an AUG codon and ending with an in-frame stop codon, with the reasoning that an initiating ribosome does not “know” how long an ORF will be. With that reasoning, all AUGs in 5’ leaders were considered as potential uORF starts: e.g. two in-frame AUGs were considered as two separate uORFs that ended at the same stop codon. This was such that in the analyses where only transcripts with 1 uORF were considered, the uORF initiation site would be unambiguous.

*Initiation Context Primary Sequence*

For analyses of the effect of primary sequence on translation initiation, the initiation context was taken to be the sequence of the 20 nucleotides (10 upstream, 10 downstream) surrounding the start (AUG) codon. To construct a position-specific scoring matrix (PSSM) representing favorable initiation contexts, initiation contexts of CDSes of transcripts lacking uORFs were compiled, with contributions weighted by their translational efficiencies. Motifs representing this PSSM were created using Weblogo 3 (Crooks et al. 2004). Other initiation contexts (around AUGs elsewhere in transcripts) were scored using the log-likelihood transform of the constructed PSSM, using a background that was the nucleotide frequency of the entire PSSM.

*Initiation Context Secondary Structure*

For analyses of the effect of secondary structure on translation initiation, the initiation context was taken to be the 10 nucleotides upstream of the AUG codon, the AUG itself, and 22 nucleotides downstream (35 nucleotides total), which is approximately the footprint of an initiating ribosome. ViennaRNA RNAfold was used to determine the ensemble free energies of these initiation contexts, at 37°C for Human and Mouse, and 28°C for Zebrafish.
**uORF Depletion Analyses**

Codon and nucleotide frequencies were determined for each position of the 5’ leader with respect to the CDS start. Tri-nucleotide bias was calculated as the observed codon frequency a given position, normalized for the expected codon frequency given the nucleotide frequencies at that same position. While this bias is not a direct measure of depletion of tri-nucleotides, it is a worst a conservative underestimate of depletion. The observed frequency of uORFs ending at positions upstream of the CDS start was normalized against the frequency of uORF ends on shuffled 5’ leader sequences (each leader was shuffled a thousand times, yielding an expected frequency of shuffled sequences). This ratio was plotted with respect to the position from the CDS start.

**Linear modelling**

Linear models were constructed using ordinary least squares regression of Z-normalized sequence features. Multiple Z-normalized sequence features were combined by weighting for their individual coefficients from the linear modelling.

**Acknowledgements**

We thank A. Murray, C. Hunter, S. Ramanathan, E. Valen, A. Subramaniam, D. Koppstein, S.E. Calvo, M. Rabani and members of the Schier lab for helpful discussions. Many computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. This research was supported by funding from the Howard Hughes Medical Institute (International Student Fellowship, G.-L. C), the National Institutes of Health (K99 HD076935 from NICHD - Eunice Kennedy Shriver – National Institute of Child Health and Human Development, A.P. and A.F.S.) and the Human Frontier Science Program (HFSP) (A.P.).
### Table A.1: Numbers of transcripts analyzed

<table>
<thead>
<tr>
<th></th>
<th>Number of transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>23756</td>
</tr>
<tr>
<td>Total with uORFS</td>
<td>8167</td>
</tr>
<tr>
<td>Total with translated uORFs</td>
<td>5580</td>
</tr>
<tr>
<td>Filtered</td>
<td>12902</td>
</tr>
<tr>
<td>Filtered with uORFs</td>
<td>6507</td>
</tr>
<tr>
<td>Filtered with translated uORFs</td>
<td>4998</td>
</tr>
</tbody>
</table>

### Table A.2: Comparing TE of transcripts without uORFs against increasing number of uORFs

<table>
<thead>
<tr>
<th>No uORFs vs.</th>
<th>KS test distance</th>
<th>KS test p-value</th>
<th>Mean percentage difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 uORF</td>
<td>0.086178</td>
<td>2.2031e-11</td>
<td>-17.95 %</td>
</tr>
<tr>
<td>2 uORFs</td>
<td>0.147505</td>
<td>5.03449e-18</td>
<td>-29.87 %</td>
</tr>
<tr>
<td>3 uORFs</td>
<td>0.152603</td>
<td>6.1665e-13</td>
<td>-33.74 %</td>
</tr>
<tr>
<td>4+ uORFs</td>
<td>0.152603</td>
<td>6.32937e-26</td>
<td>-38.71 %</td>
</tr>
<tr>
<td>With uORFs</td>
<td>0.120513</td>
<td>5.08809e-35</td>
<td>-26.55 %</td>
</tr>
</tbody>
</table>

### Supplementary Figures

All data from mES cells (Ingolia et al. 2011) unless otherwise stated.

