Visualising ribosome profiling and using it for reading frame detection and exploration of eukaryotic translation initiation.

by

Audrey Mannion Michel

Thesis in fulfilment for the degree of PhD (Science)

National University of Ireland, Cork.

School of Biochemistry and Cell Biology

October 2013

Head of School: Professor David Sheehan

Supervisors: Professor John Atkins, Dr Pavel Baranov

Contents

Declaration

This thesis is my own work and has not been submitted for another degree, either at University College Cork or elsewhere.

Signed:_____

Audrey Mannion Michel

Abbreviations

Abbreviation	Term
GWIPS	Genome Wide Information on Protein Synthesis
GWIPS-viz	Genome Wide Information on Protein Synthesis visualised
RPFs	ribosome protected fragments or footprints
ribo-seq	ribosome profiling
mRNA-seq	randomly fragmented mRNA
CDS	coding sequence
ORF	open reading frame
pORF	protein coding open reading frame
uORF	upstream open reading frame
nORF	non-upstream open reading frame
PTS	Periodicity Transition Score
CSCPD	Cumulative Sub-Codon Proportion Difference
TIS	Translation Initiation Site
uTIS	upstream Translation Initiation Site
dTIS	downstream Translation Initiation Site

Acknowledgements

I wish to take the opportunity to thank all my colleagues and collaborators who have helped me on various aspects of the work described in this thesis.

In particular I wish to thank Dr Kingshuk Roy Choudhury, Dr Andrew Firth and Dr Nicholas Ingolia who provided invaluable help and feedback for my ”dual coding” story.

Thanks also to Dr Dmitry Andreev for his suggestions pertaining to the ”translating initiation” work.

I wish to thank all my colleagues who have collaborated with me on the development and realisation of GWIPS-viz. In particular Gearoid Fox who helped get GWIPS-viz kick-started. Also thanks to Christof De Bo, Stephen Heaphy, Claire Donohue and James (aka Paddy) Mullan. A big thank you to Patrick O'Connor, who has also contributed to GWIPS-viz and with whom it is always great to share tips for analysing ribosome profiling data.

I wish to thank all members of the LAPTI and Recode labs who have made the last 3 years a very enjoyable and rewarding experience. Thank you Dr Anmol Kiran and Dr Virag Sharma for all your help when I first joined the lab. Thanks also to Dr Gary Loughran, Dr Ivaylo Ivanov, Martina Yordanova, Arthur Coakley, Dr Christophe Penno, Anjali Pai, Ioanna Tzani, Dr Ming-Yuan.

My grateful thanks to Professor John Atkins for giving me the opportunity to work on such interesting projects and for sharing his considerable knowledge on developments in the field.

Finally, my deep gratitude to Dr Pavel Baranov for all his help throughout my PhD. I may have tried his patience on occasion, but he always took the time to explain concepts in an entertaining and engaging manner. Thank you.

Abstract

Ribosome profiling (ribo-seq) is a recently developed technique that provides genome-wide information on protein synthesis (GWIPS) *in vivo*. The high resolution of ribo-seq is one of the exciting properties of this technique. In Chapter 2, I present a computational method that utilises the sub-codon precision and triplet periodicity of ribosome profiling data to detect transitions in the translated reading frame. Application of this method to ribosome profiling data generated for human HeLa cells allowed us to detect several human genes where the same genomic segment is translated in more than one reading frame.

Since the initial publication of the ribosome profiling technique in 2009, there has been a proliferation of studies that have used the technique to explore various questions with respect to translation. A review of the many uses and adaptations of the technique is provided in Chapter 1.

Indeed, owing to the increasing popularity of the technique and the growing number of published ribosome profiling datasets, we have developed GWIPS-viz (`http://gwips.ucc.ie`), a ribo-seq dedicated genome browser. Details on the development of the browser and its usage are provided in Chapter 3.

One of the surprising findings of ribosome profiling of initiating ribosomes carried out in 3 independent studies, was the widespread use of non-AUG codons as translation initiation start sites in mammals. Although initiation at non-AUG codons in mammals has been documented for some time, the extent of non-AUG initiation reported by these ribo-seq studies was unexpected. In Chapter 4, I present an approach for estimating the strength of initiating codons based on the leaky scanning model of translation initiation. Application of this approach to ribo-seq data illustrates that initiation at non-AUG codons is inefficient compared to initiation at AUG codons. In addition, our approach provides a probability of initiation score for each start site that allows its strength of initiation to be evaluated.

Chapter 1

Ribosome profiling: A Hi-Def monitor for protein synthesis at the genome-wide scale.

This chapter has been published as a review on ribosome profiling in Wiley Interdiscip Rev RNA. 2013 Sep;4(5):473-90 (see Appendix 5.1)

Ribosome profiling or ribo-seq is a new technique that provides genome-wide information on protein synthesis (GWIPS) *in vivo*. It is based on the deep sequencing of ribosome protected mRNA fragments allowing the measurement of ribosome density along all RNA molecules present in the cell. At the same time, the high resolution of this technique allows detailed analysis of ribosome density on individual RNAs. Since its invention, the ribosome profiling technique has been utilised in a range of studies in both prokaryotic and eukaryotic organisms. Several studies have adapted and refined the original ribosome profiling protocol for studying specific aspects of translation. Ribosome profiling of initiating ribosomes has been used to map sites of translation initiation. These studies revealed the surprisingly complex organization of translation initiation sites in eukaryotes. Multiple initiation sites are responsible for the generation of N-terminally extended and truncated isoforms of known proteins as well as for the translation of numerous open reading frames (ORFs) upstream of protein coding ORFs. Ribosome profiling of elongating ribosomes has been used for measuring differential gene expression at the level of translation, the identification of novel protein coding genes and ribosome pausing. It also provided data for developing quantitative models of translation. Although only a dozen or so ribosome profiling datasets have been published so far, they have already dramatically changed our understanding of translational control and have led to new hypotheses regarding the origin

of protein coding genes.

1.1 Introduction

The race for the completion of the human genome yielded a by-product that is probably more important for modern biology than the goal of the project itself – cheap and powerful technologies for sequencing DNA. These technologies shifted the focus of researchers from studying individual molecules and pathways to studying the whole composition of molecules inside the cell. However, most of the popular high-throughput techniques provide only static information on the composition of the cell. For example, proteomics approaches such as mass-spectrometry give information on the composition of a proteome, while RNA-seq captures information on the composition of a transcriptome. An assumption is used whereby the abundance of transcripts can be interpreted as a measure of transcription levels. This assumption is problematic because of the varying stability of RNA transcripts. Because of the high variability in protein molecule half-lives, inferring gene expression levels from protein abundance is even more problematic. A high concentration of a particular protein in the cell does not necessarily mean that the corresponding gene is being highly expressed at the moment of measurement.

Until recently, no simple high-throughput technique existed for measuring gene expression at the level of translation. The situation has changed with the advent of the ribosome profiling technique developed in the laboratory of Jonathan Weissman at UCSF (Ingolia et al., 2009). By providing Genome Wide Information on Protein Synthesis (GWIPS), ribosome profiling filled the technological gap existing between our abilities to quantify the transcriptome and the proteome (Weiss and Atkins, 2011) (see Fig. 1.1). It is now possible not only to detect RNA and protein molecules in the cell, but also determine which protein molecules are being synthesized in the cell at any given moment and therefore quantitatively measure the immediate reaction of the cell to a change in its internal environment.

The technology is the product of a propitious marriage of an existing methodology with massive parallelization offered by second-generation sequencing platforms (Ingolia et al., 2009). The ability of ribosomes to protect mRNA fragments from nuclease digestion has been used since the 1960s (Steitz, 1969). In ribosome profiling (see Fig. 1.2), this procedure is carried out for the entire cell lysate generating a pool of ribosome protected fragments or footprints (RPFs). Recovered footprints are converted to a format suitable for massively parallel sequencing. Analysis of the resultant sequences allows the quantification of ribosomes translating mRNAs at a genome-wide scale (Ingolia et al., 2009; Ingolia,

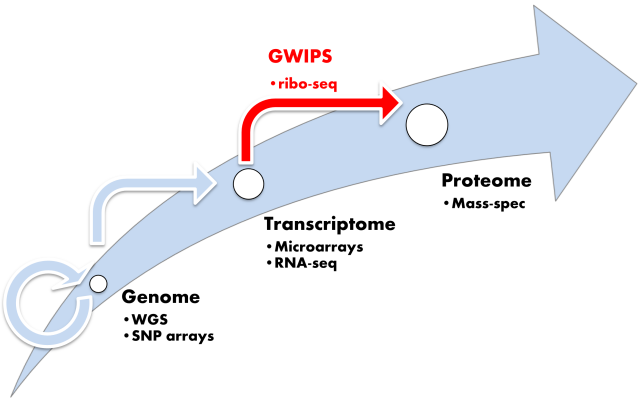

Figure 1.1: *The emplacement of GWIPS (Genome Wide Information on Protein Synthesis) and the role of ribo-seq in characterizing the molecular status of the cell.*

2010; Ingolia et al., 2012). Therefore ribosome profiling can be used for measuring gene expression at the translational level. However, this was already possible with polysome profiling where a pool of translated mRNAs is isolated from the polysome fraction of a sucrose gradient. This approach, where the abundance of transcripts in a polysome fraction is assessed either with RNA-seq or microarray techniques, has become a popular way of identifying genes whose expression is under translational control (Arava et al., 2003; Larsson et al., 2010; Genolet et al., 2008; Rajasekhar et al., 2003). The real power of ribosome profiling in comparison with such approaches is in its ability to obtain position-specific information regarding ribosome locations on mRNAs. This is very important for several reasons. The association of an mRNA transcript with ribosomes does not necessarily mean that the main open reading frame of this mRNA is translated. Ribosomes could stall on an mRNA transcript without producing a protein. Translation could occur at ORFs other than the main protein coding open reading frame (pORF).

Because ribosome profiling reveals the exact positions of ribosomes on an mRNA transcript, two major variants of the technique have been developed: ribosome profiling of elongating ribosomes and ribosome profiling of initiating ribosomes. Elongating ribosomes can be blocked with antibiotics that inhibit either translocation (e.g. cyclohexamide (Ingolia et al., 2009) and emetine (Ingolia et al., 2011)), peptidyl transfer (e.g. chloramphenicol), or by thermal freezing (Oh et al., 2011). Information on the positions of initiating ribosomes can be obtained either by the direct blocking of initiating ribosomes with specific drugs, (e.g. harringtonine (Ingolia et al., 2011) and lactimidomycin (Lee et al.,

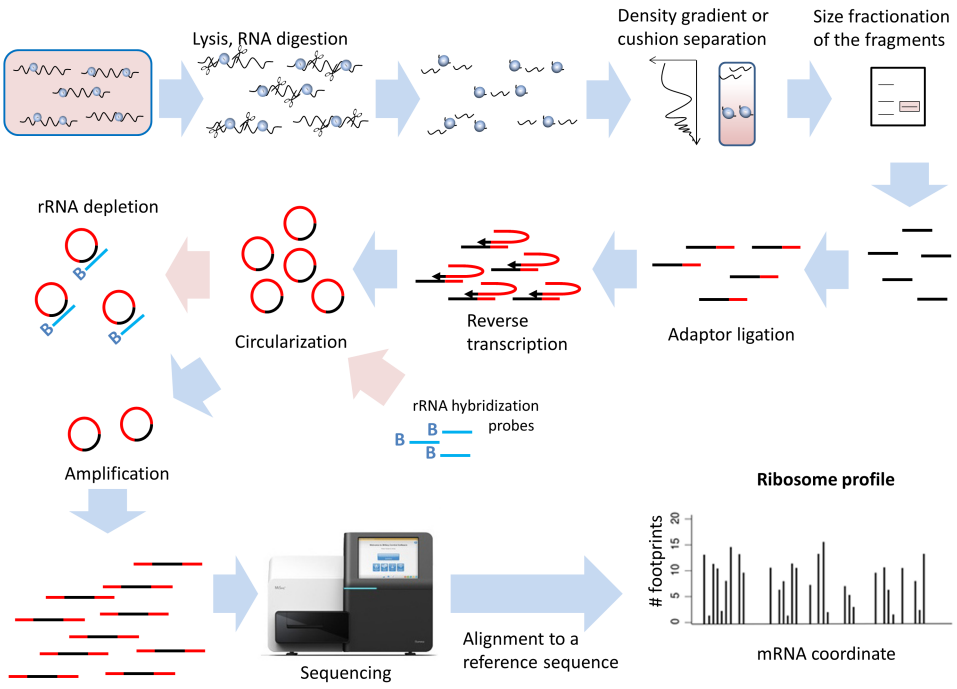

Figure 1.2: *Outline of the major steps of the ribosome profiling protocol as described in Ingolia et al. (2012). The experimental part of the protocol requires 7 days. Modifications of the protocol have been made in several other studies and commercial kits for ribosome profiling are currently available.*

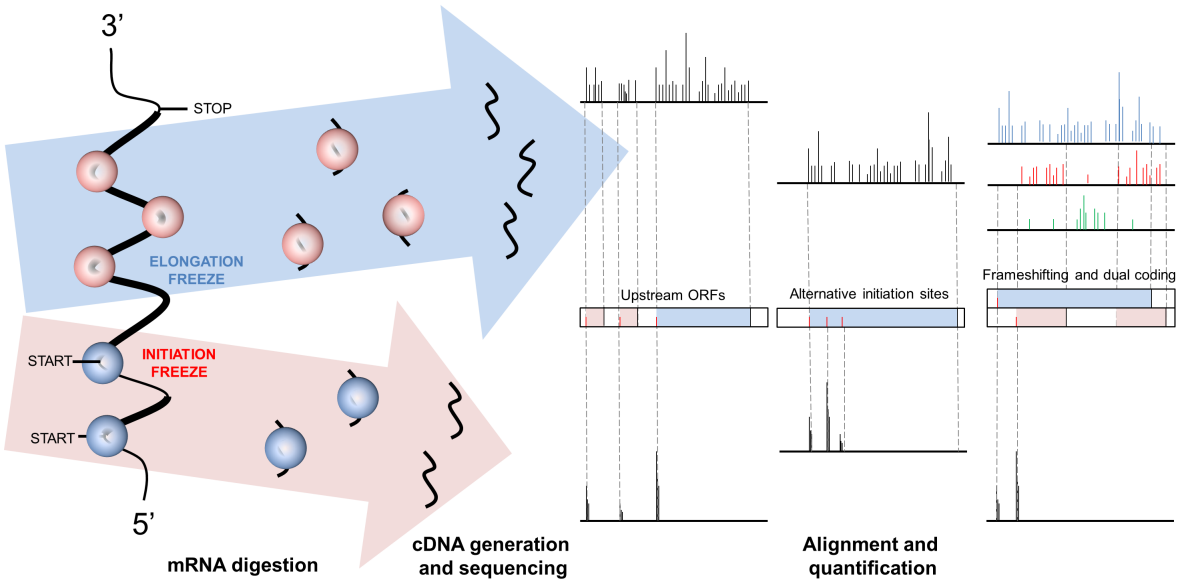

Figure 1.3: *Two main ribo-seq strategies: ribosome profiling of elongating ribosomes (top, blue arrow) and ribosome profiling of initiating ribosomes (bottom, light-pink arrow). In both cases, the freezing of ribosomes at specific stages of translation is followed by the degradation of mRNA unprotected by ribosomes and subsequent preparation of ribosome footprint cDNA libraries and their sequencing. The right-hand side of the figure illustrates how the data obtained with these ribo-seq techniques can be analysed for the identification of uORFs (shown as pink areas in the left plot), protein isoforms with alternative N-termini (middle plot), and nORFs embedded within annotated coding regions and recoding events (far-right plot).*

2012)) or by enriching elongating ribosomes near the starts by blocking them with cycloheximide following pre-treatment with puromycin that causes premature termination (Fritsch et al., 2012). Figure 1.3 illustrates how these two distinct strategies can be used for the characterization of different phenomena. For certain applications each approach has its own advantages, e.g. information on initiating ribosomes cannot be used for the detection of ribosomal frameshifting, while the detection of internal sites of initiation is impractical without this information. Often, these approaches complement each other and can be very powerful if used in parallel as has been demonstrated in a recent study (Stern-Ginossar et al., 2012). For clarity, and to emphasize the advantages of each strategy, this review is split into two main sections addressing each strategy separately.

1.2 Ribosome profiling of elongating ribosomes

The objective of using ribosome profiling is to generate a snapshot of the mRNAs that are being translated, capturing the exact locations of translating ribosomes and their densities on these mRNAs. It is imperative that the RPFs recovered from cell extracts accurately reflect the *in vivo* status of translation at the time of the experiment. Depending on the organism, the tissue and the objective of the study, the cell lysate preparation will vary. To faithfully capture elongating ribosomes in their *in vivo* translational positions, the majority of ribo-seq experiments to date have treated cells with translation elongation inhibitors to immobilize polysomes prior to cell lysis, followed by nuclease digestion. The nuclease-resistant RPFs are then recovered, converted to cDNA libraries and sequenced using massively parallel platforms (see Figs. 1.2 and 1.3).

The elongation inhibitor cycloheximide (Godchaux et al., 1967) has been used in nearly all of the elongating ribosome profiling studies carried out in eukaryotic cells to date. However, simple liquid nitrogen freezing as well as other antibiotics such as emetine in eukaryotes and chloramphenicol in bacteria have also been used (Ingolia et al., 2009; Ingolia et al., 2011; Oh et al., 2011). It is likely that the repertoire of translation inhibitors used in ribosome profiling studies will grow in the future, such as drugs that interfere with translation by stabilizing particular ribosomal conformations and thereby provide advantages for specific applications. It has been observed, for example, that the length of RPFs could be drug dependent (Ingolia et al., 2011).

For details of the ribosome profiling experimental protocol see (Ingolia, 2010; Ingolia et al., 2012) as well as the methods section of the primary research articles described in this review. In this section we will review the various applications of ribosome profiling of elongating ribosomes such as measuring differential gene expression, estimating global and local translation elongation rates and the identification of novel genes and the products of their expression.

1.2.1 Differential gene expression using ribosome profiling

The ability to detect changes in the expression of genes is essential for understanding the genetic determinants of phenotypical behaviour and the molecular response of the cell to changing conditions. For more than a decade, microarray techniques (Schena et al., 1995), and more recently RNA-seq (Mortazavi et al., 2008), have been used for measuring differential gene expression. However, the correlation between mRNA abundance and protein levels is insufficient for predicting protein expression based on mRNA concentrations (for discussion see de Sousa Abreu et al., 2009; Plotkin, 2010). Measurements of global pro-

tein and mRNA compositions have demonstrated that an important factor determining the cellular protein abundance in mammalian cells is its rate of translation (Schwanhäusser et al., 2011). As discussed in the Introduction, to obtain information on translated mRNAs, microarray and RNA-seq techniques can be applied to quantify the mRNAs bound to ribosomes by isolating the mRNAs from the polysome fractions of sucrose gradients. However, such a methodology is inaccurate. Two mRNA molecules in a polysome fraction could be translated at different rates, or not translated at all. This could occur, for example, when a ribosome is stalled on an mRNA or translation is limited to upstream open reading frames (uORFs) that often prevent translation of the main protein product ORF (removal of the monosomal fraction solves this issue for mRNAs inhibited with a single short uORF). Polysomal profiling also cannot provide information on the exact number of ribosomes on mRNAs. Since ribo-seq allows localisation of the ribosomes, this information can be assessed, therefore making it a preferential approach for differential gene expression. The very first ribosome profiling study showed a 100 fold range difference in the density of ribosome footprints across different yeast transcripts expressed at a relatively high level (Ingolia et al., 2009). The high variation in ribosome densities and the ability of ribo-seq to detect the variation, demonstrate the advantages of ribo-seq in comparison with prior approaches.

Most of the published ribosome profiling studies borrowed computational approaches from RNA-seq analysis for measuring differential gene expression levels. For a number of reasons, specifically discussed at the end of this subsection and illustrated with examples throughout the entire review, treating the density of ribosome footprints on an mRNA transcript as a direct measure of its translation may generate a number of artefacts. It is likely that specialized tools for the analysis of gene expression using ribo-seq will be developed in the future. In the meantime, however, adapting RNA-seq computational approaches is sensible for obtaining approximate information. Indeed, by using such approaches, a small number of ribosome profiling studies have already provided significant insights into certain important aspects of translational control.

