High-Throughput Assessment of small Open Reading Frame Translation in *Drosophila melanogaster*

Muhammad Ali Shahzad Mumtaz

Doctor of Philosophy

School of Life Sciences

University of Sussex

Submitted May 2016

UNIVERSITY OF SUSSEX

Muhammad Ali Shahzad Mumtaz

Thesis submitted for the Degree of Doctor of Philosophy

High-Throughput Assessment of small ORF Translation

in *Drosophila melanogaster*

Summary

Hundreds of thousands of putative small ORFs (smORFs) sequences are present in eukaryotic genomes, potentially coding for peptides less than 100 amino acids. smORFs have been deemed non-coding on the basis of their high numbers and their small size that makes it extremely challenging to assess their functionality both bioinformatically and biochemically. The recently developed Ribo-Seq technique, which is the deep sequencing of ribosome footprints, has generated significant controversy by showing extensive translation of smORFs outside of annotated protein coding regions, including putative non-coding RNAs.. Our lab adapted the Ribo-Seq technique by combining it with the polysome fractionation in order to assess smORF translation in *Drosophila S2* cells.

This thesis provides a high-throughput assessment of smORF translation in *Drosophila melanogaster* by firstly implementing complementary techniques such as transfection-tagging and Mass spectrometry methods in order to provide an independent corroboration of the *S2* cell data (Chapter 3). Secondly, the in order to expand the catalogue of smORFs that are translated, I significantly improve upon the yield and sequencing efficiency of the Poly-Ribo-Seq protocol while adapting it to *Drosophila* embryos and then implementing it across embryogenesis divided in to Early, Mid and Late stages (Chapter 4). Currently, there is still a lot of debate in the field with regards to Ribo-Seq data analysis, and various computational metrics have been developed aimed at discerning ‘real’ translation events to background noise. Chapter 5 explores some of the metrics developed and establishes a translation cut-off suitable for designating small ORFs as translated. Altogether, the improvements introduced to the protocol and my data analysis shows the translation of 500 annotated smORFs, 500 smORFs in long non-coding RNAs and 5,000 uORFs, of which only one-third of each type of smORF has previous evidence of translation. These findings strengthen the establishment of smORFs as a distinct class of genes that significantly expand the protein coding complement of the genome.

Dedication

for Shahzad A. Mumtaz (*Papa*)

Acknowledgements

I would like to thank my PhD supervisor, Prof. Juan Pablo Couso, for his guidance and support during my time in his Lab and endless meetings about translation cut-offs. I would also like to thank my co-supervisor, Prof. Simon Morley for his reassurance and support. I am grateful for the company and support of my lab mates, Inyaki Pueyo for his immense knowledge and energy, Emile Magny for his enduring optimism and encouragement. Rose Phillips for cloning the tagged constructs and fun times at the bench and Sarah Bishop for help with flywork. I would like to say thank you to past members of the lab, Julie Aspden for being a friend and a mentor and introducing Ribo-Seq to the lab and Ying Chen for help with bioinformatics analyses. I am thankful to Pedro Patraquim for his help in revising sections of this thesis and stimulating conversations. I would also like to acknowledge John Chesebro for proof reading chapters of this thesis. I am deeply grateful for the help and support of my wife and fellow PhD student Unum amin for her unwavering support, love and care during the good times and the bad. I would also like to thank her for proofreading this thesis in its entirety and passionate discussions about smORFs. Finally I would like to thank both our families for understanding, encouragement, love and support during this PhD. I would also like to extend a very special thanks to my *viva voce* examiners for reading this thesis.

Table of Contents

List of Figures and Tables

List of Abbreviations

RBF - Ribosome Bound Fragment

smORF - small Open Reading Frame

modENCODE - Model Organism ENCyclopedia Of DNA Elements

SEP - smORF Encoded Peptide

DNA - Deoxyribonucleic acid

RNA - Ribonucleic acid

mRNA - messenger RNA

CDS - Coding DNA Sequence

FB - FlyBase

FBcds - FlyBase annotated canonical CDS

FBsmORF - FlyBase annotated smORF

ncrORF - non-coding RNA ORF

lncRNA - long non-coding RNA

uORF - upstream ORF

RPKM - Reads per Base Pair per Million.

TE - Traslation Efficiency

Cvg - Coverage

AEL - After Egg Laying

Em1 - Embryonic development 0-8H AEL

Em2 - Embryonic development 8-16H AEL

Em3 - Embryonic development 16-24H AEL
tal - *tarsal-less*

Scl - *Sarcolamban*

Hemo - *hemotin*

PCR - Polymerase Chain Reaction

rRNA - ribosomal RNA

tRNA - transfer RNA

BLAST - Basic Local Alignment Search Tool

tBLASTn - translated nucleotide BLAST

cDNA - complementary DNA

hnRNA heterogeneous nuclear RNA

NGS - Next-Generation Sequencing

Ribo-Seq - Ribosome profiling

Poly-Ribo-Seq - Polysome profiling

RNA-Seq - RNA Sequencing

S2 cells - Schneider 2 cell line

UTR - Untranslated Region

IVT - *In vitro* Translation

UV - Ultraviolet

SDS-PAGE - sodium dodecyl sulfate polyacrylamide gel electrophoresis

WB - Western Blot

GFP - Green Fluorescent Protein

GST - Glutathione S-transferase

MS - Mass Spectrometry

LC - Liquid Chromatography

HPLC - High Performance Liquid Chromatography

KDa - kilo Dalton

RRL - Rabbit Reticulocyte Lysate

TnT - Transcription and Translation

ICE - Insect Cell Extract

Uhg2 - U snoRNA host gene 2

HA - Hemagglutinin

MW - Molecular Weight

MWCO - Molecular Weight Cut-Off

HRP Horseradish Peroxidase

PTM - Post-translational modifications

CHX - Cyclohexamide

NEB - New England Bio labs

UMI - Unique Molecular Identifiers

tRF - tRNA derived fragment

altORF - alternative ORF

BAM - binary alignment

SNP - single nucleotide polymorphisms

RPB - reads per base pair

FBtr - FlyBase Transcript identifier

FBpp - FlyBase peptide identifier

Chapter 1: General Introduction

The Couso lab has been focused on identifying and characterising small Open Reading Frame (smORF) genes since the serendipitous discovery of *tarsal-less* (*tal*) (Galindo 2007). *tal* is a smORF gene that was identified in a mutant screen addressing the developmental patterning of *Drosophila melanogaster* leg morphology (Galindo 2000; Galindo 2005; Galindo 2002; Couso *et al.* 1998). Mutant flies that showed a lack of tarsal segments were found and this phenotype could be mapped to the *tal* gene locus, where a putative non-coding RNA gene was annotated according to Flybase. However, our lab demonstrated that four of the five tandemly arrayed smORFs in the tal transcript give rise to smORF-encoded peptides (SEPs) of 11, 12 and 32 amino acids in length. These Tal peptides are the shortest functional peptides ever discovered in metazoans (Galindo *et al.* 2007).

Since then, it has been shown that *tal* is involved in the differentiation of the second, third and fourth tarsal sub segments in the legs of the fruit fly, and is required for embryonic development (Galindo *et al.* 2007 Pueyo and Couso 2008). This study was one of the first to highlight the fact that a SEP is translated and can perform important functions in the fruit fly (Kondo *et al.* 2007; Galindo *et al.* 2007). Therefore, if such a small peptide could determine vital structures of the adult fly as well as embryonic viability, it could be possible that there may be other previously overlooked small peptides encoded in the *Drosophila* genome.

Gene Annotation and 'the trouble with smORFs'

The definition of a 'gene' has been evolving since its original definition in the early 1900s as a unit of inheritance, to its now most widely used definition, as a segment of DNA that is transcribed into RNA and contains an Open Reading Frame (ORF). An ORF is defined as a stretch of nucleotide sequence that begins with a 'start' codon (normally ATG) and ends with an in-frame 'stop' codon (TAG, TAA, TGA) thereby potentially encoding a functional protein by the triplicate nucleotide codons between the start and stop codons (reviewed in Gerstein *et al.* (2007)). This definition of a gene was coined after the discovery of the structure of DNA by James Watson and

Francis Crick in the 1950s (Watson, and Crick 1953), followed by the development of Sanger sequencing in the 1970s (Sanger *et al.* 1977) and later the Polymerase Chain Reaction (PCR) in the 1980s (Mullis, and Faloona 1987) that enabled pioneering discoveries into the mechanisms of transcription and translation of genes. These studies culminated in the promulgation of the Central Dogma of molecular biology — DNA is transcribed into messenger RNA, which is translated into proteins that act as the functional units of genes (Crick 1970).

This definition was based on decades of research following classical forward-genetics approaches, which rely on the genetic mapping of identified mutant phenotypes from a mutagen-treated population to specific genetic loci. Over the years, this forward-genetics approach has led to the discovery of the majority of protein-coding genes, including some smORF genes with obvious phenotypes, such as the pro-apoptotic gene Reaper that encodes for a 65aa peptide (White *et al.* 1994) and *tal*, which is involved in leg development (Galindo *et al.* 2007). Furthermore, smORF genes are small by nature and therefore have a lower probability of being deleted in induced mutation screens and, if mutated, the phenotype may be attributed to regulatory regions of neighbouring, annotated protein coding genes. Previously, only 2% of the genome has been estimated to consist of protein-coding transcripts (International Human Genome Sequencing Consortium 2004), and, until recently, the majority of the genome has been believed to consist of 'junk' DNA (Ohno 1972).

The sequencing of the complete yeast genome in the 1990s (Goffeau *et al.* 1996) represented a milestone achievement in our ability to understand genes on a whole-organism scale, thus giving rise to the field of genomics. Suddenly, the identification of genes was no longer limited to phenotypic screening of mutant populations and instead could be performed *de novo* by the computational analysis of the genome sequence (reviewed in Allen *et al.* (2004)). This empowered the efforts to characterise the whole protein-coding complement in the genome, and annotation was carried out through the identification of ORFs and their subsequent scoring by homology to known protein-coding sequences and evidence of transcription.

Interestingly, these analyses uncovered the presence of hundreds of thousands of smORFs throughout the yeast genome that did not perform well in bioinformatics analyses due to their small size (Das *et al.* 1997), as discussed further in the next section. Such large numbers of genes were difficult to reconcile with what was known at the time by classical pre-sequencing genetics data, especially for a unicellular organism such as yeast, which was predicted to have only 6,000 genes (Goffeau *et al.* 1996). Consequently, an argument was put forward that smORFs occur purely by chance, as rationalised by the statistical probability of a start and stop codon occurring in the DNA sequence. Therefore, a probability based cut-off of a minimum ORF length of 300 nucleotides was arbitrarily introduced for the annotation of protein-coding genes, particularly as they outnumbered longer genes by several orders of magnitude (Fickett 1995; Basrai *et al.* 1997). Thus, of the 260,000 putative smORFs identified in the yeast genome, only four hundred that showed clear homology to known proteins were annotated, and the remaining were left out of *Saccharomyces cerevisiae* genome annotation (Fickett 1995; Das *et al.* 1997), thus setting a precedent for subsequent genome annotations in other organisms where smORFs still remained questionable (Harrison *et al.* 2002).

Targeted Bioinformatic Searches for Functional smORFs

Despite their exclusion from genome annotations, the question remained as to whether the now-recognised smORFs encode functional proteins. Therefore, with increasingly available genomic data, a number of studies aimed at computationally distinguishing functional, putatively peptide-encoding smORFs were carried out, using comparative genomics and sequence composition, in a number of different organisms (Kastenmayer *et al.* 2006; Frith *et al.* 2006; Hanada *et al.* 2007). Our lab carried out a similar study in *D. melanogaster*, using a systematic bioinformatics search for putative smORFs in conventionally non-coding regions of the *Drosophila* genome (Ladoukakis *et al.* 2011). The bioinformatics pipeline identified 4,561 putative smORFs with ATG start codons and in-frame stop codons (TAG, TAA, TGA) encoding for peptides smaller than 100 amino acids and filtering was based on conservation between *D. melanogaster* and *D. pseudoobscura* (divergence of 25–55 Myr) using BLAST (Altschul *et al.* 1990). In order to distinguish functional smORFs within this cohort, this study used evidence of transcription along with two further evolutionary conservation criteria, the ratio of non-synonymous nucleotide substitutions (Ka) that cause a change in the amino acid sequence to synonymous nucleotide substitutions (Ks) in the smORF sequence, and in addition, the conservation of sequences around the smORF in syntenic regions of the *D. pseudoobscura* and *D. melanogaster* genomes, in order to narrow this number down to

401 ‘high-confidence’ smORFs. The Ka/Ks metric identifies those smORFs whose amino acid sequences are under greater purifying selection than the nucleotide sequence, thus suggesting protein function.

This study corroborated previous studies in estimating that putatively translated smORFs form a significant component of the genome, at around 3–5%. However, BLAST and related methods measure the number of conserved residues between species, and short sequences such as smORFs are by nature unable to obtain the high conservation scores that are the accepted indicator of functionality for canonical proteins (Lipman *et al.* 2002). This bias is apparent in the results by Ladoukakis *et al.* (2011) in which smORFs below 80 amino acids in length are generally scored quite poorly, and this bias is exacerbated for even smaller ORFs below 20 amino acids. Furthermore, it is difficult to score statistically significant values for very short sequences because the number of possible nucleotide substitutions is low, such that the Ka/Ks metric loses its ability to reliably predict functional ORFs smaller than 100 amino acids. These limitations of traditional computational assessment metrics for smORFs were made apparent by testing the bioinformatics pipeline on a small pool of 25 smORFs that had previous evidence of translation through mass spectrometry (MS), as only 9 (36%) of these passed the high-confidence filters. Therefore, although this approach was useful to narrow down a starting pool of nearly 600,000 putative genomic smORFs to 400 that could be used for further investigation, the results of these searches were not conclusive in the absence of a larger pool of functionally characterised smORF genes.

The non-coding RNA revolution — blurring the lines

Gene annotation efforts were classically validated through experimental evidence of transcription through the detection of spliced transcripts in complementary DNA (cDNA) libraries. However, smORFs, which are defined as less than 300 nucleotides, were generally disregarded from these analyses as they were discarded at the computational stage and it was generally accepted that around only 5% of the genome was transcribed, of which 1–3% consists of protein coding genes and the remaining percentage accounting for classical, structural non-coding RNA (ncRNA) genes, such as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) and small nucleolar RNAs (International Human Genome Sequencing Consortium 2004).

The development of high-throughput techniques allowed the detection of transcription at a genomic scale, starting with microarray technology (Okazaki *et al.* 2002; Bertone *et al.* 2004; Carninci *et al.* 2005) and later Next Generation Sequencing (NGS) of cDNAs by RNA Sequencing (RNA-Seq) (reviewed in Wang *et al.* (2009)). Surprisingly, RNA-Seq estimates revealed that 70–90% of the genome is transcribed at some point during development (Birney *et al.* 2007; Kapranov, and St Laurent 2012; Mercer *et al.* 2012; Djebali *et al.* 2012; Hangauer *et al.* 2013). These estimates were far greater than the 1–3% of the genome attributed to annotated protein-coding genes and thus generated a lot of debate at the time as to whether these were simply sequencing artefacts and background noise (Mortazavi *et al.* 2008; van Bakel *et al.* 2010) (reviewed in Kapranov and St. Laurent (2012)). However, these findings prompted great scrutiny into studying the role of these ncRNA transcripts. In particular, a lot of attention was given to the emerging regulatory roles of small ncRNAs, such as small interfering RNAs (siRNAs) and microRNAs (miRNAs), and further studies highlighted a huge diversity within this class of RNAs with the discovery of PIWI-interacting RNAs (piRNAs) and transcription initiation RNAs (tiRNAs) (reviewed in Morris and Mattick (2014) and Cech and Steitz (2014).

Intriguingly, these studies also revealed the presence of a class of longer RNAs that are transcribed from intergenic regions that are not processed into small regulatory RNAs (such as miRNAs). Consequently, these newly discovered transcripts that were longer than 200 nucleotides and did not contain an ORF encoding for a protein longer than 100 amino acids were annotated as long ncRNAs (lncRNAs). There is an estimated 10,000 lncRNAs in mammals and 2,000 in *Drosophila*; transcripts are on average 1 Kb in length and display mRNA-like properties, such as polyadenylation, 5′-capping and splicing, and have a similar half-life within the cell (Inagaki *et al.* 2005; Guttman *et al.* 2009; Cabili *et al.* 2011; Ulitsky *et al.* 2011; Clark *et al.* 2012). LncRNAs are regulated during animal development (Rinn *et al.* 2007; Dinger *et al.* 2008a), exhibit cell-type-specific expression (Mercer *et al.* 2008; Washietl *et al.* 2014) and some show evidence for evolutionary selection (Pollard *et al.* 2006; Guttman *et al.* 2009; Washietl *et al.* 2014; Hezroni *et al.* 2015). In addition, many lncRNAs are associated with human diseases (reviewed in Wapinski and Chang (2011)). However, most of the earlier studies characterising lncRNA function have focused on their role in transcriptional regulation based on their localisation to the nucleus, which have been reviewed extensively by Wilusz *et al.* (2009) and Rinn and Chang (2012).

More recent studies have shown that most lncRNAs are not only present in the cytoplasm, but may even be enriched there (Ulitsky, and Bartel 2013; Cabili *et al.* 2015). Furthermore, lncRNA transcripts have also been shown to associate with ribosomes in a widespread manner (Wilson, and Masel 2011; van Heesch *et al.* 2014). These cytoplasmic lncRNAs have been characterised as having a variety of regulatory functions; in mRNA translation elongation, as miRNA molecular sponges (Ulitsky *et al.* 2011), translation repression, mRNA stability through RNA binding proteins (Yoon *et al.* 2012) and affecting cap-independent translation (Carrieri *et al.* 2012) (reviewed in Fatica and Bozzoni (2013)). However, these characterised lncRNAs form only a small proportion of the total population and many, if not all, lncRNAs contain smORFs that may or may not be translated. Therefore, there is great interest in ascertaining the roles that lncRNAs may play in the cell, and what proportion are truly non-coding and which lncRNAs may encode functional peptides (Clamp *et al.* 2007; Dinger *et al.* 2008b; Housman, and Ulitsky 2015).

Biochemical assessment of smORF translation and function

The fact that only 25 of a possible 4,561 putative SEPs had previously been detected by proteomics experiments highlights the limited success of traditional proteomics approaches towards assessing smORF translation. This is partly due to the fact that mass spectra are typically searched against a database of annotated proteins and the peptides that are identified are mainly used to confirm and refine gene models derived from computational work, which has generally excluded smORFs (Washburn *et al.* 2001). However, a number of targeted proteomic studies have been conducted in the past with the aim of refining computational gene annotations and identify novel genes and such methods have improved existing gene annotations and discovered new genes in a range of different organisms (McGowan *et al.* 2004; Brunner *et al.* 2007; Tanner *et al.* 2007). However, these studies have had with limited success with regard to smORF genes due to the small size and hence low detectability of SEPs (reviewed in Andrews and Rothnagel (2014)).

Recent improvements to MS technology have resulted in greater sensitivity of the technique and have enabled the discovery of novel peptides encoded by smORFs (Yang *et al.* 2011) and those generated by proteolysis (Falth *et al.* 2006; Tinoco, and Saghatelian 2011), particularly in the identification of neuropeptides (Svensson *et al.* 2003), uORFs (Oyama *et al.* 2004) and biomarkers in body fluids (Clynen *et al.* 2008). Recently, peptidomics has also been applied to characterising smORFs by using bioinformatics tools to generate a custom database consisting of peptide sequences translated from putative unannotated ORFs within the transcriptome (Slavoff *et al.* 2013; Vanderperre *et al.* 2013). This has led to the development of bioinformatics tools geared towards matching MS data to an unbiased search space by building custom databases based on the six-frame translation of putative ORFs in the genome (Costa *et al.* 2013; Andrews, and Rothnagel 2014).

Individual functional characterisation of smORFs using a reverse genetics approach, such as those published by our lab (Galindo *et al.* 2007; Magny *et al.* 2013; Pueyo *et al.* 2016), is a highly labour intensive approach due to lack of an overt phenotype and difficulty in detecting the peptide through traditional biochemical methods. For example, antibodies specifically raised against SEPs have not been successful in detecting endogenous expression of the protein. Therefore, we have had to characterise SEP function through indirect genetic methods, such as through transgenic epitope tagged SEPs to rescue mutant phenotypes. In most cases the smORF transcript was annotated as a putative lncRNA. Therefore, we had to conduct extensive research to prove that the functional unit was the translated peptide and not the RNA, through lack-of-rescue of mutant phenotype by transgenic fly lines with a frame shift mutation in the ORF. Furthermore, each of the smORFs characterised in our lab have been shown to perform highly specific roles in a diverse range of biological processes, such as leg development (*tal)*, regulation of cardiac function (*Sarcolamban*; *Scl*) and intracellular trafficking in immune cells (*hemotin*; *hemo*). Therefore, our lab has developed significant expertise in the methods used to characterise these functions, such as electrophysiological measurements and phagocytic assays. Consequently, any false positives from previous bioinformatics screens to narrow down a candidate list for future characterisation resulted in a costly waste of resources. In addition, these studies, though imperative for providing proof of principle of functional SEPs, do not address the question of the extent of smORF translation and functionality. Therefore, there was a need for a reliable high-throughput, genome-wide assessment of smORF translation in order to enable further functional studies and most importantly to establish smORF genes as an undisputed part of the proteome.

Ribo-Seq: a genome-wide assessment of translated sequences

The ribosome profiling (Ribo-Seq) technique published by Ingolia *et al.* (2009) is the deep sequencing of ribosome bound fragments (RBFs) generated by nuclease footprinting and provides nucleotide-level resolution of translated sequences. In a typical Ribo-Seq protocol, elongating mRNA-associated ribosomes are 'frozen' using a translation inhibitor, and then nuclease treatment digests unprotected mRNA to generate RBFs that are protected by ribosome binding. The RBFs (~30nt long) are then purified from the ribosomes and used for cDNA library preparation, deep sequencing and mapping to the transcriptome or the genome. Prior to the development of this technique, microarrays and later RNA-Seq have been used as the standard method for proof of transcription and estimating mRNA abundance as a proxy for gene translation. However, these methods do not take into account post-transcriptional regulation, such as at the translational level and therefore tend to have poor correlation between mRNA and protein levels (reviewed in Maier *et al.* (2009)).

Sucrose gradient fractionation of mRNAs attached to ribosomes (polysomes) combined with the techniques above, allowed the development of polysome profiling (Johannes *et al.* 1999; Arava *et al.* 2003; Qin *et al.* 2007). This technique allows quantification of polysome-bound mRNAs, which represents the fraction of transcripts under translational control as compared with a total mRNA control. The translation ratios obtained from polysome profiling experiments generally show good correlation with protein levels (reviewed in Larsson *et al.* (2013)). However, association of whole transcripts with polysomes does not provide evidence of translation of putative unannotated ORFs, such as smORFs, in lncRNAs or in the 5′ leader sequences of canonical protein coding transcripts (upstream ORFs; uORFs), the translation of which may or may not be independent of the main protein-coding ORF (reviewed in Ingolia (2014)). This is exemplified by studies using Ribo-Seq that have shown extensive localisation of lncRNAs to polysomes but are unable to elucidate their actual coding status (Wilson, and Masel 2011; van Heesch *et al.* 2014).

The main advantage offered by Ribo-Seq is the nucleotide-level resolution of ribosome-bound regions within mRNAs, as the majority of Ribo-Seq sequencing reads map exclusively to the ORFs, and the abundance measurements serve as a proxy for protein translation. Some of the key findings from the application of this technique have been reviewed extensively (Kuersten *et al.* 2013; Ingolia 2014; Jackson, and Standart 2015). Briefly, Ribo-Seq has enabled the discovery of novel protein isoforms through alternative start and stop codon usage (Ingolia *et al.* 2011; Lee *et al.* 2012; Dunn *et al.* 2013), shown the prevalence of translation initiation at non-AUG start codons (Ingolia *et al.* 2011; Ingolia *et al.* 2009; Lee *et al.* 2012) and has also revealed novel polycistronic arrangements in eukaryote transcripts (Ingolia *et al.* 2011; Aspden *et al.* 2014), including uORFs and overlapping reading frames (Duncan, and Mata 2014; Michel *et al.* 2012) and translational mechanics and codon usage, for example elongation, pausing and frame-shifting (Qian *et al.* 2012; Michel *et al.* 2012; Han *et al.* 2014; Pop *et al.* 2014). Most significantly, from the perspective of the genomics field, Ribo-Seq has for the first time provided an assessment of the translation of unannotated questionable ORFs, such as smORFs, in conventionally non-coding regions of the transcriptome (Ingolia *et al.* 2011; Aspden *et al.* 2014; Duncan, and Mata 2014; Bazzini *et al.* 2014; Smith *et al.* 2014). Ribo-Seq has significantly expanded the proteome by offering increased sensitivity through the direct measurement of ribosome association as compared to previous applications of high-throughput techniques, such as MS, that are dependent on the stability and detectability of the translated peptide.

Poly-Ribo-Seq: polysome fractionation to enrich for smORFs

Our lab used the Ribo-Seq technique to assess smORF translation in *Drosophila Schneider 2* (*S2*) cells. In order to improve the detection of smORFs, our lab developed a variation of Ribo-Seq called Poly-Ribo-Seq, which combines Polysome fractionation and Ribo-Seq. This modification to the technique allowed the fractionation of mRNAs based on the number of ribosomes bound to the transcript, thus allowing the separation of translating mRNAs into small (up to 6 ribosomes) and large polysomes (more than 6 ribosomes). The rationale behind this approach is that by excluding the large polysome fraction we deplete longer, highly translated protein-coding ORFs that are packed with many ribosomes, thus enriching for smaller ORFs. An added benefit to including the polysomal fractionation step is that it allows the selection of actively translating transcripts based on their binding by two or more ribosomes. This allowed us to address the argument put forward by Chew *et al.* (2013) that Ribo-Seq signal within lncRNA transcripts is a result of RBFs derived from of the association of scanning 40S ribosomal subunits or the non-productive assembly of the 80S ribosome complex at putative start codons, and therefore not genuine translation.

The results from our first Poly-Ribo-Seq study showed the extensive translation of smORFs in the *Drosophila S2* cell transcriptome (Aspden *et al.* 2014). Annotated smORF genes within FlyBase (FB; 228 FB smORFs) are translated at a similar proportion to canonical protein coding genes (83%), while smORFs in 5′ UTRs (2,708 uORFs) and putative ORFs within annotated lncRNAs (313 non-coding RNA ORFs (ncrORFs)) were translated at a lower level (30–34%). These are termed 'dwarf' smORFs due to their shorter median length (20 amino acids) as compared with annotated FB smORFs (80 amino acids). The widespread translation of unannotated smORFs observed in this study was corroborated by numerous other studies across several different organisms that were published around the same time (Ingolia *et al.* 2011; Duncan, and Mata 2014; Bazzini *et al.* 2014; Smith *et al.* 2014). These findings have generated a significant amount of controversy (Bánfai *et al.* 2012; Guttman *et al.* 2013; Chew *et al.* 2013) in the already ambiguous field of lncRNAs and genomics regarding the translation of these previously ignored smORFs.

Thesis Aims

The research presented here was conducted with the aim of providing a high-throughput assessment of smORF translation in *Drosophila melanogaster* by using complementary techniques, such as transfection-tagging and MS methods, in order to provide an independent corroboration of the *S2* cell data (Chapter 3). Subsequently, in order to expand the catalogue of tested smORFs, I implemented the Poly-Ribo-Seq technique across three data sets that cover all the developmental stages of *Drosophila* embryogenesis to help shed light on the regulation and developmental context of smORFs that are translated *in vivo*. In order to achieve this, I modified the Poly-Ribo-Seq protocol to significantly improve the yield and sequencing efficiency whilst adapting it to *Drosophila* embryos and then implementing it across Early, Mid and Late stages of embryogenesis (Chapter 4). Chapter 5 discusses the ever-growing debate in the field with regards to Ribo-Seq data analysis and the various computational metrics that have been developed for this, aimed at discerning ‘real’ translation events from background noise. This process is especially challenging for very small ORFs due to their small size; therefore we explored some of these metrics to establish translation cut-off in the embryo data sets to score the translation of smORFs in a high-throughput manner.

Chapter 2: Materials and Methods

Common Molecular Biology Techniques

Bacterial transformation

100µl of *Escherichia coli* DH5-α chemically competent cells were thawed on ice and 1–5µl of plasmid or ligation reaction was added to the cells and mixed by gently flicking the tube, followed by 30 minutes incubation on ice. The cells were heat shocked at 42°C for 45 seconds in a water bath and incubated on ice for a further 5 minutes, 500µl of S.O.C. Medium (Invitrogen, Cat# 15544-034) was added to the tube followed by 1 hour incubation at 37°C with shaking. The transformation mixture was then plated onto Luria–Bertani (LB) agar plates containing 50 µg/ml ampicillin and left to grow overnight at 37°C.

Colony PCR

Colony PCR (Polymerase Chain Reaction) was used to screen bacterial transformants and was carried out by picking colonies off the agar medium plates with a toothpick and dipping the toothpick in a tube containing 20µl of standard PCR reaction mix (see below). The toothpicks were then streaked onto fresh LB agar plates and left to grow overnight at 37°C. PCR products were visualised using agarose gel electrophoresis.

Agarose gel electrophoresis

Visualisation of the various RNA, cDNA and PCR products was done by agarose gel electrophoresis. Gels were prepared by combining 0.5–1.5% (w/v) agarose, based on expected product size, in 1X TBE (89mM Tris, 89mM boric acid, 2mM EDTA) and heated in a microwave to dissolve. For visualisation, 0.5µg/ml ethidium bromide was added to the liquid agarose before pouring into the gel cast. The DNA or RNA product was combined with MassRuler DNA Loading Dye (Fermentas, Cat# SM0403) at a proportion of 1 µl dye for every 5µl product, and loaded into the wells of the agarose gel alongside the MassRuler DNA Ladder Mix (80–10,000bp fragments; Fermentas, Cat# SM0403) to assess length of product and calculate approximate sample concentration. Gel pictures were taken using an Uvidoc gel documentation system (Uvitec Cambridge) and UviPhotoMW image analysis software.

PCR amplification

PCR amplification was generally performed using the *Taq* PCR Core Kit (QIAGEN, Cat# 201225) with forward (Fw) and reverse (Rv) primers at a standard dilution of 10μM. Unless otherwise stated, a standard reaction is outlined below along with a typical amplification cycle. Primers were ordered from Invitrogen Life Sciences and the melting temperature was calculated using the Oligocalc online tool (http://biotools.nubic.north Western.edu/OligoCalc.html) with the nearest neighbour method.

Component	**Volume (μl)**
PCR buffer (10x)	5
Q-Solution (5x)	10
$MgCl_2$ (25mM)	1
dNTPs (10mM each)	1
Primer Fw (10mM)	1
Primer Rv (10mM)	1
Taq (5 U/μl)	0.5
Water	29.5
Template	1
Total	**50**

- Initial DNA denaturation at 94°C, 5 minutes
- Followed by 30–35 cycles of:
  - Denaturation at 94°C, 30 seconds
  - Annealing at 3–5°C below average primer Tm, 30 seconds
  - Extension at 72°C, 30 seconds to 2 minutes (depending on target length)

- Final extended extension at 72°C, 10 minutes
- Stored at 4°C until ready to use or stored at –20/80°C

Minipreparation of plasmid DNA

Plasmid DNA was isolated from bacterial cultures using the QIAprep Spin Miniprep Kit (QIAGEN, Cat# 27106) following the instructions provided by the manufacturer. 2–5 milliliters of overnight culture were spun to pellet the cells and the remaining supernatant was poured off. The cells were lysed in an alkaline solution and cleared by centrifugation. The DNA containing supernatant was transferred to a spin column containing a silica membrane to which the DNA is adsorbed. After several cleanup washes using the kit provided reagents, DNA was removed from the membrane using the low-salt Elution Buffer provided and stored at –20°C.

Gel extraction and PCR purification

Purification of PCR reactions and extraction of DNA from agarose gel slices were carried out using the QIAquick PCR purification (QIAGEN, Cat#28106) QIAquick Gel Extraction (QIAGEN, Cat#28706) kits, following the protocol provided by the manufacturer.

Restriction enzyme digest and DNA ligation

Restriction Enzyme (RE) digestion of plasmid DNA was carried out using enzymes and buffers purchased from New England Biolabs (NEB) according to the manufacturers' instructions and visualised using agarose gel electrophoresis. RE-digested vectors and inserts were gel-purified and the vector was treated with Calf Intestinal Alkaline Phosphatase (CIP) (NEB, Cat# M0290) to prevent self-ligation. Ligation reactions were carried out using T4 DNA Ligase (NEB, Cat# M0202) as per the manufacturers' instructions, using a typical 3:1 molar ratio of insert to vector and then subjected to bacterial transformation before screening using colony PCR.

Phenol/Chloroform purification

Phenol/Chloroform extraction was performed to remove proteins from nucleic acid solutions. A mixture of phenol:chloroform:isoamyl-alcohol (25:24:1 volume ratio) at pH 4.7 (RNA) or pH 8.0 (DNA) was added in a 1:1 volume ratio to the solution and shaken for 1 minute. The sample was then centrifuged for 5 minutes and the upper (aqueous) layer was transferred into a new microcentrifuge tube, the process was repeated until no protein was visible at the interface. The aqueous layer was then further purified with an equal volume of chloroform to remove traces of phenol before precipitation.

Nucleic acid precipitation

Nucleic acid precipitation was carried out by adding 3 M NaOAc pH 5.5 or 5 M NaCl (to a final concentration of 0.3 M) and 2.5 volumes of ice-cold 100% ethanol or 1 volume of isopropanol to the RNA/DNA solution, which was then left on dry ice or put at –80°C for minimum 30 minutes up to overnight. The precipitates were centrifuged at 16,000 g for 15 minutes and the DNA/RNA pellet was then washed in 70% ethanol and dried before re-suspending in TE (Tris–EDTA) buffer or distilled water.

SDS-PAGE

SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) was carried out using the Mini-PROTEAN Tetra gel-casting, running and blotting system from Bio-Rad. Stacking (4%) and resolving (16%) gels were hand-cast using the recipe below, and Tricine (Tris–glycine) gradient gels (10–20%) were bought ready-made from Bio-Rad (Cat# 4563114). Samples were mixed with 1X Tricine Sample Buffer (Bio-Rad, Cat# 1610739) supplemented with 2.5% v/v β-mercaptoethanol and heat denatured at 95°C for 5 minutes before loading on a gel. The gels were run in 1X Tricine Buffer (10mM Tris, 10mM Tricine, 0.01% SDS; pH 8.3) from Bio-Rad at 200V until the dye front reached the bottom of the gel (~40 minutes).

16% Resolving Gel	
Component	*Volume*
ddH_2O	2.6 ml
40% Acrylamide	3.2 ml
1.5 M Tris; pH 8.8	2 ml
10% SDS	80 µl
10% APS	80 µl
TEMED	8 µl
Total	**8 ml**

4% Stacking Gel	
Component	*Volume*
ddHO	3.1 ml
40% Acrylamide	0.5 ml
0.5 M Tris; pH 6.8	1.25 ml
10% SDS	50 µl
10% APS	50 µl
TEMED	5 µl
Total	**5 ml**

Western blotting

Western blots were performed using the Trans-Blot® module wet transfer system on to 0.22µm PVDF (polyvinylidene difluoride) membrane using 1X Towbins buffer (25mM Tris, 192mM glycine; pH 8.3) with 15% (v/v) methanol and run at 100 V for 1 hour in standard transfer conditions. All membrane incubation steps were performed on an orbital shaker. Membranes were blocked in 5% non-fat milk in 1X PBS (phosphate buffered saline) with 0.1% Tween-20 (PBS-Tween) for around 1 hour. Primary antibody incubation was performed for 1–3 hours at room temperature or overnight at 4°C. Secondary horse radish peroxidase (HRP)-conjugated antibody incubation was performed for around 1 hour, followed by three 10-minute washes in PBS-Tween then 2 washes in 1X PBS. Signal was detected by treatment with ECL (Enhanced Chemiluminescence) Prime Western Blotting Detection Reagent (GE Healthcare) and autoradiography film (Amersham™ Hyperfilm™ ECL; GE Healthcare).

In vitro Translation (IVT)

Preparation of templates

Candidate small Open Reading Frame (smORF) cDNAs, CG32230 (RE56733), CG44242 (GM12693) and CG33170 (IP15859), were obtained from Berkeley *Drosophila* Genome Project (BDGP) in either pOT2 or pFLC-1 plasmids and transformed into *E. coli* DH5-α chemically competent cells. A *tal*-1A GFP pOT-2 plasmid was used as a positive control and a *luciferase* control was provided in T7 RiboMAX kit. Plasmid DNA was isolated by miniprep, linearised using restriction enzyme digest and gel purified.

FLAG-tagged and Venus-tagged smORF construct templates were amplified from Gateway destination vectors by PCR using forward primers complementary to the 5' end of the smORF transcript with a T7 promoter sequence incorporated and an SV40 universal reverse primer (for a full list of primers used see Appendix I). PCR products were visualised on an agarose gel and the correct-sized band was purified by gel extraction.

In vitro transcription

Uncapped RNA was transcribed from DNA templates in 10–20 µl reactions using the T7 RiboMAX Express Large Scale RNA Production System (Promega, Cat# P1320) according to manufacturers' instructions. The procedure for 5′-capped RNA transcription was performed using the m7G cap (NEB, Cat# S1411) in a modified protocol provided in the kit. Briefly, an RNA transcription reaction lacking GTP is initiated with 10mM of m7G cap by incubation at 37°C for 1 hour, then 100mM GTP is added to complete the reaction for a further 2 hours. Transcribed RNA from completed reactions was purified using phenol/chloroform and unincorporated deoxynucleotides (dNTPs) were removed using a Sephadex G-25 buffer exchange column (GE Healthcare; cat#27-5325-01) before ethanol precipitation and quantitation using a NanoDrop spectrophotometer (ThermoFisher).

In vitro translation

Reagents were purchased from Promega and IVT was performed using either TnT T7 Quick Coupled Transcription/Translation System (with rabbit reticulocyte lysate; RRL) or a TnT T7 Insect Cell Extract using a DNA template. *In vitro* transcribed RNA was translated using the nuclease-treated RRL System according to manufacturers' instructions. Translated proteins were labeled either by the addition of FluoroTect Green$_{\mathrm{Lys}}$ *in vitro* Translation Labeling System (with the fluorophore BODIPY-FL) and detected on the Typhoon FLA 7000 (GE Healthcare) fluorescence imager following SDS-PAGE or by using the Transcend Non-Radioactive Translation Detection Systems (with biotin label) and detected by Western blot using a streptavidin-HRP antibody.

Tagging-transfection assay and Western blotting

Drosophila Schneider 2 (*S2*) cells were cultured in Schneider's insect medium (Sigma, Cat#S9895) with 10% foetal bovine serum (FBS; v/v) and were grown to a confluent stage (2 days after splitting). The media was removed from the flask and the semi-adherent cells were then detached by tapping the flask and resuspended in fresh media. The cells were counted using a haemocytometer and diluted in new flasks. 250,000 *S2* cells in 0.5 ml medium were transferred to a 24-well plate and grown overnight; cells were transfected the next day with 1 µg of plasmid using XtremeGene HP Transfection Reagent (Roche) according to the manufacturers' instructions. The transfected cells were harvested 48 hours post-transfection, washed once with 1X PBS and pelleted before resuspension in 100 µl of 1X Tricine Sample Buffer (Bio-Rad) with 2.5% v/v β-mercaptoethanol. The samples were run on a 16% Tris-Tricine gel and Western blotting was performed using the following antibody dilutions: 1:10000 monoclonal mouse anti-FLAG M2 (Sigma-Aldrich, Cat# F1804), 1:500 monoclonal mouse anti-β-tubulin E7 (Developmental Studies Hybridoma Bank), 1:5000 monoclonal mouse anti-HA (Sigma-Aldrich, Cat# SAB1411733) and 1:10000 goat anti-mouse HRP (Santa Cruz Biotechnolgies, Cat# SC2302).

Overlapping PCR for dual-tagged dicistronic constructs

Forward and reverse PCR primers that anneal just upstream of the stop codon of ORF1 of the gene of interest were designed to include a short, gene-specific sequence to allow annealing to the construct. The gene-specific sequence was followed by one of two halves of the ~100-nucleotide-long HA-tag sequence, each half contained a small overlapping tag sequence of 12nt shared between the primers (Appendix I). The forward and reverse HA-insertion primers were used in separate PCR reactions (PCR1 and PCR2, respectively) with primers aligning to the pActin promoter (Fw – PCR2) and the SV40 sequence (Rev – PCR1). PCR1- and PCR2-produced amplicons containing partial HA sequences were gel purified and mixed together in an additional PCR reaction (PCR3) to allow the complementary DNA strands to anneal using the overlap in the partial HA-tag sequences. PCR3 was run for 10 cycles without primers to allow the polymerase to fill in the complementary ends after which the pAct Fw and SV40 Rev primers were added to amplify the whole fragment. The resulting fragment was then cloned back into the pActin plasmid using restriction enzyme digestion and ligation.

Mass Spectrometry

Differential solubilisation

25 million *S2* cells were lysed in 300 µl of urea lysis buffer (7M urea, 2M thiourea, 20mM DTT (dithithreitol) and 5X Roche EDTA-free Protease Inhibitor cocktail). The sample was precipitated by slowly dripping into 9 ml ice-cold acetone, with constant stirring at 4°C for 1 hour. The precipitated sample was pelleted with centrifugation at 19,000 g for 15 minutes. The pellet was resuspended in 1 ml of ice-cold 70% (v/v) acetonitrile with constant shaking at 4°C for 1 hour. The sample was centrifuged for 15 minutes at 19,000 g and the supernatant was flash-frozen and evaporated to dryness using a SpeedVac centrifugal evaporator (Savant Inc.). The lyophilised sample was quantified using the Bradford Protein Assay (Bio-Rad) and run on a 16% Tricine gel,

which was visualised using SilverQuest Silver Staining Kit (Invitrogen, Cat# LC6070) according to manufacturers' instructions.

Ultrafiltration

1 x 10^8 million cells in four aliquots were each resuspended in 500 µl of boiling water per tube and boiled in a water bath for 15 minutes in order to denature proteases. The samples were allowed to cool to room temperature, pooled into 2 tubes (1 ml each) and then sonicated on ice using a Branson 250 Digital Sonifier (three 20 second bursts at 40% amplitude and 0.4/0.6 seconds Pulse On/Off, with 1 minutes in between pulses).

The samples were brought to 0.25% v/v acetic acid by adding a 1:200 dilution of 50% acetic acid, and then centrifuged at 16,000 g for 25 minutes at 4°C. Most of the supernatant (leaving about 150 µl behind) was added to a new pre-chilled tube and the remaining supernatant was used to thoroughly resuspend the pellet with a pipette and spun down at 16,000 g for another 10 minutes. The supernatant from this step was then added to that from the previous step. At this point the supernatant was equally divided and applied to the membranes of two separate Vivaspin 500 centrifugal concentrators with a 30KDa molecular weight cut-off (MWCO; Sartorius, Cat# VS0121) that had been passivated using 1% (w/v) bovine serum albumin (BSA) in 1X PBS overnight at room temperature. The spin concentrators were then thoroughly washed by pipetting using 1X PBS (three rinses and two spin washes) followed by six spin washes using 0.25% acetic acid in water. The clarified cell lysate was passed through each filter until about 50 µl of sample remained; the filtrate from each concentrator was pooled together, flash frozen and then evaporated to dryness using a SpeedVac (Savant Inc.). The sample was visualised by running on a 16% Tricine SDS-PAGE gel and staining using Silver Quest Silver Staining Kit (Invitrogen, Cat# LC6070).

SDS-PAGE gel-slice

20 million *S2* cells were lysed in 100 µl of 0.075% SDS with 1X Roche EDTA-free Protease Inhibitor Cocktail. Lysis was performed by extensive pipetting followed by three freeze-thaw cycles using a 37°C water bath and liquid nitrogen with vortexing in between. The lysate was clarified at 16,000 g for 10 minutes at 4°C and the supernatant protein concentration was quantified using Bio-Rad Bradford Protein Assay. Total

protein (375 μg) was added to 50 μl of Tricine Loading buffer v/v 2.5% β-meracaptoethanol (Bio-Rad) and the volume was made up to 100 μl with Milli-Q ultrapure water (Millipore). The sample was heated at 95°C for 5 minutes and 75 μg of total protein were loaded per lane across 4 lanes per sample, on a 10-20% Mini-PROTEAN Tris-Tricine Gel (Bio-Rad, Cat#4563114). The gel was run at 100 V for 1 hour and the 5–15KDa region (as identified using the Bio-Rad Dual Xtra Ladder) for was cut out using a clean scalpel in slices of 2 lanes each that were transferred into an Eppendorf LoBind tube with 100 μl of ultrapure water.

Mass spectrometry and data analysis

The samples were sent to the Cambridge Centre for Proteomics (University of Cambridge, UK) for in-gel trypsin digestion and liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) using an Orbitrap Velos Instrument (ThermoFisher Scientific) with the following parameters: 2 missed trypsin cleavages, 25 ppm precursor mass error, 0.8 Da fragment mass tolerance, carbamidomethylation of cysteine as a fixed modification and methionine oxidation as a variable modification. The spectra were then matched against *Drosophila melanogaster* proteome using the generic Mascot algorithm. In order to search for dwarf smORFs, the spectra were also matched against a custom database consisting of all long non-coding RNA ORFs (ncrORFs) and upstream ORFs (uORFs) found in the *D. melanogaster* transcriptome as outlined in the 'Identification of novel smORFs and gene models for analysis' section below. Any hits with a single peptide match length of less than 8 amino acids or with an exact match using tBLASTn (adjusted for short sequences) against the *D. melanogaster* genome were discarded.

Generation of Poly-Ribo-Seq Libraries

Fly rearing, embryo cage and collections

Oregon-Red (Or-R) genotype flies were expanded in 250 ml polypropylene bottles containing ~50 ml of standard cornmeal fly food. Fifty adult flies were added to each bottle and left to lay eggs for a 48-hour period before the adults were transferred out of the bottle. The bottles were then kept at 25°C for about 2 weeks to allow the freshly laid eggs to develop into adults (~10 days). Between 12–16 bottles, each containing approximately 300–500 adult flies, were emptied into a large (50 cm x 30 cm x 50 cm) Perspex population cage maintained at 25°C. The flies were then left to acclimatise to the new environment for 3–5 days while being fed on a yeast-rich diet to stimulate egg-laying. Four 10 cm petri dishes containing molasses fly food were kept in the cage for embryo collection; the dishes were changed and the embryos collected every 8 hours at 9 am, 5 pm and 1 am. The embryos were harvested straight away for the 'early' 0–8 hours after egg laying (AEL) samples, or were either aged 8 hours at 25°C for the 'mid' 8–16 hours AEL samples, or aged 16 hours, for the 'late' 16–24 hours AEL samples before harvesting directly into liquid nitrogen using a rubber cell scraper (Corning) and storing at –80°C.

Sucrose gradient preparation

Sucrose solutions were prepared in polysome gradient buffer (50mM Tris pH7.5, 150mM NaCl, 10mM $MgCl_2$, 100µg/ml cyclohexamide, 1mM DTT) and different sucrose density fractions were layered on top of each other using sequential freezing with liquid nitrogen in the following order: 0.1 ml of 60%, 1.4 ml each 50%, 47%, 42%, followed 2 ml each of 34% and 38% and then 1.4 ml each of 26% and 18% sucrose (w/v). Gradients were left to thaw overnight at 4°C to allow the gradients to equilibrate.

Embryo lysis and polysome fractionation

0.75 g of frozen embryos from 2 separate collections were pooled together and ground to a fine powder using a pestle and mortar that were pre-chilled with liquid nitrogen. 3.5ml of lysis buffer (10mM Tris-HCl pH7.5, 150mM NaCl, 10mM $MgCl_2$, 2mM DTT, 1% (v/v) NP40, 0.5% (w/v) Na-DOC, 200 µg/ml cyclohexamide, 12U/ml Turbo DNase (Invitrogen), RNasin Plus RNase Inhibitor (Promega) and complete protease inhibitor (Roche)) was slowly dripped into the liquid nitrogen/embryo powder slurry with further grinding. The lysate was then transferred to a 15 ml pre-chilled Falcon tube and thawed under running water followed by a 20-minute incubation at 4°C with rotation. The lysate was spun down at 3000 g for 20 minutes using a chilled table-top centrifuge and then divided into seven aliquots of 500µL (six sucrose gradients, plus one for mRNA control). The aliquoted lysate was then clarified using centrifugation (14,000 g) for 10 minutes at 4°C. 450 µl of embryonic lysate from each aliquot was carefully loaded onto the top of six 18–60% sucrose gradients and ultra-centrifuged at 31,000 g for 4 hours. The gradients were pumped out, their absorbance at 254 nm plotted and the 2^+ ribosomes fraction collected.

Nuclease treatment and purification of ribosome footprints

The sucrose fractions from each tube were pooled together and the percentage sucrose was estimated from the polysome trace (~44%) before dilution to 10% sucrose in polysome dilution buffer (50mM Tris, 150mM NaCl, 10mM $MgCl_2$). Footprinting was performed overnight at 4°C with 1000 U RNase I (Invitrogen) per centrifuge tube and then stopped with SUPERaseIn RNase inhibitor (Invitrogen). 160 ml of the digested material was then either precipitated or concentrated down to 2 ml using 30KDa MWCO Ultrafiltration concentrators (Corning, Cat# CLS431489). 1 ml of each concentrated material was loaded onto 2 ml of a 1 M sucrose cushion (34.5%) and centrifuged for 4 hours at 70,000 g to pellet the monosomes. The pelleted material was resuspended in 1X Turbo DNase Buffer (ThermoFisher) and then DNase treated at 37°C for 30 minutes and purified using phenol-chloroform and ethanol precipitation. The sample was resuspended in formamide loading buffer (47.5% formamide, 0.01% SDS, 0.01% bromophenol blue, 0.005% xylene cyanol, 0.5mM EDTA) and heated to 80°C for 3 minutes, before loading onto a pre-run denaturing gel (8 M Urea, 10% acrylamide: bis-acrylamide (19:1)) in 1X TBE. The gel was run at 300 V for 3 hours and then stained using SYBR Gold (Invitrogen) for visualisation under UV light in order to cut out the band between the 28 and 34nt RNA markers. The excised band was shredded by centrifugation through a 0.5 ml PCR tube with holes poked at the bottom and incubated at 4°C overnight in 750 µl of gel elution buffer (20mM Tris pH7.5, 250mM NaOAc, 1mM EDTA, 0.25% w/v SDS). The eluate was precipitated with isopropanol and then resuspended in a 50 µl T4 Polynucleotide Kinase Reaction Buffer (NEB, Cat#M0201) supplemented with SUPERaseIn RNase inhibitor (Invitrogen). The completed reaction was heat inactivated and then precipitated using isopropanol.

RNA preparation

Total RNA was purified from 250 µl of embryo lysate with 1 ml of TRIzol Reagent (Invitrogen) and using 250 µl of chloroform and 0.6 ml isopropanol per ml of TRIzol. Purified RNA was resuspended in 1X Turbo DNase Buffer (Invitrogen) and DNase treated at 37°C for 30 minutes, before phenol-chloroform purification and ethanol precipitation. The RNA was then quantified using a Nanodrop spectrophotometer (ThermoFisher) and 100 µg of total RNA was used for mRNA selection using 200 µl of OligodT$_{25}$ magnetic beads from the Dynabeads mRNA Purification Kit (ThermoFisher, Cat# 61006) according to the manufacturers' instruction. The poly(A) selected mRNA was ethanol precipitated before heating (95°C for 20 minutes) in alkaline fragmentation buffer (12mM Na_2CO_3, 88mM $NaHCO_3$, 2mM EDTA pH9.3) followed by ethanol precipitation. The fragmented mRNA was size selected (50–80nt) on a denaturing gel, before gel extraction, T4 polynucleotide kinase (PNK) treatment and ethanol precipitation as described above for the ribosome footprints.