**Figure A.9**
Figures A.9 & A.10. Human (HeLa) and zebrafish (Shield Stage Embryos) versions of Figures 3.1a-d.

a. Sequence motif representing efficient initiation at CDSes is used to score initiation contexts at all ORFs. Height on vertical axis represents weighted relative entropy (WRENT).

b. Cumulative distribution of WRENT scores around AUGs at various positions in uORF-containing coding transcripts. Sequences around AUGs at the start of CDSes are more favorable for translation initiation than sequences around AUGs everywhere else on the transcript, including at uORFs.

c. Meta-profiles of predicted secondary structure ensemble free energies (in sliding 35 nucleotide window) around uORF and CDS starts. Vertical red line at -10 nucleotides indicates position where secondary structure was considered for further analyses. While profiles over human and mouse uORFs separate based on number of uORFs per transcript (more uORFs per transcript: less stable secondary structure around uORF starts), zebrafish uORFs do not.

d. Cumulative distribution of secondary structure ensemble free energies of uORF initiation contexts in transcripts with varying number of uORFs. Lower average secondary structure (higher ensemble free energy) is observed with increasing number of uORFs in human, but not zebrafish transcripts. Average secondary structure at CDS initiation contexts is indicated by green dashed line.
Figure A.11 Genome-wide distribution of uORF sequence features: uORF initiation contexts

a. Sequence motif of CDS initiation context, similarly derived as WRENT score motif, but unweighted; analogous to Kozak sequence motif.

b. Meta-profiles of predicted secondary structure ensemble free energies around CDS starts and ends, in sliding windows of various sizes. Due to sequence constraints at starts and ends of CDSes, window-size dependent features appear upstream of the CDS start and ends (the window-size bumps ending at 0 in both plots).

c. Meta-profiles of predicted secondary structure (in sliding 35 nucleotide window) over uORF and CDS ends, for transcripts with varying number of uORFs. Secondary structure around uORF ends also feature uORF-number dependence, consistent with there being selection against total amount of secondary structure in 5' leaders.

d. Comparing secondary structure around all AUGs at various positions in transcripts. AUGs at CDS starts tend to have more stable secondary structure (lower ensemble free energy) than AUGs elsewhere in the transcript.
Figure A.12
Figure A.12. (Continued) Genome-wide distribution of uORF sequence features: uORF-CDS distance a-c. AUG (green) and stop codon (magenta) trinucleotide biases in the 5' leader as a function of distance with respect to CDS start, in all 3 frames, for all 3 vertebrates (a-c). Individual positions are scatter-plotted, lines show moving trends in the individual frames. AUGs, but not stop codon trinucleotides are specifically depleted near the CDS start, in all 3 frames.

d-f. Frequency of uORF ends, as observed in transcriptomes (dark magenta), and expected from shuffled/reattached 5' leaders (light magenta), for all 3 vertebrates. Individual positions are scatter-plotted, lines show moving trends.

g. uORF ends bias (ratio of observed to expected) with respect to CDS start. uORF ends are specifically depleted in the 5' leader near the CDS, most so in zebrafish transcripts.

h-m. Individual plots for uORF end bias (h-j) and depletion (k-m), for human (h, k), mouse (i, l) and zebrafish (j, m). Individual positions are scatter-plotted, lines show moving trends.
Figure A.13. Prevalence and General Effects of uORFs

a. Histograms of number of uORFs per transcript, for human, mouse and zebrafish filtered transcripts. Approximately half of mammalian transcripts and two-thirds of zebrafish transcripts contain uORFs.