The effects of stress conditions on translation

Protein synthesis is an energetically expensive anabolic process and therefore it is expected to be sensitive to the available nutrition, in particular, amino acids. To test the ability of ribo-seq to characterize changes in protein synthesis in response to starvation, Ingolia et al. (2009) carried out ribosome profiling on yeast cells after 20 minutes of amino acid deprivation. Changes at the translational level were detected in approximately one-third of the 3,769 genes that had sufficient coverage (see examples in Fig. 1.4). For 291 genes,

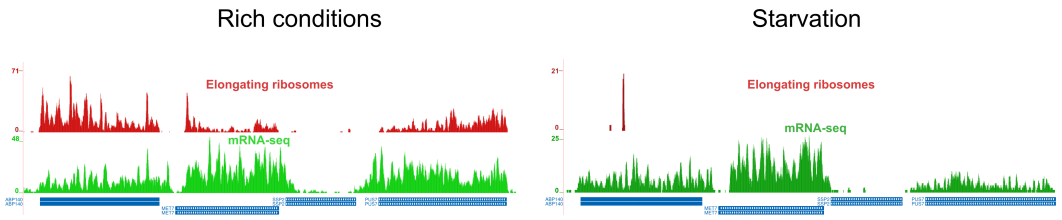

Figure 1.4: *Ribo-seq (red) and mRNA-seq (green) coverage plots for the S. cerevisiae genome locus containing ABP140, MET7, SSP2 and PUS7 genes obtained with GWIPS-viz (http://gwips.ucc.ie/) using data from (Ingolia et al., 2009). Under starvation conditions (right), ABP140, MET7 and PUS7 are transcribed, but not translated.*

up or down-regulation was found to be greater than two-fold. In particular, the translation of *GCN4* was found to increase seven-fold. While the translational regulation of *GCN4* in response to amino acid deficiency is well established and studied (Hinnebusch, 2005), this effect was not observed with a previous polysome profiling study (Smirnova et al., 2005). This example illustrates the clear advantage of ribosome profiling over polysome profiling as it allows the discrimination of mRNAs with efficiently translated coding regions from mRNAs where only the 5'UTRs are translated.

Geraschenko et al. (2012) used a similar idea to explore the translational response of *Saccharomyces cerevisiae* to oxidative stress. Yeast cells were treated with hydrogen peroxide and ribo-seq and RNA-seq were carried out in parallel 5 minutes and 30 minutes after the treatment. Many genes whose expression was altered at the transcriptional and translational level have been identified with this approach. The number of genes whose expression was changed greatly increased with the prolonged treatment: the transcript abundance of 116 genes was affected after 5 minutes and 1,497 genes after 30 minutes with similar numbers obtained for genes whose translation was altered. Interestingly, they reported several transcribed but translationally quiescent genes whose translation is activated upon oxidative stress, e.g. the *Srx1* gene which encodes sulfiredoxin. The dataset of translationally regulated genes was compared with a previous study that used polysome profiling for this purpose (Shenton et al., 2006). About 70% of translationally regulated genes found with polysome profiling were not confirmed with ribosome profiling. Geraschenko et al. (2012) argue that such a large discrepancy could be due to the inability of polysome profiling to discriminate the translation of main ORFs from regulatory uORFs.

While this review was in preparation, two more studies were published that explored translational response to heat shock (Shalgi et al., 2013) and to proteotoxic stress (Liu et al., 2013). Shalgi et al. (2013) found that 2 hrs of severe heat stress caused an accumulation of ribosomes in the first ~200nt of ORFs in mouse and human cells. Liu et al. (2013) found that proteotoxic stress in HEK293 cells resulted in elongation pausing

primarily near the site where nascent peptides emerge from the ribosomal exit tunnel. Both studies discuss the role played by chaperones in translation elongation and that early elongation pausing is triggered when chaperones are sequestered to the misfolded protein response as a result of cellular stress.

The role of miRNAs in translational control

The discovery of RNA interference (RNAi) opened up a debate regarding the potential mechanisms of translational regulation with miRNAs (see reviews Valencia-Sanchez et al., 2006; Fabian et al., 2010; Bartel, 2004). While many examples of RNAi inhibition of protein synthesis have been reported as well as cases of translational up-regulation (Vasudevan et al., 2007), the global contribution of RNAi to translational control is unclear. To address this issue, Guo et al. (2010) employed ribosome profiling in conjunction with mRNA-seq (alkaline degraded mRNA yielding fragments of a size similar to ribosome footprints) to discriminate between changes in mRNA abundance and rates of protein production caused by the expression of specific miRNAs. The experiments were carried out in human HeLa cells using exogenous miRNAs (Guo et al., 2010). Genes with at least one miRNA target site in their 3' UTRs were repressed by the addition of the corresponding miRNA resulting in fewer mRNA-seq fragments and correspondingly fewer RPFs. A very modest decrease in translational efficiency was observed for messages with miRNA target sites compared to those without. Therefore, Guo et al. concluded that, at the global level, miRNA interference affects mostly mRNA abundance with only a marginal effect on translation (Guo et al., 2010).

However, as discussed by Janas and Novina (2012), this study assessed translation and mRNA levels after 12–32 hrs, at which point only the downstream effects of miRNA function may have been observed. To study gene expression responses at earlier time points, Bazzini et al. (2012) carried out combined ribo-seq and mRNA-Seq analysis to study the global effects of a particular miRNA in zebrafish. For this purpose they focused on targets of miR-430 miRNA which is expressed at the onset of zygotic transcription and had been previously shown to promote deadenylation and degradation of maternal transcripts at 5 and 9 hours post fertilization (hpf) (Giraldez et al., 2006). The ribosome occupancy and mRNA levels of miR-430-targeted mRNAs were measured at timepoints before (2 hpf) and after (4 hpf and 6 hpf) the induction of miR-430 expression. At 4 hpf the ribosome density along miR-430-targeted mRNAs was uniformly decreased without a corresponding decrease in the mRNA. Yet 70% of the targets translationally repressed at 4hpf were deadenylated or degraded at 6hpf, suggesting that mRNA decay followed translational repression.

Stadler et al. (2012) performed parallel mRNA-seq and ribo-seq to analyze the translational changes in a set of 5 genes (lin-14 , lin-28, daf-12, hbl-1, and lin-41) which are known targets of specific miRNAs during the different stages of larval development in *Caenorhabditis elegans*. The analysis of the obtained data suggested that miRNAs interfere with gene expression by mRNA destabilization, translation initiation inhibition, and probably through other translational events during elongation.

While these studies did not end the debate regarding the role and the mechanisms of miRNA mediated translational control (Fabian and Sonenberg, 2012; Hu and Coller, 2012), they provided interesting insights into the process and demonstrate that the parallel application of ribo-seq and mRNA-seq is a powerful approach for delineating the transcriptional and translational controls of gene expression.

Characterization of the role of protein regulators of translation

mTOR is a kinase that regulates global protein synthesis by phosphorylating the protein 4E-BP whose unphosphorylated form inactivates initiation factor eIF4E whose function is to bind to the mRNA 5'-cap and initiate the assembly of the initiator ribosome complex (Dowling et al., 2010). The mTOR pathway is dysregulated in many diseases particularly in cancer, where its dysregulation is manifested by uncontrollable cell growth and overactive protein synthesis (Zoncu et al., 2011; Gingras et al., 2004). A number of genes directly regulating the mTOR pathway are well known tumor suppressors and oncogenes and it is not surprising that mTOR inhibitors emerged as potential agents for cancer therapy (Zaytseva et al., 2012).

Two recent works employed ribo-seq to study the translational regulation mediated by mTOR. Thoreen et al. (2012) carried out comparative ribo-seq analysis in mouse embryonic fibroblasts (MEFs). Treatment of MEFs with a potent mTOR inhibitor, Torin 1, resulted in the translational suppression of nearly all (99.8%) mRNAs, confirming mTOR's role as a global regulator of proteins synthesis. Hsieh et al. (2012) carried out ribosome profiling in PC3 human prostate cancer cells, where mTOR is constitutively hyperactivated, to capture changes in gene expression in response to treatment with another mTOR inhibitor, PP242. In addition to observing a global effect on translation, both studies explored a pool of mRNAs whose translation is particularly sensitive to mTOR inhibition. A 5' terminal oligopyrimidine tract (TOP) is a common feature of genes that are translationally regulated in a growth-dependent manner (Meyuhas, 2000; Bilanges et al., 2007). Hsieh et al. (2012) reported that 68% of mTOR sensitive mRNAs possess the TOP motif and 63% of such mRNAs contain a pyrimidine-rich translational element (PRTE) elsewhere within their 5' UTRs. Overall 89% of mTOR sensitive mRNAs were found to

contain either one or both motifs. Thoreen et al. (2012) were able to identify TOP or TOP-like motifs in almost the entire set of mTOR sensitive mRNAs. Therefore the presence of pyrimidine-rich sequences in 5' UTRs can be used as a strong predictor of mRNA sensitivity to mTOR inhibition. These two studies illustrate the power of ribo-seq in helping researchers to characterize cellular signalling pathways whose dysregulation is implicated in human diseases such as cancer (Gentilella and Thomas, 2012).

In a recent work focused on the characterization of the RNA binding protein LIN28A, Cho et al. (2012) used ribosomal profiling to assess LIN28A's role as a global regulator of translation. For this purpose, ribosome profiling was carried out in mouse embryonic stem cells after LIN28A knockdown. The knockdown resulted in an increased density of ribosomes on ER associated mRNAs without affecting their levels. Based on these data, Cho et al. (2012) proposed that LIN28A is a major inhibitor of translation in the endoplasmic reticulum of undifferentiated cells.

Temporal translational control

Brar et al. (2012) explored temporal changes in gene expression during meiosis in *S. cerevisiae*. Over stage-specific timepoints, ribosome profiling captured many dynamic events that occur during the progression of meiosis that were not detected with previous technologies. They found at least 10-fold variations in expression for 66% of genes. While most of these variations occur due to changes in the abundance of gene transcripts, ribo-seq also revealed pervasive translational regulation. At the global level, translation was decreased during meiosis, especially at its earliest and latest stages. Brar et al. (2012) also observed stage specific regulation in the translation of individual mRNAs matching the timing of their products known function. Figure 1.5A provides an example of stage specific translational regulation observed for the adjacent *SPS1* and *SPS2* genes. The mRNA levels for both genes showed comparable changes throughout the different stages of meiosis. Yet *SPS1*, but not *SPS2*, showed a strong temporal delay in the activation of its translation.

At the time of writing this review, Stern-Ginossar et al. (2012) published a study where temporal gene expression changes were analysed during the infection of human foreskin fibroblasts with cytomegalovirus. Measurements were made 5, 24 and 72 hours after infection. A strong temporal regulation of viral gene translation was observed with the translation of 82% of ORFs varying at least five-fold (Stern-Ginossar et al., 2012). Figure 1.5B shows a heatmap of viral ORF translation levels illustrating the temporal control of protein synthesis. Different groups of ORFs are translated at different time points with the majority switched on at the last stage.

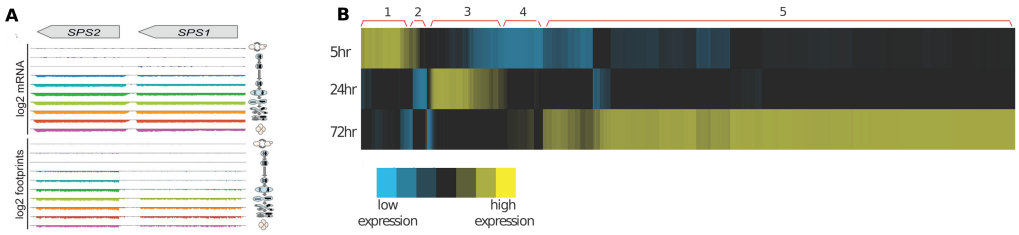

Figure 1.5: *Examples of temporal translational control. Panel (A), adapted from Brar et al. (2012) shows the expression levels of the adjacent SPS1 and SPS2 genes at different stages of meiosis in S. cerevisiae. The mRNA levels are consistent throughout all stages of meiosis. However, the ribosome profiling data for SPS1 shows strong temporal translational regulation while SPS2 does not. Panel (B), adapted from Stern-Ginossar et al. (2012), provides a heatmap of the ribosome density of viral genes clustered according to expression levels at 5, 24 and 72 hours after the infection of human foreskin fibroblasts with cytomegalovirus.*

The need for specialized computational tools for differential expression analysis using ribo-seq and RNA-seq data

Obviously two transcripts expressed at the same level but of different length would produce a different number of short reads aligning to them as the number of reads is proportional to the length of the transcript. Thus the absolute number of short reads derived from a particular transcript is usually normalized to the length of the transcript as well as to the total number of alignable reads, as in Cufdiff FKPM units (Nookaew et al., 2012). Similarly, the transcript length needs to be taken into account when measuring the relative translation of two mRNAs because the time that ribosomes would spend on the mRNAs would differ depending on the length of the translated ORF (Ingolia, 2010). Because ribosomes broadly translate mRNAs at a similar elongation rate (Ingolia et al., 2011), conversion of the absolute number of footprints into ribosome density can be used for estimating translation rates. However, this is likely to be useful only as a broad approximation because of the high variance in the time that ribosomes decode individual codons, e.g. sequence and condition dependent pausing and stalling, and also because of the complex organization of eukaryotic mRNA translation in 5' UTRs. Clearly an mRNA containing paused ribosomes is not translated as efficiently as an mRNA that is covered with fast paced ribosomes even though the density of ribosomes could be similar for both of them.

The notion that only a single ORF is translated in individual eukaryotic mRNAs and that 5' UTR stands for "Untranslated" Terminal Region, are mostly of historical interest after the discovery of functional regulatory uORFs (Morris and Geballe, 2000). The term

5' leader seems to be an adequate substitute to avoid the oxymoron "translation of 5' UTRs". The frequent occurrence of conserved AUGs in 5' leaders was revealed by phylogenetic analyses (Churbanov et al., 2005). The extensive translation of 5' leaders has been well supported by ribosome profiling studies described in this review. This implies that the ribosome density and the efficiency of the mRNA main protein product synthesis may not correlate perfectly. The ribosome footprints that originate from uORFs contribute to the overall footprint coverage of a given mRNA transcript and can affect the correct quantification of the ribosome density in a protein coding open reading frame. At a minimum it necessitates the discrimination of ribosome density in the 5' leaders from CDS regions when quantifying RPFs for protein synthesis measurements. While such discrimination would improve the assessment of the rate of main protein product ORF translation, it is unlikely to be applicable to all mRNAs because of the existence of uORFs overlapping the main ORF and also the existence of non-upstream or nested ORFs (nORFs) contained within main ORFs discovered with the analysis of published ribo-seq data (Michel et al., 2012 and Chapter 2 of this thesis). In this case, footprints aligning to the pORF do not necessarily indicate its translation. Separating footprints originating from overlapping uORFs and nORFs from footprints originating from annotated pORFs can be problematic. The use of the triplet periodicity property of ribosome profiling and the generation of sub-codon profiles (Michel et al., 2012, Chapter 2 of this thesis) can help solve this conundrum. If the ribo-seq data has well defined triplet periodicity such as in the Guo et al. study (2010), the footprints originating from ORFs in frames alternative to the pORF can be detected, thus permitting the correct quantification of pORF translation levels.

Another problem related to differential translation measurement lies in the method for normalizing translation efficiency over mRNA abundance. A change in mRNA abundance due to changes in transcription or mRNA stability would ultimately result in a corresponding change in the number of ribosome footprints. A simple approach to take this into account is to compare log ratios of ribosome densities over mRNA abundance. Hence, mRNA-seq data, generated in parallel with ribo-seq data, is used to correct for a possible contribution of differential cytosolic mRNA levels to the observed differential levels of actively translated mRNAs. However, Larsson et al. (2010) caution against using the commonly applied log ratio approach (ribo-seq levels divided by corresponding mRNA-seq levels) because log difference scores could correlate with cytosolic mRNA levels. The possible confounding effect of cytosolic mRNA levels may result in biological false positives and false negatives. As an alternative, Larsson et al. proposed analysis of partial variance (APV) as a more accurate correction method for cytosolic mRNA levels (Larsson et al., 2010). Their implementation is available in the R-package anota (analysis of

translational activity) for the analysis of differential translation using ribosome profiling datasets as well as polysome microarray or RNA-seq -based datasets (Larsson et al., 2011).

A limitation of ribosome profiling is that it allows to measure only relative changes in gene expression. Because ribo-seq does not provide information on absolute changes of translation, global suppression of translation may be misinterpreted as the activation of translation of a few unaffected genes. In RNA-seq experiments this problem is solved with the addition of synthetic RNA molecules with a different nucleotide composition (spike-in control) (Jiang et al., 2011). Han et al. ((2012)) adapted this idea by adding a synthetic 28 nt long oligonucleotide that mimics the ribosome footprint. It is desirable that standard spike-in controls will be developed and accepted by the community to allow for comparison of datasets between labs.

1.2.2 Estimating global average and local rates of translation elongation

Prior to ribosome profiling, measurements of translation elongation rates were carried out on individual mRNAs (Sorensen and Pedersen, 1991; Boström et al., 1986). To estimate the global average rate of translation elongation, Ingolia et al. (2011) used a pulse chase strategy by preventing new translation initiation using harringtonine followed by a short time for run-off elongation before adding cycloheximide. The experiments carried out in mouse embryonic stem cells (mESCs) demonstrated that ribosomes progress on mRNA transcripts at an average rate of ~5.6 codons per second (Ingolia et al., 2011). The rate of elongation is consistent across different types of messenger RNAs, independent of the length and abundance of encoded proteins. It is also uniform across the length of the coding region beyond the initial 5-10 codons. By analysing the same data using a different approach, Dana and Tuller (2012) concluded that while the average translation velocity of all genes is ~5.6 amino acids per second, the speed of elongation is slower at the beginning of coding regions and linked this observation to a decrease in the strength of the mRNA folding along the coding sequence and a decreased frequency of optimal codons in these regions, known as the "ramp theory" (Tuller et al., 2010a).

The common interpretation of ribosome profiling data is that the density of footprints at a particular location on mRNA is proportional to the time that ribosomes spend at this location. Therefore, it is possible to calculate the average density of ribosomes on specific codons to determine their relative decoding rates. All ribosome profiling studies that addressed this issue agree that there is little relationship between codon usage frequencies and their decoding rates (Ingolia et al., 2011; Gerashchenko et al., 2012; Li et al.,

2012; Stadler and Fire, 2011). This is contrary to the widespread belief that rare codons should be decoded slowly, which most likely originated from the notion that highly expressed genes have more pronounced codon usage bias (Sharp and Li, 1987). However, the lack of correlation between codon frequencies and efficiencies is not so surprising. Very early studies of translation speed and accuracy have shown that it is the availability of cognate tRNAs, rather than the frequency of codons that modulates the rate of codon decoding (Varenne et al., 1984). Jon Gallant introduced the term "hungry codon" to discriminate between the two types of codons (Weiss et al., 1988a). Several computational studies employed the data obtained with ribosome profiling to explore the relationship between codon frequencies, availability of cognate tRNAs and decoding and translation rates (Tuller et al., 2010a; Qian et al., 2012; Siwiak and Zielenkiewicz, 2010). Stadler and Fire (2011) carried out ribosome profiling in *C. elegans* in order to provide evidence in support of the hypothesis that translation is slowed down by wobble interactions between a codon and its anticodon. A discussion of ribosome profiling data in relation to codon usage can be found in a recent comprehensive review by Plotkin and Kudla (2011).

The truly unexpected observation generated by ribosome profiling was the realization that the rate of cognate tRNA selection in the A-site tRNA may not be the major factor that determines local translation elongation rates. Li et al. (2012) generated ribosome profiles in *Escherichia coli* and *Bacillus subtilis* and found that the ribosome occupancy at mRNA locations correlate with purine rich Shine-Dalgarno regions upstream of the A-site codons. The Shine-Dalgarno (SD) sequence is well known for its role in translation initiation in most prokaryotes (Shine and Dalgarno, 1975) and has previously been shown to affect elongating ribosomes (Weiss et al., 1988b). When it is located upstream of initiation codons it serves for anchoring initiating ribosomes by interacting with the complementary anti-Shine-Dalgarno (aSD) sequence in 16S rRNA. By performing a set of experiments, including ribosome profiling carried out for mRNA translated with orthogonal ribosomes (containing an altered aSD sequence), Li et al. (2012) have been able to demonstrate that SD sites indeed slow down elongating ribosomes. Under conditions of fast bacterial growth, the SD effect greatly exceeds that of particular codons (Li et al., 2012). Ingolia et al. (2011) also have been able to identify a number of ribosome pausing sites using ribosome profiles from mESCs. Although the pause sites are enriched for glutamate and aspartate codons in the A site, enrichment for particular amino acids encoded by a sequence just upstream is yet another feature that is not directly related to the identity of a codon in the A-site. Notably, both studies confirmed increased ribosome density at known sites of ribosome stalling. Figure 1.6 shows the peptide-mediated stalling at *secM* (Vázquez-Laslop et al., 2010) and *tnaC* (Seidelt et al., 2009) in *E. coli*, at *mifM* in

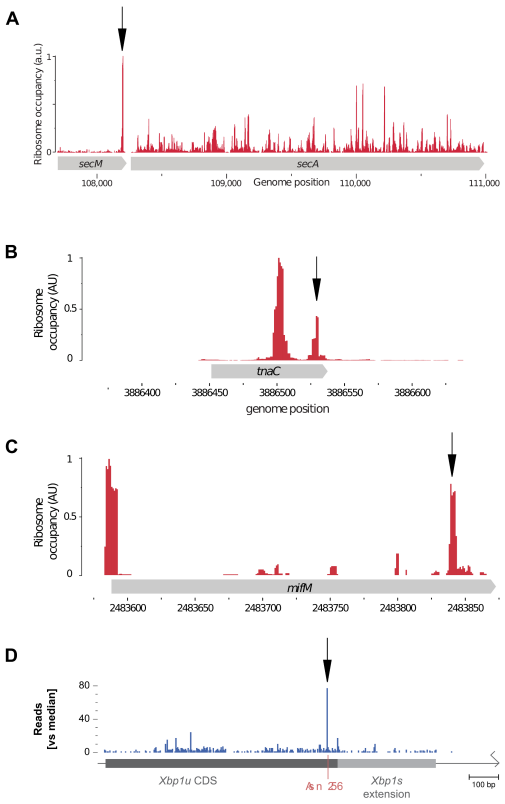

Figure 1.6: *The increased ribosome density at known sites of ribosome stalling: secM (A) and tnaC (B) in Escherichia coli; mifM (C) in Bacillus subtilis; and Xbp1 in Mus musculus (D). Panels (A-C) are adapted from Li et al. (2012) and panel (D) is adapted from Ingolia et al. (2011). Black arrows indicate the locations of known ribosome pause sites.*

B. subtilis (Chiba et al., 2011) and at *Xbp1* mRNA (Yanagitani et al., 2011) in mESCs, thus confirming the applicability of ribosome profiling for the identification of ribosome pausing sites.