Library preparation

T4 PNK treated ribosome footprints and mRNA fragments were used for library preparation using the NEBNext Small RNA Library Prep Set for Illumina (NEB, Cat# E7300) according to the manufacturers' instructions; detailed in Chapter 4. The following modifications were made to the protocol: an rRNA or tRNA depletion was performed for ribosome footprint samples (described in the next section) and the number of cycles (8–14) for the final PCR amplification step were optimised using a scaled-down reaction with one-fourth of the cDNA for all samples.

Ribosomal RNA depletion beads

Ribosomal RNA depletion was performed after 3′-adapter ligation in the library preparation protocol using subtractive hybridisation beads. In order to generate beads with single-stranded DNA (ssDNA) complementary to *Drosophila* rRNA, PCR reactions were performed to generate 500nt and 1000nt fragments of the rRNAs, using 5′-biotinylated reverse primers (Appendix I). A 5′-biotinlyated oligonucleotide that is complementary to 2S rRNA, along with the PCR products, were bound to Dynabeads MyOne™ Streptavidin C1 (Invitrogen, Cat# 65001) according to the manufacturers' instructions, and the second strand of the PCR fragments was washed away using 0.1 M NaOH. 1µl of a 50µM mixture of six rRNA depletion oligonucleotides (1–6; Appendix I) that are mixed in the ratio 2 : 3 : 2 : 2 : 3.25 : 1.75 was added to the sample along with 25µl each of the 500bp and 1kb rRNA depletion beads and 12.5µl of unhybridised streptavidin beads. The sample was heated to 70°C for 2 minutes followed by a 20-

minute incubation at room temperature with rotation. The beads were removed from the solution using a magnet and the supernatant was transferred to a new tube for a second round of depletion following the same process. The depleted footprints were ethanol precipitated and resuspended in T4 RNA Ligase2 (truncated) Buffer (NEB, Cat# M0242) supplemented with 10 (w/v) PEG-8000 and 1 µl of SUPERaseIn RNase inhibitor.

Library quantification and sequencing

Prepared libraries were size checked and quantified using the Bioanalyzer 2100 with a High Sensitivity DNA chip (Agilent). The library amount was then confirmed using a Qubit machine with a double-stranded DNA (dsDNA) HS Assay Kit (Invitrogen, Cat# Q32851). Quantified libraries were either sequenced in-house using the Illumina MiSEq machine or sent to the University of Cambridge Sequencing Facility to be sequenced on an Illumina NextSeq machine. For MiSeq samples, the library was diluted to 4nM and denatured according to the manufacturers' instructions, and 12–15pM was loaded on to a v2 (50bp-single end) or v3 cartridge (150bp-single end) (Illumina). NextSeq samples were diluted to 10nM and sent to Cambridge for 50bp-single end sequencing.

Poly-Ribo-Seq Data Analysis

Identification of novel smORFs and gene models for analysis

The EMBOSS getORF program was used to identify all smORFs (<100aa) in the *Drosophila melanogaster* transcriptome (R6.02) (from FlyBase) within 5′ UTRs of annotated protein coding transcripts and long non-coding RNAs (lncRNAs) that begin with an AUG start codon and end with a stop codon (TAG, TAA, TGA) with a minimum length of 30nt (10aa). In order exclude the possibility that the high ribosome occupancy observed in 5′ UTRs was due to signal in translated uORFs, we created a modified transcript that contained all regions except the putative uORFs for our general analyses on 5′ UTRs. Similarly, any 3′ UTR regions overlapping annotated coding DNA sequences (CDSs) were masked from the analysis. For polycistronic gene models, the 5′

UTRs were defined as the region from the start of the ORF to the beginning of the transcript or to the end of a uORF, and *vice versa* for 3′ UTRs.

FASTQ Processing and Removal of rRNA/tRNA Reads

FASTQ files of raw sequencing reads were processed using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and reads with an Illumina Q-Score below 33 were discarded followed by trimming of adapter sequences (AGATCGGAAGAGCACACGTCTGAACTCCAGTC) from all reads. Any reads shorter than 25nt in length were discarded from the analysis. The trimmed reads were matched against a pre-computed file containing all annotated rRNA and tRNA sequences in the *Drosophila melanogaster* genome (R6.02), using *Bowtie* short read aligner (Langmead *et al.* 2009).

Mapping to the *Drosophila* transcriptome and data analysis

Non-rRNA reads were matched to the *Drosophila melanogaster* transcriptome (R6.02) using the *TopHat* splice aware Read mapper (Version 2.0.11) using the *Bowtie1* algorithm and specifying a mismatch tolerance of 1nt, and only keeping uniquely mapping reads (Trapnell *et al.* 2012). *TopHat* output files (BAM files) were sorted and indexed, and reads outside the 28–34nt size ranges were filtered out using SAMtools (Version 1.2). To calculate abundance in reads per base pair (RBP), BAM files were cross referenced to BED files for each type of gene model (FB CDS, FB smORF, ncrORF and uORFs) containing the genomic co-ordinates of the 5′ UTR, ORF and 3′ UTR features, using the coverageBED command in BEDtools (Version 2.21.0). RBP values were converted to *RPKM* (reads per kilobase of transcript per million mapped reads) by multiplication with 10^3 (Kb) and dividing by the total number of reads in the BAM file in R (Version 3.2.0) (R Core Team 2015). Data analysis and manipulation was performed in RStudio (Version 0.99) (RStudio Team 2015).

Chapter 3: Independent Corroboration of Poly-Ribo-Seq Data

Introduction

As previously discussed, we observed extensive translation of smORFs in the *Drosophila* transcriptome using Poly-Ribo-Seq on *S2* cells (Aspden *et al.* 2014). These experiments revealed that annotated smORF genes within Flybase ("FB smORFs") were translated at a similar proportion to canonical protein coding genes (83%), while smORFs in 5'UTRs (uORFs) and putative ORFs within annotated lncRNAs (ncrORFs) were translated at a lower level (30-34%). These novel unannotated smORFs were termed 'dwarf smORFs' due to their much shorter median length of 23 amino acids as compared to the median length of 80 amino acids of the annotated 'FB smORFs'. Many recent studies have studies reported translation outside of annotated protein coding regions using Ribo-Seq (Duncan, and Mata 2014; Ingolia *et al.* 2011; Bazzini *et al.* 2014).

In order to develop a cut-off in the *S2* cell derived dataset, we used two metrics based on abundance of RNA, measured in Reads per Kilobase per Million Reads (*RPKM*) and coverage (cvg), which is the proportion of the ORF covered by RBF reads. Minimum *RPKM* threshold was calculated by using the 3'UTR signal in order to determine a translation cut-off value using the 90th percentile of 3'UTR background signal of standard protein coding genes annotated in *Drosophila*. Using this method, an *RPKM* value of 11.7 and coverage of 0.57 were determined, and used to define the translation of 228 FB smORFs, 313 ncrORFs and 2,708 uORFs.

N.B. Portions of this Chapter have been published in Aspden et al. (2014) Elife 3, e03528

Poly-Ribo-Seq is a new technique, and though it is now being widely used to address translation on a transcriptome wide scale, there is however still a debate as to whether the accumulation of reads outside of annotated protein coding regions is representative of meaningful translation, or whether it is spurious association of RNA with ribosomes or possibly even background signal (Chew *et al.* 2013). Therefore, there arose a need for independent corroboration of these results, which would most convincingly occur through direct detection of the peptides encoded by these transcripts, as opposed to detection at the RNA level. This led to the work described in this chapter, whereby a variety of alternative and complementary methods were utilised for the detection of smORF-encoded peptides (SEPs). Three different low to medium throughput biochemical methods were employed: *in vitro translation* (IVT), Western blot detection of FLAG-tagged SEPs in transfected *S2* cells (Tagging-transfection assay) and finally, Mass Spectrometry (MS) on small protein enriched fraction of *S2* cell lysate.

In vitro Translation

There are many application of *In vitro* translation (reviewed in Jagus and Beckler (2003)) and perhaps the most common use of IVT systems has been to assess the translation of cDNAs (Norman *et al.* 1988). RNA or DNA templates are added to cell lysate, which has been treated with a non-specific nuclease to remove endogenous RNA and DNA, in order to produce protein products that can be detected through the incorporation of radioactive isotope or biotin labeled amino acids (Pelham, and Jackson 1976). The cell-free format of this method allows the parallel, small-scale assessment of a large number of cDNA clones as plasmid extractions can be used directly in the reaction (Craig *et al.* 1992). These cDNA clones are easily available from the Berkeley *Drosophila* Genome Project (Rubin *et al.* 2000). This allows IVT to be conducted much faster than cell-based assays which require firstly the generation of expression constructs through cloning, followed by transfection in to cell culture and then finally, the detection of the peptide using Western blotting. There are now also in-gel detection methods of fluorescently labeled translation products, allowing UV-based detection of the peptides, which is much less time-consuming than previous radioactivity-based detection methods.

In addition to detecting translation of the smORFs, we also wanted to test the feasibility of using this system to correlate the different levels of translation observed in the *S2* cell Poly-Ribo-Seq data. The flexibility of the IVT system, which allows the use of a variety of different templates such as cDNA plasmids or PCR amplified cDNA templates, was advantageous especially in the case of uORFs and ncrORFs. Theoretically, using equimolar amounts of purified *in vitro* transcribed RNAs of the various smORF types, we should observe a corresponding amount of signal from smORFs with high or low levels of translation. The signal could then be quantified and normalised to the number of labeled amino acids in the sequence before correlation with Poly-Ribo-Seq metrics (Figure 3.1).

Many transcripts in the Fly transcriptome are polycistronic (containing multiple ORFs), including 4 dicistronic transcripts containing 8 previously annotated FB smORFs. By definition, uORFs are upstream of a longer annotated ORF and hence always polycistronic (reviewed in Tautz (2008) and Hayden and Bosco (2008)). Long non-coding RNA (lncRNA) transcripts tend to be littered with multiple putative smORFs and it is often difficult to ascertain exactly which smORFs are actually being translated. IVT offered the advantage of testing whether any of the putative ORFs in a given polycstronic smORF transcript produces peptides. This would require a single template and reaction, and therefore preferable compared to lower-throughput methods such as individual ORF tagging. This approach has been previously used in our lab to help ascertain the translation of the 5 tandem-arrayed smORFs within the polycistronic transcript of *tarsal-less (tal)* (Galindo *et al.* 2007), which revealed that 4 out of 5 smORFs are translated.

Tagging-transfection assay using Western Blotting

Western Blotting is a classic technique that combines the resolving power of SDS-PAGE with the specificity and sensitivity of antibody detection. We chose to use this technique, as we wanted to complement the *in vitro* translation approach by using a system that provides an *in vivo* cellular context for the production and translation of smORF peptides in *S2* cells. The *S2* cell protein samples are run on a gel, and these are immobilised onto a membrane and detected by an antibody specific to the protein being detected (Towbin *et al.* 1992). The main issue that we had to overcome for using this method is that there are not many commercially or publicly available antibodies for smORF peptides, as they are largely uncharacterised. In addition, obtaining and optimising antibodies for a large number of smORFs can be expensive and laborious. In our lab, during in-depth characterisation of three individual smORFs, we have generally found it difficult to successfully generate suitable antibodies for smORF-encoded proteins (Personal communication JI Pueyo). The only working smORF antibody is the *tarsal-less* antibody, which is only able to detect the Tal peptide when it is fused to a much larger GST tag.

This issue led us to choose to use the Gateway Cloning system from Invitrogen to generate plasmids encoding epitope-tagged fusion proteins of a subset of the different smORF classes. Other members of the lab cloned a small representative pool of 27 different kinds of S2 cell translated smORFs for testing in this assay. We have used this cloning technique extensively in our lab to generate constructs for injection into flies under the UAS/GAL4 system as well as transfection of these plasmid constructs into *S2* cells. Using *S2* cells allows the rapid transfection of several constructs at the same time. This allowed standardisation of the approach by treating each plasmid in exactly the same way and the lysate analysed by Western blotting.

The main rate-limiting and labour intensive step of this method is the cloning of the cDNA sequences, but once tagged and cloned, the constructs can be used for multiple studies, including an Immuno-staining and microscopy approach to independently corroborate smORF translation and visualise SEP subcellular localisation. Using Western blotting for the detection of SEPs would not only to provide proof of translation, but can also corroborate the predicted size of the SEP as calculated by the ORF sequence.

Investigating the translation of Dicistronic smORFs using Epitope tagging and Western Blot

Our laboratory has previously characterised three smORFs, and one interesting and notable feature of these smORF genes is their polycistronic nature. *Tarsal-less*, *Sarcolamban* and Hemotin contain multiple ORFs within the same transcript and are examples of functionally related peptides being translated from different ORFs within the same transcript (Galindo *et al.* 2007; Magny *et al.* 2013; Pueyo *et al.* 2016) . These represent a truly interesting class of genes from a translation perspective since polycistronic translation does not fit in with the standard model of translation in eukaryotes (Kozak 1978). Polycistronic transcripts have been shown to be a key feature of bacterial genomes, where it is common that a number of functionally related proteins are transcribed under the control of a single promoter (Molecular Cell Biology, Lodish 2000b). Polycistronic genes have previously been considered quite rare in eukaryotic genomes, however, bioinformatic analyses showed the putative presence of 124 polycistronic genes conserved across 12 species of *Drosophila* (Lin et al. 2007), adding to increasing evidence that polycistronic genes are a new class of genes in multicellular organisms (reviewed in Tautz (2008)). Understanding the translation of polycistronic smORFs offered an opportunity elucidate on alternative models of translation such as leaky scanning and re-initiation as seen in viruses (reviewed in Firth and Brierley (2012). Neither *tal* nor *Scl* are transcribed in *S2* cells, however we do find 4 pairs of dicistronic smORF genes within the pool of translated smORFs in our Poly-Ribo-Seq data, which were used in the tagging-transfection assay.

Peptidomics

Peptidomics is the study of native peptides and small proteins using Mass Spectrometry (MS) to provide a broad identification of the peptide pool in a sample. A typical peptidomic workflow involves protein extraction, followed by enrichment of small proteins, MS and data analysis. Recent improvements to conventional mass spectrometric methods, particularly the development of tandem MS/MS and the integration of liquid chromatography (LC), have greatly improved the sensitivity and specificity of the technique (reviewed in Yates *et al.* (2009)). Trypsin digestion is performed after protein extraction in order to fragment proteins into smaller peptides generated by cleavage at specific amino acids (Arginine and Lysine), which can then be analysed on by MS. LC-ESI MS/MS integrates nano-scale HPLC that enables the fractionation of complex mixtures of trypsin-digested peptides before they are entered into the MS gas chamber through electron spray ionisation (ESI). Tandem MS (MS/MS) allows the same peptide to be analysed twice, first as a precursor, and a second time after it has been fragmented in the MS chamber (reviewed in Schrader *et al.* (2014)). Mass spectral signatures of trypsinised peptides are matched to a database of theoretical signatures, resulting from the *in silico* translation and digestion of a previously defined library of annotated coding sequences. These methods have enabled the discovery of novel peptides encoded by small ORFs (Yang *et al.* 2011) and those generated by proteolysis (Falth *et al.* 2006; Tinoco, and Saghatelian 2011) , particularly in the identification of neuropeptides (Svensson *et al.* 2003), uORFs (Oyama *et al.* 2004) and biomarkers in body fluids (Clynen *et al.* 2008).

Recently peptidomics has also been applied to characterising small ORFs and other novel coding regions on a transcriptome-wide scale by integrating the technique with RNA sequencing. Bioinformatic tools are used to generate a custom database consisting of peptide sequences translated from putative unannotated ORFs within the transcriptome (Slavoff *et al.* 2013; Ma *et al.* 2014; Vanderperre *et al.* 2013). This has led to the development of bioinformatic tools geared towards matching MS data to an unbiased search space, as inferred from RNA sequences, allowing the building of custom databases based on six-frame translation of putative ORFs in the genome (Costa *et al.* 2013) (reviewed in Andrews and Rothnagel (2014)).

These studies suggest that the proteome is much more complex than previously assumed. Translation can be widely detected outside of annotated protein coding sequences, particularly in alternative reading frames within annotated coding regions (altORFs) as well as ORFs found in 5' and 3' UTRs. These findings have corroborated those of Ribo-Seq, which also suggests extensive translation outside of annotated protein coding regions (Aspden *et al.* 2014; Duncan, and Mata 2014; Ingolia *et al.* 2011; Ingolia *et al.* 2014; Michel *et al.* 2012). We sought to apply this high-throughput approach with peptidomics in order to complement Poly-Ribo-Seq identification of smORFs in *Drosophila S2* cells.

Enrichment of *S2* cell lysate for small proteins

As the purpose of this study was to detect endogenous smORF encoded peptides, the standard proteomic workflow was modified to enrich for small proteins. Small protein enrichment is a necessary step to develop a Peptidomics method for SEP detection as small proteins tend to be masked by the fragments of larger, more abundant proteins. The fragments of larger proteins may be generated during the protocol by degradation, or due to the standard step of trypsin digestion. Enrichment for small proteins is therefore especially important, as a large protein is likely to generate many unique fragments due to the presence of multiple trypsin cleavage sites, while a small protein, which may contain only one or two trypsin sites, would generate fewer fragments (Tinoco *et al.* 2010). My aim for these sets of experiments is to implement low and medium throughput approach towards peptide level identification of smORF genes to corroborate and complement the Poly-Ribo-Seq data. Due to the challenges in detection of SEPs, to enrich for these small peptides in *S2* cell lysate, three different techniques were tested: differential solubilisation, Ultrafiltration and SDS-PAGE.

Differential solubilisation is based on the principle that in a precipitated total protein sample, smaller proteins will re-solubilise in a 70% organic solvent due to their relatively simple structure, while larger proteins remain precipitated, as they do not re-solubilise easily. Differential solubilisation was originally developed for enrichment of peptides from small samples of serum (Kawashima *et al.* 2010) and has also been used on whole blood (Lin *et al.* 2012), but has never been used with whole cells. Differential solubilisation is a biochemical separation method, which is an advantage for the enrichment of limited samples, as the use of a physical separation material such as a membrane or a gel increase variation and losses due to non-specific adsorption of peptides.

Ultrafiltration is still a commonly used separation technique that is incorporated into almost every peptidomics workflow (Tinoco, and Saghatelian 2011). This method employs fractionation of proteins based on their permeability across a membrane of a given pore size, which is measured by KDa MWCO (molecular weight cut-off). However, the Ultrafiltration separation method has documented problems with membrane fouling and loss of material by non-specific adsorption, which contributes to variability between replicates.

SDS-PAGE is a classical technique for resolving proteins according to molecular weight and has previously been used to fractionate small proteins prior to MS (Oyama *et al.* 2004), (Slavoff *et al.* 2013). SDS-PAGE offers the advantage of accommodating large amounts of denatured protein in a soluble and stabilised gel slice that can be used directly in trypsin digestion thus minimizing protein degradation. This is the simplest approach to enrich a protein sample and offers excellent reproducibility between replicates (Ahmad *et al.* 2005).

Results

SEPs can only be detected in the IVT assay as large fusion protein

In order to detect smORF translation using IVT, three different IVT systems were used 1) a Rabbit Reticulocyte lysate (RRL) based system which can be used as a two-step *in vitro* translation reaction, using known quantities of *in vitro* transcribed and purified RNA as a template, 2) an RRL system that uses a DNA template in a coupled transcription and translation reaction (TnT) and 3) an insect cell extract (ICE) based TnT system (Figure. 3.1) (all available through Promega).

A non-radioactive method was used for detecting the *in vitro* translated peptides using the Transcend tRNA system, which uses tRNA that has been pre-charged with a labeled Lysine in the reaction. I used tRNA charged with fluorescently labeled Lysine residues in a translation reaction lacking endogenous lysine. The completed reaction can then be resolved using SDS-PAGE and the fluorescent translation products can be detected in-gel using the Typhoon imager. This approach is considerably quicker and safer than radioactive labeling. The constructs *tal*-1A GFP (previously used in our lab) and Luciferase (supplied with the kit) were used as positive controls in the assay.

Complementary DNA clones were obtained from BDGP DGC (Berkley *Drosophila* Genome Project: *Drosophila* Gene Collection) (Rubin *et al.* 2000) for three of the most highly translated FB smORFs in our *S2* cell Poly-Ribo-Seq data (CG32230, CG44242 and CG33170) as templates. The SEPs encoded by the genes have predicted sizes of 9.4, 7.4 and 8.3 KDa and contain 7, 4 and 11 Lysines, respectively, in their primary peptide sequence. These cDNA clones are available in a vector with a T7 promoter and an SV40 PolyA signal so they were technically compatible with IVT. However, after trying a number of approaches, such as using capped (m7G) and uncapped RNA in a two-step IVT assay, as well as coupled TNT reactions (Figure 3.1), detection of these cDNA cloned SEPs in the IVT assay proved unsuccessful (Results are summarised in Table 3.1). It is estimated that only 30% of the fluorescent Lysines are actually incorporated into the final IVT translation product (Promega Technical Note). Upon counting the number of Lysine residues in the amino acid sequences of the longer Poly-Ribo-Seq translated FB smORFs, it was found that only 40% of the pool of 228 SEPs contain more than 6 Lysine residues (Figure 3.2). This means that more than half of the of FB smORFs and certainly the majority of the shorter ‘Dwarf’ smORF would not be sufficiently labeled for detection using this method.

In order to overcome this issue of detection, it was decided to use tagged smORF sequences to see whether by adding the tag (FLAG or Venus) increases detection of SEPs by increasing the size and therefore the number of labeled Lysines in the peptide. In order to generate these tagged smORF templates, I used the plasmid constructs generated by our lab for the tagging-transfection assay (discussed in the next section). In order to make these constructs compatible with IVT, primers were designed to incorporate an upstream T7 promoter sequence and to align to the 5’ of the smORF transcript CDS. The reverse primer was aligned to the existing SV40 polyA signal located downstream of the tag. These primers allowed the generation of tagged cDNA templates by PCR. Using these Venus/FLAG tagged cDNA smORF templates, I was able to detect SEPs fused to the much larger 28 KDa Venus tags, which contain 20 Lysine residues, which would definitely allow for better incorporation of the fluorescent Lysines (Figure 3.3). As the FLAG-tagged SEPs could not be detected as readily, Western Blot was used to try an alternative method. Labeling FLAG is not dependent on the fluorescent labeling of Lysines, and by addition of the FLAG tag, perhaps we would increase the size of the SEP to a detectable range, but using this method there was still no evidence of signal from FLAG-tagged IVT SEPs.

It appears that these IVT strategies do not readily yield any significant progress in the detection of endogenous transcript derived SEPs, further illustrating the challenge in detecting smORF peptides using well-documented and conventional systems of peptide detection. From the work conducted here, it can be speculated that this may be an issue with the low rate of incorporated Lysines in the peptide by IVT (Promega Technical Note) or because there is masking of the peptide product by the contaminant globin band; which runs at about 10-15KDa, similar to the size of most SEPs, as the Venus and GFP tagged smORFs were able to be detected (Figure 3.3). These results make it difficult to consider whether IVT can be used as a medium-throughput method for the corroboration of smORF translation without it requiring longer additional steps required to add Venus or GFP tags to each plasmid that is tested, thus making IVT unsuitable for the objectives of this study.

Tagging-transfection Assay and Detection by Western Blot

FLAG tagged constructs design and controls

Our lab conducted a tagging-transfection assay to be used in order to verify that smORF sequences obtained from Poly-Ribo-Seq are capable of making a viable/stable peptide product. In order to carry out the tagging-transfection assay, members of the lab prepared constructs using Invitrogen Gateway cloning technology, using the smORF sequences obtained from Poly-Ribo-Seq read mapping. Primers were designed to amplify the 5'UTR and the CDS (minus the stop codon) from *S2* cell cDNA to TOPO clone into pENTR vectors. The 5'UTR was included in the test construct to provide endogenous translational context and the Kozak Sequence for the translation of the CDS. These were then shuttled into destination vectors containing a C-terminal 3xFLAG tag (pAWF) and Venus tag (pAWV) obtained from the Murphy Lab *Drosophila* Gateway vector collection (Website: https://emb.carnegiescience.edu/*Drosophila*-gateway-vector-collection) using the LR recombinase from Invitrogen. Transcription was under the control of the Act-5C promoter, which provides high levels of transcription and an SV40 PolyA signal is included at the end of the transcript (Figure. 3.4A). We initially decided to these test two different tags as 3xFLAG, which is a small tag (5KDa), can be readily detected by labeling with a specific antibody, and Venus, which is a larger, highly fluorescent tag that can be directly detected in cells without any subsequent labeling with antibodies but may overwhelm the endogenous stability/localisation of the small SEP.

As can be seen in Figure 3.3 we can detect multiple bands in the CG32230 and CG44242 lanes using IVT, the smaller band at around 25KDa may correspond to tag only signal (26KDa). These findings raised concerns that there may be significant translation from the start codon of the Venus sequence, which may influence the translation and detection of the tagged SEPs. Therefore four constructs were initially designed as positive and negative controls for the experiment. The expression vector was modified using Site-Directed Mutagenesis so that the AUG start codons of the FLAG and Venus tags were changed to a GCG codon. This step would ensure that any translation of the tag would occur from the inserted smORF test sequence and not from the start of the tag. The positive controls for the experiment used the plasmids designed above, with the 5'UTR and first few codons of a translated smORF, CG42371 (Peptide

Atlas) inserted upstream to introduce a non-plasmid AUG upstream of the GCG-FLAG and GCG-Venus sequences. These constructs were named AUG-FLAG and AUG-Venus (Figure 3.4A) and the controls were transfected and detected by Western blotting using the method outlined in Figure 3.4B.

The results in Figure 3.5 show a Western blot to assess the translation of Venus tagged smORFs and the AUG-Venus and GCG-Venus controls. 4 out of the 5 tested smORF constructs show the presence of a band at around 37KDa that is the result of the SEP-Venus fusion protein. In addition, we can observe signal in all 7 transfected samples, including the negative control GCG-Venus. The size of this band corresponds to the size of the Venus tag on its own as evidenced by AUG-Venus (26KDa) and it is present as a second lower sized band in translated smORF samples. Therefore, we concluded that the translation of Venus maybe from one of the multiple in-frame ATG or CTG start codons that are located downstream of the mutated start codon. Thus, we discontinued using the Venus tag in the tagging-transfection assay as it could interfere with the detection of the slightly larger SEP-Venus fusion proteins. These results made it apparent that the best approach forward is to use FLAG tagged smORFs with Western blotting in transfected cells as opposed using the Venus tag in IVT. However, I was unable to detect any signal from cells transfected with the AUG-FLAG construct by Western blot (Figure 3.6) even though low FLAG signal could be detected in a small proportion of cells in the microscopy based tagging-transfection experiment, which was not present in secondary antibody-only controls (personal communication with Unum Amin). The lack of signal in the Western blots may be due to the extremely small size of the peptide (5 KDa) or the poor stability of these peptides in the cell. Therefore, we used the smORF gene *Sarcolamban* (*Scl*-A) as an alternative positive control, which has proven translation and has been characterised in-depth in our lab (Magny *et al.* 2013). *Scl*-A-FLAG showed high level of signal in the assay and was reliably detected over multiple experiments thus making it a suitable positive control. An additional negative control construct was also made by cloning the sequence of a putative smORF present in a known non-coding RNA gene Uhg2 (U-snoRNA host gene 2), which does not appear as translated by Poly-Ribo-Seq.

Western Blot detection of FLAG-tagged smORFs

The tagging-transfection protocol was standardised to minimise experimental variation. This involved plating 250,000 *S2* cells/well in a 24 well plate, which were transfected with 1µg of tagged-smORF plasmid DNA 24h after plating. Cells were allowed to grow, and harvested 48hrs post-transfection. The harvested cells were pelleted and washed with 1 x PBS before being re-suspended and heat denatured in 50µl of Tricine LB supplemented with 2.5% (v/v) BME to make a stock cell lysate sample. Samples from each transfection were then run on a 16% Tricine gel in order to achieve maximum resolution of small proteins. The protein from the gels was then transferred onto a 0.2µm PVDF membrane using a wet transfer system and the membrane was cut horizontally at the 37KDa band and the lower portion of the membrane was probed with the FLAG M2 Antibody while the upper portion was probed with Anti Beta-Tubulin as a loading control.

Western blotting of the tagging-transfection assay was able to reliably detect 17 out of the18 FB smORFs that were tagged as sumarrised in Table 3.2. These included some smORFs that were below the *RPKM* (>11.7) and coverage (>0.56) translation cut-offs used in the Poly-Ribo-Seq dataset (Shaded cells), however, they had very low signal compared to loading controls. A great variation can be observed in the amount of tagged protein detected by Western blot for the different tagged smORF constructs; therefore each sample was run at least three times to confirm translation. Differing amounts of sample were loaded on the gel until clean looking bands could be seen (Figure 3.6) for each smORF. I then used this method to try and visually estimate signal strength, for example, 0.5µl of stock sample of CG33774 (high signal strength) gives a clean band with 30 second film exposure, while for CG33199 (very low signal strength) a much larger amount had to be loaded (10µl of undiluted sample) and a band could only be detected after a 5-minute exposure of the X-ray film (Figure 3.6). Overall, I was unable to glean any obvious correlation between Poly-Ribo-Seq metrics such as abundance (*RPKM*) or normalized abundance in the form of translation efficiency ($RPKM^{Ribo\text{-}Seq}/RPKM^{RNA\text{-}Seq}$) as shown in Table 3.2.

Interestingly, some smORF-FLAG constructs showed additional bands on the Western blot (Figure 3.6) that were either higher than the expected size band such as CG33170, which has a complex splicing pattern and encodes for a larger, annotated 16.2 KDa peptide (including FLAG tag) through the use of alternative initiation from in-frame upstream start codons. CG33774 also displays a second fainter band around 17 KDa but has no in-frame upstream start codon and therefore this band may arise due to post-translational modification of the peptide. CG33199 displays additional bands that run below the predicted size band at just below 15KDa, these may arise due to N-terminal proteolytic processing of the peptides only downstream initiation codon encodes for a much smaller peptide of 8.4 KDa. Altogether these results suggest that the peptides translated from these ORFs are stable and regulated through their interaction with the cellular machinery responsible for post-translation modifications.

The smORF CG32267 gene was not categorised as translated in this assay, as I was only able to detect the SEP once in a total of 5 attempts after using a very long exposure time (15 min) of the X-ray film during development. Despite this anomaly, the CG32267 transcript showed moderately high expression (*RPKM* 83.3) in the Poly-Ribo-Seq sample. In the parallel microscopy experiment, CG32267-FLAG was only detected at very low levels in a few cells and that too by shortening the post-transfection time to 30h. This suggests that the peptide made from CG32267 is either very unstable or may be toxic to the cells, but this cannot be proven since this SEP has been detected by MS and Poly-Ribo-Seq. Over-expression of the FLAG tagged SEP in *S2* cells leads to a ‘distressed’ mitochondrial morphology, suggesting that induced expression of this SEP may be detrimental to the cells (Personal Communication Unum Amin).

In addition to the FB smORFs, we also tagged 7 putative ORFs in 2 annotated lncRNA transcripts called pncr009:3L and CR30055. There is evidence of translation of these ORFs according to Poly-Ribo-Seq data. By the tagging-transfection assay followed by Western blotting, I was able to detect peptides for 4 of these 7 ORFs (Figure 3.7A). Interestingly, the ncrORF pncr009:3L-ORF4 sample showed a 15KDa band, which is significantly large than the expected size of 8.2 KDa. However, upon closer examination of the cloned sequence, we discovered an in-frame initiation codon upstream of the predicted start site, which matches the size of the 15KDa band. We also tested two uORFs, but were not able to detect these by Western blot, though by imaging in *S2* cells these tagged uORFs showed distinct expression and subcellular distribution patterns (Personal communication Unum Amin). The reason for this could be that these uORF peptides were not detected because they were either being degraded inside the cell, or perhaps they are exported outside the cell like *tarsal-less* (Pueyo, and Couso 2008). To overcome this, I tried multiple techniques to troubleshoot this problem; such as shortening the post-transfection time to 30h to limit degradation and testing

Ultrafiltration concentrated cell supernatant (50x) from these transfections, however I was still unable to detect these SEPs by Western blot. Upon analysing these results it was revealed that the main property of SEPs not readily detected by the Western blotting was their small size. Readily detected SEP-FLAG peptides have a median size of 13KDa while those not detected by Western blotting are much smaller, with a median size of 7KDa. Additionally, it has been shown that detection of very small peptides can be challenging using Western blotting as they may pass through the pores of the membrane (MacPhee 2010).

Generation of Dual tagged dicistronic smORF constructs

The smORF genes already characterised by our lab such as *tarsal-less*, *Sarcolamban* and *Hemotin*, have all been polycistronic in nature, containing 2 or more ORFs that produce stable peptides (Galindo *et al.* 2007; Magny *et al.* 2013; Pueyo *et al.* 2016). The possibility that these may represent a new subclass of smORF genes, which encode multiple functionally related peptides, makes them a very interesting avenue to follow. According to Poly-Ribo-Seq data, we found 4 dicistronic gene models encoding 8 smORFs that are translated in *S2* cells with varying signal between ORF1 and ORF2, namely CG43194/CG43210 (*Hemotin*), CG42497/CG9878, CG42371/CG15386, and CG32736/CG42308 (ORF1/ORF2). Initially, each ORF was FLAG tagged in separate constructs to be tested in the tagging-transfection assay and I was able to detect translated peptides for all of the 8 ORFs tested using Western blotting (Figure 3.7B). Any further investigation into the translational mechanisms and regulation of these dicistronic transcripts would be greatly aided by the capability to detect translation of both ORFs in relation to each other, for example with constructs containing both ORFs individually tagged. The dicistronic ORF2-FLAG constructs generated for the tagging-transfection contain the whole transcript upstream of ORF2, which includes ORF1 but it is not tagged. I decided to utilise these constructs in this system to make dual tagged constructs, which may allow us to elucidate on the translational mechanism of polycistronic genes in *Drosophila.*

Overlapping PCR was used to insert a C-terminal 3x Hemagglutinin (HA) tag at the end of ORF1 in the constructs containing ORF2-FLAG. This would allow detection of both SEPs in the same sample using one transfection (Figure 3.8). I designed forward and reverse PCR primers annealing around the end of ORF1 just upstream of the stop codon. Each primer included a short, gene-specific sequence to allow annealing to the construct, followed by almost half of the ~100nt HA tag sequence with a small amount of overlap of 12nt between the two halves of HA tag sequence in each primer. The forward (PCR1) and reverse (PCR2) HA insertion primers were used in separate PCR reactions with primers targeting the outermost sequences of the constructs, namely pActin (Fw) and SV40 (Rev) annealing outside the plasmid restriction enzyme sites. PCR1 and PCR2 produced amplicons with the partial HA sequences which were mixed together in PCR3 to allow the complementary DNA strands to anneal using the overlap in the partial HA tag sequences. PCR3 was run for 10 cycles without primers to allow the polymerase to fill in the complementary ends after which the outer primers were added in to amplify the whole fragment. The resulting fragment was then finally cloned back into the pActin plasmid using restriction enzyme digest and ligation.

This approach enabled us to detect both smORFs in the same sample (Figure 3.9). Using these dual-tagged dicistronic constructs, I was able to complete the Western blot detection of 2 out of the 4 constructs that were cloned. *Scl*-ORF1 tagged with a 3x FLAG-3xHA fusion tag was used as a positive control construct since this translated peptide can be detected in both Anti-HA and Anti-FLAG Western blots. All four dual tagged plasmids were then passed onto another member of the lab; who verified expression of all 8 ORFs in these constructs by Western blot (Personal Communication with Julie Aspden).

Mass Spectrometry

Enrichment of small proteins

Differential Solubilisation

As described in the introduction of this chapter, in order to conduct MS for the detection of SEPs, we had to employ different approaches for the enrichment of small proteins. Differential solubilisation was used to enrich for small proteins by adapting the protocol in Kawashima *et al.* (Kawashima *et al.* 2010). Differential solubilisation has previously been used for less complex samples such as serum and whole blood but not for whole cells (Figure 3.10A). The protocol began by disrupting 5-20 million *S2* cells in 3 volumes of high-urea buffer to lyse the cells and denature the proteins. The lysed sample was then passed through a 30-gauge needle syringe 10 times to mechanically shear the cells. The sample was then further clarified by centrifugation. The total protein in the lysate was then precipitated by slowly dripping the sample into 10-30 volumes of ice-cold acetone with constant stirring. This mixture was stirred for 1hr at 4°C and pelleted by centrifugation, after which the supernatant was removed. The small proteins and peptides were then solubilised by resuspending the pellet in acidified (12mM HCl) ice-cold 70% Acetonitrile (ACN) by constant stirring for 1 hour at 4 °C, and the soluble fraction was collected by centrifugation. The sample was then lyophilised in a SpeedVac and resuspended in 30 µl of water. The protein quantity in the sample was quantified using the Bradford assay, after which 15 µg of protein was loaded onto a 16% Tricine gel along with an equal amount of unfractionated cell lysate and visualised by silver staining.

In Figure 3.10B, we can see that there is a very slight enrichment of small proteins in the differential solubilisation sample despite the fact that there is negative staining of the large smeared band at the 15 KDa marker. This artefact may have been the result of possible overloading of the gel, as there was some difficulty in quantifying the protein sample in some replicates of the experiment, resulting in low readings obtained in the Bradford assay. The smearing of the dye front (Figure 3.10B) also suggests that there may have been high salt and/or DNA carryover in the sample, from the hypertonic lysis buffer used to denature the proteins. This showed that the differential solubilisation method of enrichment still required optimisation as it also showed inconsistency between experiments. Due to this, I decided to focus on the implementation of the other two enrichment methods described below.

SDS-PAGE

1-D SDS-PAGE was used to fractionate small proteins from total *S2* cell lysate. The total protein was resolved according to protein size, which was measured by comparing to an external protein standard (Bio-Rad Dual Xtra). As summarized in Figure 3.11, 25-million *S2* cells were lysed in 100µl of 0.075% SDS in 1x Roche protease inhibitor cocktail solution by 3 cycles of freeze/thaw with liquid nitrogen and a warm water bath, using vortexing in between cycles to mix the sample. The lysate was clarified by centrifugation and quantified using Bio-Rad Bradford protein reagent. 375µg of total protein was made up to 60µl in Tricine Loading buffer v/v 2.5% Beta-Meracaptoethanol (Bio-Rad) and denatured by heating at 95 °C for 5 minutes. 300µg of protein was loaded between 4 lanes (to avoid overloading the gel) along with the

DualXtra protein marker from Bio-Rad and run on a 10-20% MiniProtean Tris-Tricine Gel (Bio-Rad) for 1 hour at 100V. The region corresponding to 5-15 KDa was excised from the gel in a laminar flow hood using a clean scalpel and transferred to an Eppendorf LoBind tube with 100μl of ultrapure water. The gel slice was then sent to the Cambridge Centre for Proteomics (University of Cambridge, UK) for in-gel trypsin digestion and analysis by nano HPLC-MS/MS on an Obitrap Velos mass spectrometer (Thermo Fisher Scientific).

Ultrafiltration

The principle of Ultrafiltration is the use of a specially constructed membrane with a specific pore size to fractionate proteins by acting as a sieve according to their molecular weight (MW). Ultrafiltration spin columns are available in a wide range of MW cut-offs (MWCO). The Ultrafiltration protocol used was based on the method of Slavoff *et al.* who have used this method to identify SEPs in Human cell lines (Slavoff *et al.* 2013). Four samples of (25 million cells each) *S2* cells were spun down and flash frozen in Eppendorf LoBind tubes. These samples were individually resuspended in 0.5 mL of boiling water. The tubes were then incubated in boiling water for 15 minutes in order to deactivate any proteases present in the sample. This is an important step that is conducted to minimise proteolytic fragments of larger proteins that can contaminate the low MW fraction. The samples were then cooled to room temperature and pooled into 2 tubes containing 1 ml each and the samples were sonicated on ice using a Branson 250 Digital Sonifier. The samples were then diluted to 0.25% (v/v) acetic acid and centrifuged at 16000*g* to clarify the lysate.

For the enrichment of proteins smaller than 100 amino acids (11 KDa) the supernatant was applied to a Vivaspin 500 ultrafiltration column with a 30 KDa MWCO PES membrane. I chose to use a 30 KDa MWCO Spin column though there was the option of using a smaller 10 KDa MWCO column. Since the Flow-through was to be collected, the aim was that the larger pore size would minimise loss of sample and help aid recovery of SEPs. Smaller pore sizes are more susceptible to loss of sample through non-specific adsorption (Figure 3.12A). The sample was centrifuged until about 90% of the sample had passed through the membrane. The filtrate was then collected, flash frozen and then evaporated to dryness using a SpeedVac spin concentrator. The sample was then resuspended and run on an SDS-PAGE gel before using silver stain to confirm enrichment of proteins below 30KDa.

I observed very little protein in the gel with very faint bands below the 30KDa marker (Data not shown). Therefore, in order to reduce non-specific adsorption, Bovine Serum Albumin (BSA) was used to pretreat the membrane prior to sample loading (Application Note, Sartorius AG) The spin columns were blocked with a 1% (w/v) solution of BSA dissolved in PBS overnight at room temperature, after which they were thoroughly washed by pipetting with 1X PBS and then further washed using 0.25% (v/v) Acetic Acid in order to minimise the carryover of the BSA to the sample. The sample was then applied to the passivated spin column and visualised by silver stained SDS-PAGE using the method described above. As can be seen from Figure 3.12B, this method showed significant enrichment of proteins smaller than 30 KDa. The Ultrafiltration method was considered successful and therefore another sample was prepared the same way as described above, to send to the Cambridge Centre for Proteomics (CCP) (University of Cambridge) for trypsin digestion and MS analysis.

Flybase smORFs detected by small protein fraction Mass spectrometry

The spectra obtained from these samples were then matched against the *Drosophila melanogaster* proteome (Flybase release 5.55) using MASCOT search engine with the generic algorithm (Perkins *et al.* 1999) (Figure 3.11). Using these results, we were able to detect 64 FB smORFs in the SDS PAGE enriched sample. At this stage it was found that the Ultrafiltration sample did not run successfully due to the presence of low MW contaminants (Personal Communication with Mike Deery at CCP). As this final stage I chose not to pursue the Ultrafiltration enrichment technique any further as it was difficult to ascertain the source of the contamination. Instead, a second SDS-PAGE enriched sample was prepared and sent for MS analysis as a technical and biological repeat. The results from the second SDS-PAGE enriched sample showed the detection of 64 FB smORFs of which 53 (83%) overlapped with the first sample. Thus highlighting the fact that SDS-PAGE enrichment improves the detection of SEPs and profers a high level of reproducibility between replicates.

The total 75 SEPs detected by MS from both replicates were subjected to thresholding to filter out low confidence hits. The protein score in the MASCOT search results is a probability-based score that is calculated by summating the ions scores, which is the E-value for matching spectra to tryptic peptide sequences, for each of the peptide fragments of that protein detected by the MS experiment. This means that the longer a protein sequence is, more unique peptides are available for detection, thus resulting in a higher score. For each experiment, the program also calculates a “threshold score”, which is the minimum protein score based on a significance threshold of $p<0.05$ for ascertaining that the protein is present in the sample. In this experiment, the minimum protein score was 56. Any proteins with a score below this threshold score have a greater than 1 in 20 statistical probability of being matched by chance and there is low confidence that they are matched. These results were removed from the list of high confidence hits, resulting in a final pool of 60 FB smORFs with high confidence of translation. Of these 60 hits, only 20 had been previously detected MS according to the Peptide Atlas database (Loevenich *et al.* 2009) and almost all (59/60) of the high confidence MS hits were deemed to be translated according to our Poly-Ribo-Seq dataset (Figure 3.13A). In order to assess the stringency of our thresholding; a comparison was conducted of the various metrics of the detected SEPs before and after filtering of smORFs with a protein score less than the significance threshold of 56 (Table 3.3). This revealed that the threshold score automatically generated by the MASCOT program is too stringent as the 15 SEPs discarded by thresholding had previously been detected by MS in the Peptide Atlas experiments, therefore they must be translated. The majority of these discarded SEPs (13/15) were also deemed translated according to Poly-Ribo-Seq where we could also observe that the median *RPKM* for the filtered pool of ‘high confidence’ smORFs is significantly higher compared to the larger pool of ‘untranslated’ smORFs and this is depicted in my results. Figure 3.13B shows a comparison of the median *RPKM* of the 90 smORFs with previous peptidomic evidence (*RPKM*: 170) against the remaining 138 smORFs (*RPKM*: 48) in the list of smORFs translated according to Poly-Ribo-Seq. This observation suggests that the thresholding exercise simply selects for the more abundant peptides in the sample, as they may have a higher protein score due to a higher number of peptide fragments detected (Table 3.3). It is well known that MS tends to detect the more abundant proteins (Fonslow *et al.* 2011). However, in the absence of an alternative scoring method, we only chose to publish the peptidomic detection of the 60 ‘high confidence’ smORFs in Aspden *et al.* 2014.

Custom Database search for novel smORFs in 5'UTRs and non-coding RNAs

A custom database was designed, consisting of the 14,881 uORFs and 6,438 ncrORFs identified by our lab to search for peptide fragments arising from the novel smORF peptide sequences. These had been identified in our lab by using the EMBOSS *getorf* program to identify all putative smORFs in 5' UTRs (uORFs) and non-coding RNAs (ncrORFs) that begin with an AUG start codon and encode for a peptide longer than 10 amino acids (Aspden *et al.* 2014). Using the MASCOT program to match the spectra against this database, we were able to match the MS results to 33 uORF and 13 ncrORF transcripts. However, none of these smORFs achieved a protein score higher than the significance threshold score calculated for this experiment (56). This was not entirely unexpected as these SEPs are very small in size and are usually expressed at low levels, and the protein score is calculated by abundance and biased by the size of the protein. These hits were further examined by manually checking the peptide matches using tBLASTn (Altschul *et al.* 1990), which was adjusted for short sequences (Figure 3.14A). Any SEPs with a detected peptide fragment containing a match to another annotated gene were discarded, as were any hits with a single peptide match less than 8 amino acids in length, since searches of 7aa or less generally aligned to multiple regions in the genome using tBLASTn (unpublished observation). After this filtering step, 18 uORFs and 8 ncrORFs remained in the list of custom database detected smORFs. Upon comparison with the Poly-Ribo-Seq data, we found that many of these were not transcribed according to the RNA-Seq experiment we conducted in *S2* cells (Figure 3.14B). If these hits are removed from the total pool, we have 16 uORFs and 1 ncrORF (FBtr0309289_1) MS detected SEPs that were transcribed in *S2* cells. Of this number, 3 uORFs passed the translation cut-offs of Poly-Ribo-Seq (FBtr0307015 (66aa), FBtr0086542 (15aa) and FBtr0302277 (31aa)) while an additional 6 (FBtr0088254_5, FBtr0086542_3, FBtr0075812_4, FBtr0335396_3, FBtr0306168_3, FBtr0073546_1) had some reads and the remainder had no reads aligning to the ORF. Surprisingly, these remaining 16 uORFs had a median length of 27aa, which is significantly lower than the 45aa (5KDa) cut-off used in the enrichment step, suggesting that not all of these matches are real and may thus be false positives. Therefore, given the difficulty of determining any statistical evidence for these matches and the high probability of cherry-picking false positives, we decided not to investigate these uORFs further at the time.

Discussion

in vitro Translation

Our first attempt at detecting smORF peptide translation in a medium to high-throughput manner was by the IVT method. Despite optimisation of various techniques (RRL, TnT RRL, TnT ICE), the IVT attempt was overall not very successful at the detection of SEPs when they were provided as the endogenous transcript or tagged with the small FLAG peptide. From the results of these experiments it can be seen that there may be several reasons for IVT not being the optimal technique for verifying smORF translation. There is documented evidence from *E.coli* based IVT systems (Loose *et al.* 2007) that small proteins maybe relatively unstructured and therefore more accessible to proteases, leading to the active degradation of peptides smaller than 80aa in length. Another reason that SEPs may not be readily visualised is the Globin mRNA that is added to nuclease-treated RRL, as it is necessary for translational activity (Pelham, and Jackson 1976). Globin mRNA migrates as a bright 10 to 15 KDa band on the gel, which is the region at which most SEPs may be detected. The high level of signal from these bands may mask fainter bands in the region where SEPs would migrate to on the gel, thus hindering their detection. The final factor that may have played a role in the difficulty of SEP detection, which is insufficient signal due to inefficient incorporation of fluorescently labeled Lysine residues. It is estimated that approximately 1 in 3 lysine residues are incorporated in the IVT process (Crowley *et al.* 1993). The constructs tested in my assay contained 4, 7 and 11 lysines, and only 40% of the 228 Poly-Ribo-Seq translated FB smORFs contain more than 6 Lysine residues. This means that more than half of the of FB smORFs and certainly the majority of the shorter 'Dwarf' smORF may not be sufficiently labeled for detection using this method.

The IVT method revealed that we could only detect the SEPs when the transcript was provided as a fusion protein with the much larger Venus/GFP tag, which consists of 238 amino acids, illustrating the difficulty in detecting small peptides using this technique. This hypothesis was tested further with an experiment with 3xFLAG-tagged constructs. FLAG is a small-peptide tag (8 amino acids long) but has a very specific and efficient detection by antibody staining. Using FLAG-tagged constructs, the IVT products were followed by Western blotting, which is a highly sensitive assay and will detect even small amounts of protein by the epitope tag. This showed we are unable to detect any signal from the FLAG-tagged SEPs. Though this may have been because no product was being produced, it is more likely that these smaller tagged peptides may have been masked due to endogenous peroxidase activity of the Globin present in the sample. As previously mentioned, Globin appears on the autoradiogram in the same range as the endogenous and FLAG-tagged smORFs. Horseradish Peroxidase (HRP) conjugated secondary antibodies are used in Western blotting to detect the primary antibody (Anti-FLAG) to catalyse the formation of luminescent signal on the blot that is detected by X-Ray film. If we were to revisit these experiments, I would titrate the IVT reaction with protease inhibitor in order to test the theory of degradation of small proteins, attempt to remove the Globin contamination, or simply always use a larger Venus or GFP tagged version of the smORF transcript.

Western Blot Detection of FLAG-tagged smORFs

The Western blot approach was designed to corroborate the translation of Poly-Ribo-Seq smORFs by the tagging-transfection assay. This assay involved the transfection of tagged smORFs into *S2* cells, which were then harvested and resolved on SDS-PAGE before immuno-blotting with the highly specific and efficient Anti-FLAG M2 Antibody (Sigma). Overall, this approach was successful in providing proof of translation of the majority of the Poly-Ribo-Seq translated smORFs that were tested, including some of those that fall below Poly-Ribo-Seq metrics. However, the AUG-FLAG positive control construct, which was meant to show the translation of FLAG-only peptide from a non-tag AUG, was not readily detected. This may have been due to unstable nature of the extremely small peptide, or simply because it was too small to be retained on the PVDF membrane (MacPhee 2010). The parallel microscopy approach to detect the SEPs from the tagging-transfection assay also detected very low signal from this construct in a very small proportion of transfected cells, corroborating the difficulty of detecting this peptide (Personal communication Unum Amin). Therefore *Scl-A,* which is known to be translated and has been characterised in-depth in our lab, was used a positive control for tag translation from a smORF start codon (Magny *et al.* 2013). For all other constructs, the Western blot results correspond completely to the results from the microscopy approach.