b-d. Metagene profile of ribosome profiling reads around CDS start and end, for various ribosome profiling datasets (b. Human – HeLa cells; c. Mouse – mES cells; d. Zebrafish – Shield stage embryos). Triplet phasing is detected, as are differing 5’ and 3’ biases in ribosome profiling reads around the starts and ends (caused by differing cycloheximide treatment).

e. Distribution of uORF and CDS TEs for transcripts with one non-overlapping uORF. CDS TEs are more tightly-distributed, with higher average TE, although many uORFs have TEs exceeding that of CDSes.

f. Scatter plot of uORF TEs for transcripts with two non-overlapping uORFs. Good correlation (Pearson’s r = 0.55, p < 10^{-15}; Spearman’s rho = 0.48, p < 10^{-11}) between TEs from cognate uORFs. Red line is LOWESS fit through scatter points.

g. Distribution of uORF 1 and uORF 2 TEs for transcripts with two non-overlapping uORFs.
Figure A.14. (Continued) Human (HeLa) and Zebrafish (Shield Stage Embryos) versions of Figure 3.3

a-c. LOWESS fits of relationship between initiation context sequence (a), secondary structure (b), and uORF-end position w.r.t. CDS (c) against uORF TE, CDS TE and uORF/CDS TE. uORF repressiveness (estimated from uORF/CDS TE) increases with more favorable sequence, decreased secondary structure at uORF initiation context, and reduced uORF-CDS distance.

d. Scatter plot of linear modelling of uORF sequence features against uORF/CDS TE; red line is LOWESS fit line. Good correlation for HeLa cell data between a score that combines all measured uORF sequence features with uORF repressiveness ($r = 0.24; p < 10^{-11}$, 14.2-fold over range in uORF-repressiveness), not for zebrafish shield stage data (likely due to the lack of secondary structure influence).

e. Scatter plot of linear modelling of uORF and CDS sequence features against CDS TE; red line is LOWESS fit line. Good correlation between a score that combines all measured sequence features with CDS TE (HeLa cells: $r = 0.33, p < 10^{-20}$ over a 12.6-fold range in CDS TE; Shield stage embryos: $r = 0.30, p < 10^{-7}$ over 6.5-fold range in CDS TE).

Note: much of the zebrafish shield-stage embryo data is likely skewed by prominent expression and efficient translation of many (~100) zinc-finger proteins that have near identical initiation contexts.
Figure A.15. Individual scatter plots for Figures 3.3a-c: Quantifying effect of uORF sequence features. Red lines are LOWESS fits.

For transcripts with one non-overlapping uORF, scatter plots of:

- **a-c.** CDS TE (a), uORF TE (b) and uORF/CDS TE (uORF repressiveness) (c) against uORF WRENT score.

- **d-f.** CDS TE (d), uORF TE (e) and uORF/CDS TE (uORF repressiveness) (f) against uORF initiation context secondary structure ensemble free energy.

- **g-i.** CDS TE (g), uORF TE (h) and uORF/CDS TE (uORF repressiveness) (i) against uORF end position w.r.t. CDS.
Figure A.16.
Figure A.16. (Continued) Quantifying effect of CDS sequence features

Red lines are LOWESS fits. For transcripts with one non-overlapping uORF.

**a, f.** LOWESS fits of relationship between CDS WRENT score (a) and CDS initiation context secondary structure ensemble free energy (f) against uORF TE, CDS TE and uORF/CDS TE.