All studies where ribosome profiling is used for estimating local decoding rates require the detection of the A-site codon location. Ribosome profiling does not provide direct information on the locations of the A-site codons. It is inferred from the locations of ribosome footprints. At present there are two strategies. One, used in ribosome profiling in eukaryotes, sets an offset between the 5' end of the ribosome footprint and the expected location of the A-site codon. The offset is derived from the distance between the major density peaks for the 5'-ends upstream of the starts of main coding regions (in some studies stratified according to RPF length), (see Ingolia et al., 2009; Stadler and Fire, 2011 for details). The other, the so-called centre-weighted approach, was used for ribosome profiling in bacteria. In this case, the centre of the ribosome footprint is considered as the most probable location of the A-site, with codons adjacent to the centre also taken into account as potential A-site codons but with reduced weighting co-efficient, (see Li et al.,

2012 for details). Recently, it has been found that in bacteria, Shine-Dalgarno sequences could affect the size and symmetry of ribosome footprints (O'Connor et al., 2013), thus potentially affecting the positions of the A-sites relative to the footprint ends. To what extent this phenomenon affects the above mentioned methods of A-site codon position detection needs further investigation.

1.2.3 Selective ribosome profiling

Oh et al. (2011) introduced a procedure that they termed "selective ribosome profiling". To obtain information on ribosome-associated chaperone trigger factor (TF) targets, Oh et al. (2011) combined ribosome profiling with affinity purification of the ribosomes bound with TF, thus mapping the locations of TF bound ribosomes on *E. coli* mRNAs. They found that in the majority of mRNAs, TF binds to the nascent peptide chain after the ribosome finishes translating about a hundred codons. TF was also found to have a strong preference for binding to ribosomes translating outer-membrane protein mRNA. To study co-translational protein folding in mammalian cells, Han et al. (2012) developed the FACS (folding-associated cotranslational sequencing) technique. In this technique a specific folding is used as an affinity tag for isolating ribosomes along with protected mRNA fragments. Han et al. (2012) were able to use this technique to monitor the folding of hemagglutinin along its mRNA. Using a similar concept, Reid and Nicchitta (2012) carried out ribosome profiling after separating endoplasmic reticulum (ER) and cytosolic polysome fractions. Consequently, Reid and Nicchitta (2012) were able to identify the contribution of the two cellular compartments to global protein synthesis and found that preferential translation occurs on ER bound ribosomes. Many mRNAs encoding cytosolic proteins are loaded with ribosomes on the ER and while mRNA abundance is higher in the cytosol, the ER-localized mRNAs have a higher ribosome density. Based on their findings, Reid and Nicchitta (2012) proposed that the partitioning of mRNAs between the cytosol and ER compartments is a mechanism of post-transcriptional regulation of gene expression: while protein synthesis preferentially occurs in the ER, mRNA storage and degradation occur in the cytosol.

These three studies have demonstrated the applicability of selective ribosome profiling for studying the compartmentalization of translation inside the cell as well as for elucidating the functional properties of ribosome associated factors.

1.2.4 Identification of novel translated ORFs

The analysis of ribosome profiling data does not necessarily depend on gene annotation and thus can be used for the verification of existing gene annotations and the identification of novel non-annotated genome features such as protein coding genes or short translated ORFs. *Ab initio* annotation of genomes is particularly difficult for short open reading frames because short ORFs could exist purely by chance and information on the nucleotide composition of short ORFs may not be sufficient to discriminate coding from non-coding ORFs. Ribosome profiling provides a way to find translated ORFs irrespective of their length. Most recoded genes that require non-standard translational events, such as programmed ribosomal frameshifting, cannot be automatically identified with pure sequence analysis because of the high diversity and our poor understanding of recoding signals. Ribo-seq can be used to facilitate the discovery of novel recoded genes. It has been argued that most alternative splice isoforms may not contribute to protein synthesis (Tress et al., 2007). Identifying those that are productive is not trivial. In the following sections we discuss how ribosome profiling can provide data that can be used to discriminate translated isoforms from those that are untranslated. In addition, we review how ribosome profiling data can be used to explore the evolution of protein coding genes.

uORFs, nORFs and novel protein coding genes

Protein coding genes are usually discriminated from regulatory ORFs. While it is becoming increasingly difficult to reach agreement on a formal definition of a gene (Gerstein et al., 2007), it is colloquially used as a term for a sequence that encodes a functional protein molecule. Thus, a regulatory ORF is distinct in the sense that its translation (rather than the product of that translation) is functionally important. Clearly, the distinction is not strict. In prokaryotes, where polycistronic mRNAs are abundant, the translation of adjacent ORFs encoding functional protein products is often coupled, providing a regulatory mechanism for their co-expression. It is also possible that the translation of some short regulatory ORFs in eukaryotes may result in the biosynthesis of biologically active peptides. Ribosome profiling alone does not provide information regarding the function or importance of the translated ORF product. The distinction needs to be made based on other factors such as the organization of adjacent ORFs, phylogenetic conservation, etc. Therefore we describe the detection of regulatory ORFs and novel protein coding genes in the same section.

The very first ribosome profiling study in yeast (Ingolia et al., 2009) revealed the occurrence of extensive translational events in the 5' leaders of eukaryotic mRNAs that was

confirmed by all subsequent eukaryotic ribo-seq datasets. These translational events appeared to be very sensitive to changes in environmental conditions suggesting a regulatory role of the 5' translation (Ingolia et al., 2009; Ingolia et al., 2011; Gerashchenko et al., 2012; Brar et al., 2012). While the current ribosome profiling studies point to the existence of a large number of translated short uORFs, their identification appears to be difficult. uORF's short length, limited footprint coverage, frequent non-AUG initiation and the simultaneous translation of overlapping ORFs are among the many factors complicating the unambiguous assignment of ribosome footprints to one of several potential translated uORFs. In principle, the triplet periodicity of ribosome footprints allows the detection of the translated reading frame and this feature could help in the identification of short translated ORFs. Michel et al. (2012) (see Chapter 2 of this thesis) have demonstrated that given sufficient coverage, it is possible to use triplet periodicity for detecting the translation of reading frames alternative to the main one. The ability to predict alternatively translated frames depends on sufficient coverage, length of ORFs overlap and the relative intensity of the alternative frame translation. Despite these limitations, Michel et al. (2012) not only detected several uORFs translated at an efficiency higher than the main protein product ORF, but also ORFs with initiation codons downstream of the main ORF start codon which they termed nORFs (for non-upstream regulatory ORFs) (see Fig. 2.4).

It is as yet unclear how such nORFs could regulate the translation of main ORFs although their functional importance is supported by phylogenetic analysis. Comparative analysis of one such nORF in *NPAS2*, a gene encoding a component of the suprachiasmatic circadian clock in mammals, provides evidence for the conservation of the nORF rather than its protein sequence suggesting a role for its translation, but not for its product (Michel et al., 2012), see Figure 2.4C.

Because splicing in bacteria is uncommon, sequences of bacterial ribosome footprints can be aligned directly to genomic sequences, thus simplifying the discovery of novel protein coding genes. Strikingly the first ribosome profiling study performed in *E. coli* (Oh et al., 2011) revealed several protein coding genes that were not annotated previously despite *E. coli* K12 being one of the most extensively studied organisms with an intensively annotated genome. Hence it is evident that current sequence analyses approaches do not allow the identification of all protein coding genes based on DNA sequences even in a well-studied bacterial species and that ribosome profiling is capable of improving the situation. This was further exemplified with a recent study of Human Cytomegalovirus (HCMV) infection where ribosome profiling of elongating and initiating ribosomes increased the number of identified translated ORFs by more than a third (Stern-Ginossar et al., 2012).

Correcting annotations of existing genes and detecting protein isoforms

Ribosome profiling of elongating ribosomes has significant limitations for the analysis of initiation codons. When protein synthesis is initiated from multiple start codons, only the 5'-end start codon can be identified. Therefore, ribosome profiling of initiating ribosomes (described in section 1.3) is much more appropriate for this goal. In contrast to determining the 5' boundary of a protein coding region, ribosome profiling of initiating ribosomes provides no value for finding the 3' boundaries of coding regions. Identifying the 3' boundary of coding regions is problematic in the case of recoding events (see Atkins and Gesteland, 2010 for a compilation of reviews on Recoding). The meaning of stop codons is known to be redefined with the recoding cis-elements to either standard (stop codon readthrough) or to non-standard proteinogenic amino acids (selenocysteine and pyrrolysine insertions). In addition, in the case of programmed ribosomal frameshifting, a portion of the ribosomes shift frames at specific locations in the mRNA thus terminating at a stop codon that is out-of-frame relative to the initiator codon. Michel et al. (2012) (see Chapter 2 of this thesis) developed a method for identifying frame transitions in mRNA translation based on the triplet periodicity of ribosome profiling and demonstrated its applicability by finding known cases of ribosomal frameshifting in humans (see Fig. 2.1C) as well as a set of human mRNAs with translated overlapping ORFs. Using a similar approach, Gerashchenko et al. (2012) identified four novel cases of ribosomal frameshifting in yeast (*APE2*, *MMT2*, *URA8* and *YLR179C*). Moreover, the identified cases appear to be dependent on oxidative stress suggesting that ribosomal frameshifting plays a regulatory role in these recoded genes (Gerashchenko et al., 2012).

As suggested by Ingolia et al. (2009), the marked absence of RPFs in unspliced introns helps discriminate between alternative splice forms. When multiple isoforms exist for a given gene, ribosome profiling in conjunction with mRNA-seq, can help in the correct identification of the transcribed and translated isoform. Ribosome profiling can also be useful for discovering novel translated mRNA variants. By analysing the triplet periodicity in the ribosome profile of the human gene *C11orf48*, Michel et al. (2012) (see Chapter 2 of this thesis) found that 3' terminal exons are predominantly translated in a frame that is alternative to the predicted. More detailed analysis of available transcripts revealed the existence of an mRNA variant with an additional exon due to an alternative transcription initiation site. This shorter variant is translated in an alternative frame, resulting in dual decoding of the last three exons of *C11orf48*. The peptide generated from this additional exon has been independently detected with mass spectrometry (Oyama et al., 2007).

non-mRNA translation

Several studies have found RPFs aligning to genomic sequences that are not annotated as protein coding. Moreover, many are believed to be non-coding transcripts. This raises questions about the nature of this phenomenon, whether it reflects genuine translation and if it does, what is the function of such translation. A high proportion of the yeast non-coding genome is transcribed and these transcripts are termed Stable Unannotated Transcripts, SUTs (Jacquier, 2009). Wilson and Masel (2011) have found that over half of all SUTs are associated with ribosomes, especially at AUG codons and proposed that this type of low level non-deleterious translation may facilitate *de novo* gene birth.

Carvunis et al. (2012) extended this idea further by proposing an evolutionary model of functional genes evolving *de novo* through transitory proto-genes. Signatures of translation have been found for 1,139 of total ~108,000 unannotated ORFs (>10 codons) in *Saccharomyces cerevisiae* outside of annotated features on the same strand. To find evidence for proto-gene mediated evolution, Carvunis et al. (2012) estimated the order of ORF emergence in *S. cerevisiae* using their conservation among Ascomycota.

Evidence of translation in presumed non-coding regions in mammals has also been found. Ingolia et al. (2011) observed RPFs on >1000 large intergenic noncoding RNAs (lincRNAs) in mESCs and proposed to call them sprcRNAs for Short Polycistronic Ribosome-Associated Coding RNAs to discriminate them from lincRNAs. Lee et al. (2012) also found evidence of ribosome association with presumed non–protein-coding RNAs (ncRNAs) in HEK293 cells.

1.3 Ribosome profiling of initiating ribosomes

Although to date there have been only four published works where ribosome profiling was carried out on initiating ribosomes, we dedicate this separate section of our review to the topic. As illustrated in Figure 1.3, this type of ribosome profiling provides information on mRNA translation that cannot be captured by the profiling of elongating ribosomes. Thus we believe that such experiments will be used as frequently as the original method, and more likely used in parallel. In terms of differential gene expression, initiation is slow in comparison with elongation (unless we consider special cases like ribosome pausing) and therefore is a rate limiting step. Thus, provided that it is accurately measured, the rate of initiation of translation in most cases would be a better predictor of translation rates than the density of elongating ribosomes on mRNAs. In terms of the characterization of protein products, it is also advantageous since the data on the locations of initiation

codons can be easily interpreted to predict protein isoforms translated from different start codons. The main disadvantage of this method is its inability to provide direct information on local translation elongation rates and recoding events. Its utility for discriminating the translation of alternative splice variants is also limited.

The critical aspect of this strategy is a method for freezing initiating ribosomes. Several approaches have been used in eukaryotic systems. As yet, there have been no similar studies reported for bacteria.

1.3.1 Mapping translation initiation sites (TISs)

The first attempt to obtain a map of TISs using a direct experimental approach was made in mESCs with the drug harringtonine (Ingolia et al., 2011). Harringtonine binds to a 60S subunit and forms an 80s ribosomal complex with the initiator tRNA but blocks aminoacyl-tRNA binding in the A-site and peptide formation (Fresno et al., 1977). To identify translation initiation codons precisely, Ingolia et al. (2011) used a support vector machine (SVM) learning technique and reported 13,454 unique TISs within ~ 5000 well expressed transcripts. The majority (65%) of these transcripts contain more than one detectable TIS with 16% containing four or more sites. Extensive translation initiation at non-AUG codons was also observed, particularly upstream of annotated starts. A potential problem with this approach is that because harringtonine binds to the 60S subunit, its binding could affect the selection of initiation codons by the ribosome.

To avoid any potential selection effect of harringtonine on initiation codons, Fritsch et al. (2012) mapped TISs by enriching elongating ribosomes near start codons instead of blocking initiating ribosomes. For this purpose puromycin was used to induce premature termination of elongating ribosomes which resulted in a relative increase in ribosome density at a few codons downstream of the TISs. These ribosomes were blocked with cycloheximide prior to nuclease treatment. The identification of TISs was carried out with a machine learning technique based on neural networks yielding 7471 unique TISs in 5062 well expressed transcripts in a human monocytic cell line. Only 30% of non CDS-overlapping uORFs initiated with AUG and only 8% of CDS-overlapping uORFs initiated with AUG. This finding supports the earlier result (Ingolia et al., 2011) regarding the abundance of non-AUG initiation in 5' leaders.

To obtain TIS maps, Lee et al. (2012) used a different drug, lactimidomycin, which binds to 80S ribosomal subunits after its assembly on start codons, making any bias on the selection of start codons less likely in comparison with harringtonine. To improve the lactimidomycin TIS signal detection, initiating ribosome footprints were compared with

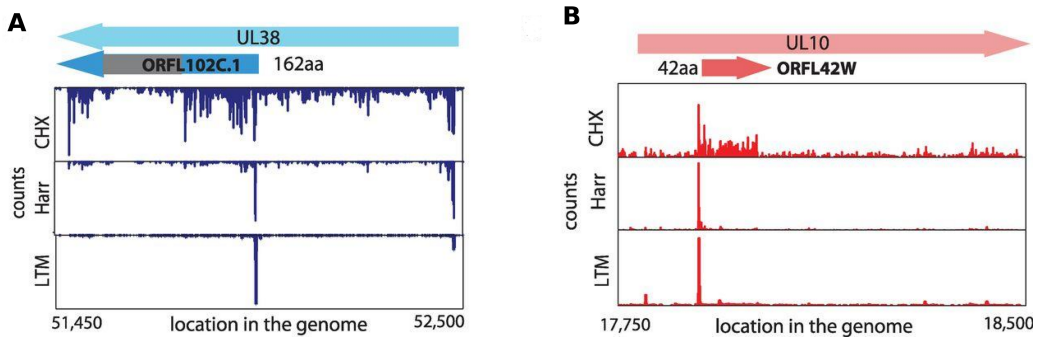

Figure 1.7: *The ribosome initiating profiles (harringtonine (Harr) and lactimidomycin (LTM)) and elongating profiles (cycloheximide (CHX)) for the HCMV genes UL38 (panel A) and UL10 (panel B) (adapted from Stern-Ginossar et al. (2012)). The two ribosome profiling approaches aided the identification of internal initiation sites in both genes, with an N-terminally truncated translation product for UL38 and a previously unknown out-of-frame translated ORF contained within the UL10 gene.*

elongating ribosome footprints generated with cycloheximide treatment carried out in parallel. From ~10,000 transcripts with detectable TIS peaks, Lee et al. (2012) identified a total of 16,863 TISs. In experiments carried out in human cytomegalovirus (HCMV) infected cells, Stern-Ginossar et al. (2012) used both harringtonine and lactimidomycin treatments and found the results comparable: >98% of the initiation sites detected using harringtonine were also detected using lactimidomycin. So although the mechanism of action of the two drugs is different, they arrest ribosomes mostly at the same locations. Stern-Ginossar et al. (2012) also generated ribosome elongation profiles of mRNAs pretreated with either cycloheximide or lysed without drug pretreatment. Together their separate profiles of initiating ribosomes and elongating ribosomes enabled the identification of hundreds of previously unidentified open reading frames in HCMV, including internal ORFs lying within existing ORFs (nORFs), short uORFs, ORFs within transcripts anti-sense to canonical ORFs and previously unidentified short ORFs encoded by distinct transcripts (see Figure 1.7).

uORFs, nORFs and novel genes

As long as no recoding events are involved in the translation of an mRNA transcript (i.e. the triplet periodicity of translation is maintained and amino acids are not incorporated at stop codons), the identification of translated ORFs can be made based on TIS detection. Moreover it is even simpler in comparison with ribosome footprints obtained with elongating ribosomes. Because ORFs overlap, it is very difficult to discriminate between the translation of a single frame and the translation of two overlapping ORFs occupying the

same transcript location. If TISs are detected with codon precision, information regarding the framing can be determined and therefore can be used for the identification of translated ORFs.

All of the studies in the previous subsection reported the existence of ORFs in different configurations relative to the main annotated ORFs with the largest proportion of them being uORFs (Ingolia et al., 2011; Lee et al., 2012; Fritsch et al., 2012; Stern-Ginossar et al., 2012). However, novel ORFs located downstream have also been detected raising questions regarding their importance (Michel et al. (2012) and Chapter 2 of this thesis).

In many cases translation initiates on very short ORFs, which are unlikely to produce functional peptides: among 751 translated ORFs in cytomegalovirus, 245 are shorter than 21 codons, 239 are in the range of 21 to 80 and only 120 are longer than 80 codons (Stern-Ginossar et al., 2012). The translation of many of these ORFs may represent gene expression noise and the products of these ORFs may have no function. They could, however, be potential targets for the host immune response and are of interest for understanding the biology of the virus.

Non-AUG translation

While initiation at non-AUG codons is frequent in many bacteria, as recently as 2010, the number of non-AUG codons identified as potential translation initiation sites in humans was small. In 2011, Ivanov et al. (2011) reported 42 novel non-AUG initiation sites which were detected with the analysis of evolutionary signatures of protein-coding sequences in the regions upstream of annotated codons. Ribosome profiling increased this number dramatically: the number of non-AUG TISs reported in the studies described here is close to a half of all TISs. In addition, non-AUG initiation occurs more frequently in uORFs. Lee et al. (2012) reported that over 74% of upstream TISs in human are non-AUG codons, often associated with short uORFs.