A good correlation was seen between the expected and observed sizes for the FB smORFs in the Western blot detected tagged smORF pool. The panel of smORFs tested showed that smORFs below the *RPKM* and coverage cut-offs used to define translation in the Poly-Ribo-Seq dataset, were translated (Table 3.2), thus illustrating the stringency of the *S2* cell Poly-Ribo-Seq cut-offs. We did not observe any correlation between Western blot signal and Poly-Ribo-Seq abundance in *RPKM*, which may be explained by the overexpression (under the Actin promoter) of recombinant constructs that are not in an endogenous context (SV40 3'UTR). We hoped that the Translation Efficiency (TE) metric, which is normalised to mRNA abundance would correlate better to Western blot signal, but this was not the case, as it does not take into account the stability or detectability of the SEPs.

The median size of the detected SEPs (13 KDa) compared to those that could not be detected (7 KDa) shows that there is a clear bias against very small proteins in the tagging-transfection assay. There were a few exceptions to the predicted size estimation where a few smORFs either ran slightly higher (CG12384), and/or showed multiple bands on the blot. This could be due to the Post-Translational Modification (PTMs) undergone by these SEPs, phosphorylation or the addition of carbohydrate chains (Walsh 2006). PTMs, hydrophobicity and amino acid usage can all affect protein migration in SDS-PAGE (Shirai *et al.* 2008). CG33199 and CG33170 were shown to have significant overlap with the mitochondrial dye (MitoTracker-Red) in *S2* cells (Aspden *et al.* 2014). This is interesting as it may explain the presence of multiple C-terminal FLAG-tagged isoforms, as peptides targeted to particular organelles are processed by proteolytic cleavage of signal peptide sequences at the N-terminal (Heijne 1990). One major difference was seen for the ncrORF pncr009:3L-ORF4 for which the expected (8.2 KDa) and observed (15 KDa) size was quite different. Upon closer examination of the cloned sequence, we discovered an in-frame initiation codon upstream of the predicted start site, highlighting the challenge of assessing smORF translation, especially for ncrORFs. This shows the value of corroborating the size of the translated SEP with the nucleotide sequence, as this mistake would have gone unnoticed without the Western blot detection of the tagged SEPs. Although we were not able to see any obvious correlation between Western blot signal and Ribo-Seq metrics, this assay was useful in providing proof of translation and assessing the stringency of the Poly-Ribo-Seq translation cut-offs. In addition, these results show the usefulness of assessing the size of the translated peptide to define the correct ORF and highlight modification and active regulation of these peptides, which can be an indicator of some function.

The construction of the dual-tagged constructs was successful and I was able to detect ORF1 (Anti-HA) and ORF2 (Anti-FLAG) in the same sample for CG32736 and CG42497 constructs by Western blot. All four dual tagged plasmids were then passed onto another member of the lab; who verified expression of all 8 ORFs in these constructs by Western blot (data not shown). These constructs are now being used for further studies to investigate the relative translation of the two ORFs by using site-directed mutagenesis to mutate the different start/stop codons, and their analysis by polysome profiling and qRT-PCR to characterise mechanism of translation of multiple ORFs from the same transcript (Personal communication JL Aspden).

Mass spectrometry

Enrichment of small proteins is most successful by SDS-PAGE

There were three techniques implemented for the enrichment of small proteins in order to perform Mass Spectrometry to independently detect SEPs from *S2* cells. The differential solubilisation method has the advantage of being a biochemical separation technique as opposed to a physical one and therefore is not susceptible loss of the sample by non-specific adsorption. Although, differential solubilisation has been used for enrichment of blood serum (Kawashima *et al.* 2010) it has previously never been used for whole cell lysate. As whole cell lysate is a much more complex sample than serum, this may explain the significant amount of high MW proteins that remained in the sample even though I was able to observe some enrichment of small proteins. It was also difficult to quantify the sample using the Bradford assay, possibly due to salt carryover from the Urea lysis buffer used to prepare the sample, but this cannot be confirmed from the detection method. Consequently, it was difficult to assess the level of enrichment of the small peptides by SDS-PAGE due to overloading of the gel and smearing of the protein-dye front as seen in Figure 3.10. The significant optimisation and desalting required for the differential solubilisation protocol meant that we would not be able to validate this enrichment technique until the MS was conducted. Therefore we decided not to use this technique.

The Ultrafiltration enriched sample showed significantly better enrichment of small peptides than the Differential-solubilisation sample. We could also observe depletion, but not complete absence, of proteins larger than 30 KDa, as compared to the unfractionated cell lysate. However, Ultrafiltration did not perform well in the MS run due to the presence of a significant proportion of low MW contaminants, even though care was taken to use high-grade reagents and maintain a clean environment during experimentation. Sample contamination has been seen to be a common problem in MS workflows particularly using Ultrafiltration (personal communication with Alan Saghathelian). It is difficult to pinpoint the source of the contaminants, which may be the atmosphere (dust) or some of the communal reagents and equipment used, as this sample was not prepared in a specialised MS lab. Finally, since Ultrafiltration may cause non-specific adsorption of small hydrophobic proteins on the membrane, it suffers from poor batch-to-batch reproducibility (reviewed in Finoulst *et al.* (2011)). Hydrophobic proteins are over represented in the *S2* Poly-Ribo-Seq translated FB smORFs (Aspden *et al.* 2014), so we may have been inadvertently excluding these using Ultrafiltration.

SDS-PAGE was the most successful of the enrichment techniques used to enrich a small peptide fraction for MS. Using SDS-PAGE is fairly straightforward, and required choosing a portion of the protein gel to excise. I chose to cut the region corresponding to 5 to 15KDa in the gel, which is roughly equivalent to 45-130 amino acids (1aa = 110Da). The excellent run-to-run reproducibility of the approach can be seen given the 83% overlap between two experimental repeats of the sample prepared from *S2* cells, and this method allowed the detection of 75 FB SEPs. The upper limit of 130 amino acids was used to allow for some proteins running higher on the gel due to PTMs, as was observed in the Western blot results. 5KDa was chosen as a lower limit as degradation products of larger proteins can overwhelm the gel region below this size. This may not have had major consequences for the identification of FB smORFs, which have a median length of 80 amino acids, but may have had an impact on the detection of dwarf smORFs, about 90% of which are smaller than 45aa (Median length: 20 amino acids = 2.2 KDa). In hindsight, this experiment may be improved by including the <5KDa region in the analysis and then removing the degradation products during the data analysis stage but it is difficult to say what impact this may have had on the overall sample as MS tends to detect the more abundant proteins in the sample. Similar studies aimed at finding SEPs that have used SDS-PAGE as an enrichment technique, divide

<15KDa region into 3 different samples of 2-5, 5-10 and 10-15KDa in order to reduce sample complexity (Ma *et al.* 2014). Ideally, we would repeat this enrichment method and divide the <15KDa gel slice into multiple samples and include the <5KDa portion as a separate sample.

Using Mass Spectrometry based approaches to corroborate translation

Overall, the peptidomics experiments showed evidence of translation of 60 FB smORFs with high confidence using the MASCOT scoring algorithm (Perkins *et al.* 1999), two-thirds of which do not have any previous proteomic evidence. I also observed the translation of a further 15 FB smORFs, 13 uORFs and 1 ncrORF that did not pass the confidence thresholds. This significant result shows the advantage of using an enrichment-based Peptidomics workflow as opposed to Proteomics, which has previously not been fruitful in the detection of SEPs (Brunner *et al.* 2007). 59 of the 60 high-confidence SEPs detected in our experiments and 90% of all SEPs detected by mass spectrometry are also translated according to the Poly-Ribo-Seq data. Peptidomics data tends to detect more abundant smORFs. Detection of very small peptides is a well-known limitation of the technique (reviewed in Chu *et al.* (2015)) and it is extremely difficult at the moment to detect dwarf smORFs by mass spectrometry due to their small size and the fact that dwarf smORFs tend to be expressed at low levels with high spatio-temporal specificity (Guttman *et al.* 2011; Guttman and Rinn 2012). Thus the process of sample fractionation on a tissue and subcellular scale and enrichment of small proteins, as was conducted on the samples here, remain a key step in the detection of novel SEPs.

In Table 3.3, we can see that of the 15 smORFs that were filtered out due to low confidence; 13 were translated according to our Poly-Ribo-Seq data and all 15 had been previously detected in published proteomic studies. This highlights the fact that the main challenge in a peptidomics approach lies in statistically matching small peptides to the resulting spectra and improving the data analysis pipeline (reviewed in Chu *et al.* (2015) and Schrader (2014)). A typical proteomics pipeline requires more than one unique peptide match for a protein to be considered translated (Wilkins *et al.* 2006) but we did not use this analysis criterion since the small size of SEPs limits both the number and size of fragments generated. The current probability-based scoring methods work well for larger proteins, but smaller proteins tend to suffer for a number of reasons. For example in the MASCOT program the protein score is the sum of all the highest ions score for each unique peptide match this means that larger protein with many unique peptide fragments score better than small peptides. This is reflected in the scoring of the Custom database where we were unable to find any statistically significant matches for dwarf smORFs based on the conventional scoring system in MASCOT matched data.

This bias could be expected, as another study targeted at identifying SEPs has relied on manually curating spectra matched to tryptic peptides and attempted to use less stringent parameters to define a peptide (Slavoff *et al.* 2013), showing that SEPs require special treatment in the post-MS analysis. Another study combining the MS and Ribo-Seq approach in Zebrafish detected a similar number of annotated SEPs (98) and only 6 novel SEPs (Bazzini *et al.* 2014). However, only 3 of the 16 manually curated uORF-SEPs detected by MS in my study are translated according to Poly-Ribo-Seq. To improve matching of low confidence hits, another approach could be to incorporate known quantities of synthetic peptides to allow a reference for protein quantification. It may be that the identification of novel SEPs could be improved by limiting the search space based on ORFs transcribed in *S2* cells (by using RNA-Seq data) thus reducing the number of false positives. This approach seems to be the current trend in the proteo-genomics field (Crappé *et al.* 2015; Ma *et al.* 2014; Koch *et al.* 2014) (reviewed in Andrews and Rothnagel (2014)) where the difficulty of detecting SEPs is highlighted and the integration of Ribo-Seq and MS is arguably the best approach forward.

For example, the proportion of SEPs with predicted trans membrane alpha helices are overrepresented in our pool of translated smORFs (Aspden *et al.* 2014) and a large proportion of the tagged smORFs studied by immuno-fluorescence microscopy co-localize to mitochondria (Personal communication Unum Amin). Therefore it may be prudent to employ a membrane or mitochondria enriched (Poston *et al.* 2013) fractionation step to identify these SEPs. In conclusion, our results show that although MS is a valuable tool for an independent assessment of smORF translation it struggles with the statistically significant matching of SEPs. Therefore Poly-Ribo-Seq is a superior method for the identification of translated smORFs due to its greater sensitivity and can be used as a guide for independent corroboration of SEP translation.

In conclusion, our results show that *in vitro* translation is not suitable for assessing smORF translation. The tagging-translation assay and WB allows us to assess the translatability of an ORF but does not correlate with its *in vivo* translation, however it is useful validating gene models through observed peptide size. Furthermore, my results show that small protein enrichment is necessary for MS detection of SEPs and the best way to enrich is SDS-PAGE. Peptidomics is a viable study to be used as a parallel and complementary technique to corroborate the translation of smORFs but when it comes to MS, lack of detection does not equate to lack of translation, as can be seen from the Poly-Ribo-Seq data. Overall these experiments were successful in the context of a medium-high throughput study of smORF translation, and they do not replace a more extensive peptidomics study. A recent study that identifies 195 novel SEPs in human K562 cells suggests that a minimum of 10 replicate MS runs in order to cover the whole variety of the proteome (Ma *et al.* 2014). Peptidomics is an excellent technique for providing proof of translation but has a tendency towards low reproducibility and is therefore not an all-inclusive representation of the proteome. Therefore, the results shown here can be used as a corroboration of the Poly-Ribo-Seq results. This data represents the first and only small-protein-enriched MS study aimed at finding SEPs in *Drosophila* and increased the catalog of translated SEPs of 75 annotated smORFs over two MS runs.

Chapter 3 Figures and Tables

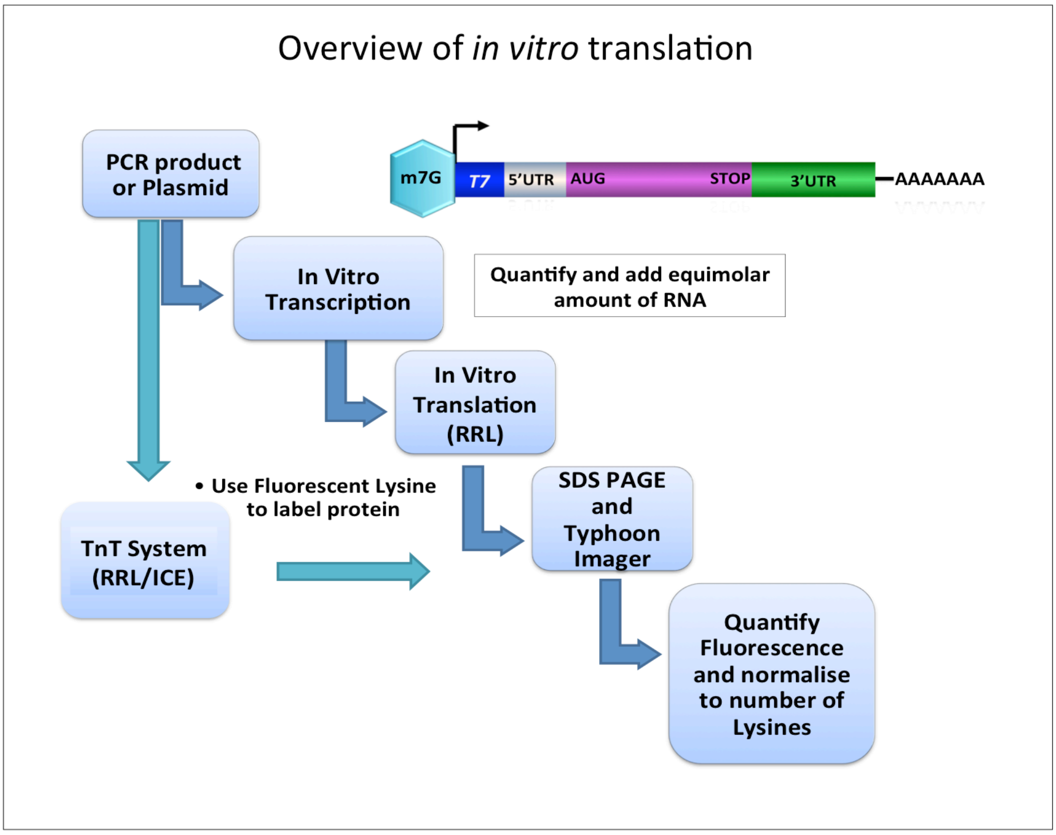

Overview of *in vitro* translation
PCR product or Plasmid
m7G
T7
5'UTR
AUG
STOP
3'UTR
AAAAAAA
In Vitro Transcription
Quantify and add equimolar amount of RNA
In Vitro Translation (RRL)
• Use Fluorescent Lysine to label protein
TnT System (RRL/ICE)
SDS PAGE and Typhoon Imager
Quantify Fluorescence and normalise to number of Lysines

Figure 3.1 Overview of *in vitro* Translation approach to corroborate translation of Poly-Ribo-Seq smORFs

In order to detect smORF translation using IVT, known quantities of *in vitro* transcribed and purified RNA were used as a template, containing an m7G cap, T7 promoter and a SV40 PolyA signal in the form of a PCR product or Plasmid of the smORF were subjected to one of the three different IVT systems: 1) A two-step *in vitro* translation reaction with the Rabbit Reticulocyte lysate (RRL) based system 2) an RRL system that uses a DNA template in a coupled transcription and translation reaction (TnT RRL) and 3) an insect cell extract (ICE) based TnT system (TnT ICE). The resulting products from these experiments were run on SDS-PAGE and viewed with a fluorescent Typhoon imager after which the fluorescence obtained from the sample was quantified and normalised to the number of Lysines in the peptide sequence of the smORF.

System	Luc	Control	Test	Comment
RRL	+	IVT Luc	-	No smORFs detected using cDNA plasmid
RRL	+	IVT Luc	-	No FLAG tagged smORFs detected by WB
TnT RRL	+	tal-GFP	-	Miniprep of tal-GFP control not detected
TnT RRL	+	tal-GFP	-	tal-GFP control detected after phenol chloroform, smORFs cDNAs not detected
TnT RRL	n/a	tal-GFP	-	tal-GFP positive, test smORFs did not work
TnT RRL	n/a	tal-GFP	+	2 of 3 test smORFs detected using Venus tag
TnT ICE	+	n/a	-	No smORFs detected using cDNA plasmid
TnT ICE	+	n/a	-	No FLAG tagged smORFs detected by Gel and Western Blot
TnT ICE	+	tal-GFP	+/-	tal-GFP plasmid not detected, very faint band for CG32230

Table 3.1 Summary of different *in vitro* Translation approaches

This table summarises the various repeats of the IVT experiments to detect the translation of smORFs. The first two attempts with the RRL-only system did not show any positive results with either cDNA or FLAG-tagged plasmids. The TnT ICE IVT system showed similar results with the same plasmids. The TnT RRL system attempts were the most successful, as the only instance of test SEP detection using the large Venus tag were apparent in the TnT RRL system, which also detects the *tal*-GFP plasmid in 3/4 attempts. The endogenous and FLAG-tagged smORF peptides are not detected using any of the three techniques, indicating that this could be an issue of size rather than unsuccessful IVT since the much larger GFP and Venus tags are detected.

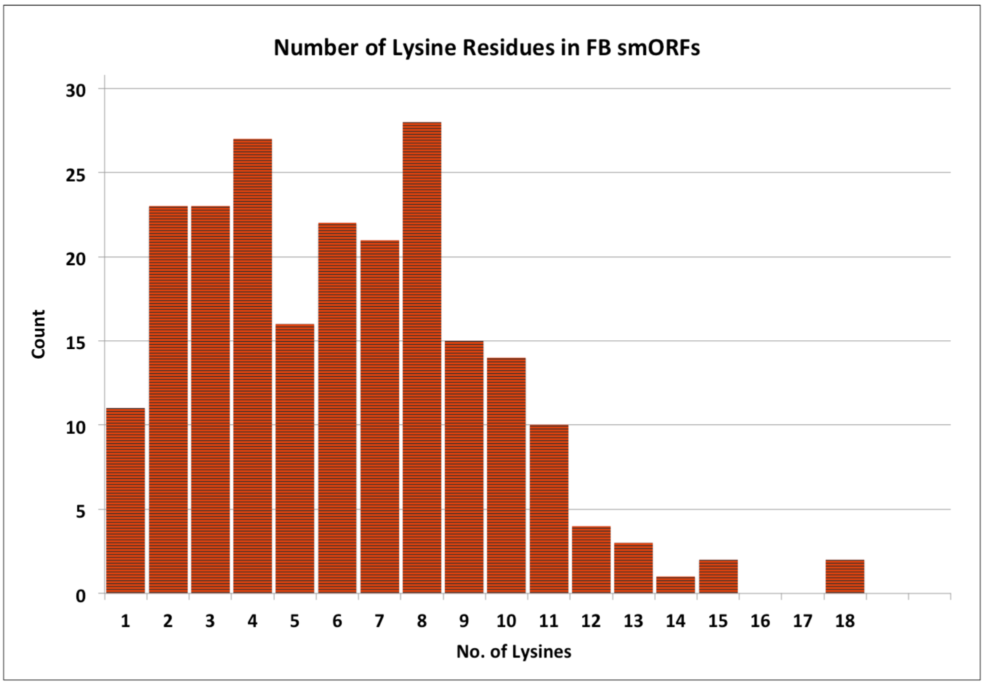


Figure 3.2 Frequency distribution of number of Lysines in *S2* cell smORFs
This graph plots the number of Lysine residues (x axis) in the peptide sequences from the 228 FlyBase annotated *S2* cell Poly-Ribo-Seq smORFs. Only 40% of the pool of 228 SEPs contain more than 6 Lysine residues and on average, one in three Lysines is labeled during the IVT reaction.

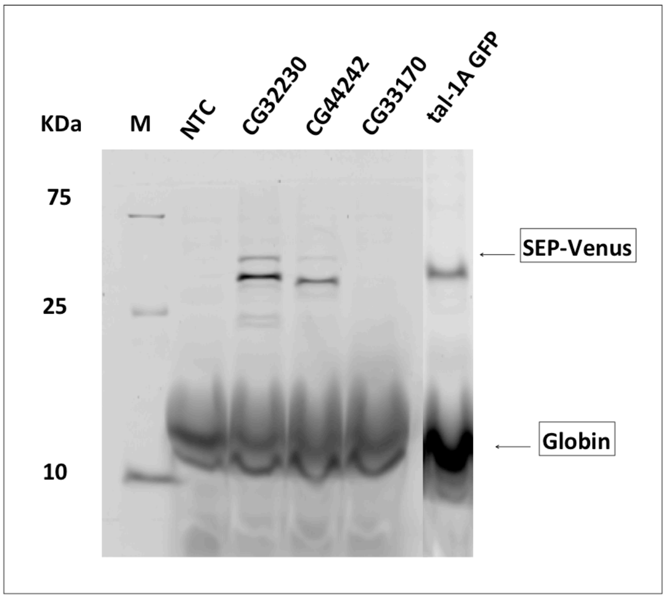


Figure 3.3 Detection of Venus-tagged smORFs in IVT assay
This figure shows the imaged SDS-PAGE gel from the successful TnT RRL IVT reaction using the Venus-tagged smORF constructs. Bands corresponding to talORF1A-GFP (positive control) CG32330-Venus and CG44242-Venus can be observed just above the 25 KDa marker. Multiple bands can be observed on the gel in the CG32230 and CG44242 samples, the lower band may be due to alternative translation or cleavage of the Venus tag. CG33170-Venus did not show a positive result in this experiment. The smear that is highlighted is Globin, which is added to the IVT reaction for maintenance of endogenous translation (Pelham, and Jackson 1976). The Globin band runs in the region of 10-15KDa, and would mask any endogenous or FLAG-tagged smORF peptide products.

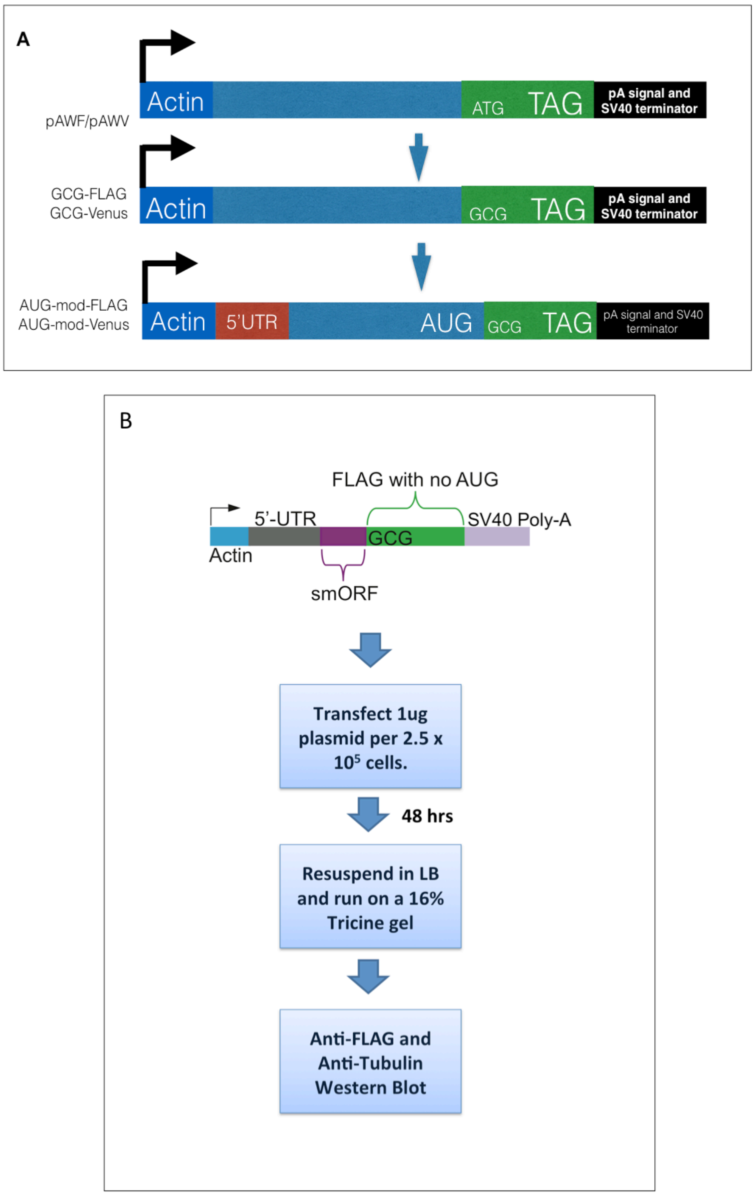

A
pAWF/pAWV
Actin
ATG
TAG
pA signal and SV40 terminator
GCG-FLAG
GCG-Venus
Actin
GCG
TAG
pA signal and SV40 terminator
AUG-mod-FLAG
AUG-mod-Venus
Actin
5'UTR
AUG
GCG
TAG
pA signal and SV40 terminator
B
FLAG with no AUG
5'-UTR
GCG
SV40 Poly-A
Actin
smORF
Transfect 1ug plasmid per 2.5 x 10^5 cells.
48 hrs
Resuspend in LB and run on a 16% Tricine gel
Anti-FLAG and Anti-Tubulin Western Blot

Figure 3.4 Overview of Tagging-Transfection Assay

A) Controls used for the Tagging-Transfection Assay were modifications of Murphy Lab Gateway Cloning expression vectors. Plasmid transcription is under control of the Actin5C promoter and the vector contains a C-terminal 3xFLAG/Venus tag sequence followed by an SV40 polyA signal. In order to eliminate expression of the tag only, the AUG start codon of the tag was changed to a GCG codon. These are termed GCG-FLAG and GCG Venus, and serve as a negative control in the experiment as there should not be translation of the tag in these constructs. The positive control AUG-FLAG/Venus construct was made by adding the 5'UTR plus the start codon of the translated smORF gene CG42371. These are termed AUG-mod-FLAG and AUG-mod-Venus

B) The 5'UTR plus the ORF sequence of the candidate smORF are cloned in-frame with the C-terminal tag in the GCG-FLAG plasmid and these were transfected in to *S2* cells. These cells are harvested 48-hours post-transfection, interrupted using LB and run on a Tricine gel. Western blot was performed and the membrane probed for FLAG M2 and β-Tubulin (loading control) to assess the presence and size of the translated SEP.

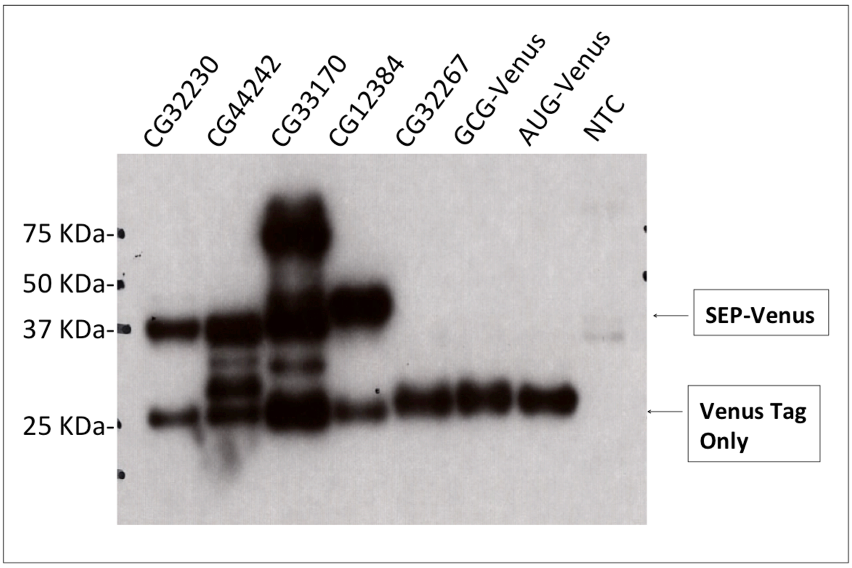


Figure 3.5 Western Blot Detection of Venus Tagged constructs
Results from Western Blotting of samples from the transfection of smORF-Venus plasmid constructs in *S2* cells probed with a cross-reactive GFP antibody. SEP-Venus fusion protein can be detected for CG32230, CG44242, CG33170 and CG12384 but not for CG32267. There is a significant presence of background in the CG33170 sample as determined by comparing with faint bands in the non-transfected control (NTC) sample. Interestingly, a 26 KDa band that corresponds to the size of the Venus tag on its own can be seen in all transfected samples including the positive control AUG-Venus. Critically, this band can also be observed in the GCG-Venus negative control sample, which suggests translation of the protein from a start codon downstream of the mutated ATG of the Venus tag.

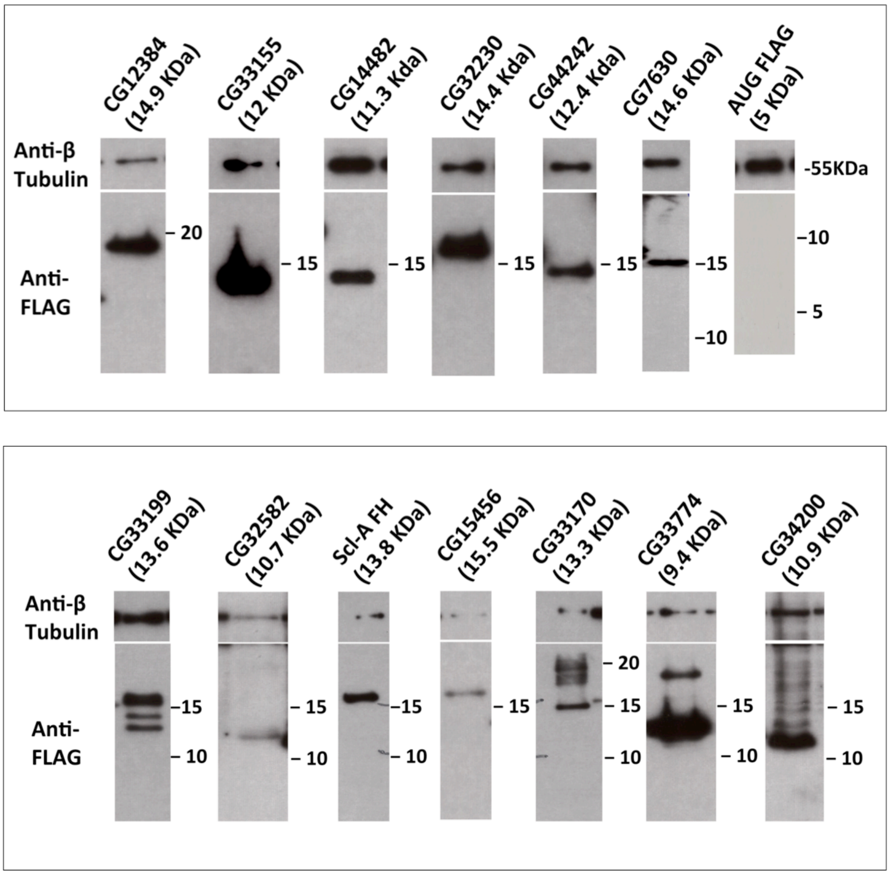


Figure 3.6 Western Blot results of Tagging-Transfection Assay

Results from Western blotting of samples from *S2* cells transfected with smORF-FLAG plasmid constructs and probed for FLAG and β-Tubulin (as a loading control). Several smORFs were tested in this assay, which were deemed as translated with high confidence in the Poly-Ribo-Seq experiments described in Aspden *et al. (Aspden et al. 2014)*. These smORFs were annotated as putatively protein-coding by FlyBase CG12384, CG33155, CG14482, CG32230, CG44242, CG7630, CG33199, CG15456, CG33170, CG33774, CG34200, CG32582 (later changed to non-coding RNA status) show translation in this experiment. CG33199, CG33170 and CG33774 show multiple bands and some smORFs run higher than their predicted size, most likely due to post-translational modifications. CG33170 has a complex splicing pattern and encodes for a larger, annotated 16.2 KDa peptide (including FLAG tag) from the use of alternative initiation from in-frame upstream start codons. CG33774 displays an additional faint band around 17 KDa but has no in-frame upstream start codon. CG33199 displays additional bands that run below the predicted size band at just below 15KDa that possibly occurs due to N-terminal proteolytic processing of the peptides or downstream initiation codon, encoding for a peptide 8.4 KDa. The positive control used was *Sarcolamban*-ORFA-FLAG (Scl-A FH) as AUG-mod-FLAG construct is not detected in these experiments despite repeat transfections but was detected in the alternative imaging method (Personal communication Unum Amin and Aspden *et al.* 2014).

CG_ID	RPKM	Coverage	Signal Strength	TE	Expected Size (KDa)
CG32230	474.29	1.00	High	3.1	14.4
CG33774	101.46	1.00	High	1.1	9.4
CG14482	527.82	1.00	Med	1.1	11.3
CG34200	291.82	1.00	Med	1.7	10.9
CG12384	180.89	1.00	Med	1.4	14.9
CG44242	134.46	0.97	Med	1.7	12.4
CG33170	74.07	0.84	Med	0.9	13.3
CG33155	29.73	0.64	Med	NA	12
CG32582	29.21	0.56	Low	2.8	10.7
CG15456	8.89	0.65	Low	0.9	15.5
CG1878	4.62	0.30	Low	3.3	11.8
CG34330	7.03	0.04	Low	0.0	10
CG7630	617.73	1.00	V.Low	1.0	14.6
CG33199	84.00	1.00	V.Low	1.2	13.6
CG32267	72.53	0.97	V.Low	1.1	10.5
CG33494	58.94	0.39	V.Low	5.0	15
CG13315	19.28	0.12	V.Low	NA	13

Table 3.2 Expression analysis of smORFs tested in the Tagging-Transfection Assay

This table summarises the strength of signal of from Western Blotting results of the transfected smORFs in *S2* cells, as compared to the metrics from Poly-Ribo-Seq results (*RPKM* and coverage and TE). This table also includes results from the transfections of 5 smORFs that fall below the cut-offs for Poly-Ribo-Seq (*RPKM*<11.7 and coverage<0.56, shaded boxes for CG15456, CG1878, CG34430, CG33494, CG13315), which all show signal in the Western Blots from repeat transfections. There is a general lack of correlation between WB signal and PRS metrics, even for Translation Efficiency (TE), which is normalised to RNASeq abundance. The final column shows the predicted peptide sizes of the SEPs.

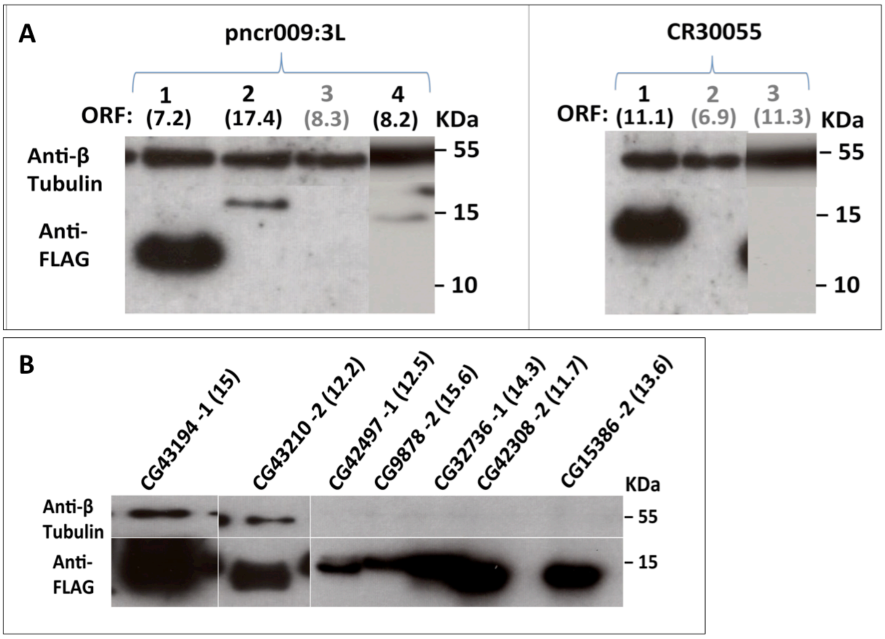


Figure 3.7 Western Blot detection of tagged smORFs found in non-coding RNA and poycistronic transcripts

A) Results from western Blotting of samples from the transfection of two polycistronic ncRNA smORF-FLAG plasmid constructs in *S2* cells probed for FLAG and β-Tubulin as a loading control. pncr009:3L ORF1 and ORF2 are the expected sizes, ORF3 is not detected whilst ORF4 runs higher (15KDa) than expected (8.2 KDa). Upon examination of the sequence it was revealed that there was an upstream, in-frame start codon, which may be the site of translation initiation of ORF4.

B) Results from western blotting of samples from the transfection of polycistronic smORF-FLAG plasmid constructs in *S2* cells probed for FLAG and β-Tubulin as a loading control. CG43194/CG43210 are dicistronic ORFs of the same transcript, as are CG42497/CG9878, CG32736/CG42308. CG15386 is ORF2 of a dicistronic smORF for which the first ORF (CG42371) was not cloned successfully. All of the transfections were successful in producing a detectable peptide from these smORFs.

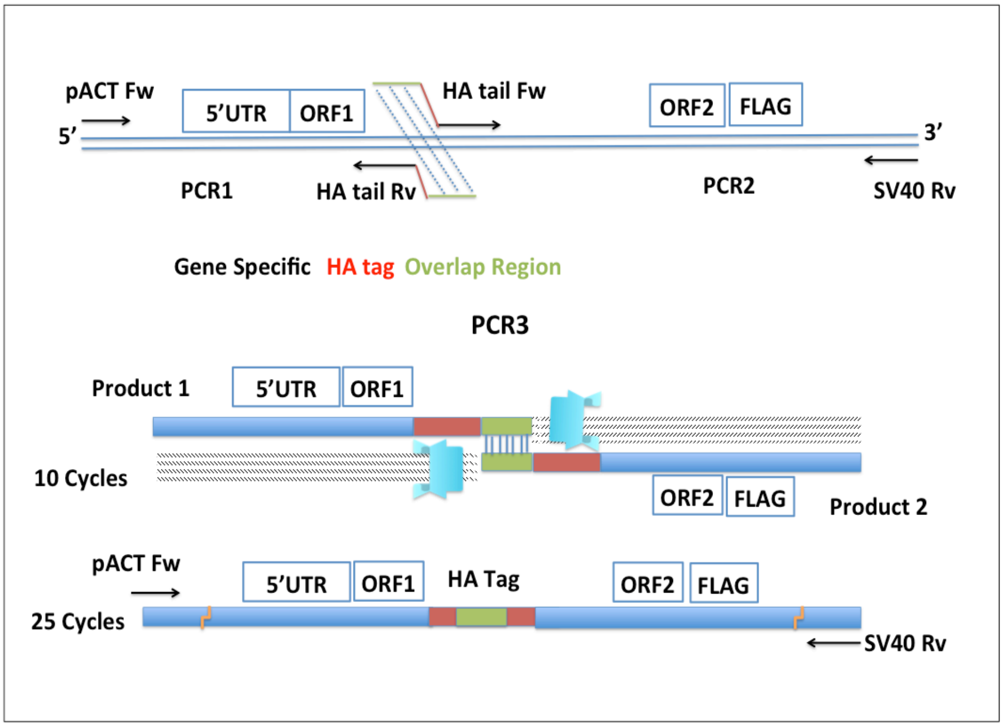


Figure 3.8 Generation of Dual-tagged polycistronic smORF constructs
Poly-Ribo-Seq in *S2* cells showed the translation of four dicistronic smORFs (also described in Figure 3.6B. These eight smORFs (CG43194/CG43210, CG42497/CG9878, CG42371/CG15386, and CG32736/CG42308 all showed varying abundance of translating mRNA and were interesting candidates to investigate as polycistronic genes are rare in eukaryotic species and usually found abundantly in prokaryotes. Using the ORF2-FLAG constructs of each of these smORFs, overlapping PCR was used to insert a C-terminal 3x Hemagglutinin (HA) tag at the end of ORF1 already present in the upstream sequences. Forward and reverse PCR primers were designed to anneal at the 5' end of ORF1 just upstream of the stop codon. These included a short, gene-specific sequence to allow annealing to the construct and approximately half of the ~100nt HA tag sequence with a 12nt overlap of HA sequence. The forward (PCR1) and reverse (PCR2) HA insertion primers were used in another PCR reaction targeting outside the plasmid restriction enzyme sites at pActin (Fw) and SV40 (Rev). PCR1 and PCR2 produced the partial HA sequences and these were mixed in PCR3 to allow the complementary overlap in the partial HA tag DNA strands to anneal. PCR3 was run for 10 cycles without primers to allow the polymerase to fill in the complementary ends and then the outer primers were added to amplify the whole fragment. The resulting fragment was cloned back into the pActin plasmid using restriction enzyme digest and ligation for subsequent transfection in *S2* cells.

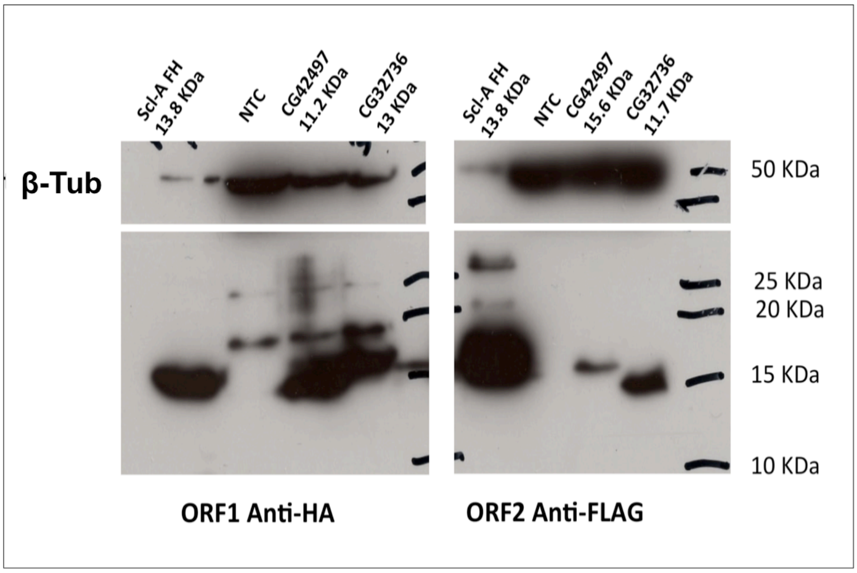


Figure 3.9 Western Blot detection of Dual-tagged dicistronic smORF constructs
From the transfections of the dual-tagged dicistronic constructs, we could observe the peptides being translated from both ORFs (ORF1-HA and ORF2-FLAG) from the same transfection. A *Scl*-ORFA construct with a 3x FLAG-3xHA fusion tag was used as a positive control construct since this can be detected in both Anti-HA and Anti-FLAG antibodies. Non-transfected cells (NTC) were used as a negative control. The two smORFs shown in this test were CG42497-HA/CG9878-FLAG and CG32736-HA/CG42308-FLAG and signal can be detected from both ORFs.

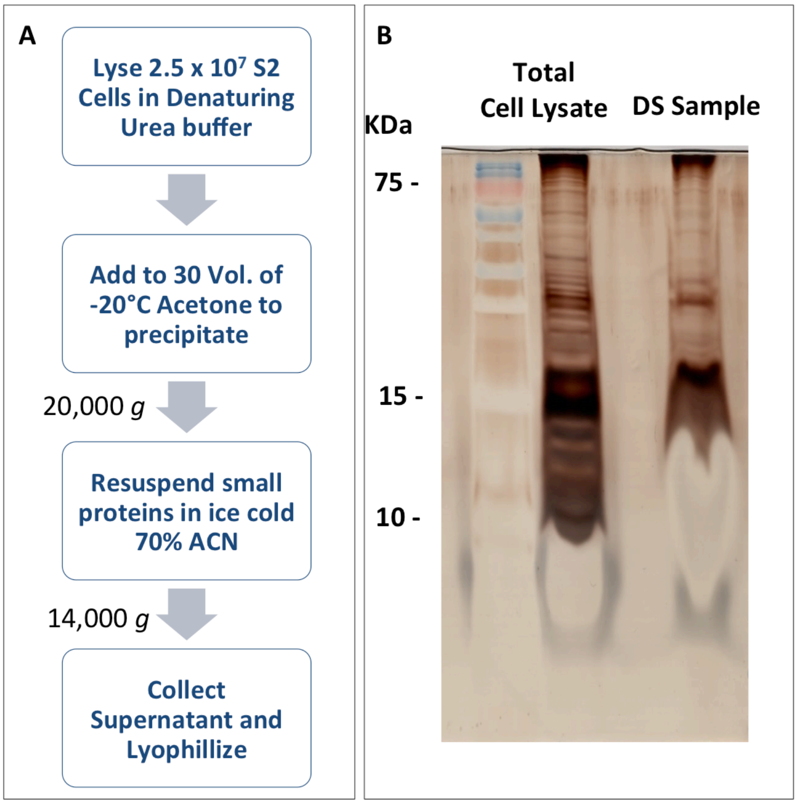

A
Lyse 2.5 x 10⁷ S2 Cells in Denaturing Urea buffer
Add to 30 Vol. of -20°C Acetone to precipitate
20,000 g
Resuspend small proteins in ice cold 70% ACN
14,000 g
Collect Supernatant and Lyophillize
B
KDa
Total Cell Lysate
DS Sample
75 -
15 -
10 -

Figure 3.10 Differential Solubilisation-based enrichment of small peptides for Mass Spectrometry

A) Overview of Differential Solubilisation enrichment process. *S2* cells are lysed in a denaturing high salt buffer, precipitated in freezing-cold acetone, and pelleted by spinning. The pellet is partially resuspended and then spun down again, this time the supernatant, containing small proteins, is collected, flash-frozen and lyophilised using a SpeedVac. The lyophilised sample is resuspended in loading buffer and run on SDS-PAGE.

B) Gel Scan of silver stained SDS-PAGE from enrichment of small proteins using Differential Solubilisation. Unfractionated cell lysate is used as a control. The DS sample shows a smearing gel front in the region where smORF encoded peptides would be found. This region appears as a negative stain. This could be due to possible salt carry over. Comparing this to the control shows that there is a possible partial enrichment of smaller peptides but they are not removed from the sample completely.

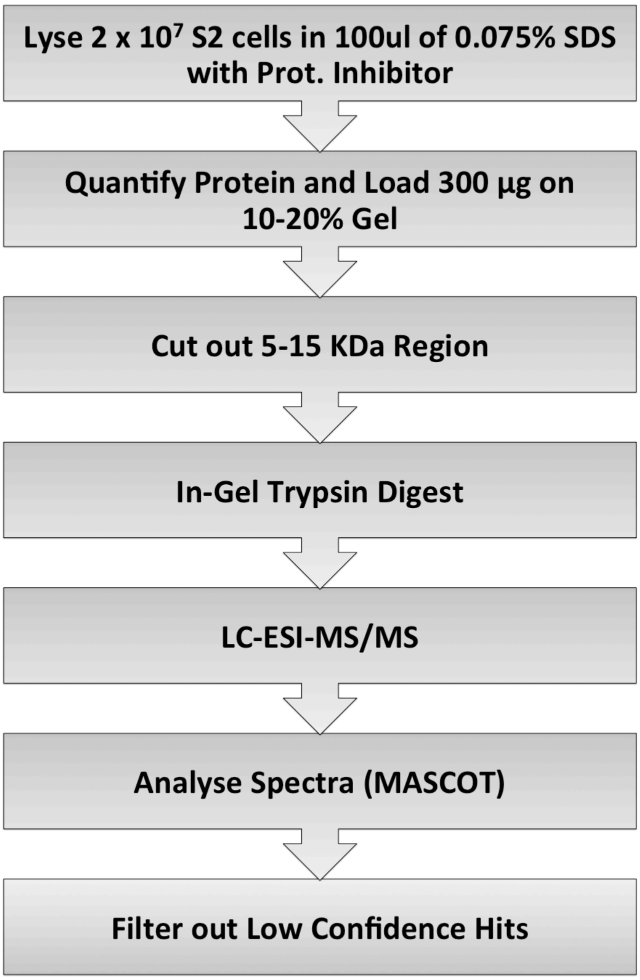

Lyse 2 x 10^7 S2 cells in 100ul of 0.075% SDS with Prot. Inhibitor
Quantify Protein and Load 300 µg on 10-20% Gel
Cut out 5-15 KDa Region
In-Gel Trypsin Digest
LC-ESI-MS/MS
Analyse Spectra (MASCOT)
Filter out Low Confidence Hits

Figure 3.11 Overview of final peptidomics workflow for the assessment of high confidence smORF peptides

Final protocol of the sample used for Mass Spectrometry study, which was enriched for small peptides using SDS-PAGE by cutting out the region corresponding from 15-5KDa. *S2* cells were lysed with 0.075% w/v SDS solution containing Protease inhibitor. The protein concentration was quantified and the sample was run on 10-20% SDS-PAGE. The region containing small peptides was cut and sent to the Cambridge Centre for Proteomics (CCP) (University of Cambridge). The sample was processed further by the CCP by an in-gel trypsin digest, followed by analysis by nano HPLC-MS/MS on an Obitrap Velos mass spectrometer. The spectra derived from the mass spectrometry were then run through the MASCOT search engine with the generic algorithm. This data was then sent back to us and filtered to remove any low-confidence hits.

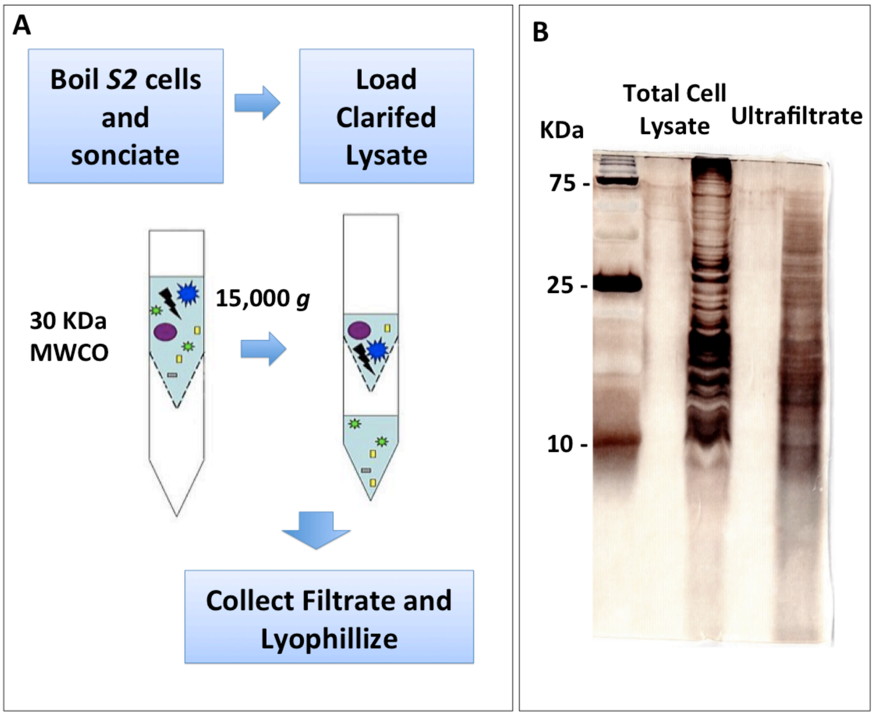


Figure 3.12 Ultrafiltration-based enrichment of small peptides for Mass Spectrometry
A) Overview of Ultrafiltration enrichment process. Ultrafiltration involves the boiling of the *S2* cells in water and this sample is sonicated. The acidified sample is then spin clarified using the VivaSpin ultrafiltration column. The flow-through, or filtrate, is collected, flash-frozen and lyophilised using using a SpeedVac.