**b-d.** Individual scatter plots of CDS TE (b), uORF TE (c) and uORF/CDS TE (d) against CDS WRENT score

**g-i.** Individual scatter plots of CDS TE (b), uORF TE (c) and uORF/CDS TE (d) against CDS WRENT score

**e, j.** uORF vs CDS WRENT scores (e) or initiation context secondary structure ensemble free energies (j). No correlation between uORF and CDS WRENT scores, small but significant correlation between uORF and CDS secondary structure ensemble free energies ($r = 0.16, p < 10^{-5}$).
Figure A.17. Linear modelling with CDS sequence characteristics

a. Scatter plot of CDS TE against CDS WRENT score, for transcripts without uORFs. While there is a decent correlation ($r = 0.16, p < 10^{-37}$, over 2.4-fold range of CDS TE), it is far from perfect, suggesting that CDS WRENT scores do not account for all the variation in CDS TE, nor do they over-fit the data.

b. Scatter plot of CDS TE against CDS secondary structure ensemble free energy. Decent correlation ($r = 0.11, p < 10^{-19}$, over 2.3-fold range of CDS TE).

c. Scatter plot of CDS TE against transcript expression. Lack of correlation between CDS TE and transcript expression levels, as might be expected.

d. Scatter plot of CDS TE against score combining CDS WRENT score and initiation context secondary structure ensemble free energy, for transcripts lacking uORFs. Significant correlation ($r = 0.19, p < 10^{-51}$) over 3.1-fold range of CDS TE, suggesting that effect of CDS WRENT score and initiation context secondary structure is somewhat additive.

e. Scatter plot of CDS TE against score combining CDS WRENT score and initiation context secondary structure ensemble free energy, for transcripts with one non-overlapping uORF. Significant correlation as well, especially given the smaller amount of data compared to transcripts lacking uORFs ($r = 0.24, p < 10^{-12}$), similar fold-change (3.4-fold), suggesting that linear model approach is valid.

f. Scatter plot of uORF repressiveness against score combining uORF and CDS sequence features. Significant linear correlation ($r = 0.27, p < 10^{-15}$) over 12.2-fold range of uORF repressiveness. By normalizing coefficients against their respective units, we can determine how much a change in WRENT score or secondary structure affects uORF repressiveness, as well as compare the relative contributions. Both uORF and CDS WRENT scores affect uORF repressiveness similarly (similar magnitudes of 1.02 and -0.715 respectively, although different signs), while secondary structure around uORF starts (1.38) has a greater effect than that around CDS starts (-0.49).
Figure A.18. Conservation of TE and uORF/CDS regulatory relationship

a-c. Scatter plot of CDS TEs of orthologs transcripts in all pairwise comparisons of human, mouse and zebrafish ribosome profiling data. Mouse-human (a), mouse-zebrafish (b), human-zebrafish (c). Individual scatter-points are colored by whether orthologs have uORFs (green: without uORFs; salmon: with uORFs), or differ by whether they have uORFs (grey). Good correlation in CDS TEs observed only between mouse and human ($r = 0.45$, $p < 10^{-294}$). Correlation is largely independent of uORF status of transcripts.

d-e. Scatter plot of ratio of 5’ leader translation to CDS translation, for mouse-zebrafish (d) and human-zebrafish (e). Correlation is good ($r = 0.38$; $p < 10^{-69}$ and $r = 0.35$; $p < 10^{-54}$ respectively), although not as good as correlation between mouse and human in Figure 4a.
Figure A.19. Conservation of uORF sequence features, correlation of divergence

a-f. Scatter plots of uORF sequence features for orthologous transcripts between mouse & zebrafish (a-c) and human & zebrafish (d-f), for initiation context sequence (a, d), secondary structure (b, e), and uORF-end position w.r.t. CDS (log magnitude transformed; c, f). Multiple uORFs were considered in this analysis, but only the uORFs with the sequence features most expected to repress downstream translation per transcript were considered. Good correlation observed between orthologous transcripts for all uORF sequence features ($r = 0.45; p < 10^{-104}$ for initiation context sequence, $r = 0.34; p < 10^{-55}$ for initiation context secondary structure, $r = 0.49, p < 10^{-122}$ for uORF-CDS distance). Individual points are colored for their divergence: red for higher in human, blue for higher in mouse, green for higher in zebrafish.

g-h. Scatter plot of linear modelling of combined divergence of uORF sequence characteristics against divergence of ratio of 5' leader to CDS ribosome profiling read density, between mouse & zebrafish (g) and human & zebrafish (h) orthologous transcripts. Very good correlations ($r = 0.59, p < 10^{-68}$ and $r = 0.51, p < 10^{-48}$) observed.


89