Protein isoforms

Figure 1.3 illustrates why ribosome profiling of initiating ribosomes is particularly suitable for the detection of alternative protein isoforms (extensions and truncations of annotated CDS). As discussed in the section 1.3 “Ribosome profiling of initiating ribosomes”, initiation at alternative sites both upstream and downstream of the annotated protein coding ORFs is pervasive. Many of these events were heretofore difficult to detect and annotate. Now, advancements can be made in gene annotations by incorporating ribosome profiling data. Figure 1.8A shows an N-terminally extended isoform of the human *RND3* gene

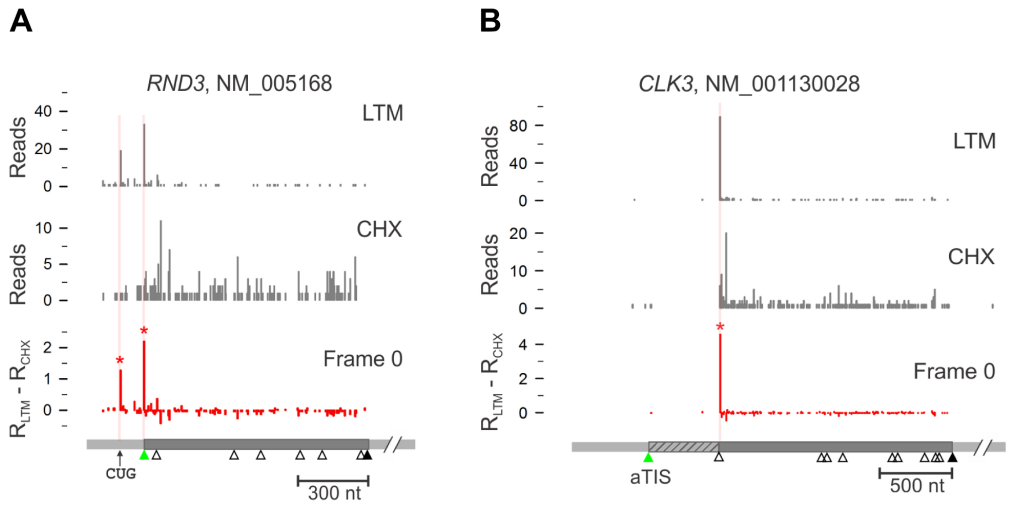

Figure 1.8: *Detection of protein isoforms with alternative N-termini. Panel (A) shows an N-terminally extended isoform of the human RND3 gene which has an in-frame CUG initiating codon. Panel (B) shows a truncated isoform of the human CLK3 gene which was found to initiate at an AUG codon downstream of the annotated AUG start codon (adapted from Lee et al., 2012).*

which has an in-frame CUG initiating codon. Figure 1.8B shows a truncated isoform of the human *CLK3* gene which Lee et al. (2012) found to initiate at an AUG codon downstream of the annotated AUG start codon. Ingolia et al. (2011) identified 570 genes with potential N-terminal extensions and 870 with N-terminal truncations in the 4,994 genes that were analyzed in mESCs. Fritsch et al. (2012) also reported 546 N-terminal protein extensions in human (regions downstream of annotated starts were not analyzed). These examples highlight the usefulness of ribosome profiling data in improving existing annotations.

1.4 Perspectives

Translation is a complex process and therefore its characterization will require the use of a combination of approaches. Ribosome profiling of elongating and initiating ribosomes was carried out in parallel in the most recent study (Stern-Ginossar et al., 2012). The combination of the two approaches benefits from the specific advantages of each method. Moreover, it is very likely that further variants of ribosome profiling will be developed in order to capture the characteristics of translation that are unattainable by the methods described in this review.

Translation is a process that is downstream of transcription and therefore it cannot be characterized accurately without information on the transcriptome. Therefore transcriptome sequencing and ribosomal profiling have to be carried out in parallel. Combined

together, RNA-seq and different ribo-seq techniques will form a universal set of tools for characterizing the molecular state of any living cell at a very detailed level. The continual reduction in cost and time of nucleic acid sequencing will ensure the accessibility of these techniques for gene expression measurements to a very wide research community. There is little doubt that the application of this suite of techniques will grow explosively. However, the ease of the data generation will demand adequate capacity to process, interpret, store, integrate and distribute the data (Nekrutenko and Taylor, 2012).

Chapter 2

Observation of dually decoded regions of the human genome using ribosome profiling data.

This chapter has been published in Genome Res (2012) 22(11):2219–2229 (see Appendix 5.2).

The recently developed ribosome profiling technique (Ribo-Seq) allows mapping of the locations of translating ribosomes on mRNAs with sub-codon precision. When ribosome protected fragments (RPFs) are aligned to mRNA, a characteristic triplet periodicity pattern is revealed. We utilized the triplet periodicity of RPFs to develop a computational method for detecting transitions between reading frames that occur during programmed ribosomal frameshifting or in dual coding regions where the same nucleotide sequence codes for multiple proteins in different reading frames. Application of this method to ribosome profiling data obtained for human cells allowed us to detect several human genes where the same genomic segment is translated in more than one reading frame (from different transcripts as well as from the same mRNA) and revealed the translation of hitherto unpredicted coding open reading frames.

2.1 Introduction

The human genome, containing slightly more than 20,000 protein coding genes (Clamp et al., 2007), generates a substantially more diverse proteome by encoding more than one protein in a single gene. The proteome is diversified through a number of molecular mechanisms that alter the sequence of the main gene product, such as alternative splicing

(Matlin et al., 2005), RNA editing (Kiran and Baranov, 2010; Wulff et al., 2011), utilization of alternative translation initiation sites (Ingolia et al., 2011; Ivanov et al., 2011) and post-translational modifications (Mann and Jensen, 2003). However, in addition to modifications of existing protein sequences, examples are known where the same genomic region codes for entirely different protein sequences. This occurs when it is decoded in alternative reading frames, a phenomenon known as dual coding. Dual coding hampers the evolutionary flexibility of nucleotide sequences (Firth and Brown, 2006; Rancurel et al., 2009). Consequently it is expected to be rare in genomes with weakly constrained size and indeed it is currently considered to be atypical. Nonetheless, comparative sequence analysis provides growing evidence that multiple instances of dual decoding do occur in humans (Liang and Landweber, 2006; Chung et al., 2007; Ribrioux et al., 2008). Here we present a method that facilitates the detection of dual decoding instances in human using data obtained by the recently developed ribosome profiling technique (Ingolia et al., 2009; Guo et al., 2010). Ribosome profiling is based on the isolation of mRNA fragments protected by ribosomes followed by massively parallel sequencing of cDNA libraries derived from the Ribosome Protected Fragments (RPFs). The technique allows mapping the locations of translating ribosomes on the entire set of mRNA molecules produced under given physiological conditions, thus providing a unique opportunity to obtain quantitative Genome-Wide Information on Protein Synthesis, GWIPS (Weiss and Atkins, 2011). This is important since protein abundance is mainly regulated at the level of protein biosynthesis (Schwanhäusser et al., 2011). The area of GWIPS is rapidly growing. Since the publication of the technique in 2009 (Ingolia et al., 2009), an increasing number of studies have been carried out using the ribosome profiling technique (Guo et al., 2010; Ingolia et al., 2011; Oh et al., 2011; Stadler and Fire, 2011; Brar et al., 2012; Reid and Nicchitta, 2012).

When RPF sequences are aligned to mRNA, a characteristic triplet periodicity can be observed for the locations of the 5'-ends of the RPFs. Such triplet periodicity was observed in ribosome profiling experiments carried out in both yeast (Ingolia et al., 2009) and human cells (Guo et al., 2010). The triplet periodicity observed in human cells (HeLa) is illustrated in Figure 2.1A. This periodicity occurs because ribosomes move not by one, but by three nucleotides, one codon at a time. As a result when RPFs are aligned to mRNA sequences, the majority of RPF 5' ends align at a specific distance from the first nucleotide of the A- site codon of the elongating ribosome. Allowing 15 nt for the distance from the decoding centre to the 5'-end of an RPF (Guo et al., 2010), the RPFs align predominantly to either the first or the third positions of the A-site codon as can be seen in Figure 2.1A. The second position has the lowest proportion of matching RPFs. Thus, the phase of

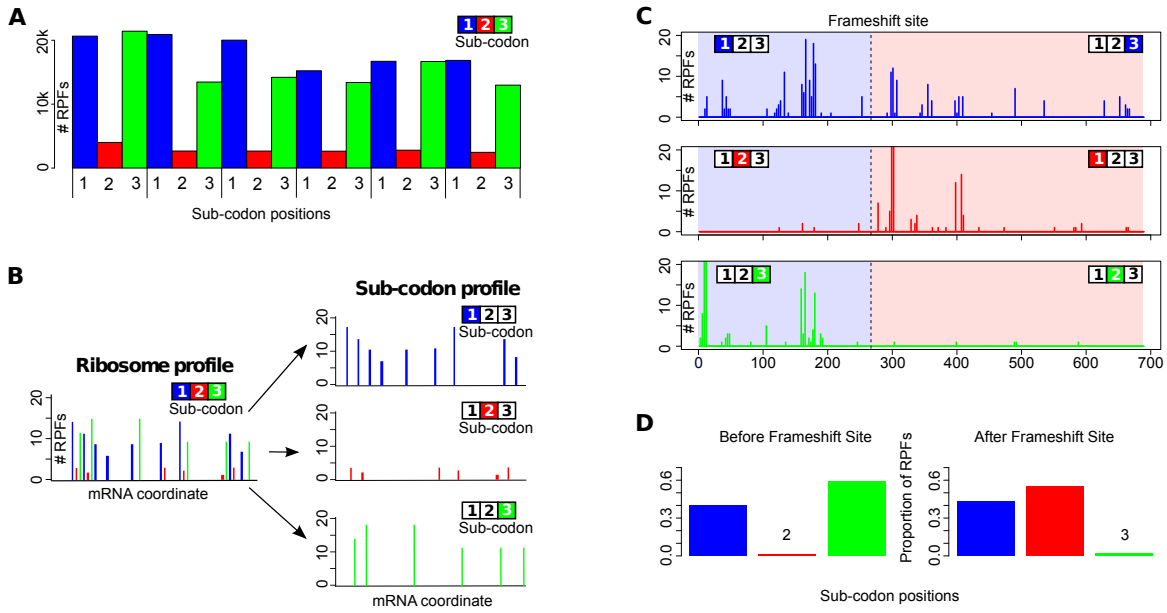

Figure 2.1: *Utilization of triplet periodicity for detecting translated reading frames. (A) A plot of the number of RPFs aligning to particular mRNA positions between the 30th and the 47th nucleotide downstream of the start codon aggregated over 6000 human RefSeq mRNAs. In each codon, sub-codon position 2 is shown as a red bar while sub-codon positions 1 and 3 are shown as blue and green bars respectively. (B) A schematic representation of the generation of a sub-codon profile from the corresponding RPF profile. Each sub-codon position (1 - blue, 2 - red, 3 - green) is shown on separated plots. (C) The absolute number of RPFs aligning to each sub-codon position is shown for the coding region of human Antizyme 1 (OAZ1) mRNA. The location of the programmed ribosomal frameshift site is indicated by a broken black line. (D) The distribution of the number of RPFs aligning to different sub-codon positions, upstream of the frameshift site (left) and downstream (right). It can be seen that the sub-codon position with the lowest RPF count shifts from the second to the third upon ribosomal frameshifting.*

the triplet periodicity can be used as a signature of one of the three potentially translated reading frames. Therefore, by analyzing the periodicity of aligned RPFs it is possible to determine the frame that is being translated.

From the aligned RPFs for each mRNA transcript, the sub-codon profile can be generated to determine the translated reading frame. A schematic representation of how a sub-codon profile is generated is given in Figure 2.1B. Sub-codon position 2 typically has the lowest number of RPFs and this feature can be used as a signature for detecting which out of the 3 reading frames is being translated. Moreover, this feature can be used for detecting shifts between reading frames, such as the one known to occur in the expression of human ornithine decarboxylase antizyme 1 (*OAZ1*) gene (Matsufuji et al., 1995). The *OAZ1* mRNA sub-codon profile is shown in Figure 2.1C. It can be seen that when RPF counts are separated by their sub-codon positions (phased relative to the start codon) (Fig. 2.1D), there is a transition between the proportions of RPFs aligning to each position. While the

second position has the lowest number of RPFs upstream of the frameshift site, it is the third position that has the lowest number of RPFs downstream of the frameshift site. This is consistent with the +1 directionality of the ribosomal frameshifting (the second coding ORF in the *OAZ1* mRNA is in the +1 frame relative to the first ORF). To find other mRNA sequences where reading phase transitions occur, we developed a computational approach for the analysis of sub-codon profiles. This method exploits the sub-codon RPF periodicity signature to identify mRNA transcripts with putative reading frame transitions. Application of this method to a number of human mRNAs for which ribosome profiling data are available allowed us to detect dually coding regions of the human genome, where the same nucleotide sequence is used to encode protein sequences in more than one reading frame.

2.2 Results

2.2.1 Periodicity Transition Score (PTS).

The most intuitive approach for determining the reading frame would be a sliding window to monitor the transition of the lowest proportion from one sub-codon position to another. However, our empirical investigation of such an approach demonstrated that it is impractical for the type of data currently generated by the ribosome profiling technique (see Supplemental Figs. 1-2 in Appendix 5.3). This is largely due to the high non-uniformity of the RPF distribution. While certain coding locations of mRNAs have a large number of aligning RPFs, the majority of mRNA coordinates have no RPFs aligning to them. In the top 10 expressed genes (Fig. 2.2A left-hand side), approximately 16% of CDS codons have no RPFs aligning to them. This increases to just over 63% when we expand the pool to the top 1000 expressed genes (Fig. 2.2 right-hand side). On the other hand, approximately 24% of the top 10 expressed genes have CDS codons where over 100 RPFs align, while less than 2% of the top 1000 expressed genes have CDS codons where over 100 RPFs align. This heterogeneity may arise from biases introduced during the experimental protocol, oligonucleotide adapter ligations to the 3' and 5' ends of short reads for cDNA library preparation introduce biases that may result in the over-representation or under-representation of some RPFs (Hafner et al., 2011), but also very likely reflects authentic features of translation. Certain locations are translated significantly slower than other mRNA locations (Ingolia et al., 2011). For example, according to the ramp hypothesis (Tuller et al., 2010a), there is an evolutionary selection for slowly decoded codons at the beginning of coding regions, resulting in a relatively higher density of RPFs (Tuller et al., 2010b). Other regions, where ribosomes move quickly, would have insufficient cov-

erage. In addition, we have observed occasional single isolated RPF peaks in sub-codon profiles for the second sub-codon position. Occurrences of such peaks in regions with otherwise no RPF coverage result in false positives. Such peaks are not necessarily artifacts of the ribosome profiling method, but may reflect authentic features of translation. In such locations, it is possible that the size of the region covered by the ribosome may differ from the average due, for example, to specific interactions with components of the ribosome inside the mRNA channel, leading to the generation of peaks that are not consistent with the average periodicity.

To overcome the problems of profile heterogeneity and local RPF length non-uniformity, we devised a different approach to assess whether a transition between frames exists in the ribosome profile of a particular mRNA. The approach is based on a sliding point where cumulative proportions of RPFs aligning with particular sub-codon positions are calculated upstream and downstream of this point as described below (see also Fig. 2.2B).

For a protein coding region between coordinate $\boldsymbol{a}$ and coordinate $\boldsymbol{b}$ we can represent a ribosome profile as an array of the number of RPFs aligning their 5' ends to a particular position, e.g. $(\boldsymbol{f_a}, \boldsymbol{f_{a+1}}, \boldsymbol{f_{a+2}} \ \ldots \ \boldsymbol{f_{b-2}}, \boldsymbol{f_{b-1}}, \boldsymbol{f_b})$. For each coordinate x within the CDS, the proportion of RPFs corresponding to a particular codon position is calculated for the upstream $\boldsymbol{q_u}$ and the downstream $\boldsymbol{q_d}$ regions as follows:

$$q_u^n(x) = \frac{\sum_{i=a}^{x} f_i^n}{\sum_{j=1}^{3} \sum_{i=a}^{x} f_i^j} \quad \textit{and} \quad q_d^n(x) = \frac{\sum_{i=x+1}^{b} f_i^n}{\sum_{j=1}^{3} \sum_{i=x+1}^{b} f_i^j} \qquad \textbf{\{1\}}$$

where $\boldsymbol{n}$ indicates a position (1, 2 or 3) within a codon relative to the first nucleotide of the start codon. We define the Cumulative Sub-Codon Proportion Difference (CSCPD) as the absolute difference between the upstream and downstream proportions:

$$CSCPD^n(x) = |q_u^n(x) - q_d^n(x)| \qquad \textbf{\{2\}}$$

The approach is advantageous in that it increases the size of the informative region while the effect of false signals generated by isolated RPFs is reduced.

The statistical confidence of the CSCPD estimation is low when $\boldsymbol{x}$ is close to $\boldsymbol{a}$ or $\boldsymbol{b}$ due to the limited number of RPFs in either the upstream or the downstream region. To account for this, we computed the CSCPD curves for each of the 1000 mRNAs with the highest number of RPFs from the Guo et al. (2010) data set and used the 95th percentile for each sub-codon position as a threshold over the length of the CDS. To address the

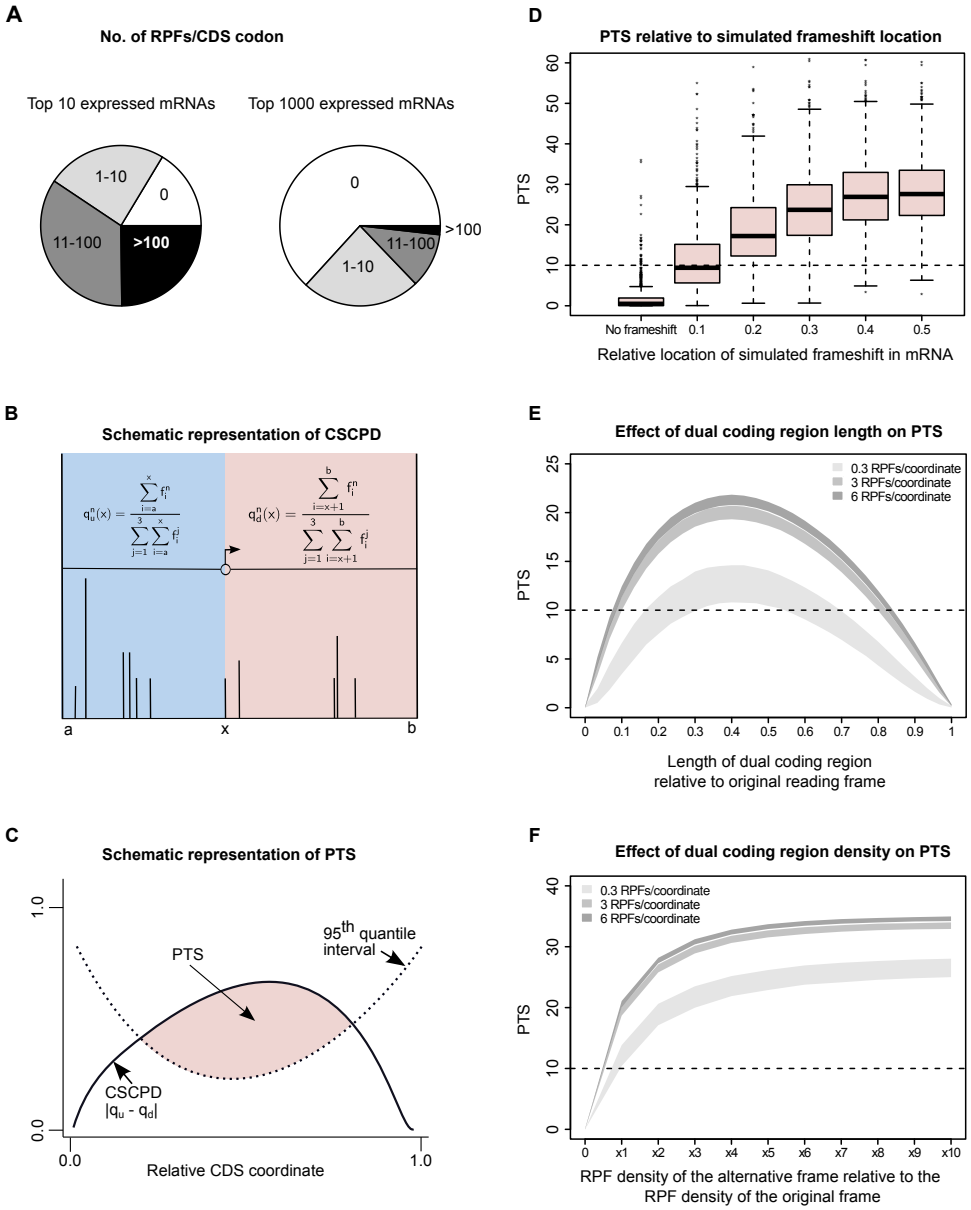

Figure 2.2: *Computational approach for detecting transitions between reading frames and its performance on simulated dual coding. (A) Segments of pie charts represent the number of CDS codons with the specific number of RPFs aligning to them for the top 10 (left) and for the top 1000 (right) most covered mRNAs from the Guo et al dataset (Guo et al. 2010). It can be seen that, even for the most RPF-covered mRNAs, many CDS codons have no RPFs aligning to them. (B) Calculation of cumulative RPF sub-codon proportion differences (CSCPD) upstream and downstream of a sliding point* ***x****. Position* ***a*** *represents the annotated CDS start while position* ***b*** *denotes the annotated CDS stop. Vertical lines represent the RPFs that align at given CDS coordinates. (C) Principle of the automated scoring scheme, Periodicity Transition Score (PTS). PTS is calculated as the area (shaded in pink) where CSCPD over the examined CDS exceeds the expected level as estimated from the 95th quantile CSCPDs of the 1000 mRNA transcripts with the highest RPF coverage. See Results section for details. (D) Boxplots representing the distributions of PTS scores (axis y) obtained for real ribosome profiles for mRNAs with artificially introduced frameshifts at different locations relative to the ends of CDS (axis x). (E) Distribution of PTS for ribosome profiles on simulated mRNAs containing simultaneously translated dual coding regions of different lengths. The simulations were carried out for three sets of mRNAs with different RPF density as indicated in the figure. The shaded areas represent the lower and upper quartile intervals for each RPF density. (F) Distribution of PTS for simulated mRNAs containing dual coding regions with varying densities of RPFs in the alternative frame. Shading is as in E.*