B) Gel Scan of SDS-PAGE from enrichment of small proteins using Ultrafiltration. Compared to Unfractionated cell lysate control, the UF sample appears to enrich for smaller proteins, as can be seen by the increased separation of peptides below 30KDa though there is some contamination of peptides larger than 30KDa in the sample.

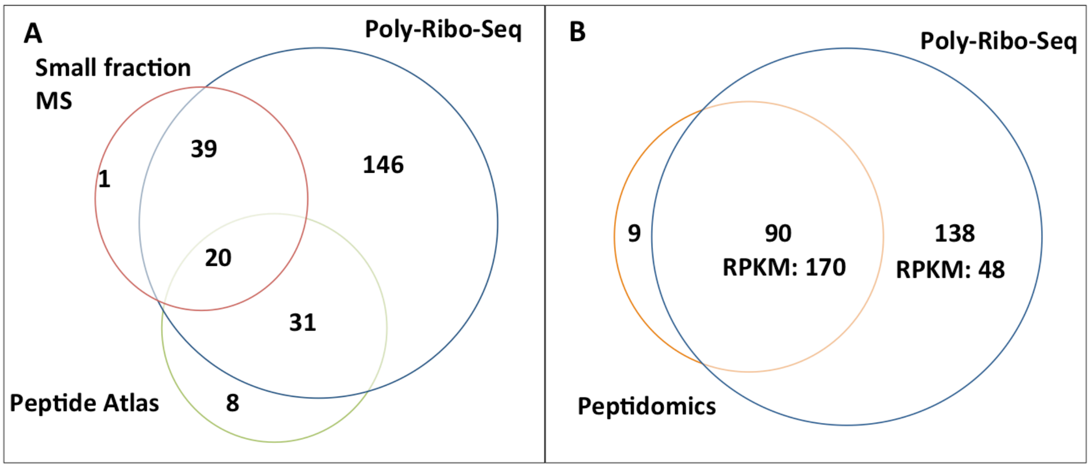
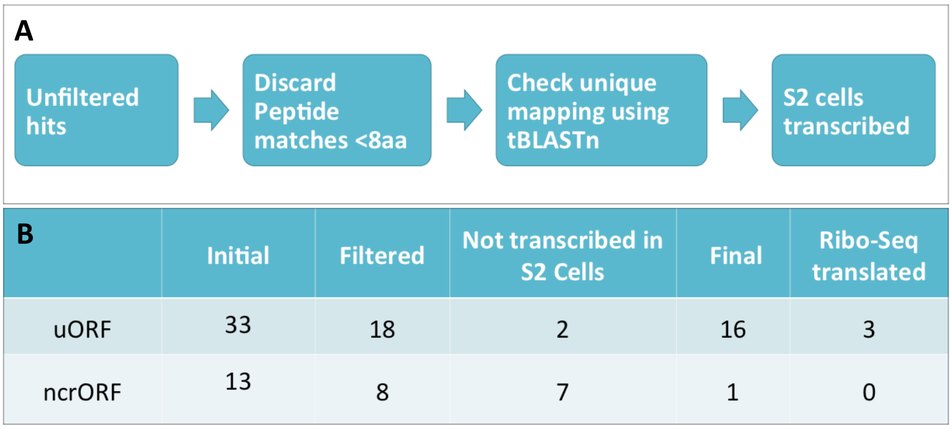


Figure 3.12 Overlap of Mass Spectrometry data with *S2* cell Poly-Ribo-Seq results
A) Venn diagram showing overlap of in-house peptidomics study with known translated SEPs. Our Mass Spectrometry experiment showed translation of 40 new SEPs for which there was previously no evidence of a translated peptide. Almost all (59/60) of the high confidence MS hits were deemed to be translated according to our Poly-Ribo-Seq dataset. This experiment corroborated the translation of 39 new SEPs predicted to be translated by Poly-Ribo-Seq. 20 of the smORFs found in our study had previous evidence of peptide detection in the Peptide Atlas database. This shows that the enrichment process for small peptides can increase the pool catalogued smORFs.

B) Venn diagram showing the median *RPKM* of the 90 smORF encoded peptides with previous peptidomic evidence compared with the remaining 138 of 228 Poly-Ribo-Seq smORFs. The *RPKM* of previously detected SEPs is much higher (*RPKM*=170) compared to those that have not been detected before (*RPKM*=48), highlighting the need for enrichment in the peptidomics study of small peptides.

Metrics	Unfiltered Hits	High Confidence
Number of smORFs	75	60
Candidates from Poly-Ribo-Seq	72	59
Hits in Peptide Atlas	48	33
Median AA Length	84	85
Conserved in Animals	46	36
Median CDS RPKM	224.3	358.1

Table 3.3 Comparison of filtered vs. unfiltered Mass Spectrometry hits from *S2* cells
This table highlights the difference between the unfiltered Mass Spectrometry hits and the ones that we deem as high confidence hits. 15 SEPs with previous Peptide Atlas evidence were discarded. The main reason for this is lack of abundance, as there if no significant difference in size (median aa length) or conservation. The *RPKM* of of the high confidence hits is higher than that of the unfiltered hits, again highlighting (as from Figure 3.12B) that the more abundant transcripts tend to be the ones that are detected and retained in Mass Spectrometry studies.


A
Unfiltered hits
Discard Peptide matches <8aa
Check unique mapping using tBLASTn
S2 cells transcribed
B

	Initial	Filtered	Not transcribed in S2 Cells	Final	Ribo-Seq translated
uORF	33	18	2	16	3
ncrORF	13	8	7	1	0

Figure 3.14 Custom Database Matching of Mass Spectrometry hits from *S2* cells

Our lab designed a custom database consisting of putative 14,881 uORFs and 6,438 ncrORFs that do not have any previous peptidomic evidence (Aspden et al 2014). As these hits could not be scored with confidence using MASCOT, these were manually filtered. Any hits with only one unique peptide match smaller than 8aa were discarded and the remaining hits were put through tBLASTn to show that they specifically map to the correct smORF genomic coordinates. Lastly, these were checked to see whether they are transcribed in *S2* cells. Using this manual curation of custom database hits, from an initial pool of 33 uORFs and 13 ncrORFs, we were left with a final pool of 16 uORFs and 1 ncrORF with peptidomics evidence in our mass spectrometry study. Of these, only 3 uORFs passed the cut-offs for translation according to the *S2* cell Poly-Ribo-Seq data.

Chapter 4: Optimization of Poly-Ribo-Seq in *Drosophila* Embryos

Introduction

Ribosome Profiling Overview

Ribosome profiling entails the sequencing of Ribosome Bound Fragments (RBFs) from mRNAs, in order to map translated sequences within the transcriptome. In a typical Ribo-Seq protocol (Figure 4.1), elongating mRNA-associated ribosomes are trapped on the mRNA by treating live cells with a translation inhibitor. The sample is then lysed, followed by nuclease treatment to digest all mRNA that is not protected by a ribosome. The digested mRNAs and cellular debris is then removed from the nuclease-treated lysate using either a column or a sucrose cushion. The RBFs are then purified from single ribosome-mRNA complexes by size selection (26-36nt) on a gel prior to cDNA library preparation, for deep sequencing and bioinformatic mapping of the sequenced product to the transcriptome. (Ingolia *et al.* 2009; Ingolia *et al.* 2012). Ribo-Seq offers a direct read-out of ribosome occupancy on mRNA at the single nucleotide level, and provides quantitative metrics directly related to the translation rate. This is far more precise than the previous technique of polysomal profiling, which simply detected which mRNAs are present in polysomes (Johannes *et al.* 1999; Arava *et al.* 2003; Qin *et al.* 2007); due to translational regulation, mRNAs may not be translated in direct proportion to their presence in polysomes, as with regulatory non-coding RNAs that associate with polysomes without being translated (Wilson, and Masel 2011; van Heesch *et al.* 2014).

As the current coding status of smORF transcripts versus that of lncRNA remains under question, Ribo-Seq can be used as a starting point to the identification of protein-coding smORFs at a genome-wide level. Though it may require the use of additional independent techniques, such as those described in Chapter III, to verify the individual characteristics of smORF encoded peptides, Ribo-Seq can help to elucidate the translational nature of smORF transcripts, identify potential unannotated START codons, and narrow down the list of biologically interesting smORFs for further study.

The Ribo-Seq protocol is challenging in the requirement of a large amount of starting material for library preparation, and the purification of ribosome bound fragments. Two sources of biological background can be identified: protection of RNA digestion by binding by proteins not constituting a ribosome (RNA-binding proteins, ribosomal subunits) or by ribosomes not engaged in translation, (either scanning, assembling, involved in nonsense-mediated decay, or simply stalled by translational regulation). The most prevalent source of background is provided by rRNA and tRNAs, which are isolated with the mRNA-ribosome complexes and can constitute up to 90% of the reads (Ingolia *et al.* 2009; Ingolia *et al.* 2012; Gerashchenko, and Gladyshev 2014). Combinations of experimental variations and bioinformatic approaches have been devised to deal with Ribo-Seq noise. Although these contaminating sequences can be depleted, they greatly reduce the sequencing efficiency as usually more than 85% of the reads are discarded before mapping to the whole transcriptome. Experimentally, after RNA digestion, Ribosome-mRNA units can be separated from other RNA-binding protein complexes using Sucrose Cushion purification (Ingolia *et al.* 2009; Ingolia *et al.* 2012). However, this purification still retains 80S monosomes that may not be engaged in productive translation, therefore Poly-Ribo-Seq has been developed to address this issue by selecting for actively translating ribosome-mRNA complexes prior to sample preparation.

Using Poly-Ribo-Seq to assess the translation of smORFs

Our laboratory developed a modification of Ribo-Seq called Poly-Ribo-Seq to tackle the problem of specifically assessing the translation of smORFs (Aspden *et al.* 2014). Poly-Ribo-Seq combines Polysome profiling and Ribo-Seq (Figure 4.1) and relies on the fractionation of mRNAs based on the number of ribosomes binding to the transcript, thus allowing the separation of mRNAs that are undergoing translation into small (2-6 Ribosomes) and large polysomes (6+ Ribosomes). The rationale behind this approach is that translating ribosomes are densely packed on to mRNAs at distances of 80-100 nucleotides in both *S.cerevisiae* and *Drosophila* polysome-mRNA complexes (Arava *et al.* 2003). What this means for smORF sequences, which are usually 1kb long with a maximum ORF length of 303 nucleotides, is that they would probably be unable to accommodate more than 3-4 ribosomes in the ORF and would therefore be enriched in the small polysome fraction. Though there may be exceptions to this hypothesis, this observation has been verified by RT-PCR for a small number of representative smORFs (Aspden *et al.* 2014).

Another benefit to adding the Polysomal fractionation step was to address the criticism made by other groups regarding published Ribo-Seq studies of whether long ncRNAs are actually translated or whether the signal is from biological artefacts. These studies explained Ribo-Seq signal within lncRNA transcripts as a result of the association of scanning-40S ribosomal subunits or the assembly of 80S non-translating ribosome complex at putative start codons or nonsense mediated decay target mRNAs (Guttman *et al.* 2013; Chew *et al.* 2013). As it would be unusual in these scenarios for more than one scanning 40S Ribosome, or non-productive 80S ribosome to be loaded on an otherwise untranslated mRNA, we chose to collect fractions of >2 ribosomes as the lower cut-off. The only disadvantage of this approach is that choosing to collect only those transcripts bound by 2 or more ribosomes may exclude the analysis of very short ORFs, such as those found in *tarsal-less* or *Sarcolamban*, which would theoretically not be able to accommodate more than one ribosome or for whom the elongation time would be shorter than that required for assembly of the translation initiation complex. However, the loss of these short ORFs and low abundance proteins is offset by the increased stringency of our protocol, as by excluding potential false positives from non-productive ribosome association and nonsense-mediated decay, we have increased confidence that the translation of the retained sample is genuine (Heyer and Moore 2016).

It is still necessary to account for a certain degree of background signal, and therefore detection of a single RBF read could also not be convincingly equated with translation. This background signal may arise from the detection of signal when there is none, as would be the case with incorrect mapping of reads or from the accidental over-amplification of specific sequences (Aird *et al.* 2011). There may also be biological background arising from RBF reads in the absence of meaningful translation such as Reads generated by non-translating ribosomes that are scanning, assembling or involved in nonsense-mediated decay or from protection of mRNA sequences by RNA-binding proteins against RNase activity (Chew *et al.* 2013; Guttman *et al.* 2013; Smith *et al.* 2014). In order to overcome these uncertainties, and to define sequences as translated with high confidence; Aspden *et al.* (2014), used the 3'UTR as a measure of background signal to define a cut-off for translation. The amount of signal for each ORF is quantified in the number of RBF reads mapping to the ORF sequence and outputted as a metric called *RPKM* (Reads per Kilobase per Million Reads) that normalises the total number of reads mapping to an ORF to the overall length of the ORF (per Kb) as well as to the total number of reads in the library (per Million Reads) (Mortazavi *et al.* 2008). The 90th percentile of the *RPKM* in the 3'UTRs (untranslated background) of annotated protein-coding transcripts was used to define an abundance threshold for translated ORFs, which was a minimum *RPKM* value of 11.7. In order to further reduce False positives arising from single, initiating or stalled ribosomes, we used a second metric based on the proportion of the ORF that is covered by RBF reads. This metric was simply termed Coverage and a minimum coverage value of 0.57 was chosen as a translation cut-off in addition to *RPKM*. This number would ensure that a 60nt smORF is covered by a 34nt read in at least two unique positions (34/60 = 0.57). Finally, in order to ensure that extremely small ORFs may not have artificially inflated *RPKM*s due length normalisation, we introduced an additional filter of a minimum of 5 Reads mapping to a translated ORF. The key findings from Aspden *et al.* (2014) were that there are 228 annotated Flybase smORFs (FB smORFs) undergoing active translation in *S2* cells. These are at a similar proportion to standard protein-coding genes (~80% of those transcribed). The median length of these 228 smORF peptides is 80 amino acids. In addition to these, there is translation of 313 ORFs in long ncRNAs (ncrORFs) and 2,708 ORFs in 5'UTRs (uORFs). These novel ORFs were termed 'Dwarf' smORFs due to their smaller size (median length 20aa). These ORFs were generally translated at a much lower proportion (~30%).

Since Poly-Ribo-Seq requires even larger amounts of starting material than Ribo-Seq due to polysome fractionation, the *Drosophila S2* cell line was an excellent tool for the development of the technique as large amounts of cells (1 x 10^8 cells) can be grown easily in tissue culture. As far as the characterisation of smORFs however, *S2* cells are far from ideal as only 60% of canonical genes (FBcds) and only a third of annotated FB smORFs and uORFs are transcribed in this cell line. The number drops even lower for the very interesting and highly novel ncrORFs as only 13% of these are transcribed in *S2* cells (Table 4.1). The low number of ncrORFs is not completely surprising as lncRNA's tend to have highly tissue and stage specific expression (Guttman, and Rinn

2012; Washietl *et al.* 2014). Therefore a cell line derived from just one type of epithelial tissue from late stage *Drosophila* melanogaster embryos (Schneider 1972) may not be the ideal system to obtain a comprehensive picture of smORF translation. Therefore, after obtaining proof of principle in *S2* cells, our laboratory decided that the next step would be to use Poly-Ribo-Seq in *Drosophila* embryos in order to extend the catalogue of translated smORFs and elucidate their translation across all the stages of *Drosophila* embryonic development.

Overview of *Drosophila* Embryogenesis and Regulation of Translation

Drosophila embryogenesis is highly complex process that is completed in a time span of just 24 hours. After this time a larva is hatched, which is complete with morphological and sensory structures required for the larval stages of growth and development. Translation regulation is a key aspect of early embryonic development in all animals and has been studied extensively in *Drosophila (Zalokar 1976)*. In particular, the first two hours after egg laying (AEL) there is absence of transcription from the zygotic genome and the key developmental processes such as establishment of the primary Antero-Posterior and Dorso-Ventral axes, are controlled purely through the translational regulation of maternal mRNA (Gilbert 1997).

Extensive studies using both Polysome Profiling and Ribo-Seq have been performed at the maternal to zygotic developmental transition (Qin *et al.* 2007; Dunn *et al.* 2013; Kronja *et al.* 2014). These studies have mainly focused on the detection of maternal and early zygotic mRNAs, and therefore do not capture the complex translational events occurring throughout embryogenesis and certainly do not assess smORF translation. For the revised Poly-Ribo-Seq experiments, we decided to sample the whole of embryogenesis for investigation. This was achieved by dividing the stages of embryogenesis into three 8-Hour windows. These were defined as Early (0-8H AEL), Mid (8-16 H AEL) and Late (16-24H AEL) Embryogenesis.

The key developmental processes occurring at these stages of embryogenesis are outlined in Figure 4.2 (Adapted from Atlas of *Drosophila* development (Hartenstein 1993)). Early embryogenesis (0-8H AEL) is characterised by the maternal to zygotic transition, followed by gastrulation, germ band elongation and the determination of the CNS. The next 8 hours (mid-embryogenesis) of development covers the process of segmentation and organ formation, including the differentiation of primordial tissues for adult organs known as imaginal discs. Additionally, gonad formation and the differentiation of neuronal cells also occur during this time. Late embryogenesis (16-24H AEL) is characterised by growth and development of musculature and movement and the fine-tuning of larval sensory and morphological structures (Hartenstein 1993).

Our own analysis of the Ribo-Seq data from Dunn *et al.* (2013) shows that almost all (93%) the smORFs translated in 0-2H embryos only represent a small fraction of transcribed genes of *Drosophila* embryogenesis. In order to gain an estimate of the number of FB smORFs transcribed throughout *Drosophila* embryogenesis, I initially analysed the publicly available modENCODE RNA-Seq data, which is generated using twelve 2-hr windows across embryonic development (Graveley *et al.* 2011). In order to compare this to to our experimental design, this data was pooled into 3 windows of 8 hours each. Figure 4.3 shows the number of FB smORF transcribed in each pooled dataset (*RPKM*>1), with numbers in parenthesis showing smORFs translated in the embryonic *S2* cell line Poly-Ribo-Seq data. From this we can observe that there are a total of 483 unique smORFs transcribed during embryogenesis, of which 211 smORFs transcribed throughout all three stages. 72% (152) of these are also translated in *S2* cells suggesting these smORF peptides may be performing more basal cellular functions required throughout embryogenesis in most cells. About 200 smORFs are transcribed exclusively in mid and late stage embryos and these have previously not been assessed by Poly-Ribo-Seq. These represent genes that may function on a more complex, multicellular level, perhaps in a tissue specific manner as organs become differentiated during the later stages of embryogenesis.

Preparation of Library for Next Generation Sequencing

To prepare RNA for NGS, the RNA must first be converted to cDNA. The sequences required for NGS primer binding are added during a process known as library preparation. In the case of RNA-Seq, intact mRNA transcripts can be converted to cDNA using oligo(dT) primers that bind to 3’ PolyA sequences or random hexamer primers and then processed for library preparation (Wang *et al.* 2009). Small RNAs such as micro-RNAs and Ribosomal Footprints lack a PolyA tail and cannot be reverse transcribed by random primers due to their small size thus leading to the requirement of adapter sequences at the 3’ and 5’ ends of the small RNA fragments. The 3’ DNA adapter, which contains a reverse transcription (RT) primer hybridisation sequence, is ligated to RNA fragment using T4 RNA ligase 2. The 5' end of the RBF is ligated to a RNA adapter using T4 RNA ligase, followed by reverse transcription into cDNA (Figure 4.4B). Library preparation requires special care, as almost each step in the process has been shown to introduce some form of bias in different sequencing datasets (reviewed in Raabe *et al.* (2014)). It has been shown that RNA ligases can introduce enzyme-based bias to the sequencing library as certain sequences and structures can be preferentially ligated, leading to over-representation in the sample (Ingolia *et al.* 2009; Zhuang *et al.* 2012). Ingolia *et al.* (2012) utilised an alternative method to dual adapter library preparation in order to minimize RNA Ligase bias by eliminating the 5' adapter ligation step. This is achieved by using a 93 nucleotide long RT Primer that incorporates the 5' adapter sequence separated by a C_{18} spacer. This method allows the single stranded cDNA to be circularised using CircLigase II (Figure 4.4A), which has been shown to reduce bias in small RNA library preparation (Ingolia *et al.* 2009; Jackson *et al.* 2014). However, in order to follow this protocol, larger amounts of starting material is required due to the multiple purification steps in the protocol and is therefore not ideal for our purpose. One way to increase the amount of library is to increase the number of cycles in the final amplification step; however, this is not generally recommended as PCR amplification can introduce its own bias due to the higher melting temperature of GC rich sequences (Aird *et al.* 2011). Furthermore over-amplification leads to many copies of the same sequence but does not improve coverage of the transcriptome; therefore I decided to test both library preparation techniques; the previously used ssDNA circularization approach developed by Ingolia *et al.* (2012) and the dual adapter ligation approach used in the NEB kit.

Chapter Aims

The aim of this series of experiments was to adapt the *S2* cell Poly-Ribo-Seq protocol (Aspden *et al.* 2014) to *Drosophila* embryos. The key challenges highlighted by the *S2* cell data generated by our lab were the requirement of large amounts of material and having enough sequencing depth to ascertain the translation of novel ORFs and specifically smORFs. I addressed the issue of the large amount input material by establishing a successful laying and harvesting protocol and optimising the different library preparation techniques. The number of usable reads was increased by improved depletion of contaminating sequences thus improving sequencing efficiency.

Results

Adapting the Poly-Ribo-Seq technique to Embryos

Despite the many advantages offered by Ribo-Seq, there are limitations that must be considered before implementation of the technique in *Drosophila* embryos. The *S2* cell protocol (Aspden *et al.* 2014) was based on the treatment of cells with Cyclohexamide (CHX) whilst the cells were 'in-culture', followed by cell lysis and small polysome fractionation. These complexes were then diluted to 10% sucrose and digested overnight at 4°C using RNase I before size selection on a polyacrylamide gel. These RBFs were then prepared for NGS using the protocol published by Ingolia *et al.* (2012) including rRNA depletion, adapter ligation conversion to cDNA and circularization. The following sections will describe the various stages of sample preparation that were considered when adapting this protocol to *Drososphila* embryos including the harvesting of tissues and their subsequent lysis, nuclease digestion (to generate RPFs), library preparation and the depletion of contaminating sequences.

Small perturbations that can occur while harvesting and lysing material for Ribo-Seq can have a large effect on the data generated from this technique. Therefore, the first step to consider was the optimisation of the egg laying and harvesting protocol for *Drosophila* embryos. This step is very important, as embryos are more difficult to collect and lyse than *S2* cells which are grown and treated in tissue culture plates under sterile conditions. In order to get a sufficient amount of embryos for the Poly-Ribo-Seq protocol, Oregon-Red (Or-R) genotype flies, which are commonly used as a wild-type fly line, were expanded in 250 ml polypropylene bottles containing ~50 ml of standard cornmeal fly food. 50 adult flies were added to each bottle and left to lay eggs for a 48-hour period before the adults were transferred out of the bottle. The bottles were then kept at 25°C for about 2 weeks to allow the freshly laid eggs to develop into adults (~10 days). The adults were discarded if they were more than 5 days old as older females have a tendency to retain fertilised eggs, which may affect the subsequent staging of the embryos (Broadie *et al.* 1992). Between 12-16 bottles, each containing approximately 300-500 adult flies, were emptied into a large (50cm x 30cm x 50cm) perspex population cage. The cage is built with a gauze window to allow changing of the plates upon which embryos are laid (Figure 4.5A). The flies were then left to acclimatise to the new environment whilst being fed on a yeast-rich diet to stimulate egg laying. This encourages the fertilised females to lay more eggs, as females on a nutrient limited diet can retain fertilised eggs (Broadie *et al.* 1992). The cage was subsequently maintained at 25°C with four 10cm petri dishes containing molasses fly food (Figure 4.5B). These petri dishes were used for embryo collection and were changed at 9am, 5pm and 1am to maintain 8-hour collection windows. Molasses fly food was preferred to the regular apple juice agar laying medium, as it was easier to remove embryos from the molasses food petri dishes and we did not see a significant difference in the amount of embryos layed. The embryos were either harvested straight away for the 0-8H AEL (early) samples or aged 8 hours in the 25°C incubator for the 8-16H AEL (mid) or 16 hours for the 16-24 H AEL (late) samples, before harvesting.

Optimisation of Embryo Harvesting and Lysis

Ribo-Seq is a technique that captures a snapshot of the rate of protein production in a sample by trapping translating ribosomes on mRNA. Therefore, it is particularly sensitive to processes that may affect the stability of mRNA-Ribosome complexes, requiring care in the regulation of temperature and stress during tissue harvesting and lysis. A translation inhibitor such as Cyclohexamide (CHX) traps elongating Ribosomes and is used to stabilise the polysomes during the lysis of samples such as yeast (Ingolia *et al.* 2009), *S2* cells (Aspden *et al.* 2014) and in this case *Drosophila* embryos. Such drug treatments can cause a pile-up of Ribosomes at the beginning of translating sequences (Gerashchenko, and Gladyshev 2014) but the benefits of using CHX outweighs the potential cost, especially when using large amounts of material as more time is required to process the samples.

Drosophila embryos are covered in tough, waterproof ‘eggshell’ layer called the chorion. For most scientific purposes, the chorion is normally removed by soaking embryos in bleach, which helps to dissolve the chorion and allow penetration of fixatives, buffers and antibodies into the embryo. The chorion presents a challenge with regards to treatment with chemical reagents since it must be removed prior to lysis of the internal tissues, which is a requirement for conducting Poly-Ribo-Seq.

While carrying out previous work in the lab, polysome gradient analysis was performed on overnight (0-16 H AEL) small-scale collections of embryos dechorionated using bleach with good results (Figure 4.6A). Embryos were harvested from the plate using room temperature water and a paintbrush into wire mesh baskets and dechorionated using 50% bleach and washed in with room temperature water. When scaling up the experiment for Poly-Ribo-Seq, which requires significantly more starting material, embryos were to be collected from multiple, large, embryo collection plates. This raised the concern that the time taken to harvest and dechorionate a large amount of carefully staged embryos may have an affect on translation as well as staging of embryonic age. An alternative method based on the cryo-lysis of flash frozen tissue has been shown to be a robust technique for isolation of polysome complexes (del Prete *et al.* 2007) and has been adapted for Ribo-Seq in *Drosophila* embryos by Dunn *et al.* (2013) using whole embryos. Therefore, I tested both lysis methods, with and without chemical dechorionation in the development of the protocol.

To test the various harvesting and lysis techniques, the dechorionated embryos were incubated for 15 minutes in 1xPBS with Cyclohexamide (CHX) and then transferred to lysis buffer in a glass homogeniser. This approach worked well for small-scale experiments (150 mg Embryos) using non-staged, fresh embryos. For the larger amount of material required for Poly-Ribo-Seq, multiple collections of embryos were flash-frozen and pooled together for use on a later day. It can be seen that this approach of dechorionation and fixing with CHX in PBS and lysed with glass homogeniser did not work particularly well with previously frozen embryos as the freeze-thawing process led to a collapse of the Polysomes (Figure 4.6B).

There was then a choice of two previously published methods for processing tissue for polysome analysis, the first method is based on using lysate prepared from fresh, bleach dechorionated embryos that is aliquoted and flash-frozen for later use (Clark *et al.* 2000; Kronja *et al.* 2014). The second method is based on the cryo-lysis of flash-frozen whole tissue (in this case, embryos with a chorion intact) and subsequently thawing in the presence of CHX containing lysis buffer (del Prete *et al.* 2007; Dunn *et al.* 2013). I decided to employ the latter method, as the embryo laying and staging protocol required the harvesting of samples at difficult times of the day for the sane researcher. As the volume of lysate that can be loaded onto a sucrose gradient run has to remain constant (6 x 0.5ml), the ratio of embryo mass to lysis buffer volume is increased during scale up of the experiment for Poly-Ribo-Seq. The results of the freeze-thaw method also may have been influenced by use of a glass homogeniser as it could be seen that there was poor mechanical disruption of embryos and thus inferior extraction of polysomes (Figure 4.6B). The cryo-lysis approach uses flash-frozen embryos which are ground to fine powder using mortar and pestle prior to addition of lysis buffer and is better suited for larger amounts of material (~1 gram). Using this approach, superior recovery of polysomes was obtained, compared to the previous protocols (Figure 4.6C). The only perceived disadvantage of using this approach is that the outer layer of the embryos is no longer removed or washed prior to lysis, which can increase the chances of contamination from yeast and food debris from the embryo collection plate. This step is omitted from the published protocol due to the fact that the whole point of the cryo-lysis approach is to harvest and freeze the large-scale embryo collections with minimum processing. In order to overcome this, the amount of yeast added to the plates was minimised and only plates with a high density of embryos were harvested in order to minimise carry-over of yeast and food debris (Figure 4.5B)

To summarise this method, a rubber cell scraper was used to remove embryos from the outer edge of the plates and these were immediately flash-frozen by dipping the cell scraper into a small container of liquid nitrogen. Harvested embryos were then stored in 2ml pre-chilled tubes and stored at -80°C until further use. Frozen embryos from two separate collections were pooled and ground using a pestle and mortar pre-chilled with liquid nitrogen. Lysis buffer containing CHX was slowly dripped into the liquid nitrogen-embryo powder slurry at volumes of 1.5ml at a time and then further ground. The lysate was then transferred to a 15ml pre-chilled falcon tube.

In order to complete cellular lysis after tissue homogenisation from cryo-lysis, the sample was thawed under running water followed by a 20 minute incubation at 4°C with rotation after which the lysate was spun down using a table-top centrifuge. Due to the nature of the sample, there was a large amount of cell debris at the bottom of the tube and a thick waxy layer of chorion floating at the top of the lysate. Care was taken to avoid transfer of these layers to a second pre-chilled falcon tube, and the sample was then divided into 7 aliquots of 500µL (6 sucrose gradients plus one for mRNA control). The lysate was clarified using high-speed centrifugation (14,000*g*) for 10 minutes at 4°C.

Polysome Fractionation

Polysome fractionation was conducted in a manner similar to that in Aspden *et al.* (2014). Initially, these experiments were started with the aim of isolating the small-polysomes fraction, which consists of 2-6 ribosomes and all attached mRNAs, with the hope that this process would enrich the sample with smORF and other small peptide encoding mRNA. The relative volume of each sucrose gradient was optimised for this by adding a larger volume of the 42% to 50% sucrose fractions, which coincides with the distance travelled by 2-6 Ribosome fractions (Figure 4.7A). However, obtaining sufficient material for the protocol was a major concern at the time so we decided to re-evaluate our decision to use only the small polysomes. The reason for this change was two-fold; retaining only the mRNA with 2-6 ribosomes would only give results from a fraction of translating ribosomes and since this data would span the whole of *Drosophila* embryogenesis and is the first such experiment of its kind, we wanted to retain as much information as possible. By including large polysomes in our library, it would help to obtain a complete picture of translation during embryogenesis. Finally, in terms of smORF discovery using small polysome enrichment, the number of additional FB smORFs (12%) detected by this method in *S2* cells data did not offset the decision to exclude the remaining polysomes as it came at the cost of translational information on canonical genes (Aspden *et al.* 2014). Furthermore, at the time we did not appreciate the fact that polycistronic ncrORFs and uORFs of longer annotated ORFs may have more than 6 ribosomes attached to the transcript.

In my experiments, the polysome fractionation step was still used to select for 2+ Ribosomes, as this biochemical purification of actively translating ribosomes increases our confidence that the observed footprints are a result of active translation particularly in the context of corroborating novel smORF translation. From the previous step of harvesting and lysis of the sample, 450µl of embryonic lysate from each aliquot was carefully loaded onto the top of an 18-60% sucrose gradient. Gradients were made using a step-by-step freezing protocol, with lighter sucrose fractions being layered over frozen sections of the heavier sucrose. The tubes were then allowed to defrost overnight at 4°C to form a natural gradient. The relative volume of the different concentration sucrose fractions was adjusted to improve resolution and achieve maximum separation between the 80S and 2 ribosome peaks. This was accomplished by adding an intermediate layer of 38% (w/v) sucrose solution between the 34% and 42% sucrose fraction as well as by increasing the length of the ultra-centrifugation step to 4 hours (Figure 4.7B) in order to achieve better separation between the 80S subunit peak and the 2+ ribosomes fraction.

Generation of Ribosome Footprints

The techniques used in the generation of Ribosome footprints by nuclease digestion have been used in early studies investigating translation mechanics of individual mRNAs using *in vitro* translation (Steitz 1969; Wolin, and Walter 1988). These studies pre-date the advent of the high-throughput assessments for which they are now employed such as RNA-Seq and microarrays but reveal important implications for assessment of studies such as ours. One of the key observations from these original ribosome footprinting experiments is that ribosome footprints are a characteristic size of around 30 nucleotides (Steitz 1969). This observation is at the heart of the Ribo-Seq technique as Ribosome-mRNA complexes are subjected to digestion using a non-specific nuclease to digest unprotected RNA and using a sucrose cushion or column, the intact, single-ribosome-mRNA complexes are purified based on size. Empirical adjustment of the digestion step is necessary for each experiment that is conducted, as too much nuclease can lead to over-digestion and non-specific degradation of rRNA. On the other hand, not adding enough nuclease means that not all the complexes are digested to single ribosomes, resulting in larger RBFs that are bound by 2 or more ribosomes. This leads to potentially useful mRNA being excluded from the sample during size selection as they fall outside the 30nt range. Sample size selection happens at two levels, first biochemically by resolving the purified RNA on a polyacrylamide gel and cutting out the relevant region (~30nt) and also bioinformatically at the data analysis stage.

For the purpose of Poly-Ribo-Seq, intact polysomes are fractionated on a sucrose gradient from the cytoplasmic lysate and then diluted prior to nuclease digestion and size selection. The sucrose concentration of the pooled polysomes is been estimated from the gradient fractionation trace (~ 44%) and the sample is diluted to 10% sucrose using Polysome dilution buffer, because high levels of sucrose can inhibit the activity of the nuclease. After dilution, Nuclease is added to the solution and incubated before the digestion reaction is stopped using RNase inhibitor (RNase I) or EGTA (MNase). The digested polysomes are precipitated using isopropanol and the RBFs are resuspended and treated with DNAse I before phenol-chloroform purification and size selection on a denaturing polyacrylamide gel to cut out the region corresponding to RBFs (28-34nt).

Choice of Nuclease

For most Ribo-Seq studies published at the time of writing this thesis, RNase I Nuclease has almost exclusively been used due to its lack of bias and precision in generating RBFs (Ingolia *et al.* 2009). We have previously used RNase I for Poly-Ribo-Seq in *S2* cells (Aspden *et al.* 2014), however, during the development of this protocol, Dunn *et al.* (2013) reported that RNAse I treatment leads to degradation of 80S ribosomal RNA complexes in *Drosophila*, instead they recommend using Micrococcal nuclease (MNase). Even though our lab has successfully generated Poly-Ribo-Seq data from RBFs using RNase I in *S2* cells (Aspden *et al.* 2014), it would be pertinent to try both nucleases to see if there is a significant difference between the digestion conditions and results of the two protocols.

Dunn *et al.* (2013) performed nuclease digestion with MNase in a small (<1mL) volume of cytoplasmic lysate at 25°C for 45 minutes, while Poly-Ribo-Seq RNase I digestion is performed overnight at 4°C on polysome fractions in a 10% sucrose solution with an approximately 100 fold greater reaction volume. 0-8 hour Embryos were treated with MNase and the results were visualised on a denaturing 15% polyacrylamide gel. Dunn *et al.* (2013) use up to 5U of MNase per microgram of RNA. I quantified the RNA in the polysome fraction from one centrifuge tube by precipitating the solution and using a nanodrop to estimate an approximate value of 19-28 micrograms of polysomal RNA. Therefore, as a starting point, I tested 100 U of MNase per tube but the sample did not appear to be sufficiently digested as could be seen by a ubiquitous smear on the gel. Subsequently, higher concentrations of MNase were used and Figure 4.8 shows the results. I tested 250 and 500 Units of MNase per tube at 25°C for 40 minutes and also overnight at 4°C, in addition to testing an excess amount (1000U) at 25°C. The digested RNA from all the different reaction conditions appeared as a smear when run on a denaturing 15% polyacrylamide gel. These results suggested that the MNase enzyme might have insufficient activity in the reaction conditions tested, which may have occurred because MNase activity is not as robust as RNase I outside its optimum salt, protein and pH conditions. RNase I digestion produced a much cleaner gel profile with a distinct 28 to 34nt band as discussed in the next section.

Therefore, rather than trying to troubleshoot digestion conditions for MNase; I decided to focus on optimisation using RNase I at this stage, which had been previously used in preparation of Ribo-Seq libraries in our lab (Aspden *et al.* 2014).

Digestion conditions using RNase I

Dunn *et al.* (2013) report that *Drosophila* ribosomes are degraded by RNase I using digestion at 25°C. In order to determine the amount of RNase to add for ribosome footprinting, I tested different amounts of RNase I per microgram of RNA under a variety of digestion conditions. The results of my preliminary testing in which 500 and 1000 units of RNase I were both tested at 4°C overnight and 25°C for one hour suggest that RNase I degradation is significantly reduced at lower incubation temperatures. Therefore I decided to use 1000U of RNase I for nuclease treatment, which is at a concentration of 35-55 U per μg of RNA and similar to the concentration recommended in the protocol published by Ingolia *et al.* (2012). Digests were performed for one hour at 25°C or overnight at 4°C and 16°C. I prepared a positive control by 'overdigesting' lysate sample with incubation of 1500U RNase I at 4 °C as well as preparing an undigested sample as a negative control. Figure 4.9 shows that across all the four digestion conditions tested, there is a fairly consistent increase in signal at the 28-34 nucleotide marked in the gel, especially compared to MNase digestion. We can also observe the presence of abundant, larger bands near the tops of the lanes in both the undigested control and RNase I samples. These large bands most likely represent the larger 18S (2Kb) and 28S (5Kb) rRNA transcripts, which begin to smear with increasing amounts of RNase I (1500U). Furthermore this smearing is greater at 25°C than at 4°C using 1000U RNase, thus corroborating the hypothesis that RNase I mediated degradation of *Drosophila* ribosomes is probably a temperature dependent phenomenon. Interestingly, we can also observe an ~ 60nt band on the gel, which is brighter at 4°C than at 25°C using 1000U of RNase I. This band has been observed in a number of other Ribo-Seq experiments and speculated to be generated by steric hindrance of the nuclease by adjacent tightly packed ribosomes (Disomes) (Dunn *et al.* 2013; Guydosh, and Green 2014). Overall, these results show that we can obtain sufficient mRNA digestion using RNase I and that RNase I mediated ribosome degradation appears to be minimised at lower incubation temperatures.

Library Preparation

Preparation of Library for Next Generation Sequencing

In order to prepare RNA for NGS, the RNA must first be converted to cDNA and combined with sequences required for primer binding during a process known as library preparation. This process required optimisation before any conclusive analysis of results could be conducted, as it has been documented that almost each step in the library preparation process can introduce bias in sequencing datasets (reviewed in Raabe *et al.* (2014)). In the case of traditional RNA-Seq, intact mRNA transcripts can be converted to cDNA using oligo-dT or random hexamer primers and then processed for library preparation (Wang *et al.* 2009). When attempting to sequence small RNA's such as micro RNAs and Ribosomal Footprints, which cannot be reverse-transcribed and also lack a PolyA tail, the process requires the addition of adapter sequences at the 3' and 5' ends of the RNA. The 3' DNA adapter contains a reverse transcription (RT) primer hybridisation sequence and is ligated to the RNA fragments using T4 RNA ligase 2. This step is followed by ligation of an RNA adapter to the 5' end of the RBF using T4 RNA ligase I. After these modifications, the sequence can then be subjected to reverse transcription into cDNA (refer to Figure 4.4B for overview).

It has been shown that RNA ligases can introduce enzyme-based bias to the sequencing library as certain sequences and structures can be preferentially ligated, leading to over-representation of these in the final sequencing sample (Ingolia *et al.* 2009; Zhuang *et al.* 2012). Ingolia *et al.* (2012) utilised an alternative method to minimise RNA ligase bias by eliminating the 5' adapter ligation step and instead using a 93-nucleotide long RT Primer that incorporates the 5' adapter sequence separated by a C_{18} spacer. This allows the single stranded cDNA to be circularized, prior to PCR amplification, using CircLigase II (Figure 4.4A) and has been shown to reduce bias in small RNA library preparation (Ingolia *et al.* 2009; Jackson *et al.* 2014). The Ribo-Seq protocol developed by Ingolia *et al.* (2012) requires a large amount of starting material to compensate for sample losses at various stages of the protocol as it has three intermediate polyacrylamide gel purification steps in the Library preparation stage (Figure 4.4). Adapting this technique for Poly-Ribo-Seq in *S2* cells had previously required significant optimisation even though we started with 100 million cells per experiment. Using this adapted method, my first attempt at preparing a Poly-Ribo-Seq library using 0.8g of Embryos (Em JA sequencing run) resulted in very low yield as I only managed to generate enough material for half a small-scale sequencing run on a MiSeq cartridge (2nM x 5ul). The Cambridge Sequencing facility, which was chosen for the final sequencing of libraries requires at least 20 times more material for sequencing on a NEXTSeq machine (10nM x 20ul). One way to increase library amount is to increase the number of cycles in the final PCR amplification step. However, this sample was generated using 13 PCR cycles in the final amplification step, which is already close to the maximum recommended 15 cycles (Ingolia *et al.* 2012), especially as PCR amplification can introduce its own kind of bias due to higher melting temperature of GC rich sequences (Aird *et al.* 2011). Over-amplification leads to many copies of the same sequence but does not improve coverage of the transcriptome; therefore I had to consider methods for increasing the yield of the protocol.

Testing the NEB Kit

In order increase the sample yield for sequencing embryonic samples; I decided to test the NEBNext Small RNA Library preparation kit (NEB, Cat No. E7300). Though the use of a dual adapter strategy, such as in this kit has previously been discouraged (Ingolia *et al.* 2009) due to the sequence and structure specific bias of the T4 RNA ligase 1 enzyme used in the 5’ adapter ligation step, the protocol has fewer gel purification steps during library preparation, which can significantly affect the yield of the protocol. The key difference between the *Drosophila S2* cells experiment Aspden *et al.* (2014) and in the use of the NEB kit is the use of a dual adapter ligation strategy as opposed to the ssDNA circularization method. We decided to test this kit in our hands, particularly to assess whether or not this bias has any significant impact on the data from the perspective of smORF discovery. Both of the Library preparation methods are outlined in Figure 4.4.

I modified the NEB protocol to include a ribosomal RNA (rRNA) depletion step after 3’ DNA adapter ligation. The rRNA was depleted from the sample using subtractive hybridisation and this step was added to the protocol at the 3’ DNA adapter ligation stage in order to avoid any carryover of the complementary rRNA DNA fragments from the beads into the library. After rRNA depletion, the sample was then precipitated using isopropanol and resuspended in NEB T4 RNA ligase 2 buffer, followed by RT primer hybridisation and 5’ adapter ligation using T4 RNA Ligase 1. 3’ and 5’ adapter dimers with no insert are a common contaminant in libraries prepared using this strategy (Head *et al.* 2014); therefore RT primer hybridisation is performed prior to 5’ ligation. This allows the RT primer to hybridise and form double-stranded DNA with any excess 3’ DNA adaptor which cannot be ligated by T4 RNA ligase I. The library is then converted to cDNA, and a quarter of the reaction is used to optimise the number of PCR cycles (between 8 and 15 cycles) before final DNA amplification of the remaining library. It is considered good practice to minimise the number of PCR cycles to control any effects of PCR bias, which is well documented in NGS experiments (Aird *et al.* 2011). As shown in Figure 4.10, over-amplification using 14 cycles of PCR, leads to the partial re-annealing of library strands through complementarity of the adapter sequences. Since these are only partially re-annealed due to differing insert sequences, these appear as smeary bands that migrate through the gel more slowly than the perfectly annealed dsDNA, which appears as a tight band around 150bp mark. Since this band could be observed using 10 cycles of PCR, the remaining library cDNA was processed with this minimum number of cycles. Primers were used to incorporate the sample barcodes and gel purification of the 150bp band was conducted. The library was then resuspended in 10μl of nuclease free water in an Eppendorf LoBind tube and serial dilutions were prepared for quantification using an Agilent Bioanalyzer with a High Sensitivity chip. By using the NEB kit, I was able to generate approximately ten times more sequencing library (34ηg, referred to as Em NEB) as compared to the 3.5 ηg generated using the protocol adapted from Ingolia *et al.* (Em JA), while also using fewer PCR amplification cycles to minimise PCR induced bias.

Correlation between Em NEB and Em JA and published datasets

In order to see whether the different library preparation methods have any effect on the abundance of Ribo-Seq signal, the reads from each run were filtered for rRNA and tRNA and the remaining reads were aligned to the *Drosophila* transcriptome using the *TopHat* program. The abundance of the aligned transcripts was estimated using *Cufflinks* (Trapnell *et al.* 2012). In order to compare the sequencing data from both 0-8H embryo libraries using the two different library preparation methods, linear regression modelling was used to compare the *RPKM* values of annotated CDS’s between the Em JA and Em NEB datasets (Figure 4.11A). We could observe a good correlation ($R^2 = 0.79$) between the values in each dataset.

In order to assess the impact of using either library preparation method specifically on the translation of smORFs, we used the 90th percentile *RPKM* and coverage values of 3'UTRs of annotated protein coding transcripts to generate translation cut-offs similar to the method used in Aspden *et al.* (2014). These cut-offs were then used to define translated smORFs in the Em JA (2.9 *RPKM*, 0.08 Coverage, cvg) and Em NEB (0.94 *RPKM*, 0.07 cvg) datasets. Compared to the values used in the *S2* cell data (11.7 *RPKM*, 0.57 cvg) these cut-offs are quite low, but they are a reflection of the relatively low sequencing depth of the embryo sequencing trial runs. Compared to the *S2* data which has ~6 million reads mapping to ORFs, these are only only 249,096 reads (Em JA) and 586,674 reads (Em NEB). Although *RPKM* measurements should technically normalise for the number of reads, it may also be that we have not fully captured the level of background (3'UTR) signal present in the samples with such shallow depth of sequencing. Nevertheless, we can observe that of the total 360 smORFs translated using these cut-offs, a significant number (281) appear translated in both datasets (Figure 4.11B).

To further validate and highlight any differences in these datasets, I compared the translation of longer Flybase annotated protein-coding genes (FB CDS) and smORFs (FB smORF) in the two 0-8H libraries, with the 0-2H Embryo data published by Dunn *et al.* (2013) as well as the *S2* cell data from Aspden *et al.* (2014) as shown in Figure 4.12. Em JA has an estimated 15,047 and Em NEB has 18,064 FB CDS translated according to the translation cut-offs described above, of which a significant proportion (81%) are translated in both datasets. Furthermore, 203 (83%) of the 244 FB smORFs translated according to our analysis of the 0-2 H Em Data are translated in both the Em JA and Em NEB datasets (Figure 4.12A). A similar proportion (86%) of annotated protein-coding genes also appear translated in all three datasets. A comparison with the *S2* cell data shows a similar trend with 190/228 FB smORFs (83%) and 11,814/13,800 FB CDS (86%) appearing translated in both the Em JA and Em NEB datasets (Figure 4.12B) These results highlight the robustness of the Poly-Rbio-Seq technique, particularly from the perspective of detecting smORF translation by the good correlation between the two library preparation methods with other previously published datasets.

It is worth noting that overall, there is a higher number of ORFs translated in the Em NEB dataset, and despite having greater depth of sequencing it also has a lower translation cut-off than those calculated for the Em JA data. This cut-off is based on the amount of signal in the 3'UTRs of annotated protein-coding transcripts. Interestingly, there is a significant reduction in the proportion of 3'UTR reads in libraries prepared using the NEB kit as shown in Table 4.2. The Em NEB library prepared using dual adapter ligation has a considerably lower proportion of reads mapping to 3'UTR sequences (4.75%) than the Em JA library. This increased our confidence in the use of the NEB kit for further experiments as the Poly-Ribo-Seq libraries prepared in our lab using the single stranded DNA circularisation protocol adapted from Ingolia *et al.* (2012) have approximately 15% 3'UTR reads in *S2* cells and 8% in the Em JA dataset. Furthermore, using the NEB kit to produce an *S2* cell library showed only 3.2% 3'UTR reads (Personal Communication J L Aspden).

Considering these results, we concluded that the NEB kit library preparation method yields a lower level of background reads. I therefore persisted in the use of the NEB kit for future library preparation. Using the NEB kit gives significantly greater yield (20 fold improvement) and has a shorter library preparation with less sample loss. Secondly, since the translation cut-offs in our current pipeline were based on noise detected in the 3'UTRs, we hoped that cleaner data with a lower proportion of 3'UTR reads would help prove the translation of even more smORFs. Finally, the two different library preparation methods did not differ significantly in terms of smORF discovery in our datasets, so we did not foresee RNA Ligase bias to have a major impact on our final data.

Depletion of rRNA and tRNA: Improving Sequencing efficiency

Although RBFs are selected by size, a significant proportion of specific short ribosomal RNA fragments co-migrate with the RBFs since they are around 30 nucleotides in length (Ingolia *et al.* 2012). Furthermore, only 5% of the total RNA in a cell is estimated to be mRNA, the bulk of it is made up of ribosomal RNA (80%) and transfer RNA (15%) (Molecular Cell Biology, Lodish 2000a). These contaminants pose a significant challenge, particularly in the case of *Drosophila*, which has a unique feature in its 5.8S rRNA; it is made up of two components the 123 nt long 5.8S RNA and the 30 nucleotide long 2S rRNA (Pavlakis *et al.* 1979). Ribosomal RNA contamination is a well-documented problem in ribosome profiling experiments, and it can often make up ~90% of the reads (Ingolia *et al.* 2009; Ingolia *et al.* 2012; Gerashchenko, and Gladyshev 2014). This proportion is not surprising, since for each ~30nt ribosome footprint, several kilobases of rRNA are co-purified. The implication of this for NGS experiments is that unless the mRNA is purified, the majority of the data consists of rRNA reads.

In RNA-seq experiments, mRNAs can be enriched in the sample by using Oligo-dT beads to capture poly-A tailed mRNAs. However small RNA's like mature micro-RNAs and RBFs cannot be purified in this manner, so the alternative in Poly-Ribo-Seq is to deplete rRNA. The rRNA depletion for Ribo-Seq is accomplished using a subtractive hybridisation approach, which entails the use of biotinylated oligonucleotide probes that are complementary to the rRNA. These bind to the rRNA sequences and can be then depleted from the sample using Streptavidin beads. However, using this method, substantial rRNA still remains in Ribo-Seq libraries (Ingolia *et al.* 2012; Dunn *et al.* 2013) and despite using multiple rounds of depletion our lab had previously only achieved at best 70% rRNA reads (Aspden *et al.* 2014). This substantial proportion of useless reads poses a particular challenge especially in terms of defining smORF translation. Due to their small size very short ORFs in lowly transcribed long ncRNAs already struggle to fulfill the requirement of a minimum of 5 reads across the ORF to be classified as translated. Therefore it was a priority to improve on the depletion of rRNA sequences and increase the number of usable reads in the dataset.

Since any rRNA that can not be depleted from the library is sequenced along with the RBFs, rRNA reads have to be bioinformatically removed from the raw data by alignment to rRNA and tRNA sequences before conducting further analysis of Ribo-Seq datasets. This greatly reduces the sequencing efficiency as usually more than 85% of the reads are discarded before mapping to the whole transcriptome. This is exemplified in Table 4.2, which shows that the *S2* cell small polysome data generated for Aspden *et al.* (2014) contained only 20 Million non-rRNA reads despite having 189 Million raw sequencing reads, mainly due to the high level of rRNA contamination. This substantial level (90%) of observed rRNA contamination occurred despite the fact that the library was depleted for rRNA using subtractive hybridisation with rRNA depletion beads. The beads contained sequences that cover the complete rRNA sequences in windows of 500 and 1000 nucleotides (Aspden *et al.* 2014). The rRNA depletion beads were made by using biotinylated reverse primers to amplify sequences complementary to rRNA sequences (See Appendix 1) from cDNA or genomic DNA and the PCR products were gel-purified and bound to streptavidin beads. The non-biotinylated complementary strands were removed by washing with 1M NaOH. Other studies have also used biotinylated oligonucleotides to perform targeted depletion of specific short sequences that are highly over-represented in Ribo-Seq libraries (Ingolia *et al.* 2012; Dunn *et al.* 2013). Our approach differed as in Poly-Ribo-Seq, the contaminating rRNA sequences have a wider distribution over the rRNA transcript so we used the PCR fragment approach to target more regions. Subsequently, the Em JA library contained 87% rRNA after one round of depletion with the rRNA depletion beads. I sought to improve on this in the preparation of the Em NEB library by using two rounds of rRNA depletion, as repeat rounds of depletion had shown a considerably lower (70%) rRNA using this strategy for the *S2* cell all polysome library. Using the NEB kit only reduced the proportion of rRNA reads in the Em NEB data by only a further 2% (to 85%) despite the additional round of depletion (Table 4.2).

Consequently, I decided to investigate whether there is any other way to further reduce the rRNA fragments by analysing the read counts for each gene in the rRNA/tRNA alignments in the Em NEB data. Interestingly, rRNA actually only accounted for 10% of the total removed reads with the remaining 75% of the reads mapping to transfer RNA (tRNA) sequences. Of these tRNA reads, half of the reads map to genes coding for Glutamine tRNA genes and another 10% to Aspartic acid tRNA genes, which are some of the more abundant tRNAs (Karaiskos et al. 2015). In order compare if this was also the case in other libraries I decided to plot these results along with the Em JA, *S2* cell all polysomes data (called *S2* MiSeq) as well as the 0-2H embryo data generated by Dunn *et al.* (2013). Figure 4.13 shows that tRNA reads formed 35% of the total reads in the Em JA run and 14% of the total reads in the *S2* MiSeq data, again with Glutamine and Aspartic acid being the most abundant tRNAs in each case. In contrast, only 2% of the total reads arise from tRNAs in the 0-2H Embryo Dunn *et al.* (2013) data, which instead has significant rRNA contamination (82%). Based on this observation, I concluded that tRNA, which is associated with polysomes, is a significant and previously overlooked contaminant in Poly-Ribo-Seq but not in Ribo-Seq Libraries (Figure 4.13).

It could therefore be speculated that the drastically lower level of tRNA in the Dunn *et al.* (2013) library is because they are removed during the sucrose cushion purification step that they employ, which is used to isolate monosomes after nuclease treatment of the lysate. Since nuclease digestion is performed on already fractionated polysomes in Poly-Ribo-Seq (Aspden *et al.* 2014) this additional purification step was skipped in this technique, therefore it is no surprise that these 70nt tRNAs are abundant in Poly-Ribo-Seq libraries as they are a major component of the translational machinery. However, the significantly greater proportion of tRNA reads in Em NEB as compared to Em JA suggests a bias for tRNA fragments using the dual adaptor library preparation approach. As tRNAs are highly structured clover shaped RNAs, this could probably be due to a structural bias of the 5' adapter T4 RNA ligase I which is used in the NEB kit but not in the other Ribo-Seq protocols. Thus in order to improve sequencing efficiency, I decided to prepare two further libraries with 0-8H embryos; one with depletion of tRNAs (tRNA Dep) and the second with a sucrose cushion purification step (SucCush) to compare the results and assess whether we can improve the number of usable reads after sequencing.

Sucrose Cushion vs tRNA depletion

As described in the previous section, I decided to optimise the depletion of the large amount of contaminating tRNAs using two methods. The first approach was using subtractive hybridisation to target the two most abundant tRNAs present in our data (tRNA^{Glu} and tRNA^{Asp}). In order to better inform the design of the targeting oligonucleotides, the number of reads against each nucleotide position in the 70nt tRNAs were plotted along the transcript. Interestingly, the results showed that >99% of the reads aligned to the first 31-33 nucleotides of the tRNA sequences. Therefore the biotinylated oligonucleotides were designed complementary to these 5' tRNA fragments (5'tRFs). The tRNA^{Glu} and and tRNA^{Asp} oligonucleotides were mixed in a 9:2 ratio of a 60μM mixture, the proportion being based on the number of reads aligning to each tRNA in the Em NEB data. 1μl of this mix was added to the sample just prior to the rRNA depletion step in the protocol and the hybridised oligonucleotides were pulled out of solution by the addition of an excess of streptavidin beads alongside the rRNA depletion beads. The library was then prepared as normal using the NEB kit and the sample sequenced on a MiSeq machine. Figure 4.13 shows the success of this approach in the tRNA Dep library, with only 5% of the total sequenced reads arising from tRNAs and a further 52% mapping to rRNA. This result gave us an unprecedented 46% of the raw reads mapping to mRNAs, which is three-fold greater than the proportion of mRNA reads in the previous embryo datasets (15%).

Despite the success of this approach, the abundance of 5'tRFs led to the question of whether there were other non-RBF sequences co-purifying with our sample. These false positive reads could potentially arise from fragments of other polysome-associated ncRNAs as well as sequences protected by RNA binding protein complexes. Since tRNAs make up only 2% of the contaminating sequences in the Dunn *et al.* (2013) data, this substantiated the use of a sucrose cushion purification step. After the RNase I digestion of fractionated polysomes, nuclease-treated polysomes in 10% Sucrose solution (160 mL) were concentrated down to 2 mL using two 50 mL Ultrafiltration spin columns (30KDa MWCO). The concentrated sample was then carefully layered on to two tubes, containing 2ml of 1M Sucrose in each. The samples were ultracentrifuged at 70,000 *g* for 3 hours in order to pellet the monosomes through the sucrose solution while smaller, less dense complexes remain in the viscous sucrose solution (Figure 4.15). The pellet was then resuspended in TRIzol to purify the RBFs before continuing with library preparation using the NEB kit. Figure 4.13 shows that the sucrose cushion purified library shows a similar profile to the Dunn *et al.* (2013) dataset with 83% of the reads mapping to rRNA and only 1% mapping to tRNAs, with the remaining 16% mapping to mRNAs. By adding a sucrose cushion purification step in the protocol, an additional observation was the higher percentage of uniquely mapping, size filtered, 28-34nt reads.

The successful incorporation of a sucrose cushion step in the protocol allowed the verification of whether the majority of the rRNA reads in the SucCush library could be attributed to a specific set of over-represented sequences as observed in other Ribo-Seq libraries prepared in this way (Dunn *et al.* 2013; Ingolia *et al.* 2012). Similarly to the analysis used for the Em NEB library, I plotted the number of reads against nucleotide position for rRNA sequences. These results showed that 75% of all rRNA reads that remained could be attributed to 6 short sequences, three each in the 18S and 28S ribosomal subunits, in similar but not identical position to those reported in Dunn *et al.* (2013) I hoped to improve the rRNA depletion further by using a combination of the rRNA depletion beads as well as a mixture of biotinylated oligonucleotides targeting these overrepresented sequences.

Table 4.2 shows that 68% of reads remained after mapping and size filtration in the SucCush data, which is higher than the 47% in the tRNA Dep dataset. This number is substantially better than what was observed the Em JA and Em NEB libraries, which were much lower with 29 and 38% uniquely mapping 28-34nt, non-rRNA/tRNA reads (Table 4.2). To determine whether this result was due to a higher proportion of unique mappers or a higher proportion of 28-34nt reads, I compared the tRNA Dep and SucCush libraries. From this, we were able to see that the proportion of uniquely mapping reads is approximately the same (~75%) between the two libraries. However the SucCush library has a higher proportion of 28-34nt reads (87%) as compared to the tRNA depletion run (65%). This is probably due to the removal of non-RBF reads that are outside the characteristic RBF size range of 28-34nt by the sucrose cushion purification step. Thus the final 8-16H and 16-24H embryo Poly-Ribo-Seq libraries were generated by incorporating the sucrose cushion step into the protocol and by performing two rounds of rRNA depletion using both the rRNA beads and biotinylated oligonucleotides mixed in a ratio as previously described. This strategy was implemented with great success to achieve an unprecedented level of rRNA depletion of our sequencing samples with 67% usable mRNA in the 16-24 H (*Em3*) Library and 42% mRNA for the 8-16 H (*Em2*) Library as shown in Table 4.3.

Overview of Final Embryo Datasets

The 8-16H and 16-24H Em Poly-Ribo-Seq libraries were prepared using sucrose cushion purification, rRNA depletion and library preparation using the NEB kit as outlined in Figure 4.14. The prepared libraries were then barcoded, pooled together, and sent to the University of Cambridge Sequencing Facility to be sequenced on an Illumina NextSeq machine. The NextSeq uses a high output run, with a maximum capacity of 400 million reads. From these runs, we obtained 178 million reads for the 8-16H embryo dataset and 111 million reads for the 16-24H embryo dataset. Using the various optimisation techniques and the improved rRNA depletion strategy, I was able to reduce the level of rRNA contamination down to 59% in the 8-16H dataset and an even lower 33% for the 16-24H libraries.

The incorporation of a sucrose cushion step also significantly improved the proportion of 28-34nt unique mapped reads at ~70% in both the 8-16H and 16-24H libraries as compared to libraries prepared without this purification (29-46%). The overall impact of these improvements introduced to the final protocol can be seen in Table 4.3 as we achieved ~45 million reads mapping to ORFs in each dataset. This number is approximately 8 times as many reads as the published *S2* cell small polysome dataset (5.9M Reads) despite the latter having more raw sequencing reads than the embryonic data sets. For 0-8H embryos, the reads from all four optimisation runs were pooled together to make a single dataset of 33.4 million reads with ~3 million reads mapping to ORFs. However, we have recently resequenced the SucCush library on a NEXTSeq machine to obtain 289 million raw reads in order to make this dataset more comparable to the 8-16H and 16-24H datasets.

Preparation of mRNA controls

RNA-Seq is normally performed as a control alongside Ribo-Seq experiments in order to be able to normalise the number of ribosome footprints across an ORF to the abundance of the mRNA transcript. This allows us to calculate a metric called Translation Efficiency (TE), which is calculated by dividing the *RPKM* of the CDS in the footprinting sample by its *RPKM* in the RNA-Seq sample. This metric can be useful to assess the translation of a particular coding region. In order to prepare RNA-Seq libraries for each of the embryo samples, 300µl of the embryo cytoplasmic lysate was purified using TRIZOL followed by DNAse treatment and phenol chloroform purification (Figure 4.14). 125µg of the purified RNA was Poly-A selected using Oligo-dT beads, and then fragmented by alkaline hydrolysis, performed by incubating at 95°C for 20 minutes in fragmentation buffer (pH 9.3) (Kronja *et al.* 2014). The fragmented mRNA was size selected using denaturing PAGE to resolve the RNA and the region corresponding to 50-100nt was excised from the gel. The size-selected mRNA fragments, which were chosen to be of a similar size as the RBFs, were then prepared for RNA-Seq using the NEBNEXT small-RNA library preparation kit before short read sequencing.

Single-End RNA-Seq was performed on 0-8 hour embryo lysate from the Em NEB sample using a Version 2 50bp MiSeq cartridge, while the 8-16H and 16-24H RNA-Seq libraries were pooled together and run on a Version 3 150bp MiSeq cartridge. The newer Version 3 cartridges can sequence reads up to 150bp long and has larger capacity (30M reads) than the Version 2 (20M reads). In order to maximise this additional capacity to produce longer reads and achieve better coverage of the transcriptome, I reduced the RNA fragmentation time and cutting out a larger portion of the denaturing PAGE (region corresponding to 50-150nt) gel for these samples. The resulting 0-8H Embryo RNA-Seq dataset had 20 million raw sequencing reads while the 8-16H and 16-24H RNA-Seq datasets had 15 million reads each (Table 4.4).

Discussion

The experiments discussed in this chapter were conducted with the primary aims of adapting the Poly-Ribo-Seq protocol to *Drosophila* embryos and improving the library preparation protocol to improve the number of usable reads for analysis. This is particularly important in the context of corroborating the translation of novel smORFs outside of annotated protein coding regions, due to skepticism about whether these smORFs are actually translated. Due to their extremely small size, novel smORFs (Median size: 69nt) cannot be accommodate the same number of ribosomes as compared to canonical protein-coding ORFs (Median size: 2076nt, Flybase). Therefore it is difficult to statistically score the translation of small ORFs due to low numbers of RBF reads mapping to them. This challenge is exacerbated for smORFs located in long ncRNAs, many of which are transcribed at very low levels (less than 0.1 copies per cell) (Kung *et al.* 2013) and in a tissue-specific manner (Washietl *et al.* 2014). For example, a recent Ribo-Seq study estimated that only 0.08% of the total reads map to long ncRNAs (Housman, and Ulitsky 2015).

Although *RPKM* abundance measurements are normalised to the size of the feature, usually to define translation an additional filter is also required to account for a minimum number of reads across the ORF (Ingolia *et al.* 2011; Aspden *et al.* 2014; Popa *et al.* 2016). Furthermore, in the analysis of Ribo-Seq data, many of the computational metrics that have been developed to discern real translation from background noise require a substantial number of reads across the ORF in order to score novel ORFs, as will be discussed in Chapter 5. The challenges in defining smORF translation are only exacerbated by the low sequencing efficiency of the library generated using the original protocol in Aspden *et al.* (2014). For example, after bioinformatic removal of contaminating sequences and filtering for 28-34nt reads, only 5.9 million (3%) of the total 189 million reads in the *S2* small-polysome sequencing run aligned to ORFs.

Therefore in the hope to increase sequencing efficiency of my experiments, I aimed to increase the proportion of usable reads sequenced. This was accomplished firstly by the optimisation of the harvesting and lysis of the embryos themselves to have enough starting material, and then biochemically by improving rRNA and tRNA depletion, incorporating a sucrose-cushion purification step and lowering the proportion of background 3’UTR reads by using the NEB library preparation kit. The NEB kit also offered another significant advantage, which was the significantly higher yield of usable material from this protocol. This addressed what was a major concern at the beginning of these experiments as we had struggled to produce library for even a single sequencing run using the previously used protocol of our lab (Em JA). With increasing application of the Ribo-Seq and NGS techniques, more and more information is now available regarding sources of bias and distortion in the data derived from such experiments. This required careful consideration of issues such as embryo lysis, nuclease treatment, Library preparation and how these steps may influence the quality of the data that we have generated. The following sections go through the various stages of the Poly-Ribo-Seq protocol that required optimisation to generate the final libraries for the three different stages of *Drosophila* embryos.

MNase vs RNase and Optimisation of Digestion

Our lab has previously used the RNase I nuclease to successfully generate ribosome footprints for the Poly-Ribo-Seq data generated from *S2* cells (Aspden *et al.* 2014), however Dunn *et al.* (2013) had previously reported that *Drosophila* ribosomes are sensitive to RNase I mediated degradation during Ribo-Seq library preparation. Instead, Dunn *et al.* (2013) recommend the use of *micrococcal* nuclease (MNase) for *Drosophila* samples to reduce the effects of ribosomal degradation; therefore it was pertinent to compare both RNase I and MNase enzymes for generating ribosome footprints from *Drosophila* embryos to understand and assess the translation of smORFs. RNase I performed much better than MNase in the digestion conditions tested in these experiments, contrary to what was reported by Dunn *et al.* (2013). This could be attributed to the differences in the digestion conditions between Ribo-Seq and the conditions required for Poly-Ribo-Seq, particularly because MNase has been known to be highly sensitive to temperature, protein content and buffer composition of the reaction (New England Biolabs Product literature, Cat. No. M0247S).

A variety of different reaction conditions were tested with RNase I, using various concentrations of the nuclease. The results from these experiments suggested that ribosome degradation can be greatly reduced by carrying out the reaction at 4°C rather than at 25°C as had been previously conducted by our lab in the *S2* cell Poly-Ribo-Seq experiments (Aspden *et al.* 2014). The use of RNase I has also been documented to show an enrichment of shorter ORFs as compared to MNase generated data (Miettinen, and Björklund 2015). Further diminishing the case for the use of MNase is the fact that it has been shown to display a preference for digesting A-T rich sequences, which can affect P-site mapping of Ribo-Seq reads (Dunn *et al.* 2013). This concern will be discussed in further detail in the next chapter. MNase has also been shown to produce a higher proportion of background 3'UTR reads as compared to RNase I in *Drosophila* and human cell lines (Miettinen, and Björklund 2015), and in fact this finding may lead one to question the precision of the abundant stop codon read-through reported by Dunn *et al.* (2013). Considering all of these issues as well as the promising results with RNase I, we chose to remain with RNase I mediated footprinting.

Library Preparation And Implications For smORF Discovery

As discussed in the results of this chapter, it has been shown that almost every step in NGS library preparation, from adapter ligation to PCR amplification, introduces some bias to the data (reviewed in Raabe *et al.* (2014)). Though the ssDNA circularisation protocol used by Ingolia *et al.* (2012) for Ribo-Seq library preparation has been shown to reduce adapter ligation bias, it does not completely eliminate it (Jackson *et al.* 2014). Therefore, I decided to try and use the dual-adapter NEB library as preliminary results (Em NEB) showed that we would be able to generate approximately 10 times more sequencing library than the Ingolia protocol (Em JA) while also using fewer PCR cycles to limit the effects of PCR based over-amplification of sequences.

Our comparison by linear regression modelling of CDS *RPKM* from the sequencing data generated by the two different library preparation methods (Em JA versus Em NEB) showed a good correlation with an R^2 value of 0.79. This value is not significantly different to the correlation we observed using the same analysis of replicate libraries in the *S2* cell sequencing datasets (R^2 0.83) (Aspden *et al.* 2014). There was also a satisfactory amount of overlap (~80%) between the smORFs translated in 0-8H embryos using the two library preparation methods. These smORFs include more than 90% of the smORFs detected as translated in each the 0-2H embryo dataset (Dunn *et al.* 2013) and the *S2* cell dataset (Aspden *et al.* 2014). This correlation supported our decision to use both RNase I and the NEB kit, and was further endorsed by independent Ribo-Seq data from 0-2H unfertilised *Drosophila* eggs generated by

Kronja *et al.* The Kronja *et al.* library was prepared from RNase I treated lysate of bleach dechorionated eggs and used a dual adapter ligation approach for library preparation (Kronja *et al.* 2014). Despite these differences in harvesting of material, nuclease and library preparation, their data shows a good correlation (R^2=0.89) with the translation efficiency (TE) of ORFs in Dunn *et al.* (2013) and also with protein levels measured by mass spectrometry (R=0.99), thus highlighting the robustness of this technique.

An unforeseen but useful consequence of switching to the NEB kit for library preparation was the significantly lower proportion of sequencing reads mapping to 3'UTRs (~5%) as compared to the 8-15% observed in libraries prepared using the single stranded DNA circularisation (Table 4.2). Since our translation cut-offs are calculated using the 90th Percentile *RPKM* of background 3'UTR signal, this result was particularly encouraging as it heavily influences the analysis and direct confidence in what ORFs are truly translated. The reason for this discrepancy is unclear, but it can be speculated that it could be due to an uncharacterised bias introduced by CircLigase or PCR bias resulting from the use of Phusion DNA polymerase. Phusion DNA polymerase is known to show amplification bias against GC rich regions (Aird *et al.* 2011). 3'UTRs are rich in A-U rich sequences (Barreau *et al.* 2005) and thus may be over-represented in libraries prepared using this enzyme due to the amplification bias against GC rich regions. On the other hand, the DNA polymerase supplied with the NEB kit is claimed to minimize such PCR bias.

One documented technique that we could use to reduce or eliminate PCR amplification based bias in future experiments is by using unique molecular identifiers (UMIs) (Kivioja *et al.* 2012). UMIs are short randomised nucleotide sequences that are incorporated in to each library fragment during adapter ligation and before PCR amplification. The addition of a unique UMI to each RBF ensures that PCR duplicated reads are only counted once at the data analysis stage thus improving gene quantification (Duncan, and Mata 2014). These randomised sequences can also be added to the ligating ends sequencing adapters in order to minimise any structural or sequence specific bias of the ligation enzymes against specific sequences (Fuchs *et al.* 2015).

tRNA Derived Fragments

Transfer RNA's are small (70nt) non-coding RNAs that have a highly conserved clover shaped structure that allows them to function as the link between mRNA and the ribosome. It was therefore expected that we observed a significant amount of tRNA co-purifying with polysomes and in our sequencing data. However, the difference in the proportion of tRNA-derived fragments (tRFs) observed between the two different library preparation methods was surprising; with more than twice the proportion of tRFs present in Em NEB (75%) sample than to Em JA (35%). The main difference between the two protocols is the use of T4 RNA ligase I for 5' adapter ligation in the NEB kit therefore, the most likely explanation for this increase in tRFs could be a structural affinity of this enzyme towards these highly structured RNA fragments. Although sequencing adapter and RNA fragment co-folding has previously been shown to play a role in influencing the efficiency of the T4 RNA ligase 2 mediated 3' adapter ligation step (Zhuang *et al.* 2012), a recent study has shown that the 5' adapter ligation step is also prone to significant structural bias (Fuchs *et al.* 2015).

More than 99% of the tRFs sequenced align to first 31-33 bases of the tRNAs, of these 50% are derived from Glutamyl-tRNA and a further 10% from Aspartyl-tRNA. The abundance of these particular tRNAs is corroborated by a recent small RNA sequencing study in *Drosophila* (Karaiskos et al. 2015). Glutamyl-tRNA is highly abundant as it performs an additional biological role as a precursor for haem (in insects and animals) and chlorophyll (in plants) biosynthesis (Jahn *et al.* 1992). 5' and 3' tRFs also have recently emerged as a new class of regulatory small non-coding RNAs reported to have a role in aging (Dhahbi *et al.* 2013) and stress response (Blanco *et al.* 2014) through a variety of post-transcriptional mechanisms (reviewed in Kirchner and Ignatova (2015) and in Wilusz (2015)). 5' tRF's in particular have been shown to modulate translation through direct inhibition of translation initiation (Ivanov *et al.* 2011; Sobala, and Hutvagner 2013) and are also within the same size range as RBFs. Therefore it is not surprising that these tRFs form an abundant portion of Poly-Ribo-Seq libraries. The difference in 5'tRF abundance between library preparation methods highlights the need to be aware of this possible bias for tRFs in small RNA libraries, particularly when it comes to correlating multiple datasets. It also helps us to develop future methods for the depletion of these potentially contaminating sequences from NGS data.

Sucrose Cushion Purification And rRNA Depletion

The Sucrose cushion purification step was incorporated in to the final protocol for as it provides us with increased confidence that the purified footprints are indeed from 80S Ribosomes. The sucrose cushion allows removal of false positive reads arising from RNA fragments protected by scanning 40S ribosomal subunits or RNA binding protein complexes as well as RNase-resistant structural sequences such as tRNA fragments. This is evidenced by the significant increase in the proportion of uniquely mapped reads retained after filtering for 28-34nt reads in libraries prepared with sucrose cushion (70%) as opposed to libraries prepared with no sucrose cushion (30-46%). Furthermore, sucrose cushion purification limits rRNA contamination to 6 specific sequences that make up 70% of the total rRNA reads in the library. This brought the rRNA contamination to a manageable number and I was able to improve the rRNA depletion step by supplementing the rRNA beads with oligonucleotides complementary to these six over-represented sequences. This allowed us to achieve an unparalleled level of rRNA depletion in the Poly-Ribo-Seq library with up to 67% of the total reads mapping specifically to mRNA in the 16-24H library. These results, coupled with the lower proportion of 3’UTR reads resulted in marked improvement in the overall proportion of usable reads at 26- 42% as compared to the 3-16% in *S2* cells, which is what we set out to accomplish by these experiments (Figure 4.15).

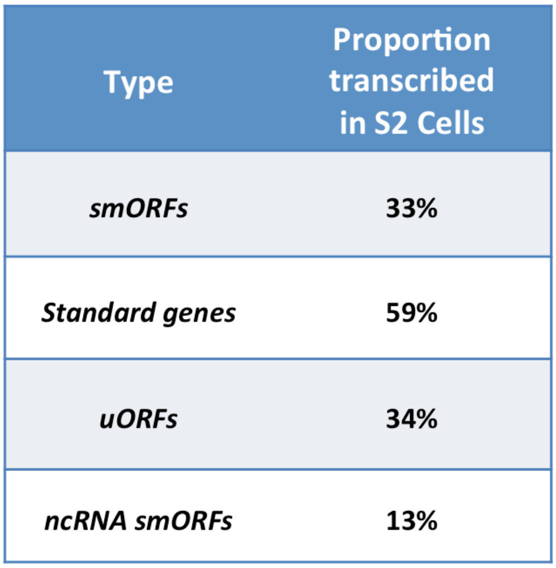
Final Library Prep And Improvements To Experimental Setup

RNA-Seq is performed as an mRNA control for Ribo-Seq in order to normalise the abundance of the footprints to the abundance of the transcript in the sample. In order to minimise sequencing and data alignment bias, the control RNA-Seq reads are kept at a similar length (~50 bp) as the RPFs. However, in trying to maximise RNA-Seq coverage using the Version 3 MiSeq cartridge, longer reads (ranging from 50-150bp) were used for the 8-16H and 16-24H libraries (Table 4.4) from which we were able to obtain 2.9 million and 4.3 million reads respectively. In hindsight we could see that these RNA-Seq runs should have been performed using the same conditions as the 0-8H sample, to enable better correlation between samples as well as between Footprints and mRNA (TE), especially since the accuracy of RNA-Seq measurements is influenced more by number of reads than read length (Li *et al.* 2010). As will be discussed in the next chapter, it is important to have similar depths of sequencing for control RNA-Seq and the corresponding Poly-Ribo-Seq libraries. We did attempt to address this issue at the data analysis step and in addition have resequenced the 8-16H and 16-24H libraries at a greater depth (40 to 50 million raw reads). In order to have absolute rather than relative quantification of individual transcripts, it may have been prudent to include a known quantity of an internal standard control transcript in order to compare transcript measurements (Wang *et al.* 2009; Han *et al.* 2014).

Other potential improvements to the experimental set up would include the sequencing of at least two technical and biological replicates for each embryogenesis stage. Libraries could have also been prepared with separate barcodes, and multiplexed on the same sequencing run as reads can be pooled later and may be more consistent for being sequenced at the same time. It would also have been ideal to have prepared a 0-8H embryo library using the final protocol for 8-16H and 16-24H samples, so that they would be sequenced at the same depth. The current 0-8H samples are the result of all the optimisation runs pooled together to form a single file with 2.9 million reads mapping to ORFs (Table 4.3). Since then we resequenced the SucCush library with 289 million raw reads, which will give approximately 30 million reads mapping to ORFs.

Overall, we were satisfied with the development of the final protocol, which has not only made the experiment quicker and less labour intensive, while also increasing the yield of the Poly-Ribo-Seq protocol. This has not only reduced the amount of starting material required but it is now possible to use this protocol to generate Poly-Ribo-Seq libraries from single organs or tissue in order to search for lowly transcribed, tissue specific smORFs. The improvements that I have made to the Poly-Ribo-Seq protocol resulted in the percentage of total reads mapping to ORFs increasing up to 42% in the 16-24H and 26% in the 8-16H embryo libraries from values as low as 3.5% to 16% in *S2* cells (Table 4.3). What this means is that even though the 16-24H embryo dataset with 110 million reads has fewer raw reads than the *S2* small polysome dataset (189 million reads); the final number of reads mapping to ORFs is approximately 8-fold greater (45 million reads) in the embryo dataset as compared to the 5.9 million reads in the *S2* cell data. This was the most important objective achieved in these experiments because even though we used quite a stringent *RPKM* and coverage based translation cut-off in the *S2* cell data, very short or lowly expressed ORFs could not be defined as translated since they did not accrue enough reads (minimum 5) across the ORF. The advantage of having a far greater number of reads will become apparent in the next chapter which discusses the analysis of this data to define smORF translation.

Chapter 4 Figure and Tables



Type	Proportion transcribed in S2 Cells
smORFs	**33%**
Standard genes	**59%**
uORFs	**34%**
ncRNA smORFs	**13%**

Table 4.1 *S2* cell transcription of *Drosophila* genome

My analysis of the *S2* cell RNA-Seq data in Aspden *et al.* (2014) shows that *S2* cells express only a subset of the total pool of genes transcribed by the *Drosophila* genome. Only 59% of standard annotated genes (FBcds) and 33% of annotated Flybase smORFs are transcribed by *S2* cells. Of all the potential uORFs found in *Drosophila*, only 34% are transcribed and the number of smORFs found in long ncRNAs (ncrORFs) is 13% of the total annotated.

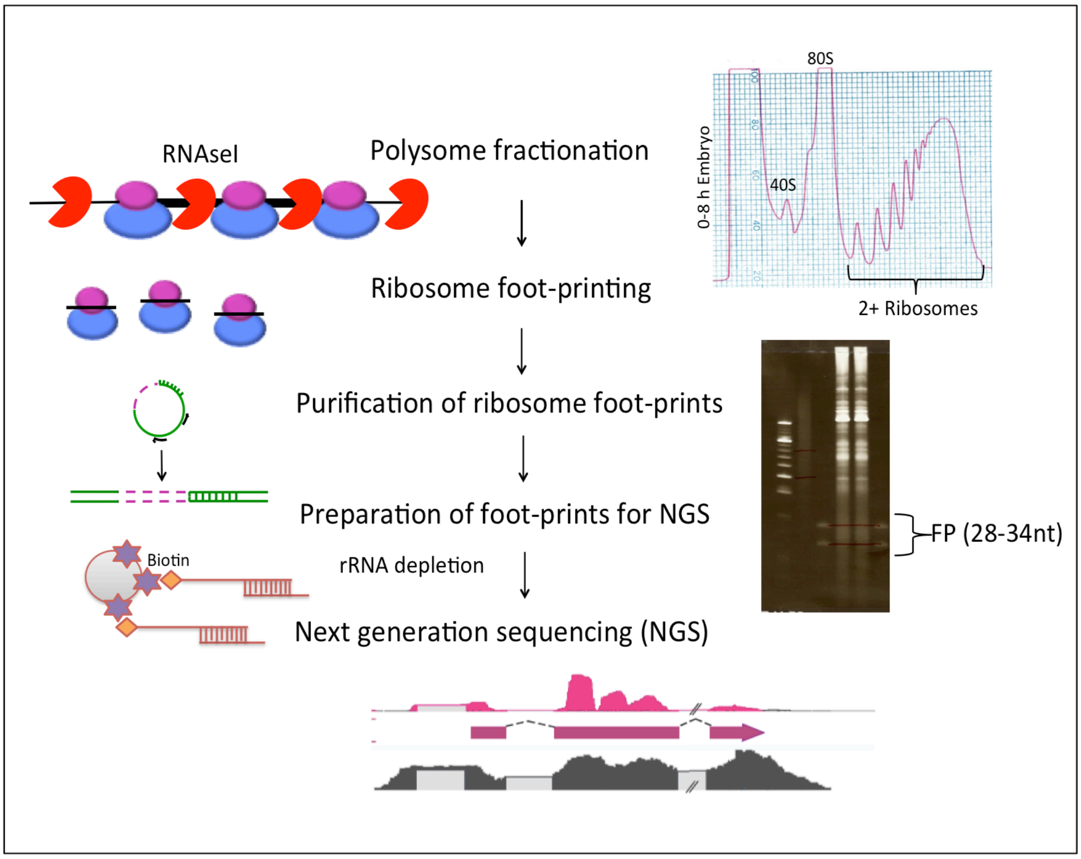


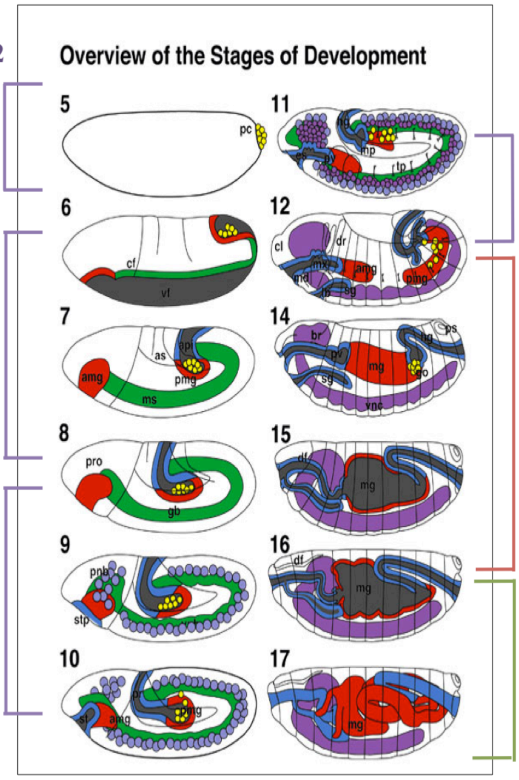
Figure 4.1 Overview of the Poly-Ribo-Seq Method
Poly-Ribo-Seq is the integration of Polysome fractionation with Ribo-Seq (Aspden *et al.* 2014). Polysomes (2+ Ribosomes) are fractionated on a sucrose density gradient to enrich actively translating ribosomes by excluding non-translating ribosomal subunits such as the scanning 40S and initiating 80S complexes in the monosome peak. The isolated polysome fraction is then subjected to RNase I treatment digest unprotected mRNA and generate ribosome footprints. These 28 to 34 nucleotide footprints (FP) are purified on a gel and then used to generate a library for Next Generation Sequencing (NGS). Ribosomal RNA (rRNA) is a major contaminant in Ribo-Seq libraries; therefore it is depleted during the library preparation stage by subtractive hybridisation using streptavidin beads. The prepared libraries can then be sequenced to visualise the mapping of ribosome footprints to the transcriptome.

0-8H AEL= Stages 1-12

Maternal RNA translation
Zygotic Genome activation
Pole cell migration

Gastrulation,
Germ-band elongation
Segment-Polarity genes

Segmentation of neuroblasts
Cephalic furrow formation
Head and Mouth features



Epidermal parasegmentation
Tracheal pit invagination
Neuroblast formation
Germ band retraction
Optic lobe invagination
Ventral closure
Segment formation

8-16H AEL= Stages 12-16

End of germ band retraction CNS and PNS differentiation
Dorsal closure
Head involution begins
Imaginal discs invaginate
Cuticle deposition begins
Dorsal epidermal segmentation
Advanced denticles visible
Shortening of the ventral nerve cord

16-24H AEL= Stage 16 - Hatch

The tracheal tree fills with air
Retraction of the ventral cord
Fine tuning of larval structures

Adapted from Atlas of Drosophila development Hartenstein, V. (1993)

Figure 4.2 Stages of *Drosophila melanogaster* embryonic development
Drosophila embryogenesis takes place in a time span of 24 hours after egg laying (AEL), during which the syncytial blastoderm develops into a 1st Instar Larva. For the purpose of our experiments, these 17 stages of varying length, have been divided into three distinct groups called Early (0-8H AEL), Mid (8-16H AEL) and Late (16-24H AEL) embryogenesis. Early embryogenesis includes embryonic stages 1-12, as can be seen from the figure (Adapted from Atlas of *Drosophila* development (Hartenstein 1993)), prior to stage 5 there are very few morphological changes in the embryo, but this time includes the expression of maternal genes deposited into the egg by the *Drosophila* female and encompass the maternal to zygotic gene expression transition. From stage 5 onwards, the first major morphological changes of patterning begin to occur through gastrulation, germ-band elongation and the early stages of CNS development. In this way it can be said that early embryogenesis is the time of the most drastic change for the embryo, as it encompasses the most stages of development. Mid embryogenesis is a time of rapid organ formation of the gut and gonads. Precursor tissues called imaginal discs which will form eyes, wings, sensory organs and appendages of the adult fly are also formed and matured into how they will remain in the larvae and the CNS and PNS are developed almost fully. During Late embryogenesis, the larva is almost ready to hatch and there is some further neuronal development and PNS formation, as well as fine-tuning of larval structures. These stages provide a context for the collection of data, as we are likely to see distinct sets of gene expression at the different times of embryogenesis.

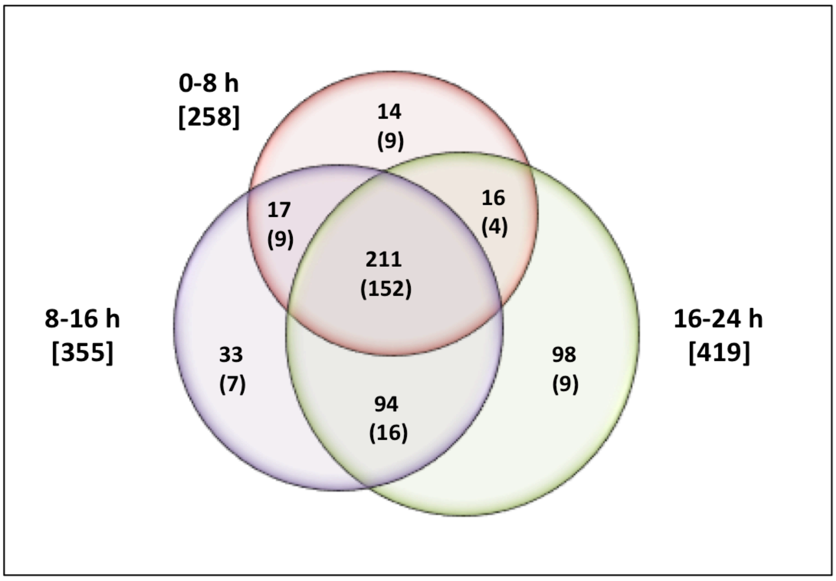


Figure 4.3 Transcribed FlyBase smORFs during Early, Mid and Late embryogenesis

The data presented in this Venn diagram are from modENCODE RNA-Seq experiments, showing the number of FB smORFs transcribed (*RPKM*>1) in each dataset that we are aiming to test. The numbers in parenthesis shows the subset of these smORFs translated in *S2* cells according to Poly-Ribo-Seq (Aspden *et al.* 2014). There are a total of 483 unique smORFs transcribed during embryogenesis, 211 smORFs are transcribed in all three stages. 152 of these are also translated in *S2* cells and these smORFs represent a large majority of all smORFs expressed in the embryo. There is little overlap (17 smORFs) between Early (0-8h) and Mid (8-16h) or Early and Late (16-24h) embryogenesis (16 smORFs) whilst there is a significant overlap between Mid and Late embryogenesis (94 smORFs). Of the 335 smORFs expressed in Mid embryogenesis, only 33 are unique to this stage, whereas of the 419 expressed in Late embryogenesis, there is a much larger number than both the other stages (98 smORFs unique to Late embryogenesis). There is very little overlap of *S2* cell Poly-Ribo-Seq hits with the smORFs found uniquely in Mid and Late embryogenesis.

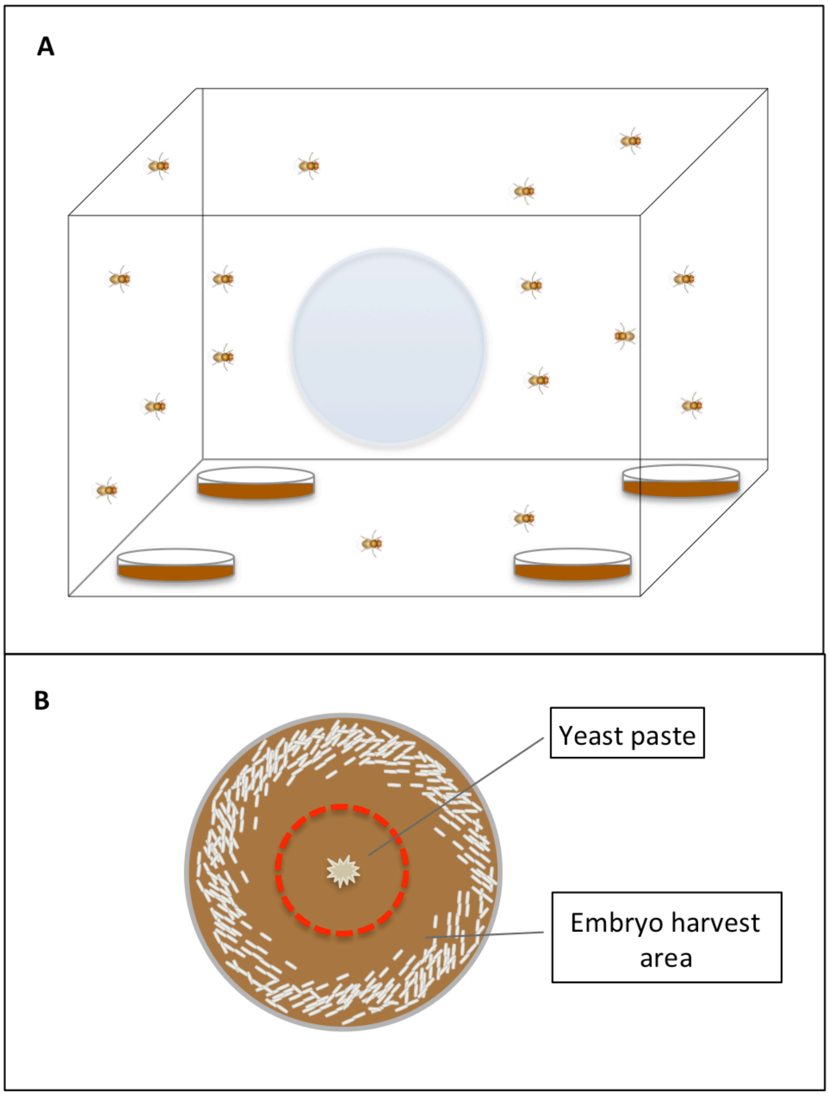


Figure 4.4 Harvesting *Drosophila* embryos

A) A Fly cage (50cm x 30cm x 50cm) was set up with 12-16 bottles each containing approximately 300-500 adult *Oregon-R* flies. The cage has a gauze window to prevent flies from escaping while allowing changing of the plates upon which embryos are laid. The cage was kept at 25°C with four 10cm petri dishes containing molasses fly food and were changed at 9am, 5pm and 1am (8 hour intervals).

B) As flies lay eggs in food for the future larvae to feed on when they hatch, Molasses fly food set into 10cm petri dishes was used to feed the flies and collect embryos. Plates were set up to contain a small amount of yeast paste in the middle. This area (circled in red) was avoided during embryo collection to minimise the carry over of yeast into the harvested sample. The embryos were either harvested immediately for the Early embryogenesis sample, or aged 8-16 hours at 25°C for Mid or Late embryogenesis samples.

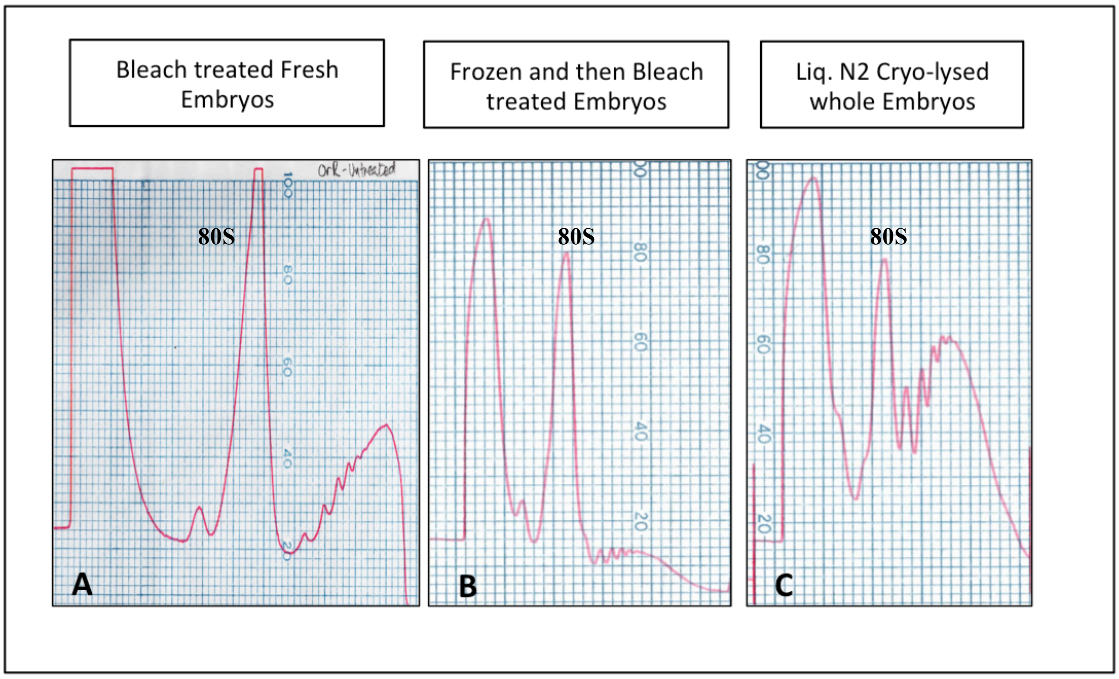


Figure 4.5 Polysome Gradient analysis of embryo lysis methods

A) Polysome gradient analysis was performed on overnight (0-16 H AEL) small-scale (150mg) collection of embryos. This sample was made from freshly harvested, bleach-dechorionated embryos that were lysed using a glass homogeniser.

B) This Polysome gradient analysis was conducted on a much larger amount of embryos, as would be required for Poly-Ribo-Seq. Embryos were collected from multiple plates and were flash-frozen and pooled together for use on a later day. On the day of the analysis, embryos were thawed, dechorionated with bleach, fixed with CHX in PBS and lysed with glass homogeniser. It can be seen from the absorbance trace that this method did not work; as there is a collapse of the Polysomes.

C) Polysome gradient analysis on embryos that have undergone Cryo-lysis. These embryos were collected, flash frozen and pulverised in lysis buffer using pestle and mortar. On the day of the experiment, the sample is thawed in the presence of CHX buffer. This method allowed the scaling up of embryo collection (0.75g of embryos used) and maintains the quality of the Polysome gradient, which can be evidenced by the ratio of the height of the 80S peak to the polysome peaks.

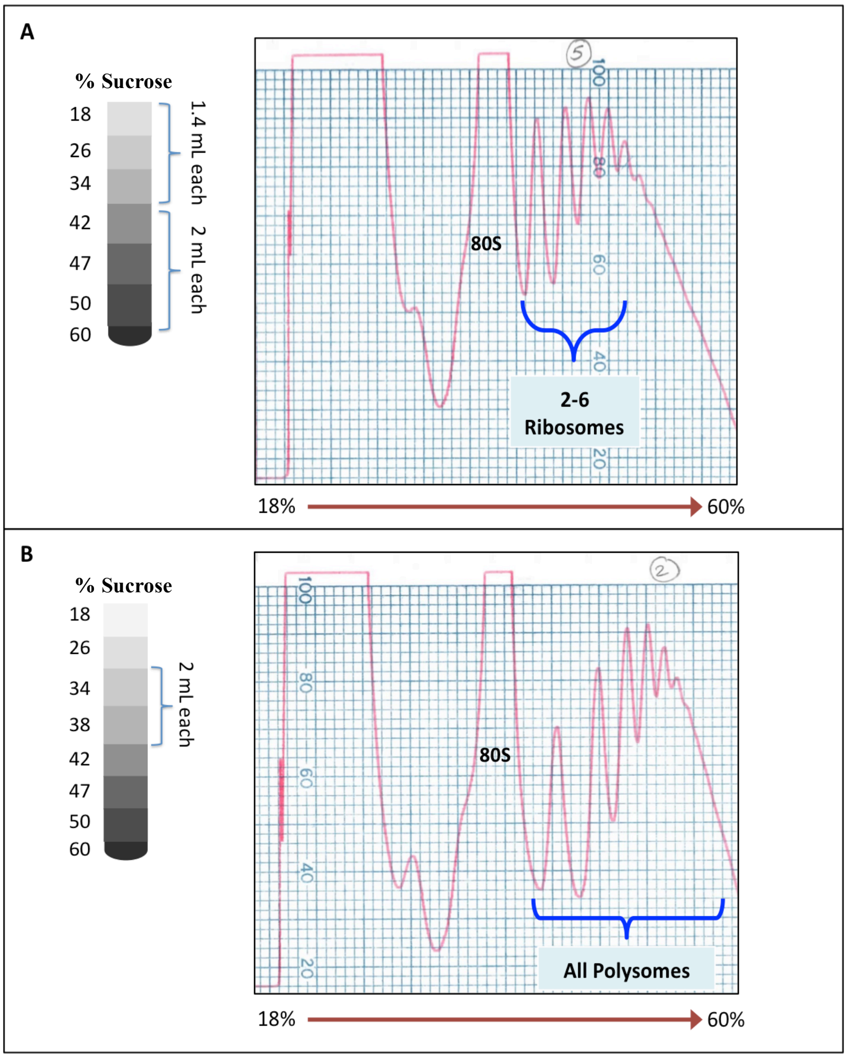


Figure 4.6 Polysome Fractionation optimisation

A) The sucrose gradient upon which the sample was fractionated to achieve maximum separation of small polysomes (2-6 ribosomes) from large polysomes, as conducted on *S2* cells in Aspden *et al.* (2014). This involved increasing the volume of the heavier 42%, 47% and 50% w/v sucrose fractions to 2ml, whist keeping the lighter fractions (18%, 26% and 34%) at a volume of 1.4ml. The corresponding Polysome gradient analysis shows separation of the 2-6 ribosomal peaks.

B) For the embryo Poly-Ribo-Seq experiments, we decided to collect all polysomes (2+ ribosomes) and the sucrose gradient pouring protocol was adjusted to achieve maximum seperation between the 80S and 2-ribosomes peak. This involved the addition of an additional sucrose layer (38%) and increasing the volume of the 34% and 38% w/v sucrose fractions. The corresponding Polysome gradient analysis shows the improved separation of the non-productive 80S peak and the 2-ribosomes peak.