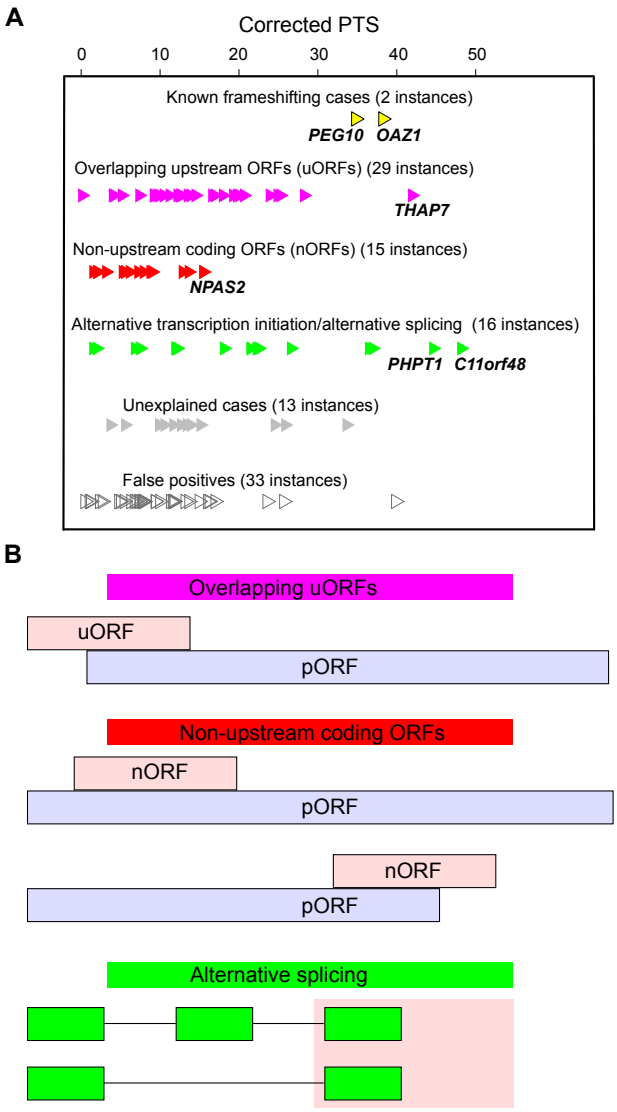

Figure 2.3: *Classification of dual coding regions. (A) Classification and PTS of 108 candidates. (B) Schematic organization of three major classes of dual coding. pORFs are shown as light blue bars and alternative frames as light pink bars. Splicing organization: green bars correspond to exons included in transcript variants and lines indicate intronic regions excised during splicing.*

differences in CDS lengths, the CDS coordinates of each mRNA were normalized into their relative positions within the CDS, where the length of the CDS is considered to be 1 (each CDS coordinate is divided by the total length of the coding region with the start taken as 0.0 and the stop as 1.0). Each CSCPD curve is evaluated at 100 equispaced normalized CDS positions between 0 and 1 using smoothing spline interpolation. A pointwise 95% confidence envelope for each sub-codon position (C1, C2, C3) was then obtained from the 95th percentiles of the 1000 CSCPDs at each normalized CDS position (see Fig. 2.2C and Supplemental Figure S3 in Appendix 5.3).

Under ideal conditions for detecting a frame transition (high coverage and uniform distribution of RPFs with the 2nd sub-codon position counts being higher than the 1st and the 3rd sub-codon position counts), there should be a local point at which the CSCPD reaches its maximum, and such a point should correspond to the location of the frame transition (see also Supplemental Material section entitled "Testing PTS on simulated dual coding sequences" in Appendix 5.3).

Thus, we used the area of CSCPD excess over the 95th percentiles for each subcodon position. The Periodicity Transition Score (PTS) is calculated as the sum of excess areas for each sub-codon position (PTS1, PTS2, PTS3) (Fig. 2.2C). An example of a PTS plot for an mRNA with a known case of programmed ribosomal frameshifting (human Antizyme 1 mRNA) is shown in Supplemental Figure S4 in Appendix 5.3.

To determine the threshold of PTS that can be used as an indicator of a frame transition in an mRNA, we calculated the PTS scores for a random 1000 mRNAs from the pool of 6000 most-covered genes (but outside of the pool of genes used for the 95th percentile calculations) and compared them with the PTS scores obtained for the same 1000 mRNAs, after introducing single nucleotide deletions to mimic translational frame transitions at different locations in the mRNA. The results of these comparisons are shown in Figure 2.2D. It can be seen that before introducing an artificial frameshift, the majority of mRNAs have a PTS below 10. Since it is expected that some of the 1000 mRNAs may have naturally occurring transitions, we decided to use a PTS of 10 as the threshold for selecting the candidates reported in this study. As can be seen from Figure 2.2D, when using a PTS threshold of 10, the potential false negative rate is higher if a reading frame transition occurs closer to either end of the coding region than if the transition occurs closer to the middle of the main reading frame. To estimate the p-value for transcripts with a PTS score of 10 and higher, we permuted the RPF densities of the 1000 most highly expressed transcripts and generated 1,000,000 artificial transcripts (see Supplemental Material in Appendix 5.3 for details). Transcripts with a PTS equal to or higher than 10, were considered as false positives. This yielded a p-value of 0.057. After removing cases where

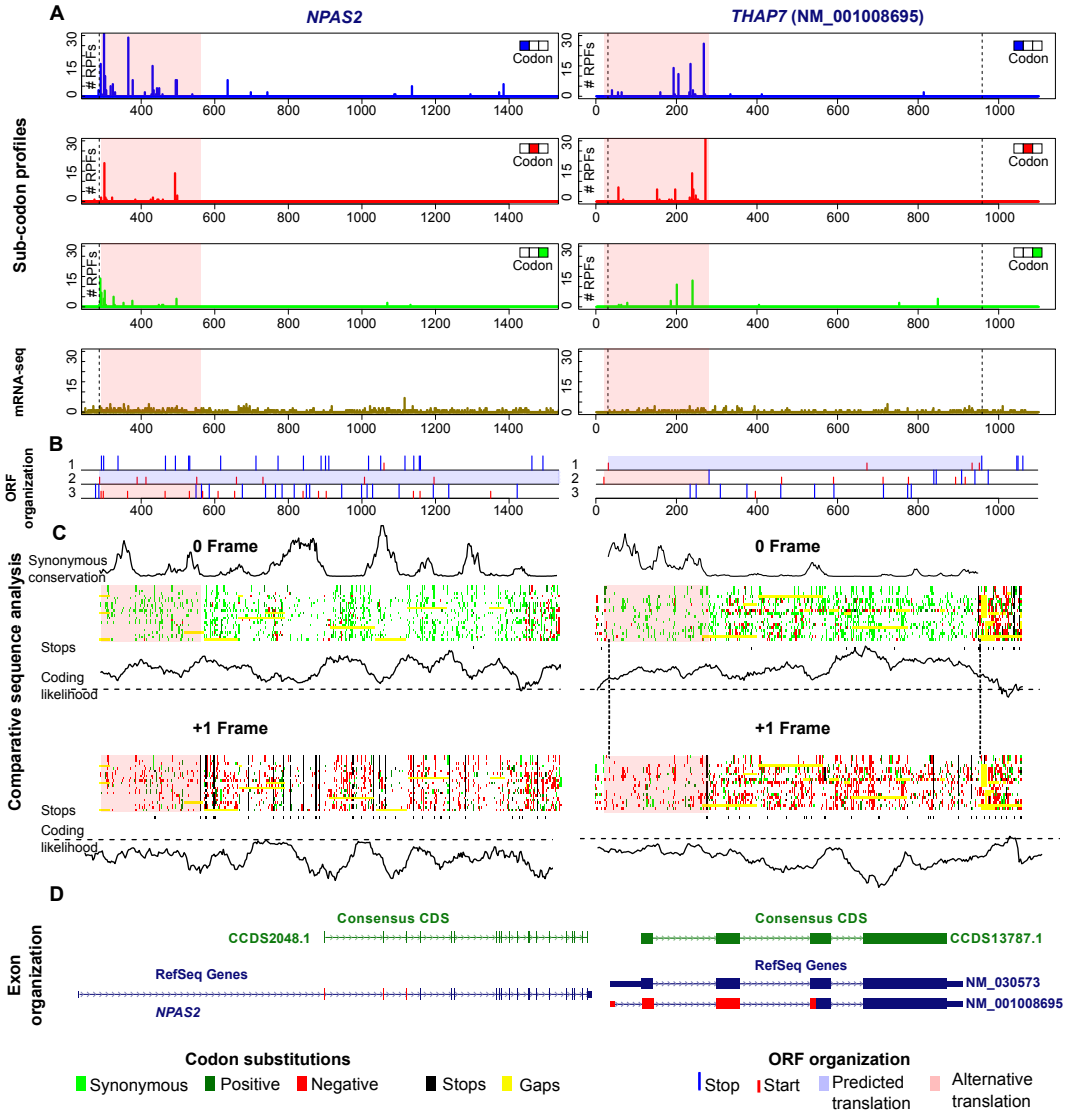

Figure 2.4: *Dual coding in NPAS2 mRNA due to the presence of a translated non-upstream ORF and in THAP7 mRNA due to the overlap of the main ORF with an uORF. (A) Sub-codon profile (top three rows) and mRNA-seq (fourth row) for NPAS2 mRNA (left, NM_002518) and THAP7 mRNA (right, NM_001008695). CDS coordinates are marked with dotted vertical lines. (B) ORF organization of NPAS2 mRNA (left) and THAP7 mRNA (right). The three reading frames are indicated as 1, 2, 3. Blue vertical lines indicate stop codons and start codons are indicated in red. Annotated CDS is shaded in light blue. The areas where translation in alternative frames is detected are shaded in light pink. (C) Comparative analysis of orthologous genomic sequences from 23 vertebrate species for NPAS2 (left) and from 19 vertebrate species for THAP7 (right). Coloured bars represent codon substitutions within multiple sequence alignments for the standard (top) and alternative (bottom) reading frames (detailed alignments are in Supplemental Figs. 121-122). Light green and dark green boxes correspond to synonymous and positive (in the BLOSUM62 matrix) substitutions respectively, red boxes correspond to negative (in BLOSUM62 matrix) non-synonymous substitutions. Gaps are shown in yellow and stop codons are in black. Stop codons are also aggregated across the entire alignment beneath each bar. Plots of coding likelihood are shown underneath the coloured bars for both reading frames as calculated with MLOGD. Synonymous position conservation for the standard translation phase (pORF) is shown above the coloured bar. (D) Exon organization of the NPAS2 locus (left) and the THAP7 locus (right). CCDS and RefSeq gene tracks from UCSC Genome Browser are shown in green and blue bars respectively. Alternatively decoded regions are indicated in red.*

sub-codon positions 1 and 3 contribute to a high PTS with no contribution from sub-codon position 2 (see section below on "Further refinements of PTS"), the re-estimated p-value for a PTS of 10 or higher, drops to 0.0084.

The above simulations of dual coding regions addresses a simple case, where the transition between alternative frames occurs at a specific location and all ribosomes continuing translation shift their reading frame. Such a situation occurs in the case of ribosomal frameshifting in *OAZ1* mRNA (Fig. 2.1C and 2.1D). However, with other examples of dual coding, certain sequence segments could be translated in two alternative frames. To explore how PTS performs on such cases of dual coding and how different features of dual coding affect PTS, we carried out additional simulations that are described in the Supplemental Material section "Testing PTS on simulated dual coding sequences" in Appendix 5.3. Figures 2.2E and 2.2F shows how PTS depends on features of dual coding such as the length of the overlapping region and the density (absolute and relative) of RPFs.

2.2.2 Further refinements of PTS.

After the PTS had been computed for a set of mRNAs for which ribosome profiles are available, it appeared that the PTS performs well in predicting mRNAs with reading frame transitions. For example, two known cases of ribosomal frameshifting (*OAZ1* (Matsufuji et al., 1995) and *PEG10* (Shigemoto et al., 2001; Manktelow et al., 2005)) had PTS scores among the highest ten.

However, a large source of false positive cases (from manual examination) was found in situations where the PTS is high due to mutual fluctuations in the proportions of RPFs corresponding to the 1st and the 3rd sub-codon positions, with the 2nd position proportion unaffected. That is, PTS2, calculated for the 2nd position alone, is low. In our experience, the profiles containing bona fide transitions (either known or artificially introduced) always result in a high PTS2, along with either an increase in PTS1 or PTS3. This is because the number of RPFs aligning to sub-codon position 2 increases as a result of the transition in the reading frame. Correspondingly, the number of RPFs aligning to either sub-codon position 1 or 3, depending on whether the alternative frame is +1 or -1, decreases. Therefore, in this work, all mRNA sequences with PTS2 lower than both PTS1 and PTS3 were removed without further analysis.

In addition, our empirical manual analysis revealed many false positives with less than 100-or-so RPF locations per mRNA. Therefore we decided to apply an additional filter - the minimal number of RPF locations required in sub-codon position 2 - to reduce the false discovery rate. We found that 12 RPF locations or more in sub-codon position 2

greatly reduced the inclusion of false positives. To reduce the effect of single high peaks in sub-codon position 2 contributing to a high PTS, we removed the highest RPF peak in position 2 and recalculated PTS for each candidate. We then used this corrected PTS to score the candidates (see Supplemental Table 1 in Appendix 5.3).

Yet further manual examination of profiles revealed that a high PTS can occur for an mRNA transcript whose profiles are inconsistent with the behavior expected in the case of a frame transition. The most prominent example of how a high PTS could be generated for an mRNA without a transition is the existence of a single isolated peak of RPFs corresponding to a second position within a codon. As discussed earlier, such peaks do not necessarily reflect fluctuations in the noise of the technique, but could be due to a systematic alteration of RPF length in a local region in a sequence dependent manner. However, irrespective of the origin of such peaks, they significantly contribute to the PTS and generate a large number of false positives in our analysis.

2.2.3 Dual coding genomic sequences

Manual evaluation of sub-codon profiles allowed us to categorize these candidates into six groups as outlined in Figure 2.3A. The functional categories illustrated in Figure 2.3B include (i) instances where dual coding occurs due to overlaps between regulatory upstream ORFs (uORFs) and main protein coding ORFs (pORFs); (ii) overlaps between pORFs and non-upstream ORFs (nORFs); (iii) transcript variants generated as a result of alternative transcription initiation or alternative splicing.

The second class of dual coding is the most surprising, as translation of uORFs and dual coding due to alternative splicing has been documented previously. Fifteen mRNAs in our set were classified as containing non-upstream protein coding ORFs (nORFs) (see Fig. 2.3A and Supplemental Table 1 in Appendix 5.3). The sub-codon profile of the top scoring nORF, neuronal transcription factor *NPAS2* (RefSeq mRNA NM_002518) is shown in Figure 2.4A (left panel) and Supplemental Figure 15 in Appendix 5.3. This candidate has RPFs aligning in an alternative ORF which is located close to the 5' UTR. Comparative analysis of the genomic sequences revealed absence of stop codons in the alternative ORF in 22 of the 23 available vertebrate *NPAS2* orthologs (see alignments in Fig. 2.4C). The exact sequence in the vicinity of the predicted start codon is CTA***ATG***G**ATG**AAG**ATG**AGAA (where ATG codons are shown in bold, the predicted pORF start codon is also in italics, and the alternative frame start codons are underlined) [for simplicity and consistency we use T to denote both uridines and thymidines here and elsewhere]. It is plausible that start codons in such close proximity to each other compete for initiation (Matsuda and Dreher,

2006) and therefore the role of the alternative ORF may be regulatory. It would be very interesting to investigate a potential relationship between such regulation and a function of *NPAS2* as a part of a molecular clock in the human brain (Reick et al., 2001). A somewhat similar situation of competing initiator ATG codons is observed in initiation factor *EIF4E2* mRNA (Refseq NM_004846) (see Supplemental Fig. 17 in Appendix 5.3). Such a competition could be regulated by changes in the stringency of start codon selection, that has been shown to be mediated by *EIF1* and *EIF5* factors (Loughran et al., 2012).

Among all nORF candidates, about one half are situated entirely within the corresponding pORF (nested nORFs), while the other half extend into the 3'-UTRs.

The largest class of dual coding genomic sequences (29 instances) corresponds to regulatory uORFs overlapping pORFs. The profile of the highest scoring uORF candidate, transcription suppressor *THAP7* mRNA, is shown in Figure 2.4A (right panel) and Supplemental Figure 50 in Appendix 5.3. A significantly higher density of RPFs is observed in the region of the uORF that overlaps the pORF. Interestingly, the highest peak of RPF density is situated near the stop codon of the uORF. Perhaps ribosomes stall at the end of this uORF in a manner similar to the well-established ribosome stalling mediated by the MAGDIS peptide encoded by the uORF in S-adenosylmethionine decarboxylase (*AMD1*) (Hill and Morris, 1993) or by a specific mRNA-binding protein as in the regulation of the *MSL2* mRNA by Sex lethal (Medenbach et al., 2011). Comparative sequence analysis of available *THAP7* orthologs from the genomes of 19 vertebrates (right panel, Fig. 2.4C) suggests that the amino acid sequence of the *THAP7* uORF evolved faster than the protein sequence encoded in the same region by the pORF frame. However, none of the sequences from other vertebrates contain stop codons within the region corresponding to the *THAP7* uORF. Moreover, the position of the uORF stop codon is almost universally conserved among the analyzed orthologs. This points to the evolutionary significance of this uORF and suggests that the significance of its translation may be mainly regulatory rather than for the production of a functional protein product. It also highlights the limitations of dual coding detection by comparative sequence analysis, since alternatively translated regions do not necessarily evolve under the same evolutionary constraints as protein coding regions.

In nearly half of the detected translated uORFs we failed to find suitable ATG codons for initiation of uORF translation. This could be either due to non-ATG initiation (Ingolia et al., 2009; Ingolia et al., 2011; Ivanov et al., 2011), incompleteness of the corresponding RefSeq mRNA at the 5'-end, or differences among alternative splice variants in the 5' UTR.

Another source of dual coding in the human genome is alternative splicing (Liang and

Landweber, 2006). Some transcript variants contain sequences originated from the same genomic loci, but in different translational phases relative to the initiation codon. An established case where the same exon is translated in two alternative frames is the *CDKN2A* (also known as *INK4a*) (Quelle et al., 1995) gene. Among our candidates we have identified 16 instances of dual coding that can be attributed to alternative splicing events or to initiation of transcription at alternative starts (see Supplemental Table 1 in Appendix 5.3). The top scoring candidate among identified cases is *C11orf48* (Fig. 2.5 and Supplemental Fig. 67 in Appendix 5.3). The majority of RPFs for *C11orf48* mRNA, are located at the end of the predicted RefSeq pORF and extend into the 3' UTR (Fig. 2.5A). Examination of mRNA-seq reads for *C11orf48* revealed that mRNA-seq density is increased in the area of the alternatively decoded region (pink area in Fig. 2.5A). This indicates that the RPFs are likely to originate from the translation of additional transcripts whose sequences are not included in the RefSeq database. Indeed, such a transcript exists among Ensembl transcripts (accession number ENST00000524958). The sub-codon profile that has been generated for ENST00000524958 and the distribution of RPFs is consistent with the CDS predicted for that transcript (right panel, Fig. 2.5A). Additional independent evidence that the area of high RPF density encodes a protein product in an alternative frame corresponding to ENST00000524958 transcript is provided by evolutionary analysis. The multiple alignment of genomic sequences corresponding to the *C11orf48* orthologs from 15 vertebrate species is shown in panel C of Figure 2.5. It can be seen that codon substitutions in the area with high RPF density are consistent with purifying selection acting on ENST00000524958 CDS which is in the +1 frame relative to RefSeq CDS. Also, it can be seen that conservation of synonymous positions in pORF codons (0 frame) are markedly elevated for the region corresponding to high RPF density. Strikingly, it can also be seen that conservation, positive coding likelihood and a lack of stop codons in the +1 frame are observed only for the short region with high RPF coverage and not for the full ORF. Moreover, Oyama et al. (2007) has detected expression of this alternative protein using mass spectrometry. Thus, the *C11orf48* locus is an example of a situation where the same genomic sequence is simultaneously translated in different frames in two alternative transcripts that co-exist in HeLa cells.