	M	1	2	3	4	5	M	6	L
Mnase (U)		1000	500	500	250	250		0	
Temp (°C)		25	4	25	4	25		4	

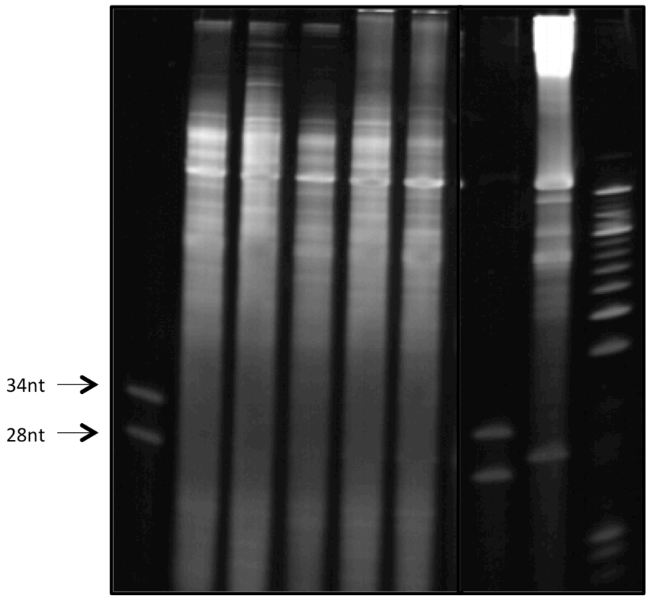


Figure 4.7 Testing MNase Nuclease for the generation of ribosome footprints
The top of this figure shows the various testing conditions used for MNase incubations to generate ribosome footprints from 0-8H *Drosophila* embryos. 200 and 500 units of MNase were tested at either 4°C (overnight) or at 25°C for 45 minutes, similar to the protocol used by Dunn *et al.* (2013). An excess of MNase (1000U) was also tested as a positive control, this sample was incubated at 25°C for 45 minutes. An undigested sample was also run, after being kept at 4°C overnight. All samples were run on a denaturing 15% polyacrylamide gel. The region of interest is determined by the RNA oligonucleotide markers in the lanes marked M to show region between 34-28nt. The different digestion conditions compared to the control in lane 6 shows that there is some digestion occurring in samples using MNase. The region of interest is however obscured by smears in all the conditions tested.

Lane	1	2	3	4		5	
RNAse (U)	1500	1000	1000	1000		0	
Temp (°C)	4	25	16	4		4	M
Time (h)	16	1	16	16		16	

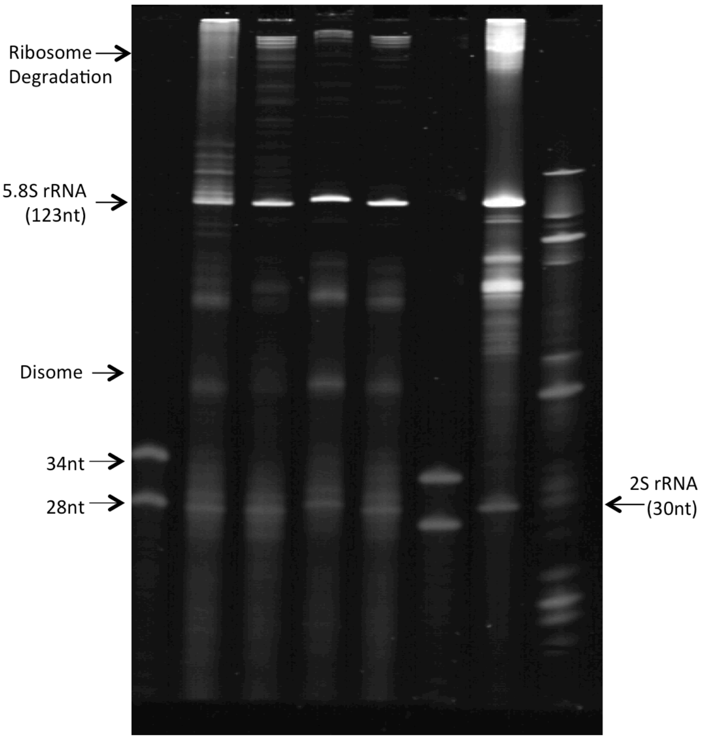


Figure 4.8 Testing RNase I Nuclease for the generation of ribosome footprints

The table at the top of this figure shows the various testing conditions for RNase I digestion using 1000U per polysome tube, similar to the established protocol in Aspden et al. (2014). Digests were performed for one hour at 25°C (Lane 2) or overnight at 4°C (3) and 16°C (4). An 'overdigested' lysate sample of incubation of 1500U RNAse I at 4 °C (1) was used as a positive control and undigested sample as a negative control. These samples were run on a denaturing 15% polyacrylamide gel. The region of interest is highlighted by the RNA markerts to show region between 34-28nt (M). The enrichment of digested sample is apparent using RNase I as compared to the undigested sample. Ribosomes appear to be degraded the most in the ‘overdigested’ sample containing 1500U of RNase I, and in decreasing amounts according to the temperature of incubation for the test concentration of 1000U as is shown by the smears at the top of the gel. The ribosomal smear is lowest for 1000U of RNase I at 4°C. The bands that appear at 123nt, which can be seen in all samples, is most likely the 5.8S ribosomal RNA (rRNA). There is the presence of a bright 30nt band in the undigested sample, which can be considered to be co-migrating with the footprints in the digested samples. This band most likely corresponds to the *Drosophila* 2S rRNA, which is 30nt in length.

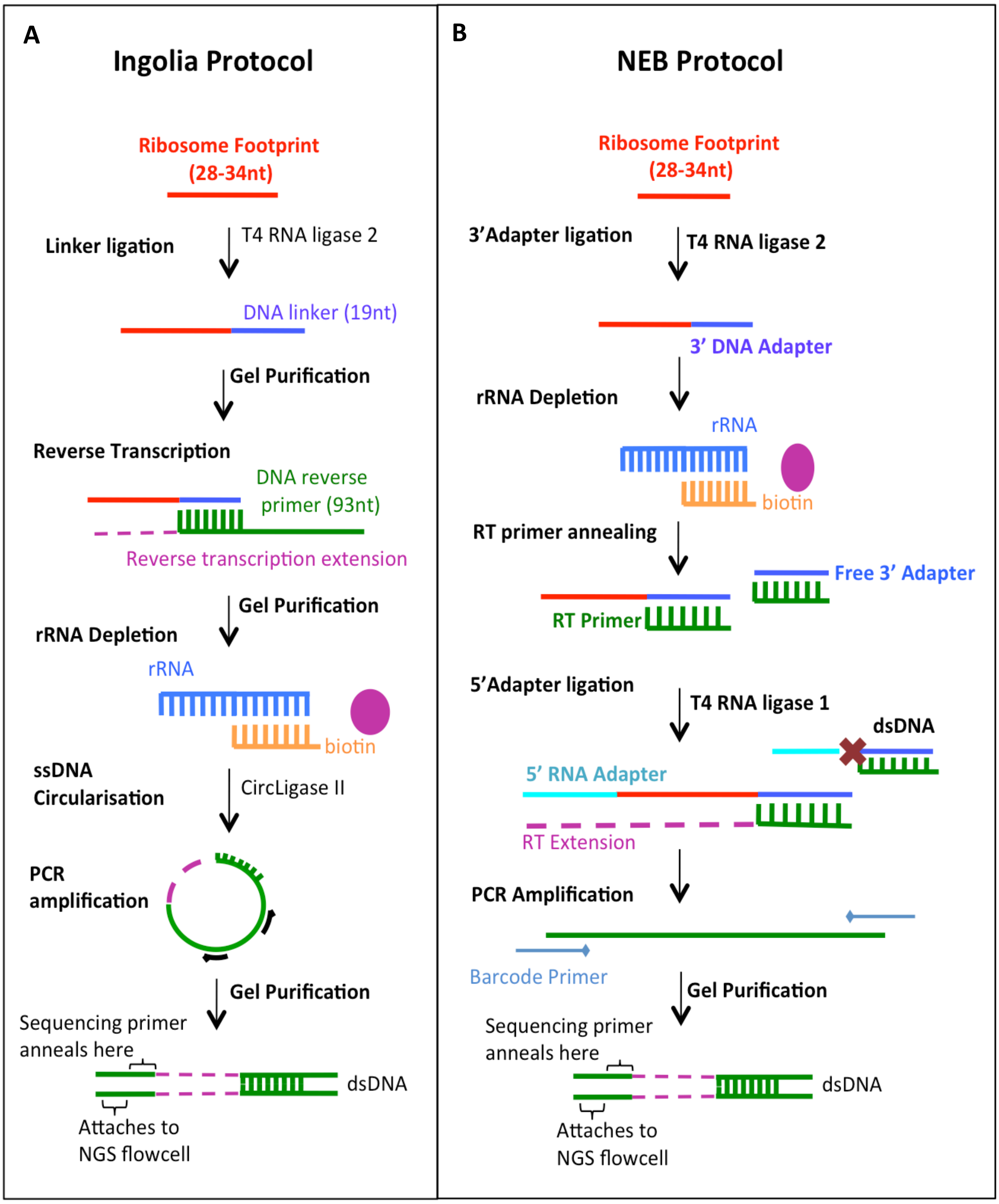

A
Ingolia Protocol
Ribosome Footprint
(28-34nt)
Linker ligation
T4 RNA ligase 2
DNA linker (19nt)
Gel Purification
Reverse Transcription
DNA reverse
primer (93nt)
Reverse transcription extension
Gel Purification
rRNA Depletion
rRNA
biotin
ssDNA
Circularisation
CircLigase II
PCR
amplification
Gel Purification
Sequencing primer
anneals here
dsDNA
Attaches to
NGS flowcell
B
NEB Protocol
Ribosome Footprint
(28-34nt)
3'Adapter ligation
T4 RNA ligase 2
3' DNA Adapter
rRNA Depletion
rRNA
biotin
RT primer annealing
Free 3' Adapter
RT Primer
5'Adapter ligation
T4 RNA ligase 1
dsDNA
5' RNA Adapter
RT Extension
PCR Amplification
Barcode Primer
Gel Purification
Sequencing primer
anneals here
dsDNA
Attaches to
NGS flowcell

Figure 4.9 Library Preparation methods for Poly-Ribo-Seq

A) This figure highlights the main differences between the Ribo-Seq library preparation protocol by Ingolia *et al.* (2012) and the dual-ligase NEB protocol. In summary, the Ingolia *et al.* protocol uses a 3' DNA linker (adapter sequence - 19nt) ligated to the footprints using T4 RNA ligase 2. The sample is then gel purified (1) after which a 93 nucleotide long RT Primer that incorporates the 5' adapter sequence separated by a C_{18} spacer. After this, another gel purification is performed (2) and biotinylated beads complementary to rRNA are used to deplete rRNA. After this the single stranded cDNA is circularised using CircLigase. The final primers used for PCR amplification contains barcode identifiers and the third gel purification step is performed after which the sample can be sequenced. The 'Ingolia protocol' aims to minimize RNA Ligase bias by eliminating the 5' adapter ligation step. Due to the 3 gel purifications during this protocol, this method requires a lot of starting material.

B) The NEB method employs the use of adapter sequences at the 3' and 5' ends of the RNA footprints (RBFs). The 3' DNA adapter contains a reverse transcription (RT) primer hybridisation sequence and is ligated to the footprint using T4 RNA ligase 2. RT primer hybridisation is performed prior to 5' ligation. This allows the RT primer to hybridise and form double-stranded DNA with any excess 3' DNA adaptor which has not been ligated by T4 RNA ligase I.The library is then converted to cDNA, and a quarter of the reaction is used to optimise the number of PCR cycles before final DNA amplification using primers were used to incorporate the sample barcodes. At this stage the only gel purification in this protocol is conducted and the region corresponding to the 150bp band is processed for sequencing.

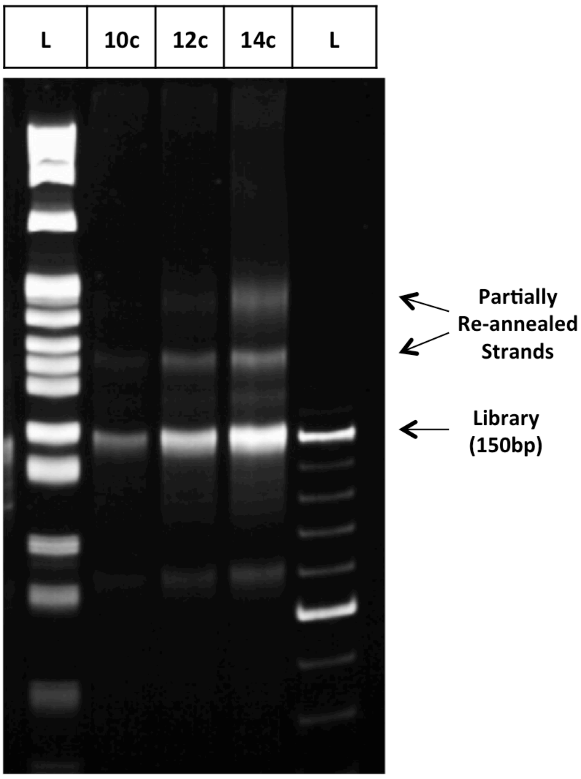


Figure 4.10 Optimisation of PCR cycles for Library amplification

A portion of the library prepared for 0-8H embryos using the NEB kit was run for 10, 12 and 14 PCR cycles to optimise for the correct number of cycles to amplify the entire sequencing libraries for *Drosophila* embryo Poly-Ribo-Seq. The library can be seen to run at 150bp. There is over amplification of the library with 14 PCR cycles as can be seen from the slower migrating smear from partially re-annealed strands which have been amplified due to complementarity of the adapter sequences. For the 0-8H NEB sample, 12 cycles were chosen as this showed a clear band without any smeared partially re-annealed strands and for all future library preparations this number was between 10-12 PCR cycles.

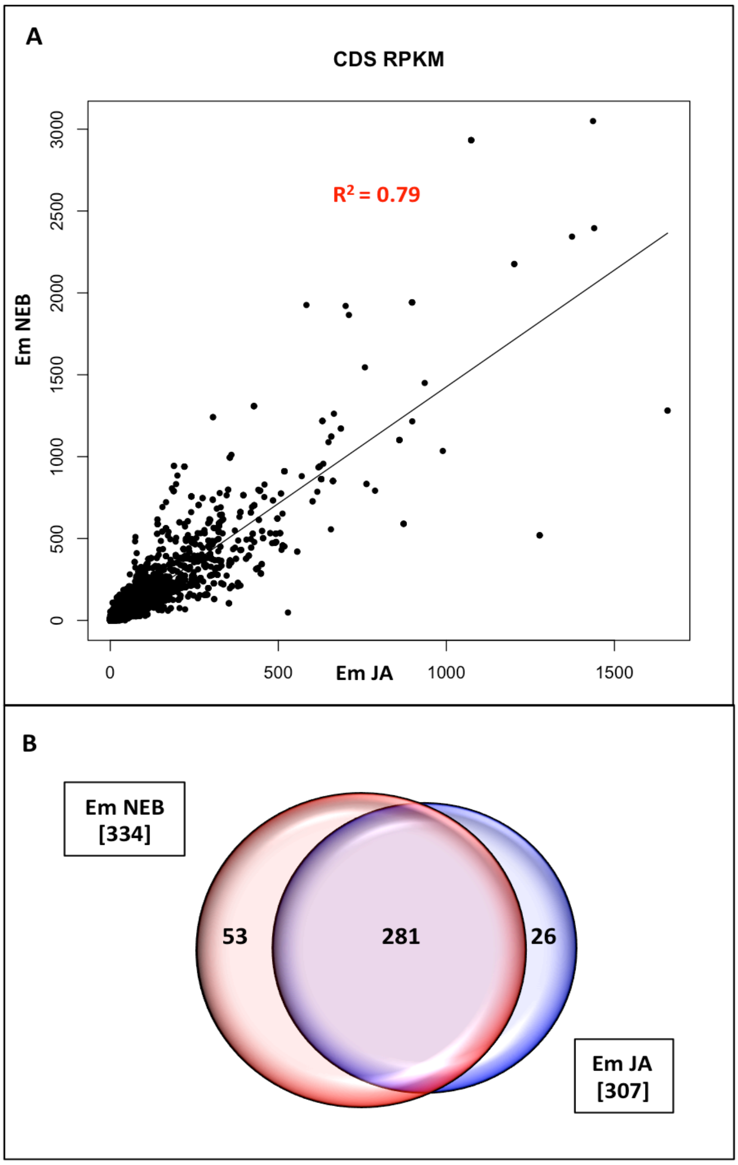


Figure 4.11 Correlation between results of Em JA (old) and Em NEB (new) library preparation methods

A) Linear regression modelling of annotation protein-coding CDS *RPKM*s of the Em JA and EM NEB library preparation methods. From the analysis, we can observe good correlation (R^2 = 0.79) between the *RPKM* values of coding sequences in each dataset. This indicates that there is no significant difference between the data that is generated between the two library preparation methods.

B) The number of FB smORFs translated in both Em JA and Em NEB samples was determined using the *S2* cell Poly-Ribo-Seq bioinformatics pipeline. The cut-offs for these datasets are Em JA: 2.9 *RPKM*, 0.08 Coverage and Em NEB: 0.94 *RPKM*, 0.07 Coverage. The Venn diagram shows that there is good overlap between the smORFs deemed as translated using the two library preparation methods. The Em NEB data shows translation of 53 smORFs that are unique to this method, as there is lower background signal in 3'UTRs in this sample.

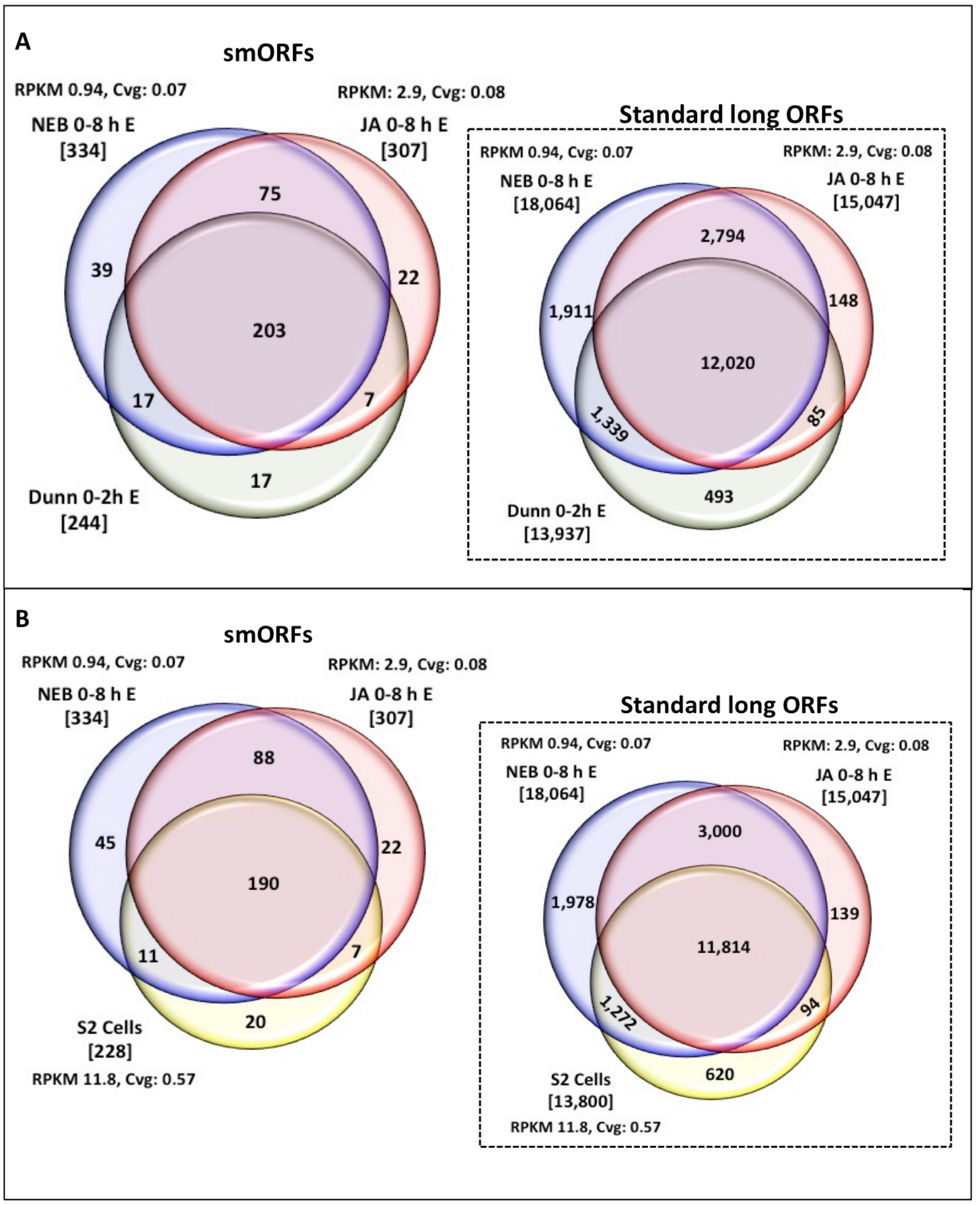
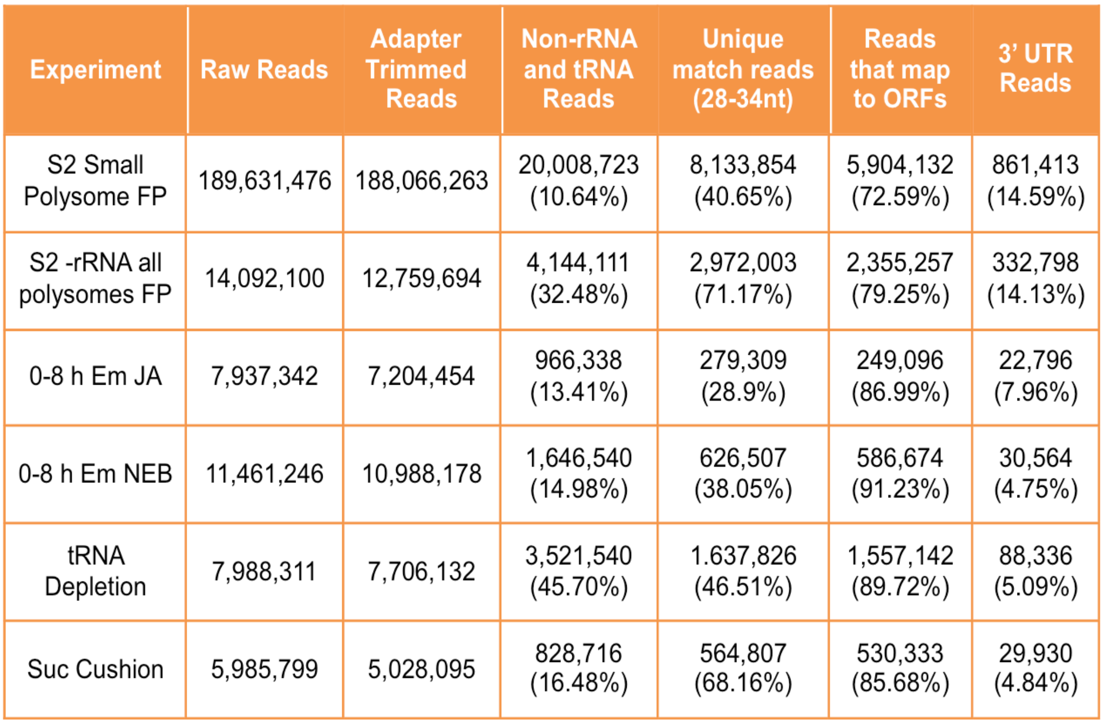


Figure 4.12 Correlation between results of Em JA, Em NEB and published *Drosophila* Ribo-Seq data

A) Using the *S2* cell cut-offs from Aspden *et al.* (2014), we re-analysed the data from 0-2H embryos published by Dunn *et al.* (2013). The Venn diagrams presented here show the overlap between the 0-8HEm JA, Em NEB and Dunn *et al.* data. For smORFs detected at early embryogenesis there is good overlap between all the data sets with the most overlap occurring between our in-house library preparations. The same can be said for standard long ORFs (inset panel). The NEB kit tends to perform better than all the other library preparations with more smORFs and long ORFs passing cut-offs due to the cleaner preparation.

B) The Venn diagrams presented here show the overlap between the 0-8H Em JA, Em NEB and *S2* cell Poly-Ribo-Seq data from Aspden *et al.* (2014). There is a surprisingly high proportion of overlap between *S2* cells and embryo smORFs, as can be expected as the former is an embryonic epithelial cell line. The comparison of standard long ORFs (inset) is similar to that of panel A, with a majority of longer ORFs being translated in all three data sets.



Experiment	Raw Reads	Adapter Trimmed Reads	Non-rRNA and tRNA Reads	Unique match reads (28-34nt)	Reads that map to ORFs	3' UTR Reads
S2 Small Polysome FP	189,631,476	188,066,263	20,008,723 (10.64%)	8,133,854 (40.65%)	5,904,132 (72.59%)	861,413 (14.59%)
S2 -rRNA all polysomes FP	14,092,100	12,759,694	4,144,111 (32.48%)	2,972,003 (71.17%)	2,355,257 (79.25%)	332,798 (14.13%)
0-8 h Em JA	7,937,342	7,204,454	966,338 (13.41%)	279,309 (28.9%)	249,096 (86.99%)	22,796 (7.96%)
0-8 h Em NEB	11,461,246	10,988,178	1,646,540 (14.98%)	626,507 (38.05%)	586,674 (91.23%)	30,564 (4.75%)
tRNA Depletion	7,988,311	7,706,132	3,521,540 (45.70%)	1.637,826 (46.51%)	1,557,142 (89.72%)	88,336 (5.09%)
Suc Cushion	5,985,799	5,028,095	828,716 (16.48%)	564,807 (68.16%)	530,333 (85.68%)	29,930 (4.84%)

Table 4.2 Depletion of rRNA to improve sequencing efficiency

This table summarises the results of the sequencing data from the various Poly-Ribo-Seq experiments conducted by our lab. Column 2 shows the Raw reads from the experiment, which are trimmed for adapter sequences (Column 3) from which rRNA and tRNA reads are removed bioinformatically. These are shown as a number of reads and proportion of adapter-trimmed reads (Column 4). Column 4 shows the number and proportion of non-rRNA and tRNA reads which are map uniquely to the transcriptome, which means that they unambiguously map to only one position in the genome. The next column (5) shows the number of reads and proportion of uniquely mapped reads that align specifically to ORFs. Column 6 is the total number of uniquely mapped reads that align to 3'UTRs. The *S2* cell small polysome footprint (FP) and *S2*-rRNA depletion all-polysomes FP were sequenced on the Illumina HiSeq and MiSeq machines, respectively, which is reflected in the number of raw reads which are present in the output data. The 0-8H embryo samples (Em JA, Em NEB, tRNA Dep and SucCush) were all sequenced on the MiSeq machine. From this work flow we can see that from rRNA depletion using beads alone, in embryos we do not achieve better than 15% of usable reads following the Aspden *et al.* protocol for *S2* cells. Using tRNA depletion, we achieve a substantial improvement with almost 46% of usable reads. The sucrose cushion experiment again reverts to approximately 16% of usable reads.

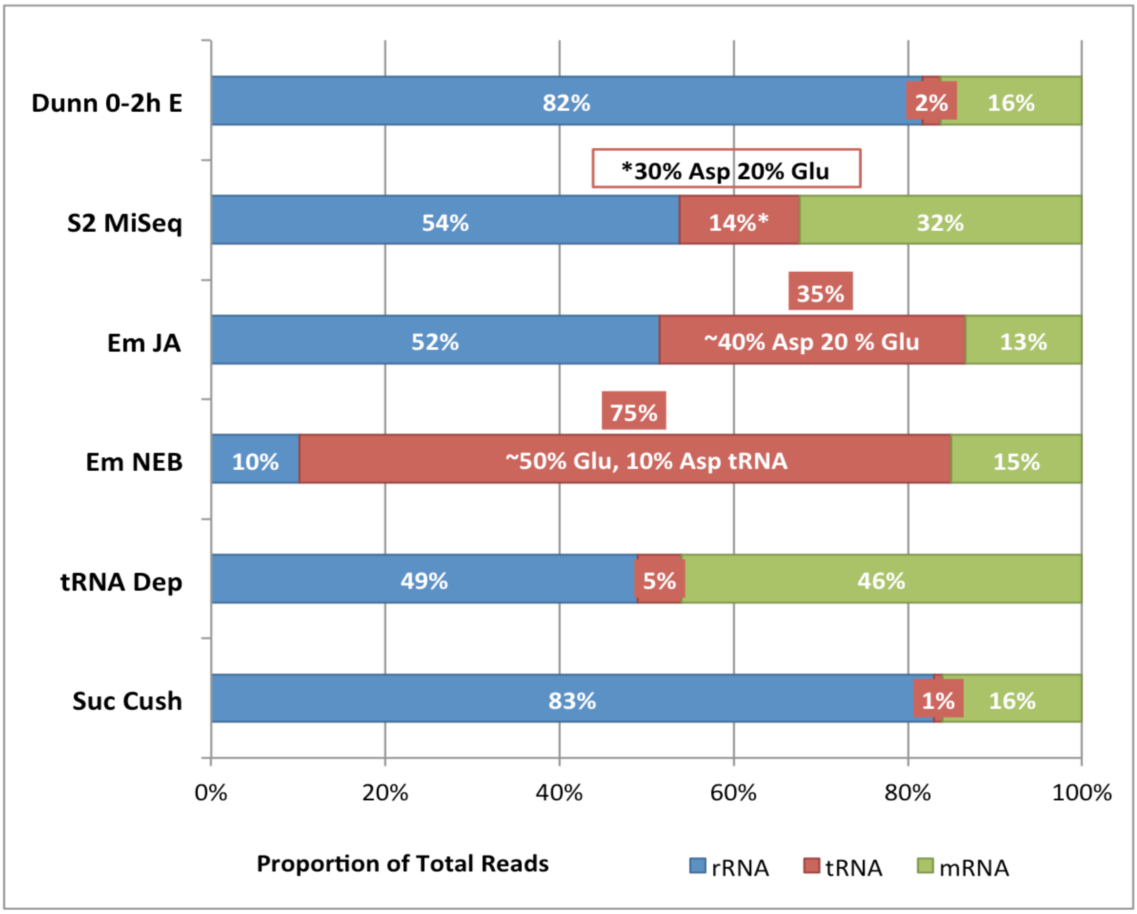
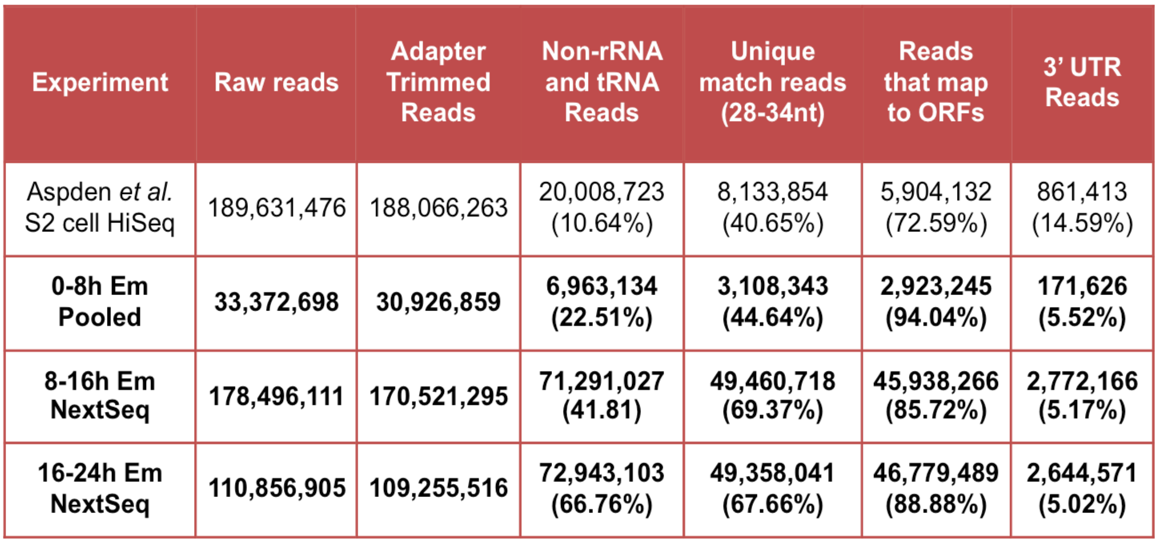


Figure 4.13 Proportion of depleted rRNA and tRNA reads from various datasets

The graph presented here illustrates the proportion of adapter-trimmed reads that are aligned to rRNA, tRNA and mRNA. To compare the various library preparation methods, the Dunn *et al.* (2013) 0-2H embryo sample, as well as the *S2* cell rRNA depletion sample are included. As can be seen from the numbers, one the main contaminants in *S2* cell Poly-Ribo-Seq is actually tRNA (14% of all reads, of which 50% are represented by tRNA^{Asp} and tRNA^{Glu}). The 0-2H Dunn *et al.* embryo sample's main contaminant is actually rRNA (82%) and tRNA represents only 2%. Of the 85% rRNA/tRNA contamination in Em JA, based on the method of Aspden *et al.*, 35% is attributed to tRNA, of which 40% is tRNA^{Asp} and 20% is tRNA^{Glu}. The Em NEB sample has an overwhelming proportion of tRNA contamination (75% of total 85% contamination by rRNA/tRNA) and the proportion of tRNA^{Glu} and tRNA^{Asp} is reversed with 50% tRNA^{Glu} and 10% tRNA^{Asp}. Using subtractive hybridisation to deplete the tRNA^{Glu} and tRNA^{Asp} reduced total tRNA contamination down to only 5% of all reads in the tRNA Dep embryo sample. We also attempted the Dunn *et al.* (2013) protocol using the sucrose cushion (SucCush) as the 0-2H embryos only had 2% tRNA contamination. The 0-8H embryo SucCush sample showed a similar reduction of tRNAs (1% of all reads). However, this sample has a massive proportion of rRNA contamination (83% of all reads), which we decided to tackle with the rRNA depletion beads used in the Aspden *et al.* (2014) protocol.



Experiment	Raw reads	Adapter Trimmed Reads	Non-rRNA and tRNA Reads	Unique match reads (28-34nt)	Reads that map to ORFs	3' UTR Reads
Aspden *et al.* S2 cell HiSeq	189,631,476	188,066,263	20,008,723 (10.64%)	8,133,854 (40.65%)	5,904,132 (72.59%)	861,413 (14.59%)
0-8h Em Pooled	**33,372,698**	**30,926,859**	**6,963,134 (22.51%)**	**3,108,343 (44.64%)**	**2,923,245 (94.04%)**	**171,626 (5.52%)**
8-16h Em NextSeq	**178,496,111**	**170,521,295**	**71,291,027 (41.81)**	**49,460,718 (69.37%)**	**45,938,266 (85.72%)**	**2,772,166 (5.17%)**
16-24h Em NextSeq	**110,856,905**	**109,255,516**	**72,943,103 (66.76%)**	**49,358,041 (67.66%)**	**46,779,489 (88.88%)**	**2,644,571 (5.02%)**

Table 4.3 Final Sequencing Results of Embryo Poly-Ribo-Seq

This table presents the final results of all three stages of Poly-Ribo-Seq conducted on embryos. The Aspden *et al.* (2014) *S2* cell results are included for comparison. The individual column treatments are described in Table 4.2. The 0-8H embryo sample now includes all of the 0-8H experiments (Em JA, Em NEB, tRNADep and SucCush) for a total of over 33 million raw reads. Of these, approximately 23% are non-rRNA and tRNA reads, of which 45% are uniquely mapped to the. The fact that 94% of these can be mapped to ORFs, while only 6% are found in the 3'UTRs shows that these experiments are quite successful in the removal of background in these non-coding regions, which is a significant improvement from the Aspden *et al.* (2014) protocol which results in almost 15% of reads mapping to 3' untranslated regions. The 8-16H and 16-24H samples were run on the NextSeq and have considerably higher numbers of Raw reads (110-178 million reads) From these samples, which were made following the final optimised protocol, we can see a marked increase of non-rRNA and tRNA reads (42-67%) of which almost 70% of reads uniquely map to the transcriptome. Up to 89% of these map to ORFs, and signal in the 3'UTR remains close to 5%.

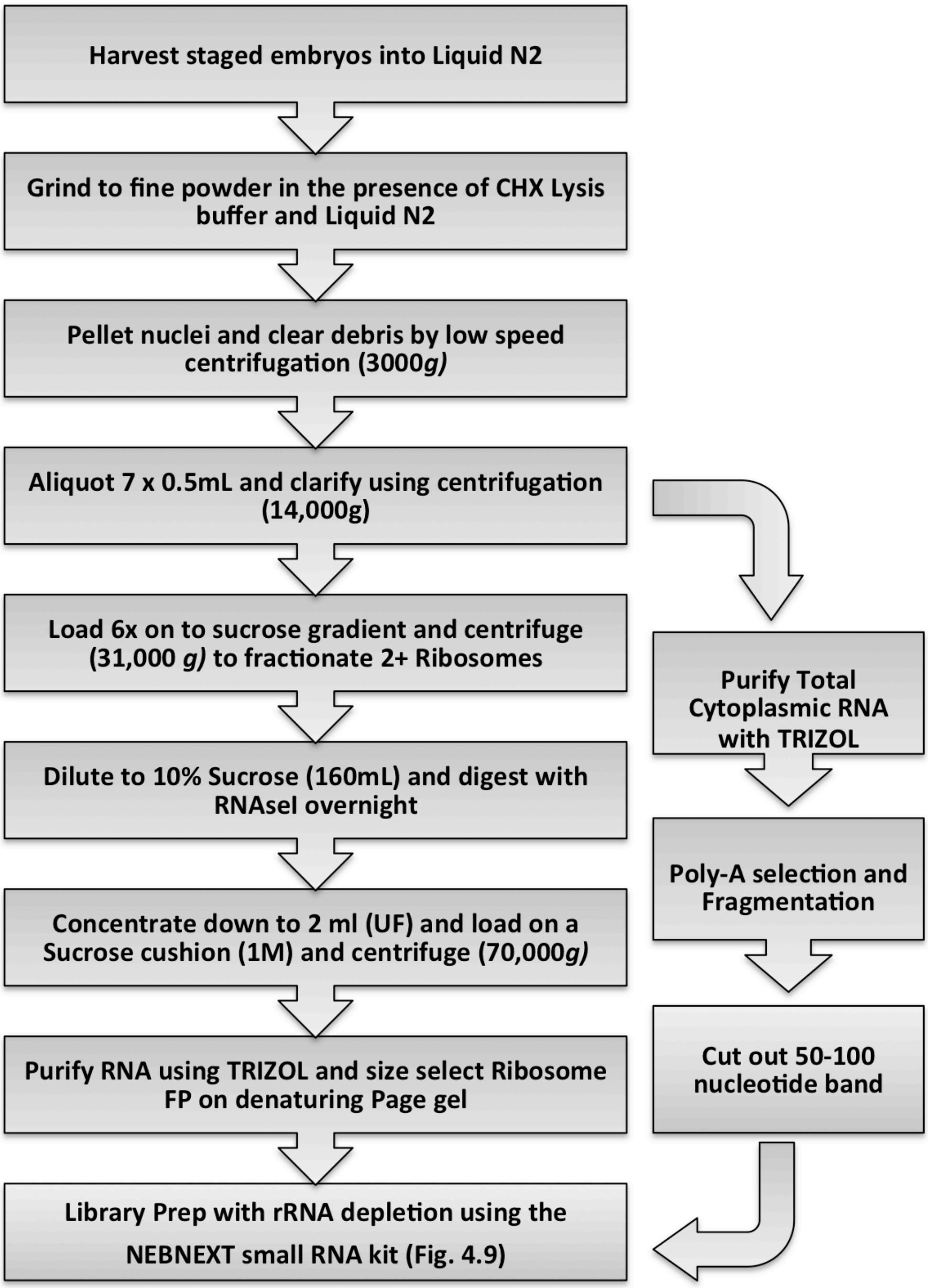
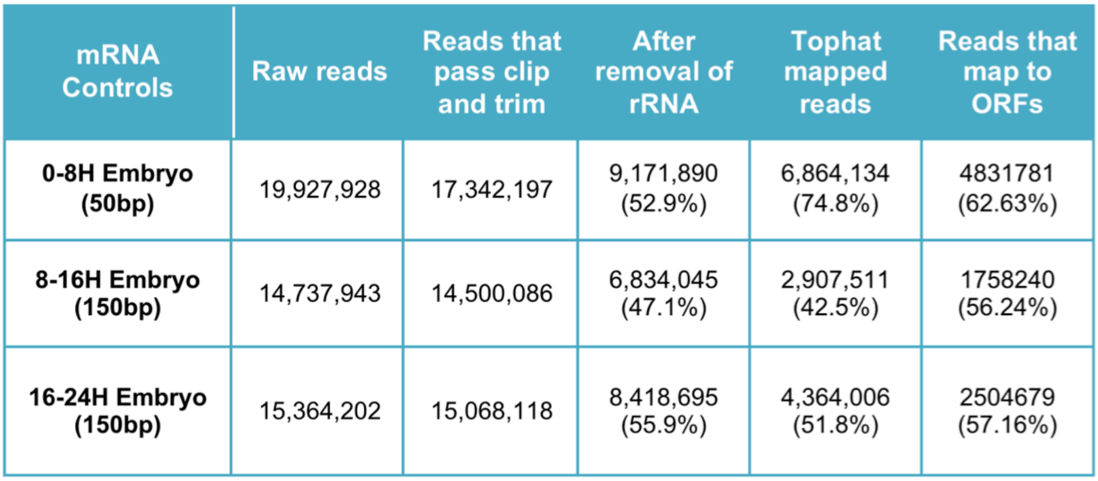

Harvest staged embryos into Liquid N2
Grind to fine powder in the presence of CHX Lysis buffer and Liquid N2
Pellet nuclei and clear debris by low speed centrifugation (3000*g)*
Aliquot 7 x 0.5mL and clarify using centrifugation (14,000g)
Load 6x on to sucrose gradient and centrifuge (31,000 *g)* to fractionate 2+ Ribosomes
Dilute to 10% Sucrose (160mL) and digest with RNAseI overnight
Concentrate down to 2 ml (UF) and load on a Sucrose cushion (1M) and centrifuge (70,000*g)*
Purify RNA using TRIZOL and size select Ribosome FP on denaturing Page gel
Library Prep with rRNA depletion using the NEBNEXT small RNA kit (Fig. 4.9)
Purify Total Cytoplasmic RNA with TRIZOL
Poly-A selection and Fragmentation
Cut out 50-100 nucleotide band

Figure 4.14 Overview of final Poly-Ribo-Seq Workflow

This figure represents the optimised Poly-Ribo-Seq protocol, used exclusively for 8-16H and 16-24H embryo samples in this thesis, and optimised on the different 0-8H embryo samples. Compared to previous protocols, this includes the method for harvesting and lysis of tissue samples (as opposed to cells only), the addition of a Sucrose Cushion step, the use of the NEB library preparation kit and two rounds of rRNA depletion.



mRNA Controls	Raw reads	Reads that pass clip and trim	After removal of rRNA	Tophat mapped reads	Reads that map to ORFs
0-8H Embryo (50bp)	19,927,928	17,342,197	9,171,890 (52.9%)	6,864,134 (74.8%)	4831781 (62.63%)
8-16H Embryo (150bp)	14,737,943	14,500,086	6,834,045 (47.1%)	2,907,511 (42.5%)	1758240 (56.24%)
16-24H Embryo (150bp)	15,364,202	15,068,118	8,418,695 (55.9%)	4,364,006 (51.8%)	2504679 (57.16%)

Table 4.4 RNA-Seq results of mRNA controls

RNA-Sequencing is performed on Poly-A selected, fragmented mRNA from each sample that is subjected to Poly-Ribo-Seq in order to normalise Ribo-Seq footprint abundance to the abundance of mRNA in the sample. This table outlines the processing of the RNA-Seq data on mRNA controls for each stage. Similar to the Poly-Ribo-Seq results, the RNA-Seq data must be processed and filtered to obtain the final reads which map to ORFs, in order to perform ratio calculations of Ribo-Seq/RNA-Seq to illustrate the abundance of the ORFs that are undergoing translation. This process entails subjecting the raw reads of the sample (Column 2) to pass clip-and-trim, which removes adapter sequences and any reads below a ‘Q’ score of 33, which is the standard for RNA-Seq (Column 3). Ribosomal-RNA reads are then removed (Column 4) and the remaining reads are run through TOPHAT, which aligns them to the transcriptome (Column 5). Finally, reads that are mapped to ORFs are calculated (Column 6) for further analysis of Poly-Ribo-Seq data.

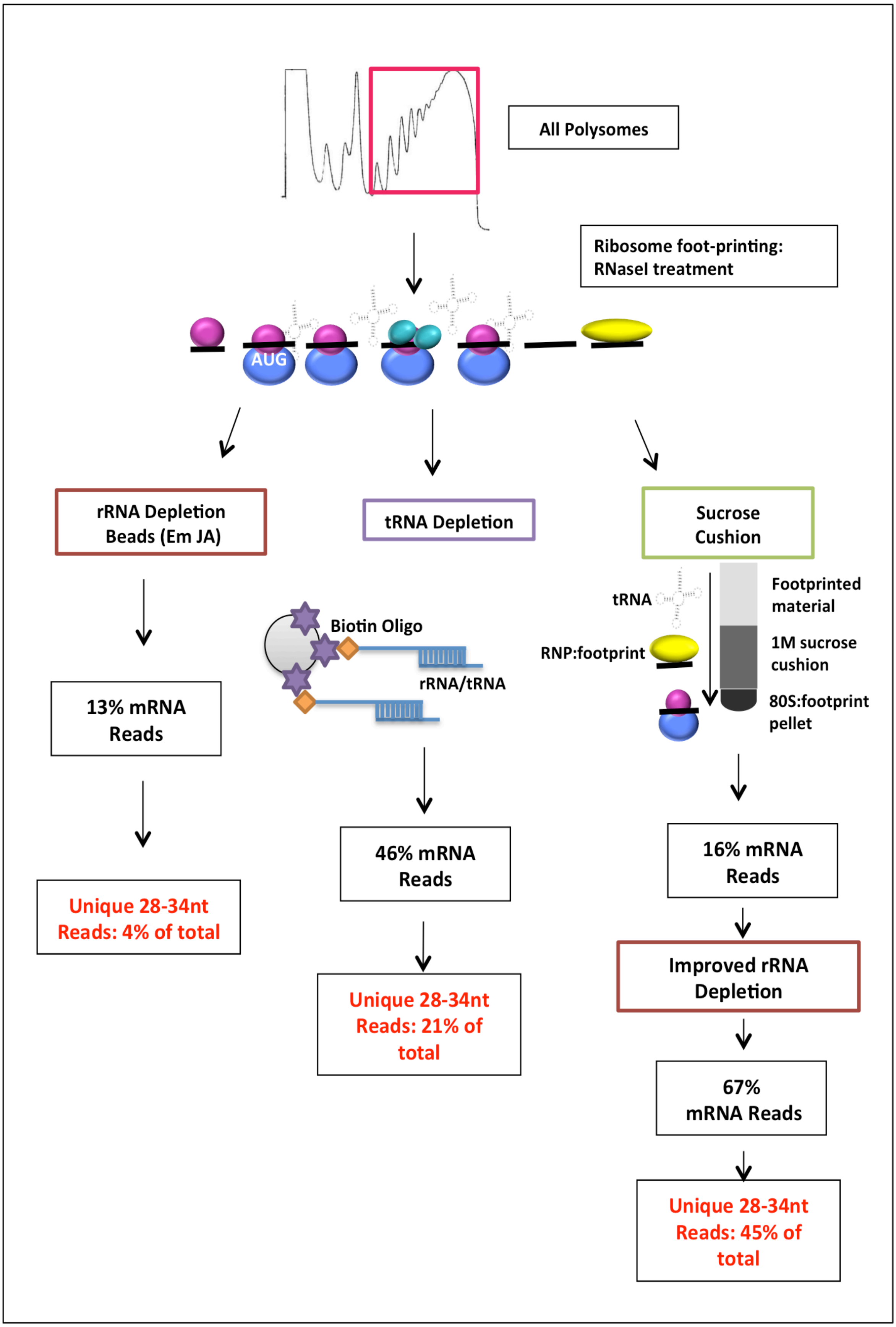

All Polysomes
Ribosome foot-printing:
RNaseI treatment
AUG
rRNA Depletion
Beads (Em JA)
tRNA Depletion
Sucrose
Cushion
13% mRNA
Reads
Unique 28-34nt
Reads: 4% of total
Biotin Oligo
rRNA/tRNA
46% mRNA
Reads
Unique 28-34nt
Reads: 21% of
total
tRNA
RNP:footprint
Footprinted
material
1M sucrose
cushion
80S:footprint
pellet
16% mRNA
Reads
Improved rRNA
Depletion
67%
mRNA Reads
Unique 28-34nt
Reads: 45% of
total

Figure 4.15 Overview of Poly-Ribo-Seq improvements described in Chapter 4

This figure shows the various methods that were tried in order to optimise and adapt Poly-Ribo-Seq to be used on *Drosophila* embryos. Firstly, to retain more material for performing the experiment, All polysomes were used instead of 2-6 (small) polysomes. By observing the presence of 6 specific rRNA sequences, we were able to significantly improve rRNA depletion as well, which when combined with the Sucrose Cushion method, gave us unprecedented proportions of usable reads to enable the assessment of smORF translation. The addition of a Sucrose Cushion allowed almost complete removal of the large proportion of tRNA reads which were present in the NEB library and when combined with two rounds of rRNA depletion, improved the result from the initial protocol 11-fold, from 4% unique 28-34nt reads in Em JA (following the old Protocol to 45% in the optimised version of Poly-Ribo-Seq.

Chapter 5: Analysis of Poly-Ribo-Seq Data and Defining smORF Translation

Introduction

Annotation of smORFs and long non-coding RNAs

Next Generation Sequencing (NGS) has provided us with the capability to sequence genomes in a high throughput manner and has given rise to the field of genomics (Margulies *et al.* 2005). Protein-coding genes were defined previously by the presence of Open Reading Frames (ORFs) and therefore encoding functional proteins. Computational examination of the first sequenced eukaryotic genome (*Saccharomyces cerevisiae*) uncovered the presence of hundreds of thousands of small ORFs (smORFs) through out the genome (Basrai *et al.* 1997), in numbers that were several orders of magnitude higher than the 6000 protein-coding genes that had been previously estimated (Goffeau *et al.* 1996).

Due to the extremely large numbers of smORFs, an argument was made that ORFs smaller than 300 nucleotides are aberrations that have a high statistical probability of randomly occurring in the genome. The small size of smORFs and smORF-encoded peptides makes it extremely challenging to characterise their translation and function, both bioinformatically and through traditional biochemical means (as discussed in Chapter 3). Consequently the majority of smORFs, which do not have experimental evidence of function or homology with other protein-coding genes, were simply discarded from genome annotations despite the fact that several examples of functional SEPs did exist at the time (Andrade *et al.* 1997); thus setting a precedent for future genome annotations in other organisms.

Recent RNA-Seq estimates have revealed that up to 85% of the mammalian genome can be transcribed (Djebali *et al.* 2012; Hangauer *et al.* 2013), which is far greater than the currently estimated 3% that accounts for annotated protein-coding genes (Kapranov, and St Laurent 2012). Most examples of smORF translation and functionality have either been based on obvious homology to other protein coding genes or through serendipitous discovery of mutant phenotypes. Candidate smORFs for in-depth functional characterisation in our lab have also been chosen based on a previous indication of translation or function (Galindo *et al.* 2007; Magny *et al.* 2013; Pueyo *et al.* 2016). Therefore, in the absence of biochemical data or a systematic assessment of smORF translation, functional smORFs are traditionally dismissed as the exception rather than the rule, and are therefore not considered as a class of protein-coding genes despite their high numbers in eukaryotic genomes (Mumtaz, and Couso 2015).

The discovery of non-coding RNAs has revolutionised our understanding of the diverse regulatory functions that can be performed by RNAs through a variety of mechanisms (reviewed in Cech and Steitz (2014)). Long-ncRNAs (lncRNA), which are defined as being longer that 200nt in length, have previously been shown to perform a veriety of functions. From our lab's point of view, they are particularly interesting as they are capped, spliced and polyadenylated like mRNAs (Guttman, and Rinn 2012), in the same size-range as smORF transcripts and are currently estimated to number about 10,000 in mammals and 2,000 in *Drosophila* (Young et al. 2012). The majority of the earlier research on long-ncRNAs has focused on their role in transcriptional regulation (Guttman, and Rinn 2012) as they were initially believed to be mainly localised to the nucleus. However recent studies have shown that they are also present in the cytoplasm displaying inclusively, in some cases, expression enrichment in cytoplasm (Ulitsky, and Bartel 2013) and association with ribosomes (Wilson, and Masel 2011; van Heesch *et al.* 2014). Some known non-coding functions of cytoplasmic lncRNAs include roles in mRNA translation elongation, regulation of miRNA levels through titration (molecular sponges) (Ulitsky *et al.* 2011), as well as regulatory effects on mRNA stability and cap-independent translation (reviewed in Fatica and Bozzoni (2013)).

Nevertheless, most (if not all) lncRNAs contain smORFs many of which appear to be translated as shown by ribosome profiling (Ingolia *et al.* 2011; Crappé *et al.* 2015; Duncan, and Mata 2014; Aspden *et al.* 2014). Furthermore, the functionally characterised examples of lncRNAs form a small proportion of the total number of lncRNAs detected and it is unclear how many of these novel transcripts that are annotated as putative non-coding RNAs are actually mis-annotated smORF transcripts because they do not contain ORFs that are longer than 100 codons. As such, there is great interest in ascertaining the role(s) lncRNA genes may play in the cell, and whether the functional unit is the RNA transcript or a translated peptide or both (reviewed in Housman and Ulitsky (2015)).

Ribo-Seq and translation outside of annotated protein coding regions

Approximately half of the long ncRNAs in mouse embryonic stem cells have ribosome profiling reads that can be mapped to them (Ingolia *et al.* 2011). The same can be said for an abundant pool of long ncRNAs in flies (~34%) (Aspden *et al.* 2014) and in zebrafish (~14%) (Bazzini *et al.* 2014), though the proportion is lower for the latter. Some of these lncRNAs may be acting as regulatory molecules at the RNA level (see above), but others contain smORFs that have been shown to be translated by other methods (Galindo *et al.* 2007; Magny *et al.* 2013; Aspden *et al.* 2014; Bazzini *et al.* 2014; Duncan, and Mata 2014; Slavoff *et al.* 2014) and definitely produce peptides with biological functions (reviewed in Andrews and Rothnagel (2014) and Saghatelian and Couso (2015)). Although the proportion of lncRNAs defined as translated varies depending on the organism, technique and analysis, overall it can be said that a significant proportion of lncRNAs show the presence of Ribosome Bound Fragments (RBFs) within the multiple putative ORFs located in lncRNAs. These tend to be very short ORFs with a median SEP length between 20 to 30 amino acids and along with uORFs (Median length 20aa) are referred to as dwarf smORFs (Bazzini *et al.* 2014; Duncan, and Mata 2014; Aspden *et al.* 2014; Ji *et al.* 2015).

Ribosome profiling has enabled the deep study of the translatome, highlighting its spatio-temporal regulation. The extensive translation of small ORFs has been shown by an evergrowing number of Ribo-Seq publications in the last few years. These studies describe that this translation is not limited to just ORFs in non-coding RNAs and 5'UTRs of canonical protein-coding transcripts (uORFs), but they can also be found in 3'UTRs (dORFs), as well as within overlapping longer annotated protein coding sequences in an alternative reading frame to the canonical CDS (altORFs) (Michel *et al.* 2012) (reviewed in Ingolia (2014)). The most well known function of uORFs is regulation of translation of the longer downstream ORF through the recruitment and/or stalling of scanning ribosomes. This ribosomal association was not thought to result in actual translation of biologically relevant uORF peptides (reviewed in Barbosa *et al.* (2013)). However, along with translation evidence from ribosome profiling studies, the translation of uORFs (Oyama *et al.* 2004) (Chapter 3) and many of other novel smORFs has been corroborated by peptide detection studies using mass spectrometry and tagging methods (Slavoff *et al.* 2013; Crappé *et al.* 2013; Vanderperre *et al.* 2013; Ma *et al.*

2014; Aspden *et al.* 2014) showing that the mapping of Ribo-Seq reads to these regions does result in the production of detectable peptides.

The extensive translation detected by Ribo-Seq outside of annotated protein-coding regions has generated discussion and controversy regarding the level of background signal from ribosome profiling, and whether the RBFs that are detected are an indicator of biologically relevant protein production (Guttman, and Rinn 2012; Bánfai *et al.* 2012; Chew *et al.* 2013; Ingolia *et al.* 2014). The smORFs in lncRNAs display fewer Ribo-Seq and RNA-Seq reads (due to their very short length and low abundance), which could perhaps indicate that this is 'noisy' molecular interaction with little biological relevance due to low levels of actual translation. Even with the application of ribosome profiling and the clear mapping of RBFs to smORFs, there is still a lingering lack of consensus within the field as to whether these results indicate biologically relevant translation (Ingolia 2014; Brar, and Weissman 2015; Ji *et al.* 2015) even though several examples of functional SEPs encoded by putative smORFs in lncRNAs have been published (Galindo *et al.* 2007; Magny *et al.* 2013; Pauli *et al.* 2014) and are reviewed in Saghatelian and Couso (2015)).

In order to elucidate upon the aforementioned controversy, experimental refinements on Ribo-Seq, such as Poly-Ribo-Seq (as previously discussed) and the addition of various translation inhibitors have been introduced to the protocol to decrease the effects of non-translating ribosome association with mRNA as well as background from established non-translating regions (Ingolia *et al.* 2011; Aspden *et al.* 2014; Popa *et al.* 2016). The use of harringtonine and similar drugs allow trapping of ribosomes at start codons during the transition between the initiation and elongation phases of mRNA translation. This technique enables the identification of uORFs in up to 60% of canonical coding genes, alternative translation initiation sites, such as 5'-terminally extended or downstream internal AUG codons found within many ORFs, as well as any non-AUG start codons (Ingolia *et al.* 2011). However, defining novel coding sequences from ribosome profiling data still remains the fundamental challenge in characterising this phenomenon of widespread translation (Bazzini *et al.* 2014; Chew *et al.* 2013; Guttman *et al.* 2013; Ingolia *et al.* 2014; Ji *et al.* 2015) (reviewed in Brar and Weissman (2015)).

This chapter will provide an introduction to some of the computational metrics that have been developed in order to discern meaningful translation from Ribo-Seq noise. Overall, these metrics should allow for the robust identification of canonical translated sequences as well as smORFs. However, the translation of many dwarf smORFs remains controversial due to their small size and hence low number of reads mapping to these novel ORFs (reviewed in Mumtaz and Couso (2015)).

This chapter details the analysis performed on the Poly-Ribo-Seq data generated from *Drosophila* embryos (from Chapter 4) in order to ascertain the translation of smORFs using the bioinformatics metrics introduced in this section. Particularly, there is a focus on highlighting the differences between canonical protein coding sequences and the very short ‘dwarf’ smORFs and their impact on the adjustments of computational scoring of translation based on Ribo-Seq Data. The translative properties of the different kinds of ORFs expressed across *Drosophila* embryonic development are also described. Finally, an improved cut-off for translation is described, in view of comparison to our previously published Poly-Ribo-Seq data pipeline (Aspden *et al.* 2014).

Results

Identification of novel smORFs

We used the EMBOSS getORF program to identify all smORFs in *Drosophila* 5'UTRs and lncRNAs that begin with an AUG start codon. A minimum ORF length of 30nt (10aa) was used to remove extremely small ORFs, which are extremely likely to appear in the *Drosophila melanogaster* genome by chance. Using this tool, we identified 21,431 unique novel smORFs, located in the 5'UTRs of 9031 transcripts and belonging to 3,621 genes. We also identified 24,878 unique ORFs in 2,548 lncRNA transcripts, arising from 2,159 lncRNA genes. In this work, and for the purpose of stringency and the sake of simplicity, I do not analyse smORFs that occur in a different frame of translation within annotated protein coding regions nor at smORFs located in 3'UTRs (reviewed in Ingolia (2014)).

Ribo-Seq Data Preparation and Mapping of Sequencing Reads

Although standard sequencing analysis programs can be used for mapping RBFs to the reference transcriptome, the process is more challenging due to the nature of the Ribo-Seq libraries. As discussed in the previous chapter, Ribo-Seq data contains a large number of rRNA/tRNA reads that have to be removed prior to analyses. Secondly RBFs are very short in length; therefore it can be difficult to map them to the reference genome without losing a significant proportion of reads to pseudogenes and repetitive sequences (~25% RBF reads, Table 5.1). Furthermore, it is difficult to obtain isoform-level information, as there are fewer exon-junction spanning reads as compared to the level found in paired-end RNA-Seq (Ingolia 2014).

Sequencing data is typically released as a FASTQ file containing the raw sequencing reads. This file is then processed using the FASTX command line toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), in order to discard low quality reads and trim library preparation adapter sequences from the sequencing reads (outlined in Figure 5.1). The trimmed reads are then matched against a pre-computed file containing annotated rRNA and tRNA sequences for our reference genome. This is accomplished using *Bowtie*, a short read aligner algorithm (Langmead *et al.* 2009). Following this step, the remaining non-rRNA reads are mapped to the *Drosophila melanogaster* genome using Bowtie and any unmapped reads are then aligned to the transcriptome using *TopHat*, a splice junction mapper (Trapnell *et al.* 2012). We chose a conservative approach to our analysis, deciding to discard any reads that map to more than one region in the genome and only keep uniquely matched reads; this allowed a more accurate quantification of the data by discarding reads mapping to pseudogenes or repetitive sequences. Additionally, a mismatch of 1nt per read was permitted to allow for single nucleotide polymorphisms (SNPs) and sequencing errors. The *TopHat* program outputs a machine-readable binary alignment (BAM) file, which contains the mapping information for each read and can be visualised using a Genome Browser. The BAM file is then filtered to exclusively retain reads that are within the 28-34 nucleotides length range using the SAMtools program (Li *et al.* 2009), corresponding to the estimated sizes of RBFs (Ingolia *et al.* 2009).

Quantifying Ribo-Seq signal and Normalisation methods

After the mapping of unique reads, abundance of transcripts can be estimated by matching the BAM file against a BED file, which is a text file containing the annotation and genomic co-ordinates of each type of feature (5'UTR, ORF and 3'UTR) in a transcript. In our pipeline we use 3 different BED files containing: 1) the Flybase annotated protein-coding transcripts (FBcds and FBsmORFs), 2) the putative ORFs identified in non-coding RNA transcripts (ncrORFs) and 3) ORFs located in the 5'UTRs of annotated protein coding genes (uORFs).

For polycistronic gene models such as uORFs and ncrORFs, the 3'UTR is defined as the region downstream of the ORF until the start of the downstream ORF and *vice versa* for 5'UTRs. The *Cufflinks* algorithm is used to output a data frame with the estimated abundance in Reads Per Kilobase per Million (*RPKM*) rather than using an absolute count of sequencing reads. The *RPKM* (Reads – with single end sequencing) or *FPKM* (Fragments in Paired-end sequencing) metrics were initially developed for RNA-Seq to measure the relative abundance of each transcript within the dataset. *RPKM* abundance measurements are normalised for length, as longer transcripts will generate more fragments and therefore, more reads that can align to it.

Unlike RNA-Seq, RBF reads are mainly localised to coding sequences due to the fact that the reads are not generated by a random fragmentation of RNA transcripts; rather, each read represents a translating ribosome. Therefore RBF read counts are also normalised for length due to the fact that longer CDS's are able to accommodate a larger number of translating ribosomes and thus generate more reads. Due to the fact that sequencing provides a random sampling of all the fragments present in the library, read counts are also normalised to the total number of sequencing reads in the sequencing dataset. Therefore the *RPKM* metric calculates the normalised, relative abundance of each feature within the sample (Mortazavi *et al.* 2008). Thus, more complexity of the initial mixture used to generate the sequencing library lowers the overall *RPKMs* of individual transcripts. This is a key technical aspect, particularly when looking at translational regulation of samples placed under stress conditions, where a global reduction in translational levels may be misinterpreted as evidence for increased levels of translation of a few genes (Wagner *et al.* 2012).

Normalisation with respect to length is a necessary step, however, it is not ideal for very short features such as ncrORFs. Normalising by feature length does not have a large impact when comparing between similarly sized transcripts or canonical CDS's, but has its limits when we compare features that vary vastly in their length. For example the average CDS length (2.076 bp) of a canonical protein coding transcript (http://flybase.org/static_pages /docs/release_notes.html) must have ~35 times more reads than an average ncrORF (60 bp) in order to attain the same *RPKM* value within a given dataset due to normalisation per Kb. Therefore variation of even a few reads mapping to an ncrORF can have a large effect on the *RPKM* value whilst it may not have a large impact on a longer ORF which will accrue many more reads for the same value.

Bioinformatic tools such as *Cufflinks*, which are designed with more typical RNA-Seq analyses in mind *e.g.* assigning reads to differentially expressed transcripts (Trapnell *et al.* 2012), typically assume longer features and read lengths than in the analysis. One such example is the effective length correction option, which is default to *Cufflinks* and is meant to correct for fewer reads originating from the ends of transcripts (known as the "edge effect" (Trapnell *et al.* 2009)). Therefore in our case, where we inquire about the transcription and translation of very short features, the edge may represent a large proportion of the feature, leading to inflated *RPKM* values, thus this option was disabled in our data analysis. Furthermore, *Cufflinks* uses a probabilistic model to distribute reads between individual transcript isoforms (Trapnell *et al.* 2009). The short length of Ribo-Seq reads mean that there are fewer isoform-specific and exon-exon boundary spanning reads, making it difficult to obtain isoform-specific information (Ingolia 2014). The implication of this for our dataset was that the already few reads that may be aligning to an ncrORF are further divided between multiple transcripts, making it more difficult to assess the translation of that ORF. Therefore *RPKM* was calculated separately for the 5'UTR, ORF and 3'UTRs of a transcript.

Alternatively, the average read count of aligned reads can be calculated using BEDTools, which calculates *coverage* in *Reads Per Base pair (RPB)* (Quinlan, and Hall 2010). The RPB metric corresponds to the number of times each nucleotide of a given genomic feature is covered by experimental sequencing reads, which is then averaged for all nucleotides in the feature in question (CDS/ORF, 5'UTR or 3'UTR) to give a *RPB* value for that feature. This allows for normalisation of the metric by length, and removes bias based upon length as a longer feature will generate more fragments and hence have a larger number of reads aligning to it. The only problem with the *RBP* metric is that it does not normalise to the sequencing depth of the sample, so the *Em1* dataset (with ~3 Million reads) will have a much lower *RBP* coverage than the *Em2* and *Em3* datasets (with ~50 Million reads - see Table 5.1), therefore making them difficult to compare. As such it is important to normalise the *RBP* value to the total number of reads in each dataset by calculating *RBPM*, or *Reads per Base Pair per Million*.

In contrast using *Cufflinks* to estimate abundance in *RPKM*, a consequence of using BEDtools to calculate abundance based on coverage is that each read is calculated multiple times for each overlapping transcript; which can affect the accuracy of abundance measurements. Since RBF reads mainly map to coding sequences, alternative transcript models are not relevant to abundance estimations. The FlyBase annotation of the reference sequence in *Drosophila,* where each transcript of a gene is given a unique Flybase transcript ID (FBtr) and each peptide produced from a transcript is given a corresponding unique Flybase polypeptide ID (FBpp) can lead to inflation of ORFs being perceived as translated. For example, there are 19,915 unique FBpp ID's transcribed at the *Em1* stage but these correspond to only 14,466 unique polypeptide sequences as determined by their CDS coordinates. This was even more relevant for lncRNAs genes with multiple transcripts, as many putative ncrORFs were annotated per transcript and the 3,078 unique ORF ID's actually encode for only 1,807 unique peptides. Therefore, in order to improve the representation of the data and minimise inflation of read counts, I decided to exclusively count unique genomic CDS co-ordinates as separate peptides (Figure 5.2).

Secondly, even after normalisation, *RBPM* numbers are generally very low which can make it an inconvenient measure for cross-sample comparisons. Thus we decided to use *RPKM*, which is the most commonly used abundance measurement for RNA-Seq data and makes it easier to compare with the *S2* cell and other datasets. In order to maximise the number of reads to unique ORFs, we calculated *RPKM* by multiplying the *RBP* values obtained using BEDtools by 1000 (Kilobase) and dividing the resulting value by the total number of reads in the BAM file.

Considerations for Sequencing Depth

RNA-Seq

Sequencing coverage is an important consideration for NGS experiments. Coverage of sequencing is calculated by multiplying Read Length by the Number of Reads divided by the estimated size of the transcriptome (Read length x Number of Reads / Size of transcriptome) (Illumina Technical Note). This metric was developed for the analysis of whole-genome sequencing results and is based on the reasoning that sequencing reads are not distributed evenly over an entire genome, simply because the fragmented genome is sampled in an independent and random manner (Lander, and Waterman 1988). Therefore, due to the effects of sampling, some regions will be sequenced less frequently than the average, while other regions will be over-represented in the same sequencing data. In RNA-Seq experiments, this matter is complicated by the considerable difference in abundance of different transcripts. Therefore it is generally recommended to sequence at a depth of around 10x coverage of the transcriptome, depending on the application of the sequencing (Illumina Technical Note). For example, an in-depth study aimed at discovery of novel transcripts or transcript isoforms may wish to sequence well beyond 10x coverage whereas a study aimed at looking at general transcriptional changes to more abundant transcripts, like those in a mutant or before/after treatment, can be satisfied with a lower sequencing coverage.

Our control RNA-Seq libraries were sequenced in-house using an Illumina MiSeq machine, which generally has a much lower output than what is now considered a high depth of sequencing (Table 5.2). The *Em1* dataset was sequenced on a Version 2 cartridge, which has a maximum capacity of 15M reads of up to 50bp in length. The *Em2* and *Em3* mRNA libraries were pooled together and run on a Version 3 MiSeq cartridge which has a maximum capacity of 25 M reads up to 150bp in length (Median Read length: ~70bp) reads. Figure 5.3 shows all the of the read length distributions from RNA-Seq and Poly-Ribo-Seq for each experiment. These sequencing runs yielded 2.9M (*Em2),* 4.4M *(Em3)* and 6.9M (*Em1*) uniquely mapping reads to the transcriptome (Table 5.2). As the estimated size of the *Drosophila melanogaster* transcriptome is 61Mbp (O'Neil, and Emrich 2013), the *Em1* and *Em3* datasets contain sequencing coverages of 5.7x and 5.0x, respectively, while the *Em2* dataset has 3.3x coverage. These figures reveal that the level of sequencing is lower than recommended into the transcription of lowly transcribed transcripts such as lncRNAs, which would be better resolved at 10x sequencing coverage. This is also particularly relevant in the case of dwarf smORFs that generate a relatively small number of reads due to their small size and may be discarded as un-transcribed despite signal in Ribo-Seq dataset. As detailed above, a low number of final reads leads to inflated *RPKM* values due to the normalisation by sequencing depth step, as each read carries more weight. All of these considerations can affect TE, which is calculated by dividing Ribo-Seq *RPKM* by the *RPKM* in the RNA-Seq, as discussed in further detail in the TE section below. Therefore, in order to deal with these issues, we have now sequenced the *Em2* and *Em3* mRNA control libraries at a higher depth, with 40M and 50M raw reads respectively. These datasets have not yet been analysed and therefore have not been presented in this chapter.

Poly-Ribo-Seq

Techniques such as Ribo-Seq that sequence specific, translated regions of the transcriptome require even greater depth of sequencing due to enormous variation in translational levels. Again, this is exacerbated if the biological question is meant to address small ORFs that generally struggle to accrue enough reads for statistical analysis of translation as has been previously discussed. In cases such as ours, 30-100x sequencing coverage is generally recommended (Illumina Technical Note). At the time of data analysis, the *Em2* and *Em3* Poly-Ribo-Seq datasets had ~50 Million reads mapping to the transcriptome, but a much lower number of ~3 Million reads for the *Em1* sample, which is comprised of the reads from 4 different optimisation runs pooled together (see Chapter 4).

We estimated the size of the *Drosophila melanogaster* translatome to be 41Mbp by multiplying the number of protein-coding exons by their average size as provided in the FlyBase release notes (http://flybase.org/static_pages /docs/release_notes.html). Using this estimate and an average RBF length of 31nt, I calculated the *Em2* and *Em3* datasets as having 35x coverage while the *Em1* dataset has only ~2.3x coverage, which is significantly lower than the other two. This means that we cannot use the same rigorous metrics to *Em1* as those applied to the other embryo datasets. In order to address this issue, I have now sequenced the sucrose cushion-prepared library from Chapter 4 on a NEXTSeq machine, which is expected to yield approximately 35M unique mapping reads, making *Em1* comparable to *Em2* and *Em3* datasets with approximately 25x sequencing coverage. Unfortunately these results are still underway at the time of writing this thesis so cannot be included for analysis.

Analysis of Embryo data using the published *S2* Data analysis pipeline

Defining Transcription in the RNA-Seq Data

We initially used the data analysis pipeline employed in Aspden *et al.* (2014) on the embryo datasets, which was implemented by a bioinformatician in the host lab, Dr. Ying Chen, to determine the number of translated ORFs in *Drosophila S2* cells. The adapter sequences were trimmed from the raw reads and then aligned to *Drosophila melanogaster* tRNA and rRNA sequences as described above. The remaining non-rRNA/tRNA reads were then aligned to *Drosophila melanogaster* genome using Bowtie short read aligner. Finally, unaligned reads that span mRNA splice junctions, were aligned to the transcriptome using *TopHat* a splice-aware short read aligner. The *RPKM* abundance metric was estimated using the *Cufflinks* program and *coverage* was calculated using BEDtools for the 5' UTR, CDS and 3'UTR of each transcript.

The transcriptional levels were determined in parallel, using the RNA-Seq dataset, which was analysed in the same way as described above for the Poly-Ribo-Seq data. We used an *RPKM* value of greater than 1 to define transcribed features, a cut-off that has been extensively used in published RNA-Seq data analysis (Aspden *et al.* 2014). Importantly, only ORFs that were defined as transcribed were used in further analyses that interrogated translation. However, upon closer inspection of the *Em1* data, we were able to observe some ORFs that were defined as untranscribed even though they displayed Ribo-Seq signal. This was due to the fact that our original pipeline uses an *RPKM*>1 cut-off across the ORF to determine transcription. This issue is again particularly problematic for very short novel ORFs that form a very small proportion of a transcript and therefore it is not surprising that a short ORF (median ~70nt) occurring in a much longer lncRNA (average size ~1 Kb) may not accrue any reads even though the transcript is expressed. Since RNA-Seq reads should be evenly distributed across the transcript and do not show a bias towards coding sequences as in Ribo-Seq, we decided to use the *RPKM*>1 cut-off across the entire transcript rather than the ORF to determine transcription.

Defining Translation based on the 90th percentile of 3'UTR *RPKM*

The datasets outputted by the analysis pipeline were analysed in R using RStudio, and the 90th percentile value of the *RPKM* and *coverage* for the 3'UTRs of annotated protein-coding transcripts, which can be considered background signal, was calculated as a cut-off for translation. However, the *coverage* metric is generally used in conjunction with *RPKM* measurements to score translated ORFs (Bazzini *et al.* 2014; Aspden *et al.* 2014). This metric describes the proportion of an ORF that is covered by RBF reads; the greater the coverage for a given ORF, the greater the likelihood that is translated, thus excluding from the analysis false positives that result from, for example, a pile-up of stalled, non-translating ribosomes at the start codon. In the Aspden *et al.* (2014) analysis of *S2* cells, we employed a *coverage* cut-off of 0.57 for novel ORFs which was calculated as maximum RBF length/median dwarf smORF length (34nt/60nt), meaning that the ORF should be covered by RBFs in more than one unique position.

The caveat of the *coverage* metric, however, is that it also depends on sequencing depth and sampling of the individual ORFs; as such, it is not suitable for inadequately sequenced stages. Consequently, given the lower number of reads in the *Em1* dataset (3M Reads), it is not practical to use this metric in this study. For example, Figure 5.4 shows that the *Em1* stage has a broad distribution of coverage for transcribed ORFs as it does not have abundant sequencing depth, however *Em2* and *Em3* transcribed ORFs have a tight distribution around the maximum coverage of 1.0 as they were sequenced at a high depth with 50 Million Reads each. Therefore coverage was not considered for further data analysis presented in this chapter, as there was such good depth of sequencing that use of a Coverage metric was deemed redundant.

The 90th Percentile of the *RPKM* values of 3'UTRs of annotated protein coding genes transcribed in each dataset were used to calculate translation cut-offs, which should represent untranslated features and therefore background signal. Given that there was a significantly lower percentage of 3'UTR reads (~5%) in Embryos as compared to the *S2* cell data (~14%), it was surprising to discover that the cut-off values generated were significantly higher in embryos (*RPKM* cut-offs *Em1*=15.19, *Em2*=14.79, *Em3*=10.58) than the value used in the *S2* cell dataset (*RPKM*=11.7). This was due to the fact that the dataset that was used to generate the *RPKM* translation cut-offs for the *S2* cell data was based on all annotated transcripts regardless of whether they were actually expressed in *S2* cells.

When we adjust the *RPKM* cut-off based on the 3'UTR signal of transcripts that were expressed explicitly in each stage, as that would be a more representative sample to use as a 'background' reading, the cut-offs are further inflated for the embryo datasets (*RPKM* cut-offs *Em1*=21.26, *Em2*=18.03, *Em3*=14.73). Interestingly, this formula retroactively applied to the *S2* cell data makes the *S2* cell *RPKM* cut-off three times higher than the lowest depth of sequencing sample (*RPKM*=62.12). Using these higher *RPKM* cut-offs I only observed translation of half of annotated protein-coding transcripts at each Embryo stage. This shows the impact of readjusting the *RPKM* cut-off to represent the expression of transcribed genes according to RNA-Seq data. We also recalculated for each data set, including *S2* cells, the number of transcribed ORFs based on an *RPKM*>1 across the whole transcript, not just the ORF, as was done for the *S2* cell data. These results also strongly indicated the need to create a more rigorous and reliable translation cut-off using *RPKM*.

In order to do this, factors such as gene models that may include overlapping CDS's and 3'UTRs based on alternative transcript isoforms were brought into consideration. For example, a gene may encode for a truncated protein isoform that is generated through alternative splicing and polyadenylation. In this hypothetical case, reads originating from the isoform encoding the full-length protein may be attributed to the 3'UTR of the transcript encoding the truncated isoform, due to identical sequence. This reads may thus be mis-attributed and therefore contributing to higher 3'UTR signal, which may increase the rate of false negatives. In order to address this issue, all 3'UTR sequences that overlap an annotated CDS were masked from the analysis, and this analysis was used to generate new *RPKM* translation cut-offs. This analysis brought down the cut-offs of the Embryo data (*RPKM* cut-offs *Em1*=18.76, *Em2*=16.70, *Em3*=12.72), but the *S2* cell *RPKM* cut-off remained very high (*RPKM*=58.96). These results are summarised in Table 5.3.

Using this 90th Percentile of chopped/masked 3'UTR *RPKM* cut-offs and only counting unique genomic CDS co-ordinates as separate peptides, I observed translation of 52-56% of canonical protein CDS's and 55-67% of annotated smORFs in embryos. ncrORFs were translated at an even lower level at only 1-6% and uORFs were translated at between 14-28% of transcribed. The effect of the new cut-off for the *S2* cell data is quite drastic compared to those published in Aspden *et al.* (2014), out of the 11,214 canonical protein FBcds's initially deemed translated with the old cut-off, only 3,287 remain (25.4% of transcribed). The number of FB smORFs are lowered to 173 from the 228 described in the paper (63.6% of transcribed), ncrORFs are lowered to 97 of the initial 313 (4.25% of transcribed) and uORFs are 524 from 2,708 (8.89% of transcribed). Therefore I considered alternative methods to use the 3'UTR based *RPKM* cut-off.

Categorisation of ORF Translation based on abundance

Since the 90th percentile cut-offs were more stringent than our initial expectations, I decided to relax the stringency of this cut-off by using the 90th, 80th and 70th percentiles as metrics to confer a high (>90th%), medium (80th to 90th %) and low (70th to 80th%) confidence of translation to each class of ORFs and to classify translated ORF based on their relative abundance. Table 5.4 is a master table that shows the cut-offs generated for each dataset (Grey) and the numbers of FBcds (Green), FBsmORFs (Blue), ncrORFs (Red) and uORFs (Purple) translated using theses cut-offs in each of these categories. The numbers in parenthesis show the proportion of transcribed ORFs translated in each category which, when added up give a more plausible representation of translation in the sample. This data corroborates the conclusion reached in Aspden *et al.* (2014) that FB smORFs are highly translated at a proportion similar to canonical protein-coding genes. These results also highlight a striking feature of lncRNA translation, which is that ncrORFs in the *Em2* stage are translated at a much lower combined proportion (4%) as compared to *Em3* (19%) and *Em1* (12%) stages. Surprisingly this proportion is half of what is observed in the *S2* cell line (8%), suggesting that this may be a result of active translational regulation at this stage as discussed in further detail later in this chapter.

These results can also be used to separate highly transcribed and translated genes and even gave clues to the measure of translational activity. Figure 5.5 shows the overlap between the three stages of embryogenesis for high (90th percentile) and medium (80th percentile) confidence translated annotated protein coding genes, smORFs and ncrORFs. For annotated protein coding genes and smORFs, the high confidence subset seems to mainly consist of genes that are translated at a high level or ubiquitously across embryogenesis with 43% (FBcds) and 53% (FB smORFs) being translated across all 3 embryonic stages, however there is significant stage-specific translation of each type of ORF, as well as between subsequent stages. The 80th-90th percentile translated genes appear to be mainly temporally restricted in their expression and translation as transcripts belonging to this medium confidence category are largely stage-specific, with only 9% FBcds, 0.5% FB smORFs and 1% of ncORFs translated across all 3 embryonic stages. The highest proportion of these medium-confidence ORFs are translated by each stage with little overlap between temporal stages.

The 90th percentile numbers and proportions of all ORFs translated throughout embryogenesis is quite low when compared with the patterns of transcription (Figure 5.6), where we observe 78.5% FBcds are transcribed throughout embryogenesis and 76% of FB smORFS are transcribed throughout embryogenesis, whilst only half are being translated. Also, though almost 46% of ncrORFs are transcribed throughout embryogenesis, just a third of these are being translated with high confidence. This can be interpreted as a lag, regulation or restriction of translation of the ORFs transcribed, not only of canonical protein-coding genes, but for all ORFs, in a similar fashion.

Overall, these findings led us to reconsider the value and weight of using the 3’UTR as a basis for generating cut-offs for three reasons: the observation that proportion of 3’UTR reads may differ based on the library preparation method used - as discussed in Chapter 4; the discovery of translated ORFs within 3’UTRs (Slavoff *et al.* 2013; Vanderperre *et al.* 2013; Ingolia *et al.* 2014) and lastly, the discovery of abundant stop codon read-through in *Drosophila* (Dunn *et al.* 2013), thus implying that some of the signal observed in 3’UTRs may arise from genuine translation events and not background. However this reconsideration of basing background signal and *RPKM* cut-offs on reads in the 3’UTR has only come to light after the extensive calculations and results that have been shown here.

Translation Efficiency

The *RPKM* metric does not differentiate between intensive translation of a moderately abundant transcript and low translation of a very abundant transcript. Furthermore, *RPKM* is only a broad approximation of translation in Ribo-Seq, as this is a snapshot of a dynamic system *i.e. RPKM* values for two mRNAs could be similar but one of them would not be translated as efficiently as the other if it has a pile-up of paused ribosomes. Thus, it becomes clear that it is important to normalise Ribo-Seq signal to transcript abundance. This process generates a metric known as Translation Efficiency (TE), which is the ratio of $RPKM^{Ribo\text{-}Seq}/RPKM^{RNA\text{-}Seq}$ for a given CDS and unlike *RPKMs*, this metric pinpoints genes that are undergoing translation in a regulated manner. TE is considered a good metric to evaluate the translatability of an ORF due to the normalisation of Ribo-Seq signal by mRNA abundance, as it allows for better correlation between datasets, and highlights any mRNAs that are translationally regulated.

TE performs quite well for inspecting the overall translation status of datasets (Kronja *et al.* 2014). However as TE is a ratio and doesn't factor in a the actual abundance of the ORF being scored, only CDS’s with a minimum number of reads in both the Ribo-Seq and RNA-Seq datasets are used (between 10–50 reads) in order for TE calculations to be statistically tractable. As shown in Figure 5.7, the TE for *Em1* and *Em3* CDS is higher than the TE of their 5'UTR and 3'UTR. The overall TE of CDS’s in annotated protein-coding transcripts at the *Em2* stage is significantly lower than at the *Em1* and *Em3* stages and also displays a lower degree of variation. Oddly, this value is comparable to the TE of 5’UTRs at *Em2*, which is higher than the TE of 5’UTRs in *Em1* and *Em3*. Our RNA-Seq libraries were prepared using PolyA selection, so there exists a bias towards the 3’ ends of the transcripts, particularly due to the capture by Oligo-dT beads of 5’ to 3’ partially degraded transcripts with their 3’ ends intact. The decreased signal of *Em2* CDS and increased signal of *Em2* 5'UTRs could be explained by greater translation regulation of these transcripts at this stage as discussed in greater detail later in this chapter. However, this difference is more likely overstated due to the low sequencing coverage of the *Em2* RNA-Seq dataset. In our RNA-Seq datasets, between 32-38% of the reads map to the 3’UTRs (Median length 630 nt) while only 4-7% of the reads map to the 5’UTRs (Median length: 340nt). This means that the low TE values for 3’UTR and the higher TE values in 5’UTRs may be exaggerated (Figure 5.7);

thus introducing a significant bias for uORFs, which are located in the 5’ ends of the transcripts, and therefore have lower RNA-Seq signal.

TE works well to assess translation of regular protein coding transcripts where the coding sequence comprises the majority of the transcript, however it is not as effective for discerning the translation of Dwarf smORFs such as ncrORFs (Median: 23aa) and uORFs (Median: 19aa), which may not accrue RNA-Seq reads as they form a very small proportion of the transcript. For example, the median length for all annotated protein coding sequences is ~1,500nt, corresponding to approximately 70% of the transcript (Median: 2,050nt). However, the median length of an ncrORF is 69 nt which forms only 4% of a ncRNA transcript (Median: 1,750nt). Coupled with the fact that long non-coding RNA transcripts tend to be expressed at low levels it is not surprising that a large proportion (~30% of transcribed) of the ncrORFs have either an artificially inflated TE value due to very low *RPKM* in the mRNA data or cannot be calculated due to a division by zero in the absence of reads. This is especially relevant in our datasets where the RNA-Seq libraries (3-8M Reads) are not sequenced at a similar depth to the PRS libraries (50M reads). Therefore I decided that TE is not a suitable metric for use in this analysis and instead looked at other metrics such as framing.

Framing and P-Site Mapping

Some read lengths predominantly map to one frame of translation (Ingolia *et al.* 2011; Bazzini *et al.* 2014). By mapping the P-site or the 5'-most end of the 28-34nt read, it should become apparent at which frame the ribosome is translating (Figure 5.8). Translating Ribosomes are associated with triplet codons in the mRNA sequence while amino acids are added concomitantly to the translating peptide through their association with tRNAs. tRNAs first bind to the A-site (Acetyl tRNA) in the ribosome, where codon recognition takes place, before the peptidyl-tRNA bound at the P-site transfers the translating peptide to the amino acid of the A-site tRNA. At this point the outgoing empty tRNA exits the ribosome through the E-site, while the peptidyl tRNA is displaced to the P-site. For translation to occur, the P-site should display a tri-nucleotide periodicity that reflects the codon-to-codon translocation of a ribosome. Given the length constraint of RBFs (28-34nt), the distance of the P-site to the 5’ end of the RBF can be estimated for each read length. This can be achieved by measuring the most prevalent distance between the 5'end of the read and the 'U/T' of the start codon where reads tend to synchronise. In a 28nt read, this would occur at approximately -12nt as illustrated in Figure 5.8.

This estimation allows the mapping of the P-site of each RBF with single nucleotide resolution, and hence determine 1) codon *coverage* by counting the fraction of codons that harbour putative ribosome P-sites in the same frame of translation as the ORF being scored (Bazzini *et al.* 2014), 2) the number of reads in frame for a given ORF (i.e. *framing*; see below), and 3) new or poorly defined start and stop codons. P-site assignments are however only an estimation and can suffer from various technical issues; for example, there is no consensus as to what should be the exact length of the RBFs used, and various studies have used different values within the 25-36nt range (Ingolia *et al.* 2014; Lareau *et al.* 2014; Bazzini *et al.* 2014; Aspden *et al.* 2014). Framing has not been previously reported or observed in Ribo-Seq data in *Drosophila* (Kronja *et al.* 2014; Dunn *et al.* 2013; Aspden *et al.* 2014), possibly due to ribosome degradation occurring in these experiments as discussed in Chapter 4. RBF length and P-site offset must reflect the degree of packing and folding of the mRNA that gets 'tucked away' inside/around the ribosome and protected from digestion, and this could be species-specific and sensitive to the biochemical specificities of the experimental setup. As such, P-site assignment can only be ascertained with a good degree of certainty in subsets of reads that display a narrow length range. Additionally, framing becomes less clear in the middle of long ORFs, away from the start and stop codons where reads pile-up and are synchronised.

Nevertheless, this triplet phasing nature of RBFs can be used to define a framing score for ambiguous or novel ORFs and make their translation more convincing. The framing score is calculated as the proportion of reads across the ORF that have their P-site (or 5' end) in frame with the ORF (Bazzini *et al.* 2014). This metric, developed by Bazzini *et al.* (2014) was based on the observation that 85% of their Ribo-Seq reads could be mapped to one reading frame in zebrafish. For this metric to be statistically significant, at least 50% of reads in a given read length are required to display a preference for any one specific frame. The most prominent advantage of the Framing score has been in highlighting the translation of alternative ORFs (altORFs) that overlap annotated ORFs, by differentiating the source of RBFs (Michel *et al.* 2012). Although a Framing based metric would be the most unambiguous signature of translation, it shares the limitations of P-site assignment. Also as only a subset of read lengths exhibit preference for a particular reading frame, which vary between species (Ingolia *et al.*

2009; Bazzini *et al.* 2014), it is difficult to predict how this would turn out for any given data set.

For our data, the Ribo-SeqR package in R (Chung *et al.* 2015) was used in order to visualise the framing and map the P-Sites of the RBF reads in the Embryo datasets. The software is provided with a *FASTA* file containing all annotated mRNA transcripts and annotates all ORFs beginning with an ATG and an in-frame stop codon. It then maps the 5' end of all reads relative to the frame of translation in these putative ORFs to calculate the framing for each RBF read length (28-34nt). Figure 5.9 shows the output of the framing function for each embryo dataset, where it can be observed that the RBF reads in our data do not show a clear preference or bias to any particular reading frame (0, +1, +2).

In order to map the P-site in these reads I calculated the offset from the 5' of the read to the start and stop codons (where reads tend to synchronise) for each ORF. Figure 5.10 shows the output of the *PlotCDS* function in order to visualise the results (*Em3* used as an example). This is plotted for each read length (28-34nt) and it can be seen according to the offset for each read length that, unlike data generated in Yeast and Zebrafish (Bazzini *et al.* 2014; Smith *et al.* 2014), there is no one unambiguous offset value that can be used to map the P-sites in *Drosophila*. Reads greater than or equal to 30nt begin to show a preference, either for Fram 0 or Frame 2, but this also varies between datasets and read lengths >30nt.

The P-site offset values for each read length are summarised in Figure 5.10B for each data set. These showed that even though an unambiguous offset value can be calculated for some read lengths in each dataset, ultimately this analysis is insufficient because this limits the number of reads that can be used for further analysis, as for example, 30nt reads would only account for a small subset of the total number of reads. Thus it is difficult to establish a framing score with any confidence of translation for ORFs which already have few reads attributed, such as very small ncrORFs, because the issue would be compounded by lack of the number reads necessary for a robust statistical analysis. Nevertheless, as shown in Figure 5.10B (values highlighted in bold), we estimated the offset required to map the P-site of individual Read lengths in our embryo datasets using RiboSeqR. These 5' offset values were chosen by visual examination of the data represented by Figure 5.10A to decide the predominant 5' offset value.

As highlighted in Figure 5.11 using Actin5C as an example, it can be seen that framing can reveal the translation of the longer canonical ORF (Blue), but also identifies multiple small ORFs (Red), based on a minimum threshold of 50 in-frame reads, that overlap the annotated CDS in the very highly transcribed and translated *Actin 5C* (*Act5C*) transcript. This data highlights a problem of P-site mapping in the absence of clear framing. Due to the ambiguity in mapping P-sites, it is difficult to discern whether these reads represent genuine translation of alternative reading frames of overlapping ORFs or are rather due to heterogeneity in the digestion of RBF 5' ends. The use of Ribo-SeqR could be a potential source of the problem, as this program utilises all putative ORFs annotated by the *FindCDS* function for framing and P-site mapping calculations, which may contribute to noise within the framing data. Therefore, we looked at framing and the 5' ends of Reads relative to start and stop codons only for annotated protein coding regions. Dr. Ying Chen Eyre-Walker performed this analysis using a custom Perl script, and this resulted in a very similar conclusion to my parallel analysis carried out using the Ribo-SeqR package (Figure 5.10-5.11). These results strongly suggest that framing is not observed in *Drosophila* Ribo-Seq datasets and that the distribution of P-sites is usually bimodal, mapping to two alternative mRNA positions, an observation concurrent with previously published results from others, as well as in our Laboratory (Dunn *et al.* 2013; Aspden *et al.* 2014).

Using Mass Spectrometry Data to establish a cut-off for translation

Since Mass Spectrometry (MS) is the current benchmark technique for providing high throughput evidence of translation, it was possible that a comparison of Ribo-Seq data to MS could provide a source of establishing a cut-off. We had noticed in the preliminary work for Aspden *et al.* (2014) that Poly-Ribo-Seq *RPKM* roughly correlates with MS detection. Therefore we looked again at the Poly-Ribo-Seq abundance estimates (*RPKM*) of all of the FB annotated protein-coding genes that were detected by the MS experiments conducted on *S2* cells (Chapter 3). The *S2* cell "All Polysomes" MiSeq dataset was used in this comparison as it was most similar to the *Em* datasets and also because the enrichment (2-6 ribosomes) in the larger "Small Polysomes" *S2* cell dataset may have affected the abundance estimates of proteins larger than 100 amino acids, as these were deliberately excluded from the sample to enrich for smORFs (Aspden *et al.* 2014). Using this data, I determined that the lowest

RPKM for a MS-detected peptide was *RPKM* 7.54 for CG34015-PA, which is a 139 amino acid protein. CG34015 is an experimentally verified protein-coding gene that is highly conserved from insects to humans, has peak expression in early embryogenesis and a low level of expression in *S2* cells according Flybase RNA-Seq data. This showed the proof of principle that it was possible to reverse the process of establishing a cut-off, *vis-à-vis* using MS to retrospectively show how much Ribo-Seq abundance leads to meaningful translation that results in a detectable protein.

Interestingly, this value is close to the 15th percentile of the *RPKM* of coding sequences of all annotated protein-coding sequences transcribed in *S2* cells (7.52). Considering that approximately 85% of annotated protein-coding genes are translated with high confidence using either an *RPKM* and *coverage* cut-off in *Drosophila S2* cells (Aspden *et al.* 2014) or using instead a *framing*-based ORF score approach in Zebrafish embryos (Bazzini *et al.* 2014), it can therefore be reasoned that it would be better to use the 15th percentile of the *RPKM* of the CDS of all annotated protein coding genes transcribed in each dataset, so that the cut-off generated reflects the transcriptional complexity at each embryonic stage (outlined in Figure 5.12).

A new *RPKM* cut-off based on the 15th percentile of Translated ORFs

Since *RPKM* is a relative measure of RNA abundance within a sample and is thus dependent on within-sample complexity, it not an ideal measure to compare between different libraries (Wagner *et al.* 2012). Therefore an *RPKM* value of 7.54 may not represent the same level of translation in the embryo datasets as it does in the *S2* dataset, particularly as there are more transcripts transcribed in the embryo stages. This is particularly important as an 8-hour window of embryonic development is expected to be a much more transcriptomically diverse sample than an *S2* cell line, and would thus include genes that are transcribed in tissue specific manner (spatially constrained) or have tightly regulated translation at specific stages (temporally constrained), leading to higher variance in *RPKM* values.

Therefore, using this new cut-off, the final number of translated ORFs in embryo datasets was determined using the 15th Percentile of the *RPKM* values for the CDS of all annotated protein-coding genes transcribed at each stage *Em1* (3.32), *Em2* (4.80) and *Em3* (4.30). (Figure 5.13A) Using this new metric, we show translation of 493 FB smORFs, 488 ORFs in lncRNA's (ncrORFs) and 4,839 uORFs across the whole of embryogenesis, of which only about 30% of each type have previous evidence of translation in the published *S2* cell dataset (Aspden *et al.* 2014). In contrast, 67% of the canonical protein coding genes translated across embryogenesis are also translated in *S2* cells, supporting the hypothesised specialised and stage specific translation and, by extension, function of smORFs.

In embryos we see again that the overall pattern of translation differs significantly between 'Dwarf' smORFs and the longer annotated protein-coding genes (including FB smORFs, as discussed earlier). About 65.6% of the longer canonical proteins (FBcds) (Figure 5.13B) and 66.3% Flybase annotated SEPs (FB smORFs) (Figure 5.13C) are translated during all three embryo datasets, while 15.8% of ncrORFs and 25.6% of uORFs are translated throughout embryogenesis. Amongst the ncrORFs, 37% are specific to stage *Em1* and a further 20% are specific to the *Em3* stage. The translation of uORFs is more prevalent at stage *Em3*, with 23% of uORFs specific to that stage and 81% of all uORFs being translated in *Em3* embryos.

The *Em2* stage dataset has a significantly lower level of ncrORF translation (4%) when compared to stages *Em1* (15%) and *Em3* (13%). What is also clear is that there is a distinct decrease in ribosomal association of lncRNAs in *Em2* embryos even though the RNA-Seq data shows that they are highly transcribed at this stage with 80% of all long ncrORFs transcribed overlapping this stage of which 20% are exclusive to the *Em2* stage (Figure. 5.6D). This suggests that there is distinct translational regulation of lncRNAs through development; where translation of ncrORFs is limited during mid embryogenesis phylotypic stage when body patterns are being established and translation of genes is evolutionary constrained (Domazet-Lošo, and Tautz 2010).

Figure 5.14 shows the density plot of the log2 *RPKM*s for FBcds, smORFs and ncrORFs transcribed at each stage. The vertical lines represent the highest and lowest cut-offs tested in this analysis. We can observe that there is a distinct distribution of *RPKM* values for each type of smORF plotted. The log2 of *RPKM* density distribution goes from wider to restricted of FBcds>FB smORFs > uORFs > ncrORFs.

Finally, I compared the 90th percentile FBcds 3'UTR *RPKM* cut-off with the 15% FBcds CDS *RPKM* cut-offs in regard of number of reads for the smallest ORF size (30nt) in our datasets (Table 5.5). The 3'UTR-based *RPKM* gives a substantial 20-25 reads (*Em2* and *Em3*) per 30bp, which is very stringent when compared to other datasets (Ingolia *et al.* 2009; Dunn *et al.* 2013; Aspden *et al.* 2014; Popa *et al.* 2016). Whilst the 15th Percentile based lower cut-offs still give some 6.5-7.5 reads per every 30bp of an

ORF, which is more stringent than the minimum 5 reads across the ORF (regardless of size) used in Aspden *et al.* (2014). *RPKM* cut-offs may not represent a clear yes/no binary decision for real translation, but do provide a certain level of confidence or number of reads which can equate to high confidence of translation. Thus by the reasoning described here, it can be said, with high confidence, that these ORFs are translated.

Discussion

This chapter provides insights into the assessment of smORF translation during *Drosophila* embryogenesis and discusses the different computational metrics that have been recently developed for the characterisation of ORF translation in Ribo-Seq Data. The aim of this chapter was to improve upon the original data analysis pipeline developed for the *S2* cell data (Aspden *et al.* 2014). This analysis of the embryo Poly-Ribo-Seq data also details the suitability of various metrics used for the proof of translation of very short features such as uORFs and ncrORFs. The aims of this chapter were addressed by replicating that data analysis pipeline with the new data from the Poly-Ribo-Seq of the three embryo stages and, then evaluating a number of metrics that were used in the *S2* cell pipeline to establish translational cut-offs in order to filter out background noise. Similar to our Lab's *S2* cell Poly-Ribo-Seq results (Aspden *et al.* 2014), I have shown here that a cut-off based on abundance works best in *Drosophila* datasets. The method by which this conclusion was achieved for embryos was significantly different to the data analysis pipeline of *S2* cells.

Translation Efficiency is not suitable for the analysis of very small ORFs

Translation Efficiency (TE) is a useful metric for comparing translated ORFs between datasets by normalising Ribo-Seq signal in the ORF to the abundance of the transcript. It certainly works very well for longer canonical ORFs that make up the majority (~70%) of the transcript and highlights genes that may be undergoing translation regulation. However, it does not perform particularly well for ncrORFs, which typically account for only 4% of a lncRNA transcript. As such, a robust TE calculation is only really possible when there are enough reads mapping to ORFs in both RNA-Seq and Ribo-Seq datasets and therefore it is important to have a comparable total number of reads in the parallel experiments. This is important due to the fact that typically only 0.1% of Ribo-Seq reads map to lncRNAs and only about 2% in RNA-Seq data (Housman, and Ulitsky 2015).

In our datasets, the *Em2* and *Em3* RNA-Seq libraries were under-sequenced (3M and 4.3M mapped reads) and therefore TE calculations were unsuitable for discerning smORF translation. Furthermore, even in the presence of a sufficient number of Reads, TE is not particularly useful at determining the translational status of novel ORFs as it cannot, in the absence of a minimum number of reads threshold, differentiate between a highly transcribed and translated CDS that has an *RPKM* of 500/300, or an ncrORF with an *RPKM* of 0.5/0.3,. Furthermore, ratios of very low numbers can be easily distorted, particularly when the sample is not sequenced to sufficient depth and each read carries more weight due to *RPKM* normalisation as discussed previously.

TE was a useful metric for looking at the overall translation status of each dataset and highlighting the overall decreased translation at *Em2* stage observed through the distribution of *RPKM* values and the lower proportions of dwarf smORFs translated. With the further sequenced RNA-Seq datasets, it should be interesting to see differences in TE at different developmental stages in more detail, particularly of transcripts that are expressed at multiple *Em* stages, but only translated at specific ones. Furthermore TE could also be used to look at the effect of the translation of an uORF on the translation of the downstream ORF by comparing the TE of the downstream ORF across different stages.

Patterns of transcription and translation and the stage specific regulation of lncRNA translation

The calculation of FBcds 3'UTR *RPKM* based on High (90th Percentile), Medium (80th-90th Percentile) and Low (70th to 80th Percentile) confidence thresholds for translation were not used to define translational status, however, their analysis did highlight some interesting patterns of translation across embryogenesis. By separating translated sequences into the different populations of smORFs (FB smORFS, ncrORFs, uORFs) based on their *RPKM* abundance and plotting the overlap of the ORFs between embryonic stages, I was able to highlight their stage specific translation.