The situation where two alternative transcripts co-exist and are translated at the same time is not always the case. We also found situations where only one transcript is present in the cell under the given conditions. Such an example is *PHPT1* mRNA (6th top in Supplemental Table 1 in Appendix 5.3) that is illustrated in Figure 2.6 and Supplemental Figure 60 in Appendix 5.3. The *PHPT1* gene contains four exons. Two mRNA transcripts are known for this gene: NM_001135861 contains all four exons and encodes isoform 2

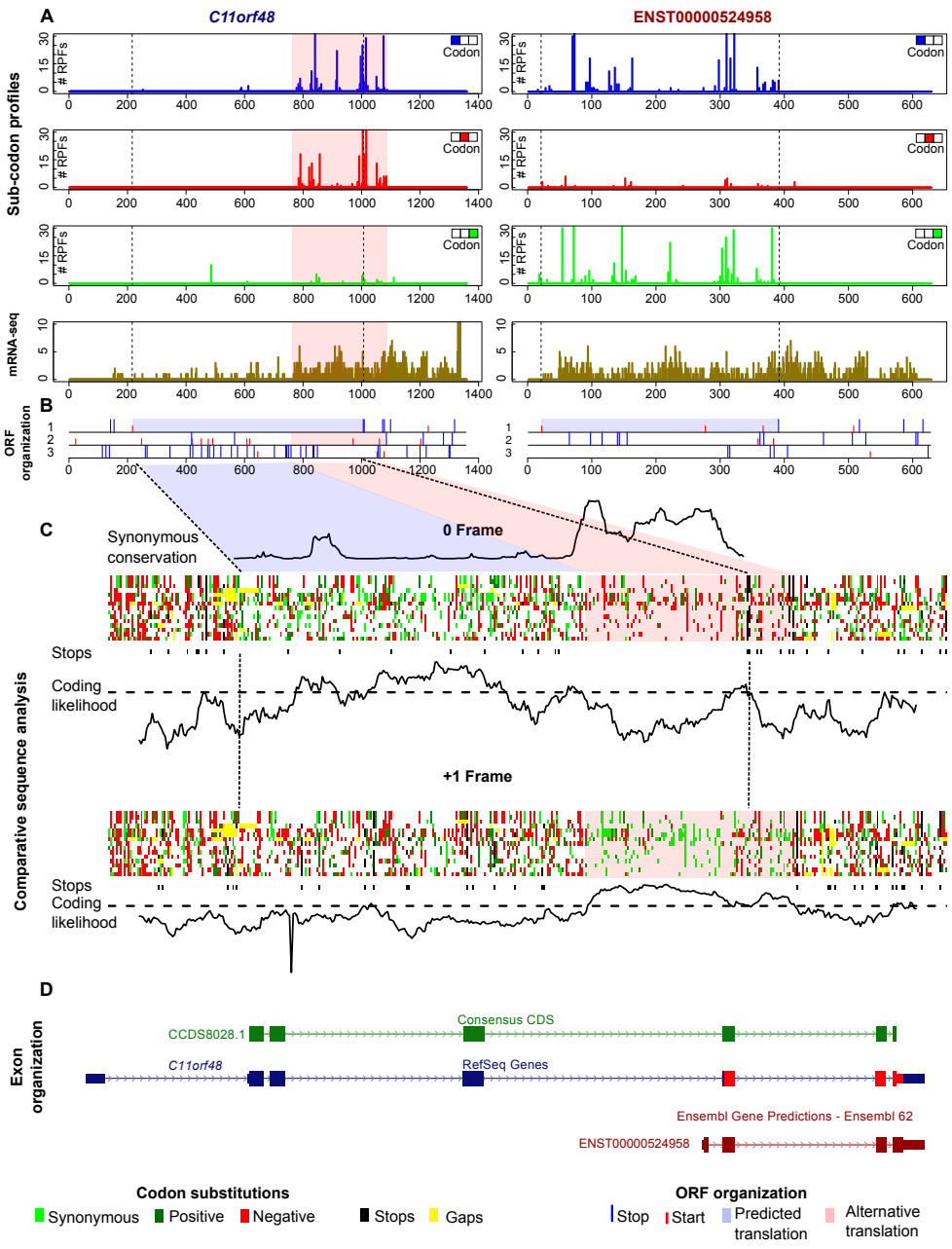

Figure 2.5: *Dual coding in C11orf48 locus. (A) Sub-codon profile and mRNA-seq for RefSeq mRNA NM_024099 (left) and predicted Ensembl transcript ENST00000524958 (right). (B) ORF organization of NM_024099 mRNA (left) and ENST00000524958 (right). (C) Comparative sequence analysis of corresponding genomic alignments from 15 vertebrate species for RefSeq mRNA NM_024099. (D) Exon organization of the C11orf48 locus. For detailed description see Figure 2.4 legend. The higher density of mRNA-seq reads for NM_024099 (fourth row panel A, left) in the shaded pink area indicates that RNA-seq reads are being generated from an additional transcript variant corresponding to Ensembl transcript ENST00000524958. In panel C, it can be seen that for most of the predicted CDS, codon substitutions are consistent with RefSeq CDS predictions (the area is greener in the zero-frame). However, for the pink shaded area, substitutions are consistent with protein coding evolutionary signatures in the +1 frame. It can be seen that the coding likelihood for the +1 frame exceeds the threshold in the area of dual decoding. The conservation plot of synonymous codon positions, shown above the 0 frame, shows that conservation of synonymous positions is significantly higher in the shaded pink area. This is consistent with the purifying selection acting on protein coding sequences in two frames in this region.*

of *PHPT1* while NM_014172 lacks the third exon and encodes isoform 3 of *PHPT1*. As a result of exon skipping, the 3'-terminal exon is positioned in different reading frames relative to the initiation codon in these two transcripts (Fig. 2.6D). The ribosome profile was initially built for the transcript with the longest isoform (see Methods in section 2.4). However it produced a high dual coding score because RPFs at the 3' end of the CDS originate from the alternative transcript where this region is in a different frame. The analysis of RNA-seq fragments (Fig. 2.6A) shows the lack of fragments corresponding to the skipped exon, thus suggesting that only the short transcript is expressed in HeLa cells.

Although we were able to identify many dual coding regions, a number of mRNAs with a high PTS are false positives. Ribosome profiles of about a third of all candidates produced high scores for reasons other than dual translation. The most prominent example (4th top in Supplemental Table 1 in Appendix 5.3) is a profile for the dystrophin *DMD* mRNA (RefSeq NM_004010). The sub-codon profile for this mRNA (Supplemental Fig. 90 in Appendix 5.3) is inconsistent with dual decoding and scored highly due to the limitations of our computational technique (see Supplemental Discussion in Appendix 5.3 for details). In addition to the 33 false positive candidates, 13 candidates have sub-codon profiles that suggest dual coding but dual coding cannot be explained by their ORF organization. These unexplained cases are discussed in the Supplemental Discussion in Appendix 5.3.

2.3 Discussion

Our work demonstrates the applicability of the ribosome profiling technique for the detection of translated reading frames in human mRNAs. This allowed us to identify a number of genomic loci that are being translated in more than one frame. An immediate simple question raised by this study is how many dually decoded regions are in the human genome. A primitive extrapolation of the number of cases identified among 6,000 genes would indicate approximately 1%. However, this is clearly an underestimate for the following reasons. Firstly, RPF coverage for the majority of analyzed mRNAs is lower than what is required for detecting such regions. Secondly, the method allows dually coded regions to be detected only if the alternative frame has RPF coverage comparable to, or higher than, that of the standard frame (see Figs. 2.2E and 2.2F and also "Testing PTS on simulated dual coding sequences" in the Supplemental Material in Appendix 5.3). It is reasonable to expect that there are many cases where an alternative frame is translated less efficiently than the standard one. More sensitive statistical techniques coupled with deeper ribosome profiling are needed for the detection of such cases. Thirdly, ribosome profiling

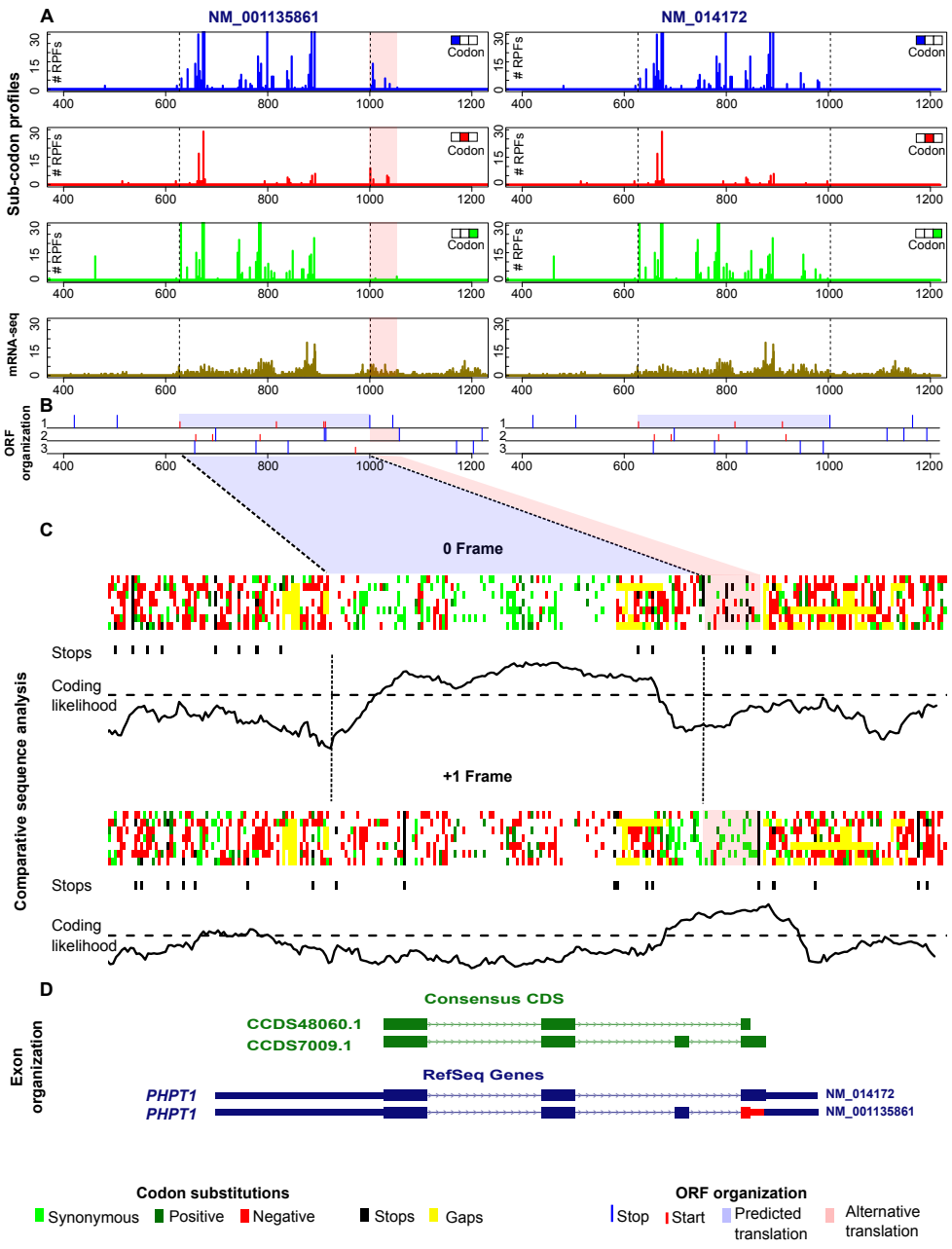

Figure 2.6: *Dual coding in alternatively spliced PHPT1 exon. (A) Sub-codon profile and mRNA-seq for PHPT1 mRNA variant NM_001135861 (left) and variant NM_014172 (right). (B) ORF organization for NM_001135861 (left) and NM_014172 (right). (C) Analysis of codon substitutions within the multiple alignments of orthologous genomic sequences for NM_001135861. (D) Exon organization of the two PHPT1 mRNA variants. For notations see Figure 2.4 legend. Sub-codon profiles for variant NM_001135861 (panel A, left) which is the longest isoform (see Methods), indicate that while the translated frame is the same as the CDS for most of the CDS region (low RPFs density for the second [red] position), the sequence is translated in the +1 frame relative to the CDS frame at its end and downstream (pink shaded area). In addition, there is an evident gap in translation in the sub-codon profile and mRNA-seq just prior to the pink shaded area which corresponds to the third exon in PHPT1 mRNA variant NM_001135861 (panel D). As a result, the fourth exon in the NM_001135861 mRNA is in an alternative frame relative to the CDS start codon. Codon substitution analysis of multiple sequence alignments (panel C) is consistent with the dual decoding of the 5'-end of the fourth exon. Synonymous and positive non-synonymous substitutions are predominant in both the zero and +1 frames in the locations where RPFs are found.*

experiments were carried out under particular conditions. Dual decoding is likely to be regulated. Therefore, only the standard frame may be translated under particular experimental conditions. Finally, it is likely that dual coding is more prevalent in low expressed genes, since highly expressed genes are optimized for efficient translation and their coding sequences are too restrained to accommodate additional coding information. The pool of genes analyzed in this study, however, is limited to highly expressed genes.

This is supported by comparison of our list of dual coding candidates with sets of genes that have been predicted as dual coding in previous studies. We have been able to identify reading frame switches in two (*OAZ1* and *PEG10*) out of six known cases of programmed ribosomal frameshifting in humans (Bekaert et al., 2010), three of which did not have sufficient coverage by RPFs. Our method did not detect the well-established examples of dual coding: *GNAS*, *XBP1* (Calfon et al., 2002) and *CDKN2A* (also known as *INK4a*) (Quelle et al., 1995), as dual coding. In the case of *GNAS* (Nekrutenko et al., 2005), the dual coding isoform was not expressed under the conditions of Guo. et al. (2010). The *XBP1* gene produced a high PTS, but failed to pass our additional filters (see Methods, section 2.4). The longest isoform of *CDKN2A*, which encodes the 16INK4a protein, had a low PTS and consequently did not appear in our final list of candidates. Among the 40 genes predicted in (Chung et al., 2007) study, transcripts from 12 are part of the 6000 pool that we have used. Three of these ARF-containing genes (*DNMT3A* , *BBX* and *RBAK*) have a PTS > 10 but were removed from our final list of 108 genes by subsequent filters explained in Methods (section 2.4). Comparative sequence analysis of 29 mammalian genomes revealed 19 (12 sense and 7 anti-sense) novel dual coding gene candidates of which 6 (sense) were among the 6000 transcripts to which our method was applied. (Lin et al., 2011). One of them, *UBE2E2* (NM_152653), was identified as dual coding in our study. The discrepancies between our dataset and previous predictions do not invalidate either predictions or our approach. It is possible that the translation of an alternative frame in previously described candidates does not occur in HeLa cells or occurs at a rate that is insufficient to be detected by our method.

The present study demonstrates that the coding and translational landscape of the human genome is more sophisticated than previously appreciated. Further development of high-throughput approaches for studying translation, combined with the growing power of comparative sequence analysis, provides an opportunity for obtaining quantitative information on the versatility of decoding and translation at the whole-cell level.

2.4 Methods

2.4.1 Analysis of 6,000 human mRNAs.

The predictive power of PTS depends on the number of RPFs that can be aligned to an mRNA sequence (see Supplemental section "Testing PTS on simulated dual coding sequences" in Appendix 5.3). Therefore, we restricted our analysis to a limited set of mRNAs for which a comparatively high number of RPFs are available. We used the 6,000 mRNA sequences with the highest RPF coverage from the dataset that has been reported by Guo et al. (2010). We found that RPF coverage lower than that of the top 6,000 is insufficient for statistically reliable determination of reading frames. Over eight hundred mRNA profiles scored above the PTS threshold. One hundred and eight of these passed the filters described in Results, section 2.2. This set of mRNA profiles was analyzed manually using additional information in order to explain the predicted frame transitions for each case. The analysis involved manual examination of the ORF organization and examination of the sub-codon profiles for the entire mRNA (as opposed to the analysis of just the previously annotated CDS region for the calculation of the PTS) in conjunction with "naked mRNA" profiles (RNA-seq). The analysis of the entire mRNA profile is required for those cases where an alternatively translated region corresponds to an ORF that overlaps the 5' or 3' end of the previously predicted CDS. In such situations we expect RPFs to also occur outside the CDS region. Such a distribution can then be used as additional evidence that the high PTS results from dual translation. We also expect RPF sub-codon proportions to be in accordance with the reading frame of the new ORF in the region in which it does not overlap with the main CDS. In addition, for a subset of cases, we examined multiple sequence alignments of corresponding genomic regions. The regions that are translated in alternative frames are expected to evolve under purifying selection with the ratio of non-synonymous to synonymous substitutions being significantly below 1 for the amino acid sequences encoded in alternative frames. Also the regions of dual coding are expected to have elevated conservation at synonymous codon positions, since synonymous positions in one reading frame are non-synonymous in the other. The details of the 108 mRNAs analyzed are given in Supplemental Table 1 in Appendix 5.3. For each mRNA sequence we provide individual sub-codon profiles and plots of ORF organization (Supplemental Figs. 9 to 116 in Appendix 5.3).

There are several mechanisms that can be responsible for the dual decoding of the genomic regions identified in this work. The method described does not allow discrimination between these mechanisms. Our classification of individual cases into mechanistic categories is based on external information that we have obtained from public bioinformatics

resources and therefore the validity of our predictions relies on what is available in those resources. For example, for a single mRNA variant we may observe the predicted CDS to be overlapped by a translated upstream ORF and therefore we would classify such a case as the translation of a uORF overlapping the main CDS. However, we cannot exclude the possibility that the ribosome profiles were derived from a different unreported splicing variant where the regions involved are joined together in such a way that the start codon of the uORF appears in frame with the previously predicted CDS and therefore all RPFs have been generated from a single ORF.

2.4.2 Generation of individual mRNA ribosome profiles.

Short sequence reads (corresponding to RPFs) generated during ribosome profiling experiments in HeLa cells (Guo et al., 2010) were obtained from the NCBI Gene Expression Omnibus (Edgar et al., 2002) (accession GSE22004). The mRNA sequences for the top 6,000 genes from quantification files were downloaded from the NCBI Refseq database (Pruitt et al., 2009) in fasta format in January 2011. Guo et al. (2010) quantification files comprise single RefSeq mRNA references for each gene where, for which genes with multiple isoforms, the longest isoform is chosen. To maximize the total number of RPFs for each gene, short reads from all available experiments in HeLa cells (SRR057511, SRR057512, SRR057516, SRR057517, SRR057521, SRR057522, SRR057526, SRR057529, SRR057532) were aggregated. The aggregated RPFs were then aligned, using the Bowtie short read aligner software package (Langmead et al., 2009). A seed region of the first 25 nucleotides at the 5'-end was used according to the method described by as in Guo et al. (2010). However, we allowed zero mismatches in the seed region.

Individual mRNA sub-codon profiles and ORF plots were then generated using custom Python and R scripts and the Biostrings package from the Bioconductor library (Gentleman et al., 2004). It has been shown previously that the distance between the 5'-end of an RPF and the position of the anticodon in the ribosomal A-site is about 15 nt (Guo et al., 2010). Therefore, to generate sub-codon profiles, each RPF was assigned to the mRNA coordinate corresponding to the 15th RPF nucleotide from the 5'-end. Thus sub-codon profiles represent the locations of the A-sites of the translating ribosomes.

The CSCPD and PTS were computed using custom scripts in R according to the algorithms described in Results, section 2.2.

2.4.3 Comparative sequence analysis

Multiz (Blanchette et al., 2004) multiple alignments for vertebrate species were obtained from the UCSC Genome Browser (Fujita et al., 2011) and visualized with the aid of a cgi script (kindly provided by Mike Lin, CSAIL, MIT) and additionally processed with custom R scripts. Sequences containing long consecutive gaps ($\geq$50 codons) were removed prior to the analysis. The coding likelihood for annotated CDS frames and alternative frames was quantified using MLOGD as described previously (Firth and Brown, 2006). Conservation at synonymous codon positions in annotated CDSs was computed as described previously (Firth and Atkins, 2009). Full MLOGD and synonymous substitution conservation plots for the examples described in the Results (section 2.2) are shown in Supplemental Figures 117 to 120 in Appendix 5.3.

Data Access

The R scripts for computing the CSCPD and PTS are provided at the end of the Supplemental Material in Appendix 5.3 and are also available on http://lapti.ucc.ie/bicoding.

Acknowledgments We are grateful to Mike Lin (MIT) for providing us with the cgi script for visualization of codon substitutions in multiple alignments and Avril Coghlan (UCC) for the R script for generating ORF plots.

Chapter 3

GWIPS-viz: Development of a ribo-seq genome browser

This chapter has been accepted for publication in Nucleic Acid Research Database issue.

We describe the development of GWIPS-viz (`http://gwips.ucc.ie`), an online genome browser for viewing ribosome profiling data. Ribosome profiling (ribo-seq) is a recently developed technique that provides Genome Wide Information on Protein Synthesis (GWIPS) *in vivo*. It is based on the deep sequencing of ribosome protected messenger RNA (mRNA) fragments which allows the ribosome density along all mRNA transcripts present in the cell to be quantified. Since its inception, ribo-seq has been carried out in a number of eukaryotic and prokaryotic organisms. Owing to the increasing interest in ribo-seq, there is a pertinent demand for a dedicated ribo-seq genome browser. GWIPS-viz is based on the University of California Santa Cruz (UCSC) Genome Browser . Ribo-seq tracks coupled with mRNA-seq tracks are currently available for several genomes: human, mouse, zebrafish, nematode, yeast, bacteria (*Escherichia coli* K12, *Bacillus subtilis*), human cytomegalovirus and bacteriophage lambda. Our objective is to continue incorporating published ribo-seq datasets so that the wider community can readily view ribosome profiling information from multiple studies without the need to carry out computational processing.

3.1 Introduction

Ribosome profiling is based on the isolation of messenger RNA (mRNA) fragments protected by ribosomes followed by massively parallel sequencing of the protected fragments or footprints. This allows the measurement of ribosome density along all mRNA transcripts present in the cell providing genome-wide information on protein synthesis

(GWIPS) *in vivo* (Weiss and Atkins, 2011). The ribosome profiling technique, also known as ribo-seq, was first carried out in *Saccharomyces cerevisiae* (Ingolia et al., 2009). Since the original publication, the technique has been carried out in many organisms including *Homo sapiens* (Guo et al., 2010; Stadler and Fire, 2011; Reid and Nicchitta, 2012; Lee et al., 2012; Fritsch et al., 2012; Shalgi et al., 2013; Liu et al., 2013; Loayza-Puch et al., 2013), *Mus musculus* (Guo et al., 2010; Ingolia et al., 2011; Lee et al., 2012; Thoreen et al., 2012; Shalgi et al., 2013), *Danio rerio* (Bazzini et al., 2012), *Caenorhabditis elegans* (Stadler and Fire, 2011; Stadler et al., 2012), *Saccharomyces cerevisiae* (Brar et al., 2012; Gerashchenko et al., 2012), *Escherichia coli* (Oh et al., 2011; Li et al., 2012), *Bacillus subtilis* (Li et al., 2012), human cytomegalovirus (Stern-Ginossar et al., 2012) and bacteriophage lambda (Liu et al., 2013).

To date, there have been two main strategies of ribosome profiling: ribosome profiling of initiating ribosomes and ribosome profiling of elongating ribosomes. For a review on the usages and advantages of each approach, please see (Michel and Baranov, 2013, Chapter 1 of this thesis).

The majority of published studies using ribosome profiling provide the raw sequencing data in NCBI's Sequence Read Archive (SRA) (Shumway et al., 2010). In addition, most published ribosome profiling experiments have corresponding naked mRNA controls, where total mRNA is randomly degraded to yield fragments of a size similar to ribosome protected fragments. For simplicity here we refer to it as mRNA-seq. mRNA-seq is carried out under the same experimental conditions. It helps to take into account the differential abundance of mRNA between experimental conditions and to monitor technical biases associated with cDNA library generation and sequencing.

Owing to the increasing popularity of the ribo-seq technique, the number of ribosome profiling experiments is expected to increase dramatically in the near future. However, the visualization of ribosome profiling data in a browser first requires pre-processing and aligning the raw sequencing reads. As with any type of next-generation sequencing data (NGS), demands are placed on biomedical researchers in terms of time, data storage, computational knowledge and prototyping of computational pipelines (Nekrutenko and Taylor, 2012). Web-based integrative framework tools such as Galaxy (Blankenberg et al., 2010) provide centralized platforms for researchers to carry out NGS alignment pipelines. However, because of decreasing costs, the coverage depth of ribo-seq and corresponding mRNA-seq data is continually increasing resulting in ever larger datasets. Consequently the computational resources required to process such data and the computer memory required to store such data may not be available to many biologists. Indeed, the time required to download, pre-process and align the raw data may be the most limiting factor of all for

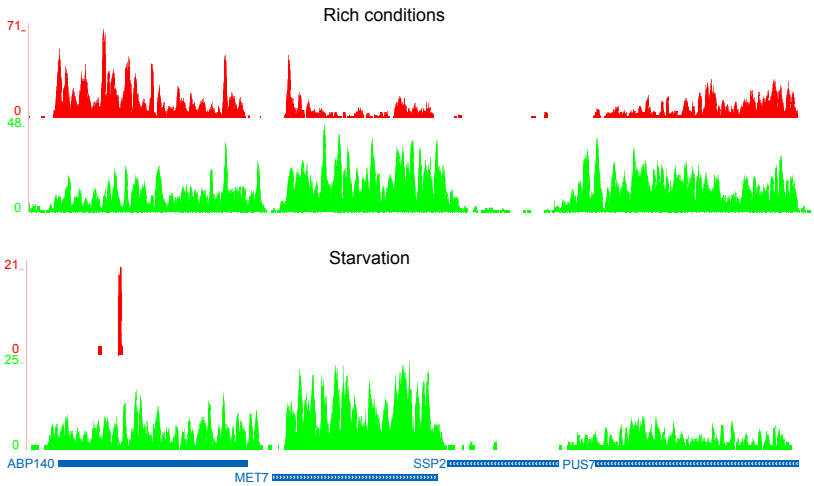

Figure 3.1: *Observing differential translation in GWIPS-viz. Ribo-seq (red) and RNA-seq (green) coverage plots for the S. cerevisiae genome locus containing ABP140, MET7, SSP2 and PUS7 genes from (Ingolia et al., 2009). Under starvation conditions (bottom panel), ABP140, MET7 and PUS7 are transcribed, but not translated.*

time-poor researchers.

To address these issues, we introduce GWIPS-viz (`http://gwips.ucc.ie`), a free online browser that is pre-populated with published ribo-seq data. The aim of GWIPS-viz is to provide an intuitive graphical interface of translation in the genomes for which ribo-seq data are available. Users can readily view alignments from many of the published ribo-seq studies without the need to carry out any computational processing. GWIPS-viz is based on a customized version of the University of California Santa Cruz (UCSC) Genome Browser (`http://genome.ucsc.edu`) (Meyer et al., 2013). Ribo-seq tracks, coupled with mRNA-seq tracks, are currently available for human, mouse, zebrafish, nematode yeast, two bacterial species (*Escherichia coli* K12 and *Bacillus subtilis*) and two viral genomes (human cytomegalovirus and bacteriophage lambda).

3.2 Usage

In GWIPS-viz, users can search for their gene(s) of interest in the genome(s) for which ribo-seq data are available and view a snapshot of the gene's translation under the conditions of the experiment. Ribosome coverage plots (red) and mRNA-seq coverage plots (green) display the number of reads that cover a given genomic coordinate. Figure 3.1 provides coverage plots for the *S. cerevisiae* genome locus containing *ABP140*, *MET7*, *SSP2*, and *PUS7* from (Ingolia et al., 2009) and illustrates how differential translation can be viewed in GWIPS-viz.

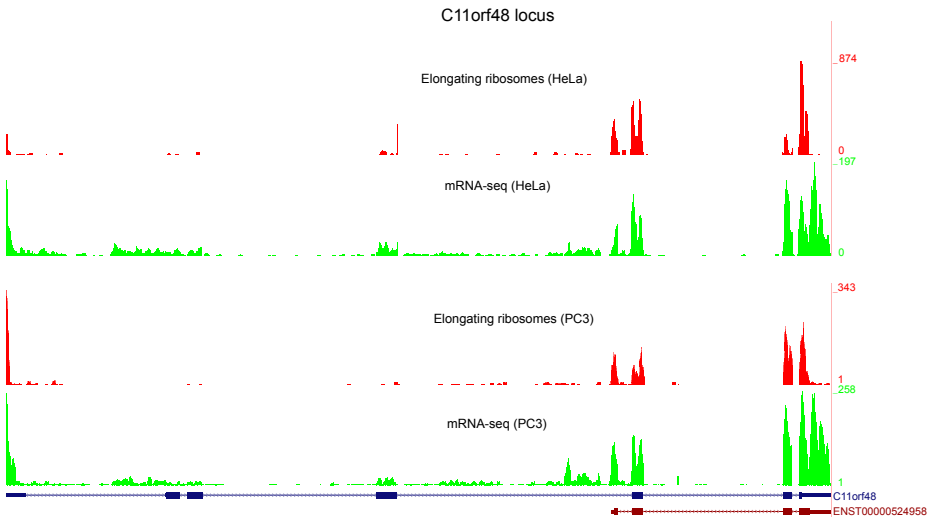

Figure 3.2: *Comparing profiles from independent studies. Data from different studies and different organisms can be compared in GWIPS-viz. The C11orf48 locus in the human genome is shown where translation of an ENSEMBL transcript (brown bars) not annotated in RefSeq (blue bars) has been identified in HeLa cells (Michel et al., 2012). As can be seen, translation of the Ensembl transcript occurs in both HeLa (Guo et al., 2010) and human PC3 cells (Hsieh et al., 2012).*

Users can visually identify which isoform(s) of a gene is transcribed and translated and also compare translation of the gene between different ribo-seq studies. For example, Figure 3.2 provides a comparison of two ribo-seq datasets obtained in different tissue-cultured human cells, HeLa (Guo et al., 2010) and PC3 cells (Hsieh et al., 2012). It can be seen that translation of a non-Refseq ENSEMBL transcript, reported based on the analysis of HeLa cell data (Michel et al., 2012), is observed in both datasets.

For the eukaryotic datasets, ribosome profiles display the number of footprint reads at a particular genomic coordinate that align to the A-site (elongating ribosomes) or P-site (initiating ribosomes) of the ribosome, depending on the study. For the prokaryotic datasets, a weighted centred approach (Oh et al., 2011) is used to indicate the positions of ribosomes. Figure 3.3 shows ribosome profile densities in a region of the *E. coli* genome that includes the gene *dnaX* (b0470). The ribosome density is scaled relative to the maximum density present within the displayed genomic segment. As a result, in the zoomed segment allowing visualization of neighbour genes (top), *dnaX* appears as lowly expressed. However, at a range covering only the *dnaX* locus, it can be seen that nearly all codons in the dnaX mRNA are covered with footprints. Moreover the coverage is sufficient to allow visual detection of decreased ribosome density downstream of the site of programmed ribosomal frameshifting which is known to cause about 50% of translating ribosomes to terminate

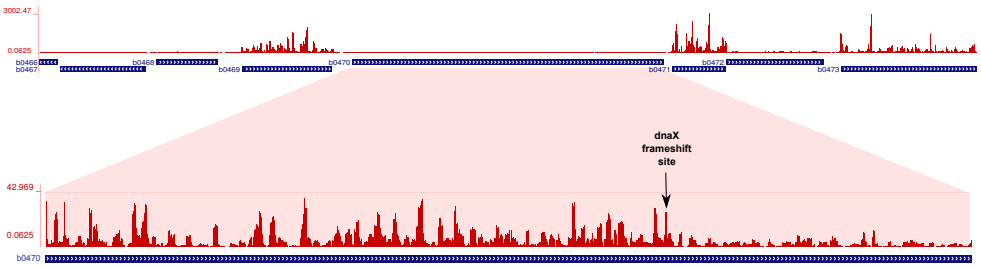

Figure 3.3: *Ribo-seq data for the dnaX locus in the E.coli genome. The top panel corresponds to a segment containing neighbouring genes. The bottom panel contains the dnaX coordinates only. The displayed ribosome density is scaled relative to the maximum density within the selected region. The position of the programmed ribosomal frameshifting site in dnaX is indicated with an arrow.*

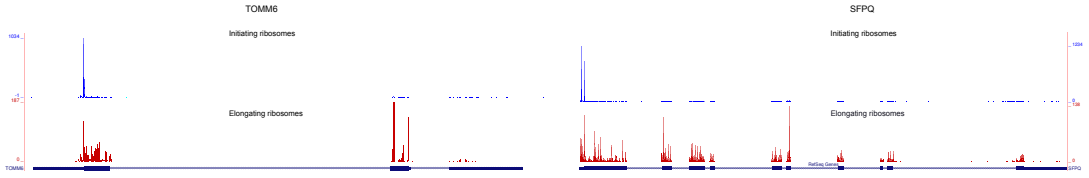

Figure 3.4: *Combining profiles of initiating and elongating ribosomes. Profiles of initiating (blue) and elongating (red) ribosomes generated in Human HEK 293 cells (Lee et al., 2012). Locations of elongating and initiating ribosomes are consistent with the annotated coding region of the TOMM6 gene (left). However, ribosome profiles of the SFPQ gene points to the existence of an additional start codon (stronger peak) upstream of the annotated start codon (weaker peak).*

prematurely (Larsen et al., 1997; Tsuchihashi and Kornberg, 1990).

Figure 3.4 provides an example of how ribo-seq tracks for elongating and initiating ribosomes can be compared. The example illustrates the data obtained in Human HEK293 cells (Lee et al., 2012) mapped to *TOMM6* and *SFPQ* genes. The latter gene apparently uses two sites of translation initiation for its expression.

3.3 Database design and implementation

GWIPS-viz is a customized version of the UCSC Genome Browser (Kent et al., 2002) version 269, and runs on Ubuntu Linux version 12.04.1, with Apache version 2.2.22 and MySQL 5.2.24. Static HTML and CSS files of the UCSC Genome Browser were downloaded from `http://hgdownload.cse.ucsc.edu/` and rehosted on our local server, while C source code for the CGI executables was downloaded and compiled using gcc 4.6.3. Selected parts of the MySQL databases were downloaded from the UCSC browser

for the majority of organisms included in GWIPS-viz (see Supplemental Fig. 125 in Appendix 5.4 for a schematic representation of the architecture of the GWIPS-viz browser).

Our partial mirror of the UCSC Genome Browser hosted on our server displays tracks for human (hg19), mouse (mm10), *S. cerevisiae* (sacCer3), zebrafish (danRer7), *C. elegans* (ce10), *E. coli* K12 (eschColi_K12), *B. subtilis* (baciSubt2), human cytomegalovirus (human herpesvirus 5 strain Merlin (HHV5)) and bacteriophage lambda (NC_001416) assemblies. Although several genome assemblies are available for many of the organisms, we chose to include only the most recent genome assembly for each organism.

Because the goal of GWIPS-viz is to be a browser for ribo-seq data, rather than a mirror of the UCSC browser, some of the functionality of the UCSC browser was removed in order to streamline the interface of GWIPS-viz. For example, the 'clade' menu in the genome selection menu was removed. In the browser window, the link "UCSC" was added in the top bar to allow the user to view the current genome position in the UCSC browser.

Depending on the organism, certain tracks were retained from the UCSC browser (Kent et al., 2002) and were consolidated into one group called 'Annotation Tracks'. Examples include RefSeq (Pruitt et al., 2012), Ensembl (Flicek et al., 2013), CCDS (Harte et al., 2012), Conservation (Pollard et al., 2010), RepeatMasker (Smit et al., unpublished data, www.repeatmasker.org), Mouse ESTs (Benson et al., 2004), SGD genes (Cherry et al., 2012), tRNA genes (Chan and Lowe, 2009).

Ribo-seq and mRNA-seq tracks were added by incorporating the outputs of our RNA-seq unified mapper (RUM) (Grant et al., 2011) alignment pipeline into the MySQL database. These tracks are divided into groups by publication and data type (ribo-seq and mRNA-seq). Tracks generated from uniquely mapping reads are colour coded according to their experiment type (elongating ribosome footprints are red, initiating ribosome footprints are blue and mRNA-seq reads are green).

3.3.1 Raw sequencing data retrieval

Published Ribo-seq and mRNA-seq datasets are downloaded from the NCBI Sequence Read Archive (SRA) (Shumway et al., 2010) and converted to FASTQ format using the fastq-dump utility (SRA Handbook citation, not in PubMed). Data from replicate experiments are consolidated into one dataset so as to have one browser track for each experimental condition.

3.3.2 Alignment pipeline

As there are no specific tools as yet for aligning ribo-seq data, RNA-seq tools are used in our pre-processing and alignment pipeline (see Supplemental Fig. 126 in Appendix 5.4 for a schematic representation of the alignment pipeline used for GWIPS-viz ribo-seq and mRNA-seq data).

Depending on the study, adaptor linker sequence or poly-(A) tails are trimmed from the 3' ends of reads using Cutadapt version 1.1 (Martin, 2011). Trimmed reads shorter than 25 nucleotides are discarded.

Contamination from ribosomal RNA (rRNA) may account for a significant proportion of the raw reads even after depletion by subtractive hybridization during the experiment. Hence it is desirable to remove rRNA reads from the dataset before performing alignments in order to increase the proportion of informative sequences and improve alignment efficiency. To detect reads that are the result of rRNA contamination, trimmed reads are aligned to rRNA sequences using Bowtie (Langmead et al., 2009). Bowtie version 0.12.8 is run using the -v option allowing three or fewer mismatches between the read sequence and the reference (rRNA) sequence. All reads that align to rRNA are discarded.

In most eukaryotes, a proportion of ribosome footprints will span splice junctions, i.e. the read will span the 3' end of one exon and the 5' end of another. There is the added complexity that ribo-seq reads are typically ~30 nucleotides in length. Hence the short-read alignment program needs to be capable of aligning reads of ~30nt across splice junctions. We use the RUM, (current version 2.0.5_05) (Grant et al., 2011). RUM handles splice junctions by using the short read aligner Bowtie (Langmead et al., 2009) to align sequence reads to both the genome and transcriptome and merging the results, before attempting to map remaining unaligned reads using another existing aligner, BLAT (Kent, 2002).

Owing to the relatively short lengths of ribosome footprint reads, a read may align to two or more distinct genomic locations due to sequence similarity. RUM outputs information separately for uniquely mapping reads and non-uniquely mapping reads (reads that align to several positions in the genome). Currently we provide tracks of uniquely mapping reads only in GWIPS-viz.

RUM's output files include a SAM alignment file showing the alignment(s) for each read, files giving the span of the alignment in genomic coordinates (RUM_Unique and RUM_NU) and coverage files (RUM.cov and RUM_NU.cov) listing the depth of coverage of reads across the genome.

The coverage files generated by the RUM alignment, RUM_Unique.cov and RUM_NU.cov, are in four column bedGraph format. The bedGraph data are converted into bigWig for-

mat, an indexed binary format that results in higher performance (Kent et al., 2010).

Ribosome profiles are generated from the RUM_Unique and RUM_NU files by obtaining the number of footprint reads whose 5' ends align at a given genomic coordinate (with an offset of 12nt designating the ribosome P-site for initiating ribosomes or 15nt for the ribosome A-site for elongating ribosomes).

3.4 Future Plans

We plan to expand the existing repertoire of ribo-seq tracks by integrating publically available ribosome profiling experiments as they become available.

GWIPS-viz currently displays the positions of the ribosomes mapped to the reference genomes. In the case of eukaryotic organisms that extensively use RNA splicing, visualization of ribosome positions in GWIPS-viz could be problematic due to a large number of long introns. Therefore, visualization of ribosome positions mapped to individual RNA transcripts is among our top priorities.

We currently provide ribo-seq and mRNA-seq tracks of uniquely mapping reads only. In the future, we wish to provide a differential display that will incorporate non-unique mapping reads (mapping to two or more locations in the genome) with uniquely mapping reads.

We also aim to provide access to the Galaxy platform from within GWIPS-viz so that researchers who generate their own ribo-seq experimental data can pre-process and align their data with the tools provided within Galaxy and then view the alignments in GWIPS-viz.

In addition, we aim to design a track specifically for the UCSC Genome Browser which will display whether a region is translated or not (one global track per genome for which ribosome profiling data exists). If a user is interested in further details of the data (cell type or tissue, particular condition, specific density profile), they can be found in GWIPS-viz where individual tracks for each experiment are provided.

Our overall objective is to continuously improve the service we provide in GWIPS-viz. As GWIPS-viz is under intensive development, some of the features described in this article could become outdated soon. Hence we encourage users to post their questions, comments and feedback on the GWIPS-viz forum. Furthermore, as ribosome profiling is a relatively recent technique that is still evolving and undergoing optimization, we provide forums for discussing the experimental protocol itself, its applications and analysis of the data. In this way, GWIPS-viz will not only be a centralized repository to visualize ribosome profiling data, but its forums will encourage researchers to actively engage in

the establishment of quality standards for ribosome profiling which will be of benefit to the community in general.

Acknowledgement

We are grateful to the UCSC Genome Bioinformatics Group for the permission to modify and rehost the browser.

Bibliography

Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 100(7):3889–94.

Atkins, J.F. and Gesteland, R.F. (2010). Recoding : expansion of decoding rules enriches gene expression. Springer: New York.

Bartel, D.P. (2004). MicroRNAs : Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes. Cell 116:281–297.

Bazzini, A.A., Lee, M.T. and Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science 336(6078):233–7.

Bekaert, M., Firth, A.E., Zhang, Y., Gladyshev, V.N., Atkins, J.F. and Baranov, P.V. (2010). Recode-2: new design, new search tools, and many more genes. Nucleic Acids Research 38(Database issue):D69–74.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2004). GenBank: update. Nucleic acids research 32(Database issue):D23–6.

Bilanges, B., Argonza-Barrett, R., Kolesnichenko, M., Skinner, C., Nair, M., Chen, M. and Stokoe, D. (2007). Tuberous sclerosis complex proteins 1 and 2 control serum-dependent translation in a TOP-dependent and -independent manner. Molecular and cellular biology 27(16):5746–64.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. and Miller, W. (2004). Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. Genome Res 14(4):708–715.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] Chapter 19(January):Unit 19.10.1–21.

Boström, K., Wettesten, M., Borén, J., Bondjers, G., Wiklund, O. and Olofsson, S.O. (1986). Pulse-chase studies of the synthesis and intracellular transport of apolipoprotein B-100 in Hep G2 cells. The Journal of biological chemistry 261(29):13800–6.

Brar, G.a., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T. and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science 335(6068):552–7.

Calfon, M., Zeng, H., Urano, F., Till, J.H., Hubbard, S.R., Harding, H.P., Clark, S.G. and Ron, D. (2002). IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. Nature 415(6867):92–6.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., Brar, G.A., Weissman, J.S., Regev, A., Thierry-Mieg, N., Cusick, M.E. and Vidal, M. (2012). Proto-genes and de novo gene birth. Nature 652:1–66.

Chan, P.P. and Lowe, T.M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic acids research 37(Database issue):D93–7.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S. and Wong, E.D. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research 40(Database issue):D700–5.

Chiba, S., Kanamori, T., Ueda, T., Akiyama, Y., Pogliano, K. and Ito, K. (2011). Recruitment of a species-specific translational arrest module to monitor different cellular processes. Proceedings of the National Academy of Sciences of the United States of America 108(15):6073–8.

Cho, J., Chang, H., Kwon, S.C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y.K. and Kim, V.N. (2012). LIN28A is a suppressor of ER-associated translation in embryonic stem cells. Cell 151(4):765–77.

Chung, W.Y., Wadhawan, S., Szklarczyk, R., Pond, S.K. and Nekrutenko, A. (2007). A first look at ARFome: dual-coding genes in mammalian genomes. PLoS computational biology 3(5):e91.

Churbanov, A., Rogozin, I.B., Babenko, V.N., Ali, H. and Koonin, E.V. (2005). Evolutionary conservation suggests a regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes. Nucleic acids research 33(17):5512–20.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K. and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences of the United States of America 104(49):19428–33.

Clements, J.M., Laz, T.M. and Sherman, F. (1988). Efficiency of Translation Initiation by Non-AUG Codons in Saccharomyces cerevisiae. Molecular and cellular biology 8(10):4533–6.

Dana, A. and Tuller, T. (2012). Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. PLoS Computational Biology 8(11):e1002755.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Molecular bioSystems 5(12):1512–26.

Dowling, R.J.O., Topisirovic, I., Alain, T., Bidinosti, M., Bruno, D., Petroulakis, E., Wang, X., Larsson, O., Selvaraj, A., Kozma, S.C., Thomas, G. and Sonenberg, N. (2010). mTORC1-mediated cell proliferation, but not cell growth, controlled by the 4E-BPs. Science 328(5982):1172–1176.

Edgar, R., Domrachev, M. and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research 30(1):207–10.

Fabian, M.R. and Sonenberg, N. (2012). The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. Nature structural & molecular biology 19(6):586–93.

Fabian, M.R., Sonenberg, N. and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. Annual review of biochemistry 79:351–79.

Firth, A.E. and Atkins, J.F. (2009). A conserved predicted pseudoknot in the NS2A-encoding sequence of West Nile and Japanese encephalitis flaviviruses suggests NS1' may derive from ribosomal frameshifting. Virology journal 6:14.

Firth, A.E. and Brown, C.M. (2006). Detecting overlapping coding sequences in virus genomes. BMC bioinformatics 7:75.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J.P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. and Searle, S.M.J. (2013). Ensembl 2013. Nucleic acids research 41(Database issue):D48–55.

Fresno, M., Jiménez, a. and Vázquez, D. (1977). Inhibition of translation in eukaryotic systems by harringtonine. European journal of biochemistry / FEBS 72(2):323–30.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J. and Brosch, M. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome research 22(11):2208–18.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D. and Kent, W.J. (2011). The UCSC Genome Browser database: update 2011. Nucleic acids research 39(Database issue):D876–82.

Genolet, R., Araud, T., Maillard, L., Jaquier-Gubler, P. and Curran, J. (2008). An approach to analyse the specific impact of rapamycin on mRNA-ribosome association. BMC medical genomics 1:33.

Gentilella, A. and Thomas, G. (2012). Cancer biology: The director's cut. Nature 485(7396):50–51.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology 5(10):R80.

Gerashchenko, M.V., Lobanov, A.V. and Gladyshev, V.N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proceedings of the National Academy of Sciences .

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome research 17(6):669–81.

Gingras, A., Raught, B. and Sonenberg, N. (2004). mTOR signaling to translation. Current topics in microbiology and immunology 279:169–197.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J. and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science (New York, N.Y.) 312(5770):75–9.

Godchaux, W., Adamson, S.D. and Herbert, E. (1967). Effects of cycloheximide on polyribosome function in reticulocytes. Journal of molecular biology. 27(1):57–72.

Grant, G.R., Farkas, M.H., Pizarro, A., Lahens, N., Schug, J., Brunk, B., Stoeckert, C.J., Hogenesch, J.B. and Pierce, E.a. (2011). Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM). Bioinformatics (Oxford, England) 27(18):2518–2528.

Guo, H., Ingolia, N.T., Weissman, J.S. and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466(7308):835–840.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T.G., Nusbaum, J.D., Morozov, P., Ludwig, J., Ojo, T., Luo, S., Schroth, G. and Tuschl, T. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA (New York, N.Y.) 17(9):1697–712.

Han, Y., David, A., Liu, B., Magadán, J.G., Bennink, J.R., Yewdell, J.W. and Qian, S.B. (2012). Monitoring cotranslational protein folding in mammalian cells at codon resolution. Proceedings of the National Academy of Sciences of the United States of America 109(31):12467–72.

Harte, R.A., Farrell, C.M., Loveland, J.E., Suner, M.M., Wilming, L., Aken, B., Barrell, D., Frankish, A., Wallin, C., Searle, S., Diekhans, M., Harrow, J. and Pruitt, K.D. (2012). Tracking and coordinating an international curation effort for the CCDS Project. Database : the journal of biological databases and curation 2012:bas008.

Hill, J.R. and Morris, D.R. (1993). Cell-specific Translational Regulation of S- Adenosylmethionine Decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame. 268(1):726–731.

Hinnebusch, A. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. Annual review of microbiology :59:407–450.

Hsieh, A.C., Liu, Y., Edlind, M.P., Ingolia, N.T., Janes, M.R., Sher, A., Shi, E.Y., Stumpf, C.R., Christensen, C., Bonham, M.J., Wang, S., Ren, P., Martin, M., Jessen, K., Feldman, M.E., Weissman, J.S., Shokat, K.M., Rommel, C. and Ruggero, D. (2012). The translational landscape of mTOR signalling steers cancer initiation and metastasis. Nature 485:55–61.

Hu, W. and Coller, J. (2012). What comes first: translational repression or mRNA degradation? The deepening mystery of microRNA function. Cell research 22(9):1322–1324.

Ingolia, N.T. (2010). Genome-Wide Translational Profiling by Ribosome Footprinting. Methods Enzymol. 470:119–142.

Ingolia, N.T., Brar, G.a., Rouskin, S., McGeachy, A.M. and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature Protocols 7(8):1534–1550.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science (New York, N.Y.) 324(5924):218–23.

Ingolia, N., Lareau, L. and Weissman, J. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell :1–14.

Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F. and Baranov, P.V. (2011). Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic acids research 39(10):4220–34.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nature reviews. Genetics 10(12):833–44.

Janas, M.M. and Novina, C.D. (2012). Not lost in translation: stepwise regulation of microRNA targets. The EMBO journal 31(11):2446–7.

Jiang, L., Schlesinger, F., Davis, C.a., Zhang, Y., Li, R., Salit, M., Gingeras, T.R. and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21(9):1543–51.

Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Research 12(4):656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, a.M. and Haussler, a.D. (2002). The Human Genome Browser at UCSC. Genome Research 12(6):996–1006.

Kent, W.J., Zweig, a.S., Barber, G., Hinrichs, a.S. and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics (Oxford, England) 26(17):2204–7.

Kiran, A. and Baranov, P.V. (2010). DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics (Oxford, England) 26(14):1772–6.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44(2):283–92.

Kozak, M. (1987a). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. Molecular and cellular biology 7(10):3438–45.

Kozak, M. (1990). Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proceedings of the National Academy of Sciences of the United States of America 87(21):8301–5.

Kozak, M. (1995). Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. Proceedings of the National Academy of Sciences of the United States of America 92(15):7134.

Kozak, M. (1997). Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. The EMBO journal 16(9):2482–92.

Kozak, M. (1987b). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Research 15(20):8125–8148.

Kozak, M. (1989). Context Effects and Inefficient Initiation at Non-AUG Codons in Eucaryotic Cell-Free Translation Systems. Molecular and cellular biology 9(11):5073.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. Gene 299(1-2):1–34.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3):R25.

Larsen, B., Gesteland, R.F. and Atkins, J.F. (1997). Structural probing and mutagenic analysis of the stem-loop required for Escherichia coli dnaX ribosomal frameshifting: programmed efficiency of 50%. Journal of molecular biology 271(1):47–60.

Larsson, O., Sonenberg, N. and Nadon, R. (2010). Identification of differential translation in genome wide studies. Proceedings of the National Academy of Sciences of the United States of America 107(50):21487–92.

Larsson, O., Sonenberg, N. and Nadon, R. (2011). anota: Analysis of differential translation in genome-wide studies. Bioinformatics (Oxford, England) 27(10):1440–1.

Lee, S., Liu, B., Huang, S.X., Shen, B. and Qian, S.B. (2012). PNAS Plus: Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proceedings of the National Academy of Sciences :1–9.

Li, G.W., Oh, E. and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature .

Liang, H. and Landweber, L.F. (2006). A genome-wide study of dual coding regions in human alternatively spliced genes. Genome research 16(2):190–6.

Lin, M.F., Kheradpour, P., Washietl, S., Parker, B.J., Pedersen, J.S. and Kellis, M. (2011). Locating protein-coding sequences under selection for additional , overlapping functions in 29 mammalian genomes :1916–1928.

Liu, B., Han, Y. and Qian, S.B. (2013). Cotranslational Response to Proteotoxic Stress by Elongation Pausing of Ribosomes. Molecular Cell :1–11.

Loayza-Puch, F., Drost, J., Rooijers, K., Lopes, R., Elkon, R. and Agami, R. (2013). P53 Induces Transcriptional and Translational Programs To Suppress Cell Proliferation and Growth. Genome Biology 14(4):R32.

Loughran, G., Sachs, M.S., Atkins, J.F. and Ivanov, I.P. (2012). Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. Nucleic acids research 40(7):2898–906.

Manktelow, E., Shigemoto, K. and Brierley, I. (2005). Characterization of the frameshift signal of Edr, a mammalian example of programmed -1 ribosomal frameshifting. Nucleic acids research 33(5):1553–63.

Mann, M. and Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. Nature biotechnology 21(3):255–61.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17:10–12.

Matlin, A.J., Clark, F. and Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. Nature reviews. Molecular cell biology 6(5):386–98.

Matsuda, D. and Dreher, T.W. (2006). Close spacing of AUG initiation codons confers dicistronic character on a eukaryotic mRNA. RNA 12:1338–1349.

Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J.F., Gesteland, R.F. and Hayashi, S. (1995). Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell 80(1):51–60.

Medenbach, J., Seiler, M. and Hentze, M.W. (2011). Translational control via protein-regulated upstream open reading frames. Cell 145(6):902–13.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer, T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D. and Kent, W.J. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research 41(Database issue):D64–9.

Meyuhas, O. (2000). Synthesis of the translational apparatus is regulated at the translational level. European journal of biochemistry / FEBS 267(21):6321–30.

Michel, A.M. and Baranov, P.V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley interdisciplinary reviews. RNA .

Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F. and Baranov, P.V. (2012). Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res 22(11):2219–2229.

Morris, D.R. and Geballe, A.P. (2000). Upstream Open Reading Frames as Regulators of mRNA Translation 20(23):8635–8642.

Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5(7):1–8.

Nekrutenko, A. and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet. 13(9):667–672.

Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P. and Makova, K.D. (2005). Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. PLoS genetics 1(2):e18.

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M. and Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic acids research 40(20):10084–97.

O'Connor, P.B.F., Li, G.W., Weissman, J.S., Atkins, J.F. and Baranov, P.V. (2013). rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics (Oxford, England) 29(12):1488–91.

Oh, E., Becker, A., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R., Typas, A., Gross, C., Kramer, G., Weissman, J. and Bukau, B. (2011). Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor InÂ Vivo. Cell 147(6):1295–1308.

Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T. and Sugano, S. (2007). Diversity of translation start sites may define increased complexity of the human short ORFeome. Molecular & cellular proteomics : MCP 6(6):1000–6.

Peabody, D.S. (1989). Translation Initiation at Non-AUG Triplets in Mammalian Cells. J Biol Chem. 264(9):5031–5035.

Plotkin, J. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nature reviews Genetics 12(1):32–42.

Plotkin, J.B. (2010). Transcriptional regulation is only half the story. Molecular systems biology 6(406):406.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome research 20(1):110–21.

Pöyry, T.A.A., Kaminski, A. and Jackson, R.J. (2004). What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? Genes & development 18(1):62–75.

Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic acids research 40(Database issue):D130–5.

Pruitt, K.D., Tatusova, T., Klimke, W. and Maglott, D.R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. Nucleic acids research 37(Database issue):D32–6.

Qian, W., Yang, J.R., Pearson, N.M., Maclean, C. and Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. PLoS genetics 8(3):e1002603.

Quelle, D.E., Zindy, F., Ashmun, R.a. and Sherr, C.J. (1995). Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. Cell 83(6):993–1000.

Rajasekhar, V.K., Viale, A., Socci, N.D., Wiedmann, M., Hu, X. and Holland, E.C. (2003). Oncogenic Ras and Akt signaling contribute to glioblastoma formation by differential recruitment of existing mRNAs to polysomes. Molecular cell 12(4):889–901.

Rancurel, C., Khosravi, M., Dunker, a.K., Romero, P.R. and Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. Journal of virology 83(20):10719–36.

Reick, M., Garcia, J.A., Dudley, C. and McKnight, S.L. (2001). NPAS2: an analog of clock operative in the mammalian forebrain. Science (New York, N.Y.) 293(5529):506–9.

Reid, D.W. and Nicchitta, C.V. (2012). Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. The Journal of biological chemistry 287(8):5518–27.

Ribrioux, S., Brüngger, A., Baumgarten, B., Seuwen, K. and John, M.R. (2008). Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC genomics 9:122.

Rogozin, I.B., Kochetov, A.V., Kondrashov, A., Koonin, E.V. and Milanesi, L. (2001). Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics 17(10):890–900.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature 473(7347):337–42.

Seidelt, B., Innis, C.A., Wilson, D.N., Gartmann, M., Villa, E., Trabuco, L.G., Becker, T., Schulten, K., Steitz, T.A. and Beckmann, R. (2009). Structural insight into nascent polypeptide chain-mediated translational stalling. Science 326(5958):1412–1415.

Shalgi, R., Hurt, J., Krykbaeva, I., Taipale, M., Lindquist, S. and Burge, C. (2013). Widespread Regulation of Translation by Elongation Pausing in Heat Shock. Molecular Cell :1–14.

Sharp, P.M. and Li, W.H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research 15(3):1281–1295.

Shenton, D., Smirnova, J.B., Selley, J.N., Carroll, K., Hubbard, S.J., Pavitt, G.D., Ashe, M.P. and Grant, C.M. (2006). Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. The Journal of biological chemistry 281(39):29011–21.

Shigemoto, K., Brennan, J., Walls, E., Watson, C.J., Stott, D., Rigby, P.W. and Reith, A.D. (2001). Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting. Nucleic acids research 29(19):4079–88.

Shine, J. and Dalgarno, L. (1975). Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. Eur J Biochem 230:221–230.

Shumway, M., Cochrane, G. and Sugawara, H. (2010). Archiving next generation sequencing data. Nucleic acids research 38(Database issue):D870–1.

Siwiak, M. and Zielenkiewicz, P. (2010). A comprehensive, quantitative, and genome-wide model of translation. PLoS computational biology 6(7):e1000865.

Smirnova, J.B., Selley, J.N., Sanchez-Cabo, F., Carroll, K., Eddy, A.A., McCarthy, J.E.G., Hubbard, S.J., Pavitt, G.D., Grant, C.M. and Ashe, M.P. (2005). Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. Mol Cell Biol. 25(21):9340–9349.

Sorensen, M.A. and Pedersen, S. (1991). Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. Journal of molecular biology 222(2):265–280.

Stadler, M., Artiles, K., Pak, J. and Fire, A. (2012). Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. Genome research .

Stadler, M. and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. Rna .

Steitz, J.A. (1969). Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. Nature 224(5223):957–964.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., Hein, M.Y., Huang, S.X., Ma, M., Shen, B., Qian, S.B., Hengel, H., Mann, M., Ingolia, N.T. and Weissman, J.S. (2012). Decoding Human Cytomegalovirus. Science 338(6110):1088–1093.

Thoreen, C.C., Chantranupong, L., Keys, H.R., Wang, T., Gray, N.S. and Sabatini, D.M. (2012). A unifying model for mTORC1-mediated regulation of mRNA translation. Nature 485(7396):109–13.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Olason, P.I., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R.A., López, G., Sadowski, M.I., Watson, J.D., Fariselli, P., Rossi, I., Nagy, A., Kai,

W., Stø rling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S.E., Reymond, A., Birney, E., Brunak, S.r., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D.T., Lengauer, T., Orengo, C.A., Patthy, L., Thornton, J.M., Tramontano, A. and Valencia, A. (2007). The implications of alternative splicing in the ENCODE protein complement. Proceedings of the National Academy of Sciences of the United States of America 104(13):5495–500.

Tsuchihashi, Z. and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proceedings of the National Academy of Sciences of the United States of America 87(7):2516–20.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I. and Pilpel, Y. (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141(2):344–54.

Tuller, T., Waldman, Y.Y., Kupiec, M. and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. Proceedings of the National Academy of Sciences of the United States of America 107(8):3645–50.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J. and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. Genes & development 20(5):515–24.

Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984). Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. Journal of molecular biology 180(3):549–576.

Vassilenko, K.S., Alekhina, O.M., Dmitriev, S.E., Shatsky, I.N. and Spirin, A.S. (2011). Unidirectional constant rate motion of the ribosomal scanning particle during eukaryotic translation initiation. Nucleic acids research 39(13):5555–67.

Vasudevan, S., Tong, Y. and Steitz, J.A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. Science (New York, N.Y.) 318(5858):1931–4.

Vázquez-Laslop, N., Ramu, H., Klepacki, D., Kannan, K. and Mankin, A.S. (2010). The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. The EMBO journal 29(18):3108–17.

Weiss, R., Lindsley, D., Falahee, B. and Gallant, J. (1988a). On the mechanism of ribosomal frameshifting at hungry codons. Journal of molecular biology 203(2):403–410.

Weiss, R.B. and Atkins, J.F. (2011). Translation Goes Global. Science 334(6062):1509–1510.

Weiss, R.B., Dunn, D.M., Dahlberg, A.E., Atkins, J.F. and Gesteland, R.F. (1988b). Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in Escherichia coli. The EMBO journal 7(5):1503–7.

Wilson, B.A. and Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. Genome biology and evolution 3:1245–52.

Wulff, B.E., Sakurai, M. and Nishikura, K. (2011). Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. Nature reviews. Genetics 12(2):81–5.

Yanagitani, K., Kimata, Y., Kadokura, H. and Kohno, K. (2011). Translational pausing ensures membrane targeting and cytoplasmic splicing of XBP1u mRNA. Science (New York, N.Y.) 331(6017):586–9.

Zaytseva, Y., Valentino, J., Gulhati, P. and Evers, B. (2012). mTOR inhibitors in cancer therapy. Cancer letters 319(1):1–7.

Zoncu, R., Efeyan, A. and Sabatini, D. (2011). mTOR: from growth signal integration to cancer, diabetes and ageing. Nature reviews Molecular cell biology 12(1):21–35.