Annotated FB smORFs follow a pattern of translation similar to canonical protein-coding sequences, with the about two-thirds of abundant smORFs translated throughout embryogenesis, though some are still stage-specific. In contrast, and as can be expected, lower abundance smORFs in the 'Medium confidence' category have mainly stage specific translation. In a sample as complex and dynamic as developing embryos, these lower abundance smORFs may be those ORFs that are undergoing spatio-temporal regulation of gene expression. However, even the High confidence category, ncrORFs fail to follow this pattern as the majority (86%) of these *loci* still display stage-specific translation, perhaps highlighting the tissue and stage specific nature of lncRNAs (Washietl *et al.* 2014; Cabili *et al.* 2011).

These translational patterns can only be partially explained by transcription patterns, which show that at any given time during embryogenesis, about 75% of annotated protein coding genes (and by extension uORFs) and smORFs are being transcribed. The number is much lower for ncrORFs found in lncRNAs (42%) transcribed throughout embryogenesis. This phenomenon could be explained by the fact that the cut-off for scoring translation (minimum *RPKM* between 3.5-5) is more stringent as compared to the cut-off used for scoring transcription (*RPKM*>1). However the specific regulation of lncRNAs is highlighted by the overall distribution of *RPKM* values and consequently, by the low number (125) and proportion (4%) of ncrORFs translated in the *Em2* stage despite high levels of transcription (80% of embryonic ncrORFs) during 8-16H AEL. These observations suggest an even greater translational specificity of lncRNAs and could also explain the apparent stochastic nature of ncrORF translation.

Although each library was prepared with embryos collected separately on 2 different days and despite the generally good correlation between different *Em1* optimisation runs (Chapter 4), the experimental setup could be improved by having two biological and technical replicates for each dataset. This is particularly important for a reliable assessment of the translation of Dwarf smORFs, which generally have low levels of signal and whose translation appears to show an element of stochasticity. It would also beneficial to have spike-in controls of known quantities of a synthetic RNA sequence to enable the absolute quantification of Ribo-Seq signal in terms of copies per cell/embryo.

Framing is not observed in *Drosophila* Poly-Ribo-Seq datasets

Despite best efforts, we were not able to observe any clear preference for any one particular frame of translation in coding sequences using the RiboSeqR package (Chung *et al.* 2015). This result was similar to what we have previously observed in *S2* cell data (Aspden *et al.* 2014). Framing analysis could have been evaluated more thoroughly using a subset of the most highly transcribed/translated annotated protein coding sequences to, first, develop the metric and second, implement it as a scoring method for novel ORFs (as has been performed in Popa *et al.* (2016)) or by splitting the P-site coverage between two positions (as performed by Dunn *et al.* (2013)). This heterogeneity may be due to the unique structure of *Drosophila* Ribosomes (Pavlakis *et al.* 1979) as discussed in Chapter 4.

Recent studies across multiple organisms have used different ranges of read lengths that can be attributed to RBFs, and the exact length distribution can vary between samples. This is possibly due to differences in nuclease treatment conditions (Ingolia *et al.* 2012) and/or due to distinct ribosome conformations (Lareau *et al.* 2014). Therefore, it may be prudent to consider a wider range of RBF sizes (26-36nt) in future data analyses, in order to increase the number of reads to make the subject of framing more feasible in terms of the data that is lost due to retaining only the reads of a size for which a P-site can be determined.

Qualifying data in terms of absolute number of reads

S2 cells used the 90th Percentile 3'UTR signal to establish *RPKM* cut-offs for assessing the translation of ORFs. For a variety of reasons, this was found to be unsuitable for a much more complex sample like Embryos. Using a retrospective approach, which entailed the use of Mass Spectrometry data to dictate the abundance of reads required to produce a detectable peptide, I established the use of a translation cut-off based on the 15th Percentile *RPKM* of all annotated protein-coding sequences that are transcribed at each stage.

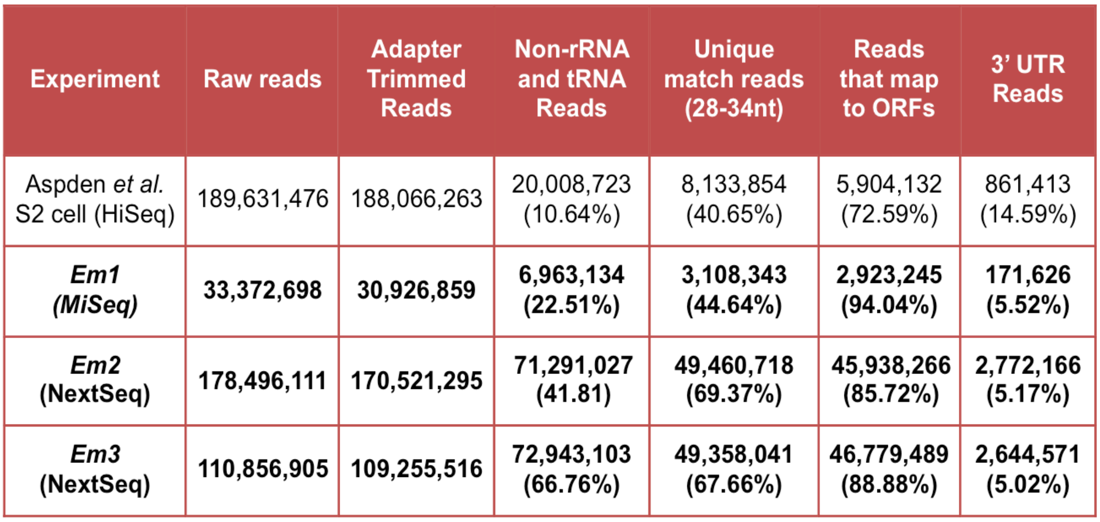
My results show that there is clear presence of ribosome footprints in a large number lncRNAs and uORFs. The data analysis and improved sequencing shows substantial mapping of more than 6 reads per every 30nt in the *Em2* and *Em3* datasets as a minimum of the novel coding sequences. Comparisons with other datasets generated within our lab shows that this is a significant improvement as compared to the *S2* cell dataset, which would have only 2 reads every 30nt if using equivalent cut-offs based on the 15th Percentile *RPKM* of annotated protein-coding sequences.

This highlights that the association of these ncrORFs to ribosomes and hence their translation is a robust phenomenon. Other papers that have used minimum number of read counts as way to define translation. Ingolia *et al.* (2009) and Dunn *et al.* (2013) used 128 reads across canonical protein coding sequences in replicate datasets to defer a high confidence of translation Ribo-Seq experiments. However, these studies were not specifically tackling the issue of scoring extremely small ORFs and could therefore use a higher number of reads to establish a cut-off based on minimum reads. Recent studies specifically targeting the translation of novel smORFs and in particular ncrORFs have used a minimum of 4 RBF reads (Popa *et al.* 2016) to score translation, which is even lower than the figures used in our embryo analysis.

Given that the median length of dwarf smORFs is around 60nt, I compared my results with the number of ORFs in lncRNAs using a cut-off of a minimum of 10 reads mapping to the ORF (5reads x 30nt x 2). This was accomplished using an online database of small ORFs detected in published Ribo-Seq datasets (www.sorf.org Accessed: September 2015) (Olexiouk *et al.* 2016). Using the minimum 10 read threshold, the number of translated ncrORFs is 124 in human, 81 in mouse and 52 in fruit flies, which are much lower than the number of ncrORFs translated in my embryo datasets (488). Taken together, these results indicate that these sequences must be translated, which appears to be the general consensus in the field (Popa *et al.* 2016; Mumtaz, and Couso 2015; Ji *et al.* 2015; Brar, and Weissman 2015; Cech, and Steitz 2014; Smith *et al.* 2014; Ingolia *et al.* 2014; Ruiz-Orera *et al.* 2014; Aspden *et al.* 2014; Cohen 2014; Bazzini *et al.* 2014). Whether this assessment of lncRNA translation actually results in biologically relevant function is an argument outside the scope of this discussion as it still remains a highly debated topic in the field, and will be discussed in further detail in the final chapter of this thesis.

With the new 15th Percentile *RPKM* cut-off, we determine that the embryo data shows definite translation of hundreds of annotated FB smORFs, as well as unannotated ncrORFs and thousands of uORFs, all from within transcripts that are undergoing active translation, using as a proxy their association with polysomal complexes (2+ ribosomes). As in the Aspden *et al.* data, these smORFs can be further separated on the basis of their size, with novel ‘dwarf’ smORF in lncRNAs and 5’UTRs having a median peptide length of 23 and 19 amino acids, respectively, while the annotated FB smORFs are longer, with a median size of 80 amino acids. Therefore, we have succesfully expanded the catalogue of smORFs using a high-throughput assessment of translation, which will hopefully allow the further characterisation of their functions in *Drosophila melanogaster* and other organisms.

Chapter 5 Figures and Tables



Experiment	Raw reads	Adapter Trimmed Reads	Non-rRNA and tRNA Reads	Unique match reads (28-34nt)	Reads that map to ORFs	3' UTR Reads
Aspden *et al.* S2 cell (HiSeq)	189,631,476	188,066,263	20,008,723 (10.64%)	8,133,854 (40.65%)	5,904,132 (72.59%)	861,413 (14.59%)
Em1* *(MiSeq)	**33,372,698**	**30,926,859**	**6,963,134 (22.51%)**	**3,108,343 (44.64%)**	**2,923,245 (94.04%)**	**171,626 (5.52%)**
***Em2* (NextSeq)**	**178,496,111**	**170,521,295**	**71,291,027 (41.81)**	**49,460,718 (69.37%)**	**45,938,266 (85.72%)**	**2,772,166 (5.17%)**
***Em3* (NextSeq)**	**110,856,905**	**109,255,516**	**72,943,103 (66.76%)**	**49,358,041 (67.66%)**	**46,779,489 (88.88%)**	**2,644,571 (5.02%)**

Table 5.1 Summary of Sequencing data of Poly-Ribo-Seq in Embryos

This table gives an overview of the final datasets of all three stages of Poly-Ribo-Seq conducted on embryos. The Aspden *et al.* (2014) *S2* cell results are included for comparison. The 0-8H embryo sample (*Em1*) dataset includes all of the 0-8H sequencing runs (Em JA, Em NEB, tRNADep and SucCush) on the MiSeq, for a total of over 33 million raw reads. The 8-16H (*Em2*) and 16-24H (*Em3*) samples were run on the NextSeq and have considerably higher numbers of raw sequencing reads (110-178 million reads)

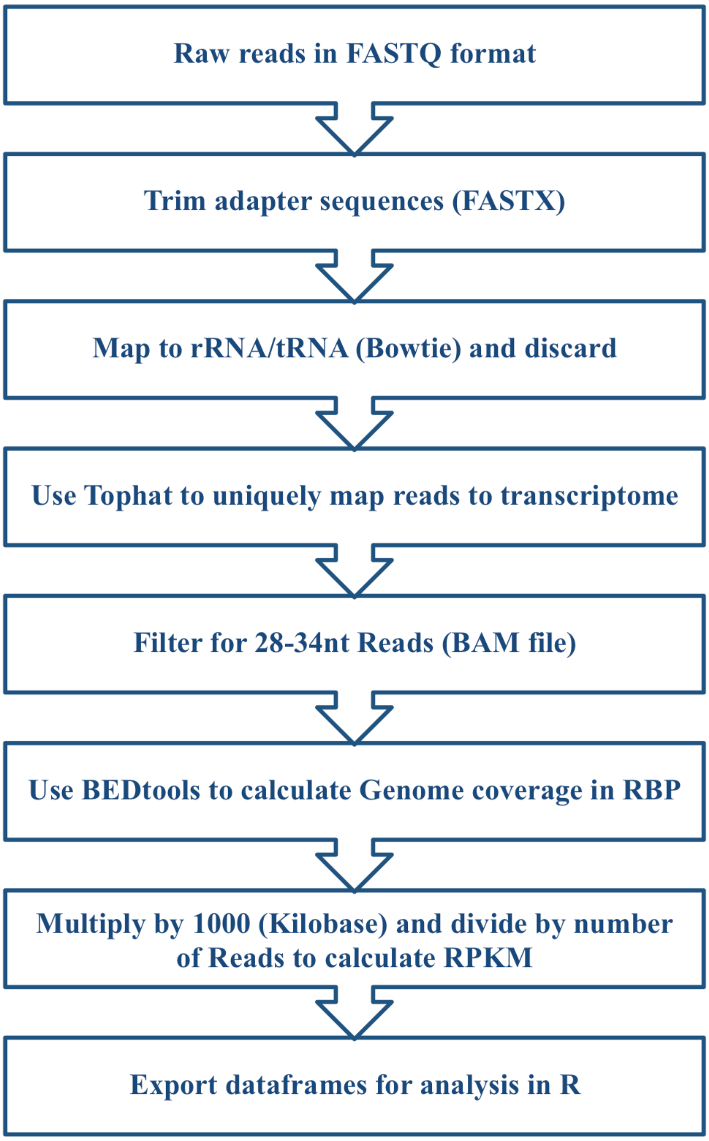

Raw reads in FASTQ format
Trim adapter sequences (FASTX)
Map to rRNA/tRNA (Bowtie) and discard
Use Tophat to uniquely map reads to transcriptome
Filter for 28-34nt Reads (BAM file)
Use BEDtools to calculate Genome coverage in RBP
Multiply by 1000 (Kilobase) and divide by number of Reads to calculate RPKM
Export dataframes for analysis in R

Figure 5.1 Overview of Poly-Ribo-Seq data analysis pipeline

This figure outlines the steps involved in the processing of raw sequencing data and the mapping of sequencing reads to the transcriptome for Poly-Ribo-Seq. After a sequencing run, the data output is in the form of a FASTQ file, which is processed into FASTX to discard low quality reads and the trimming of 5’ and 3’ adapter sequences. The Bowtie short read aligner is then used to discard annotated rRNA and tRNA sequences. Unmapped reads are run through a splice junction mapper called TopHat that uses the Bowtie algorithm to map reads to the reference transcriptome. Reads that map to more than one region in the genome are discarded to retain only uniquely matched reads. The abundance of these reads in each transcript was then calculated using BEDtools, which outputs these estimates in Reads per Base Pair (RPB) that are then converted to *RPKM*. All resulting data from this pipeline is then exported to R for further analysis.

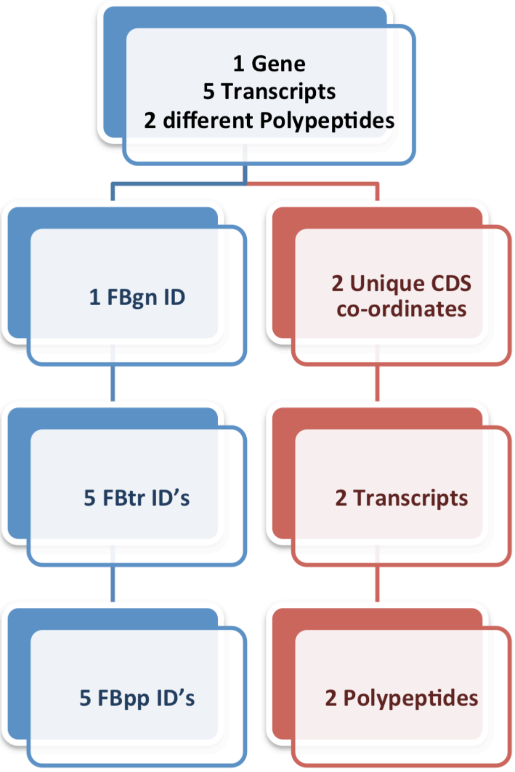
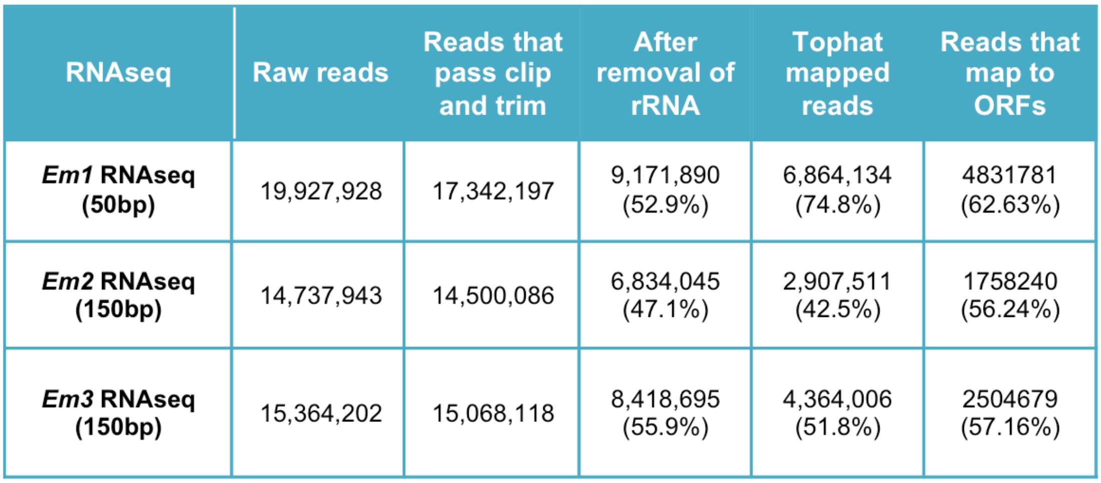


Figure 5.2 Counting ORFs by unique genomic coordinates

Since RBF reads map almost exclusively to coding sequences, alternative transcript models are not relevant to abundance estimations. This figure shows what would happen for 'hypothetical gene X' if we calculated ORF counts by the FlyBase annotation (Blue). Gene X has one FBgn ID but from our data we can see that there are two Unique CDS coordinates and therefore two transcripts in our data even though there are 5 FBtr ID's attributed to this gene. This means that rather than attributing reads towards 5 annotated peptides (FBpp), the improved method increases accuracy of the number of translated peptides by counting each ORF by its unique genomic coordinates revealed by the data as opposed to the FlyBase annotations.



RNAseq	Raw reads	Reads that pass clip and trim	After removal of rRNA	Tophat mapped reads	Reads that map to ORFs
***Em1* RNAseq (50bp)**	19,927,928	17,342,197	9,171,890 (52.9%)	6,864,134 (74.8%)	4831781 (62.63%)
***Em2* RNAseq (150bp)**	14,737,943	14,500,086	6,834,045 (47.1%)	2,907,511 (42.5%)	1758240 (56.24%)
***Em3* RNAseq (150bp)**	15,364,202	15,068,118	8,418,695 (55.9%)	4,364,006 (51.8%)	2504679 (57.16%)

Table 5.2 RNA-Seq results of Embryo mRNA controls

RNA-Sequencing is performed on Poly-A selected, fragmented mRNA from each sample that is subjected to Poly-Ribo-Seq in order to normalise Ribo-Seq footprint abundance to the abundance of mRNA in the sample. This table outlines the processing of the RNA-Seq data on mRNA controls for each stage. Similar to the Poly-Ribo-Seq results, the RNA-Seq data must be processed and filtered to obtain the final reads which map to ORFs, in order to perform ratio calculations of Ribo-Seq/RNA-Seq to illustrate the abundance of the ORFs that are undergoing translation. The RNA-sequencing runs yielded 2.9M (*Em2*), 4.4M (*Em3*) and 6.9M (*Em1*) uniquely mapping reads to the transcriptome (Tophat mapped reads).

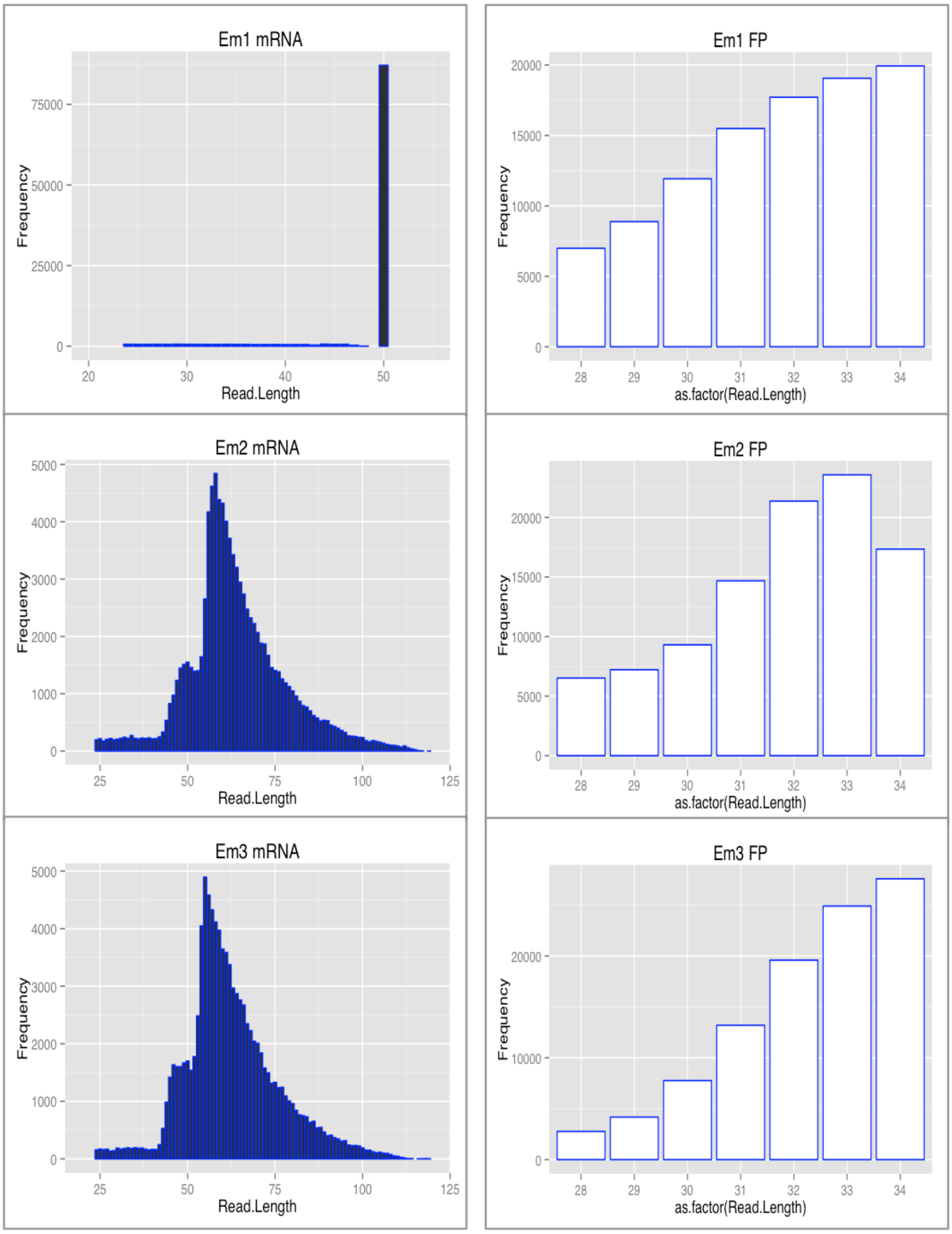


Figure 5.3 Read-length distributions for RNA-Seq and Poly-Ribo-Seq data

The RNA-Seq read lengths are shows on the left and the Poly-Ribo-Seq read lengths are shown on the right for each of the three data sets generated. *Em1* (0-8H) RNA-Seq was conducted on a Version 2 cartridge, which has a maximum capacity of 15M reads of up to 50bp in length, which is the case for most of the read lengths in this sample. The *Em2* and *Em3* mRNA libraries were pooled together and run on a Version 3 MiSeq cartridge which has a maximum capacity of 25 M reads up to 150bp in length, and though we can see that there is a wider distribution of read lengths in these samples, the median Read length for *Em2* and *Em3* is ~70bp. The Footprinting reads from Poly-Ribo-Seq of each data set are quite different in frequency of 28-34nt reads. *Em1* FP show a wider distribution of frequency across all read lengths, with the most reads being 34nt, similar to *Em2* FP read length distribution which has the most reads at 33nt. *Em3* has the most polarised read-length distribution with fewer small reads (28-31nt) and is represented mostly by 32-34nt reads.

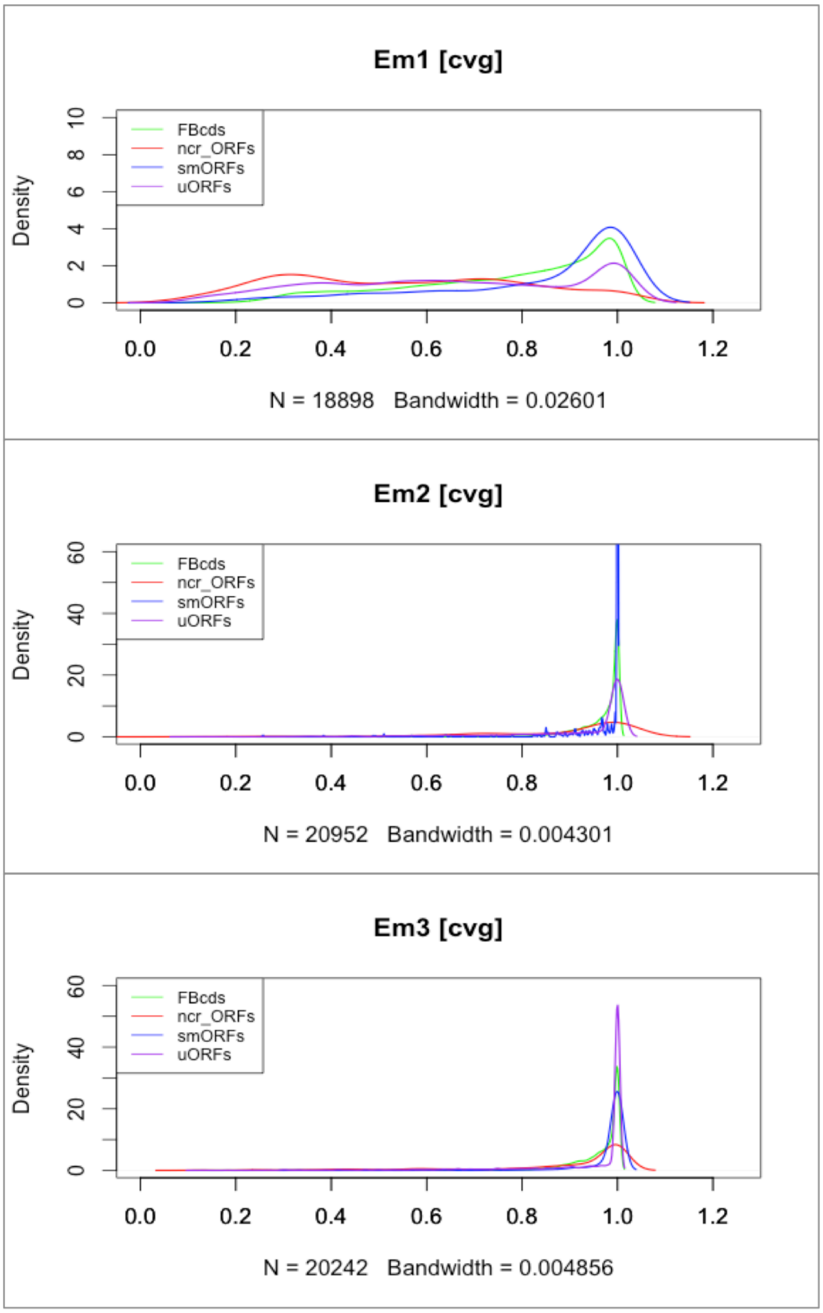
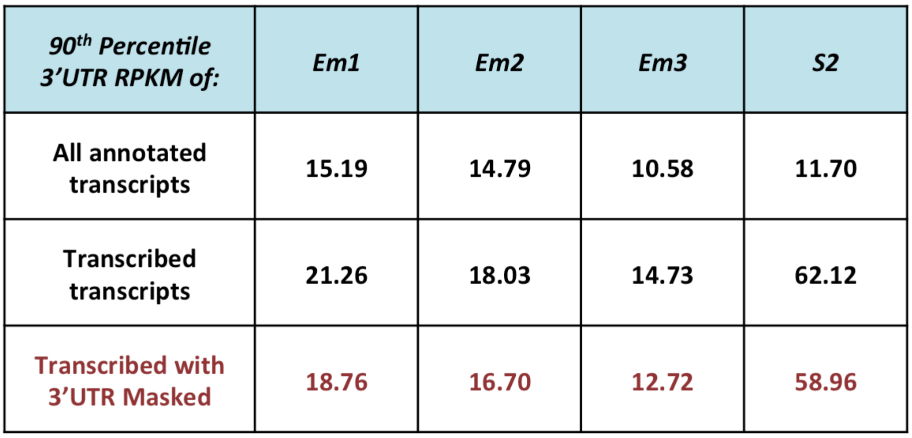


Figure 5.4 Distribution of Read Coverage in Embryo Poly-Ribo-Seq data sets

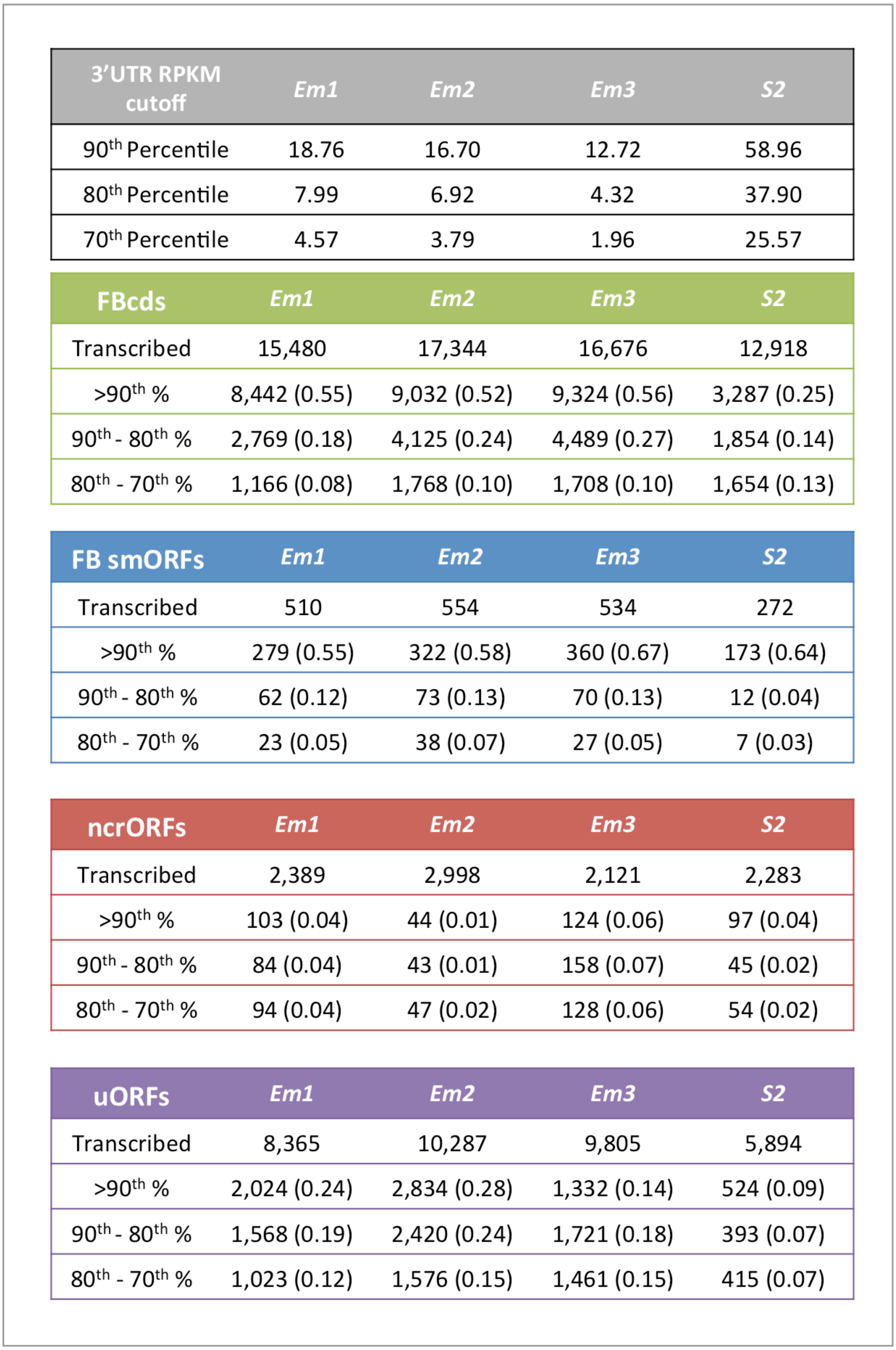
The Coverage metric describes the proportion of an ORF that is covered by RBF reads. The more coverage there is for an ORF, the greater chance there is that it is genuinely translated, however coverage is dependent on sequencing depth. As can be seen from the Coverage density plots shown in the graph for each ORF category (FB cds-green smORFs-blue, ncrORFs-red and uORFs-purple), *Em1* stage has a broad distribution of coverage for transcribed ORFs, with a slight peak at a Coverage of 1.0 for FB cds, smORFs and uORFs. *Em1* ncrORFs however have a lower average coverage (around 0.3). For *Em2* and *Em3* transcribed ORFs have a tight distribution around the maximum coverage of 1.0 as they were sequenced at a high depth.



90th Percentile 3’UTR RPKM of:	*Em1*	*Em2*	*Em3*	*S2*
All annotated transcripts	**15.19**	**14.79**	**10.58**	**11.70**
Transcribed transcripts	**21.26**	**18.03**	**14.73**	**62.12**
Transcribed with 3’UTR Masked	**18.76**	**16.70**	**12.72**	**58.96**

Table 5.3 Cut-offs of Embryo Poly-Ribo-Seq calculated by the *S2* cell Poly-Ribo-Seq pipeline

The 90th Percentile of the *RPKM* values of signal found in 3’UTRs of annotated protein-coding genes transcribed in each dataset were used to calculate translation cut-offs, as conducted for *S2* cells in Aspden *et al.* (2014). The *S2* cell cut-off was based on all-annotated transcripts, regardless of their expression in *S2* cells, and this gave a cut-off of 11.7 *RPKM* in *S2* cells. Though there was significantly less background in the 3’UTR of embryo data-sets, the *RPKM* cut-off for embryo datasets using ‘All transcripts’ were higher than the *S2* cell cut-off, except for *Em3* which is just under (second row). The next row shows 90th percentile 3’UTR cut-off based exclusively on transcripts expressed (Transcribed) at each embryo stage and in *S2* cells. Suddenly, the *S2* cell cut-off jumps to 62.12 while the embryo datasets cut-offs remain at lower values. The most stringent cut-off, shown in the last row, is established by calculating the cut-off by of masking 3’UTR and using only those transcripts expressed (Transcribed with 3’UTR masked). This further lowers the *RPKM* cut-off for embryo datasets but the same cut-off calculation for *S2* cells results in a very high number (58.92). The results of using the new ‘Transcribed with 3’UTR masked’ cut-offs on the translation of ORFs is shown in the next table (Table 5.4)



3'UTR RPKM cutoff	Em1	Em2	Em3	S2
90th Percentile	18.76	16.70	12.72	58.96
80th Percentile	7.99	6.92	4.32	37.90
70th Percentile	4.57	3.79	1.96	25.57

FBcds	Em1	Em2	Em3	S2
Transcribed	15,480	17,344	16,676	12,918
>90th %	8,442 (0.55)	9,032 (0.52)	9,324 (0.56)	3,287 (0.25)
90th - 80th %	2,769 (0.18)	4,125 (0.24)	4,489 (0.27)	1,854 (0.14)
80th - 70th %	1,166 (0.08)	1,768 (0.10)	1,708 (0.10)	1,654 (0.13)

FB smORFs	Em1	Em2	Em3	S2
Transcribed	510	554	534	272
>90th %	279 (0.55)	322 (0.58)	360 (0.67)	173 (0.64)
90th - 80th %	62 (0.12)	73 (0.13)	70 (0.13)	12 (0.04)
80th - 70th %	23 (0.05)	38 (0.07)	27 (0.05)	7 (0.03)

ncrORFs	Em1	Em2	Em3	S2
Transcribed	2,389	2,998	2,121	2,283
>90th %	103 (0.04)	44 (0.01)	124 (0.06)	97 (0.04)
90th - 80th %	84 (0.04)	43 (0.01)	158 (0.07)	45 (0.02)
80th - 70th %	94 (0.04)	47 (0.02)	128 (0.06)	54 (0.02)

uORFs	Em1	Em2	Em3	S2
Transcribed	8,365	10,287	9,805	5,894
>90th %	2,024 (0.24)	2,834 (0.28)	1,332 (0.14)	524 (0.09)
90th - 80th %	1,568 (0.19)	2,420 (0.24)	1,721 (0.18)	393 (0.07)
80th - 70th %	1,023 (0.12)	1,576 (0.15)	1,461 (0.15)	415 (0.07)

Table 5.4 High-Medium-Low confidence of translation based on percentiles

The 90th, 80th and 70th percentiles are used as metrics to confer a high (>90th%), medium (80th to 90th %) and low (70th to 80th%) confidence of translation to each class of ORFs. The cut-offs generated were calculated by the 'Transcribed with 3'UTR masked' method as shown in the previous table, for each embryo dataset and the new 90th percentile cut-off for *S2* cells.

Using these new cut-offs, the numbers of translated FBcds (Green table), FBsmORFs (Blue table), ncrORFs (Red table) and uORFs (Purple table) are shown for each data set: *Em1*, *Em2*, *Em3* as well as newly calculated values of *S2* cell 90th percentile ORFs for comparison. The numbers in parenthesis show the proportion (out of 1) of transcribed ORFs that are translated in each category. This data shows that when this stringent *RPKM* cut-off is used for high-confidence 90th% calculations, about 25% of FBcds are translated in *S2* cells (25%), while these remain between 52-56% in embryos. FB smORFs numbers are consistent between all data-sets, (55-67%) as are ncrORFs (1-6%). The translation of uORFs is between 14-25% for Embryo datasets and 9% in *S2* cells.

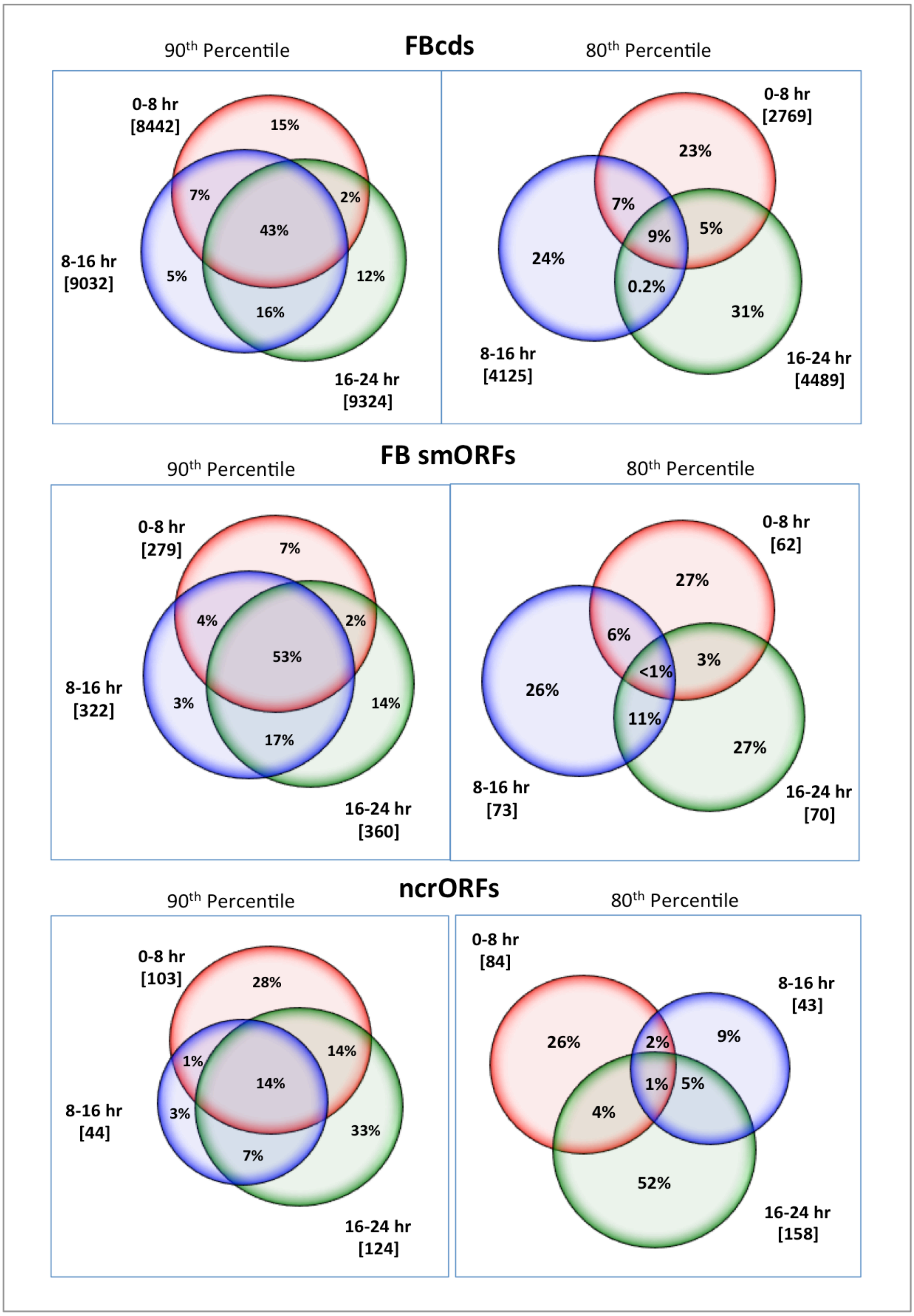

FBcds
90th Percentile
0-8 hr
[8442]
15%
7%
2%
43%
8-16 hr
[9032]
5%
12%
16%
16-24 hr
[9324]
80th Percentile
0-8 hr
[2769]
23%
7%
5%
9%
24%
0.2%
31%
8-16 hr
[4125]
16-24 hr
[4489]
FB smORFs
90th Percentile
0-8 hr
[279]
7%
4%
2%
53%
8-16 hr
[322]
3%
14%
17%
16-24 hr
[360]
80th Percentile
0-8 hr
[62]
27%
6%
3%
<1%
26%
11%
27%
8-16 hr
[73]
16-24 hr
[70]
ncrORFs
90th Percentile
0-8 hr
[103]
28%
14%
1%
14%
8-16 hr
[44]
3%
33%
7%
16-24 hr
[124]
80th Percentile
0-8 hr
[84]
8-16 hr
[43]
26%
2%
9%
1%
5%
4%
52%
16-24 hr
[158]

Figure 5.5 Overlap between three stages of Embryogenesis for 90th percentile and 80th percentile *RPKM* cut-offs

The number of translated FBcds, FB smORFs and ncrORFs are shown for each stage in brackets, and each stage is represented by a colour (*Em1* Red, *Em2* Blue, *Em3* Green). The percentage of each translated set is shown inside the Venn diagram. For annotated protein coding genes and smORFs, the high confidence (90th percentile) there is 43% overlap of all FBcds and and 53% overlap of smORFs expressed throughout all three stages, which is much lower (14%) for high confidence ncORFs. Though there is a high amount of overlap for FBcds and FB smORFs in all sets, there is a significant number of stage-specific translation of each type of ORF, as well as between subsequent stages, i.e. 0-8H and 8-16H overlap: 7% FBcds, 4% FBsmORFs, 3% ncrORFs; 8-16H and 16-24H overlap: 16% FBcds, 17% FBsmORFs, 7% ncrORFs. Interestingly, ncrORFs overlap more between 0-8H and 16-24H (14%) than FB cds and FB smORFS (2%). There is a drastic decline in the amount of overlap between all stages of medium confidence (80th percentile) FBcds (9%), FB smORFs (<1%) and ncrORFs (1%). Medium confidence translated ORFs have highly stage-specific expression, with quite low levels of overlap temporally.

A

Transcribed	FBcds	FB smORF	ncrORF	uORF
Em 1	15,480	510	2,389	8,365
Em 2	17,344	554	2,998	10,287
Em 3	16,676	534	2,121	9,805

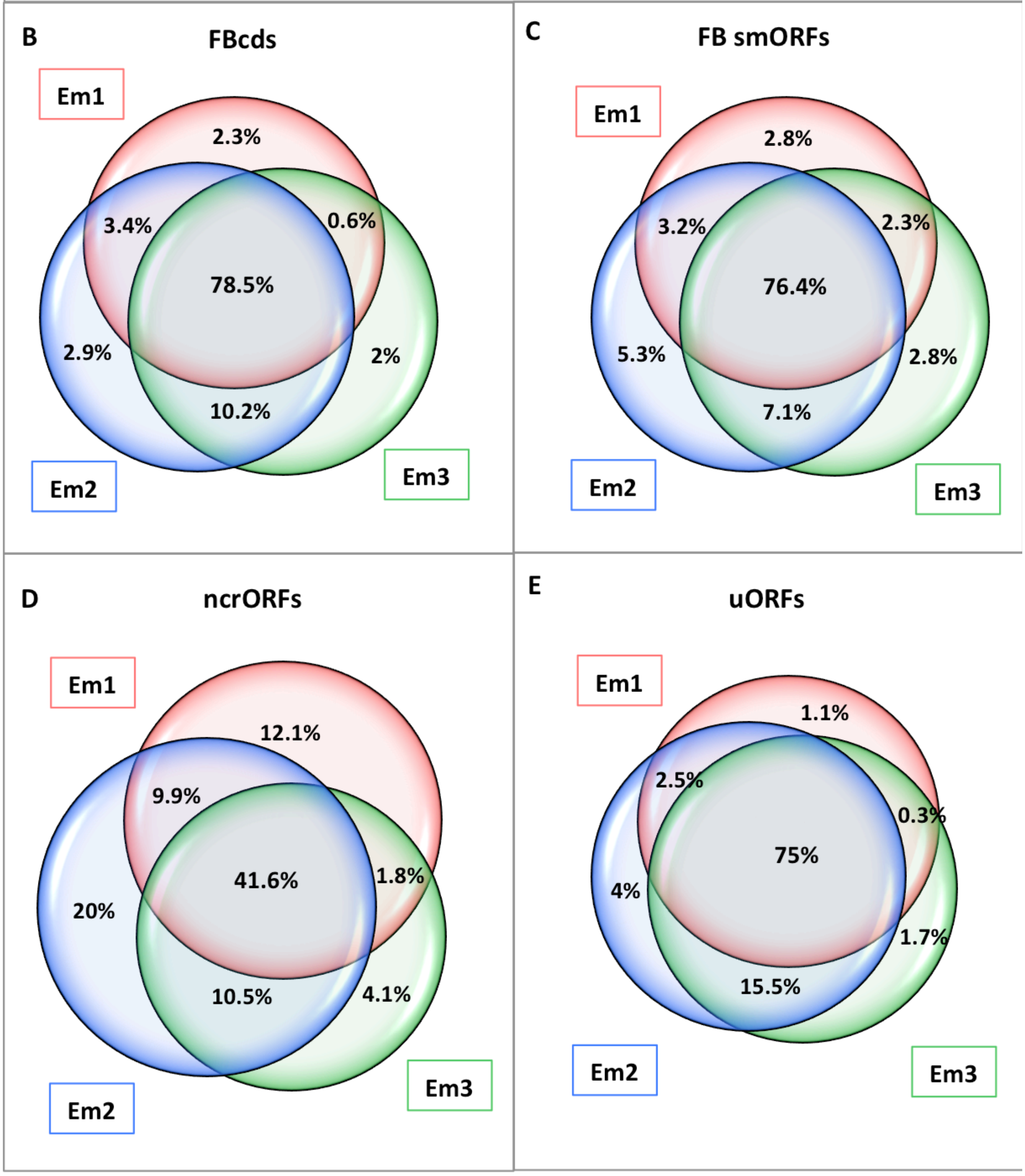


Figure 5.6 Numbers of transcribed ORFs and Patterns of Transcription across Embryogenesis

A) The number of transcribed FBcds, FB smORFs, ncrORFs and uORFS are shown for each stage in the and each stage is represented by a colour (*Em1* in Red, *Em2* in Blue and *Em3* in Green).

B) Venn diagrams of percentages in each data set of the total FBcds transcribed across embryogenesis. 78.5% of all Fbcds are transcribed throughout embryogenesis

C) Venn diagrams of percentages in each data set of the total FB smORFs transcribed across embryogenesis. Similar to FBcds, 76.4% of all annotated smORFs are transcribed in all three stages

D) Venn diagrams of percentages in each data set of the total ncrORFs transcribed across embryogenesis. 41.6% of all ncrORFs are transcribed throughout embryogenesis and 20% are exclusively transcribed at *Em2*. There is also significant overlap between *Em1*/*Em2* (9.9%) and *Em2*/*Em3* (10.5%).

E) Venn diagrams of percentages in each data set of the total uORFs transcribed across embryogenesis. Similar to the FBcds and FB smORFs, 75% of all uORFs are transcribed throughout embryogenesis.

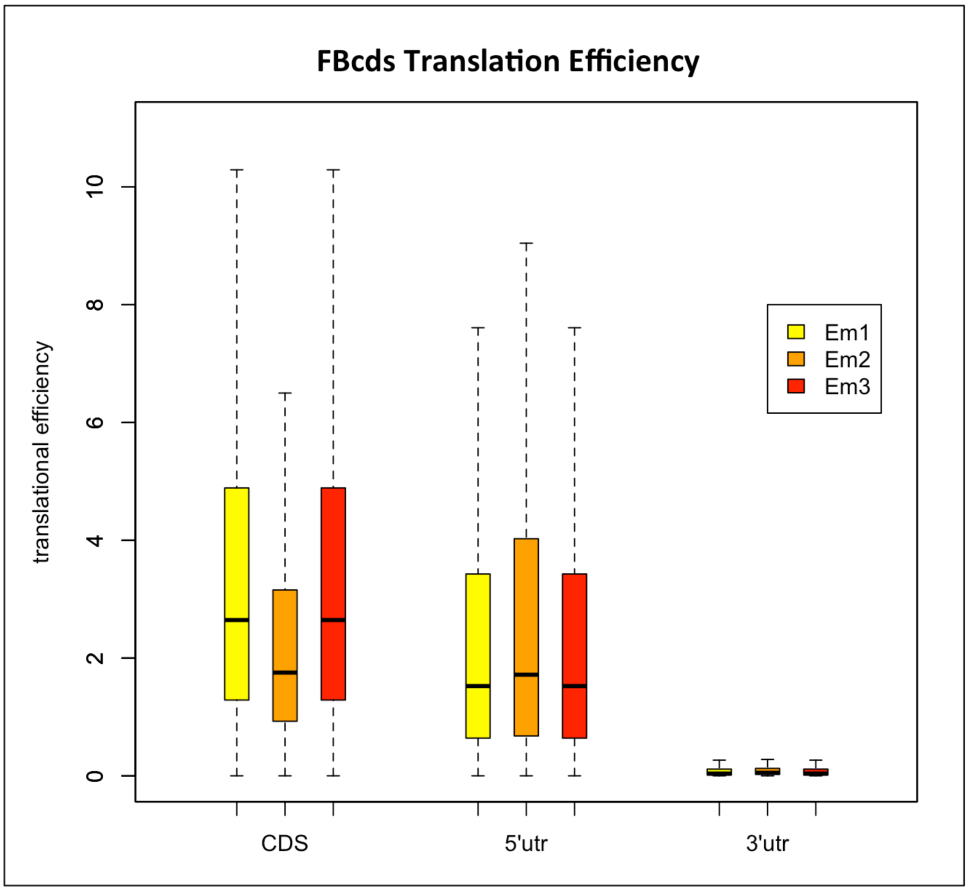


Figure 5.7 Translation Efficiency in Different Embryo Datasets

These box-and-whisker plots show the Translation Efficiency (TE) across the different features of expressed FBcds transcripts in each dataset. The box represents the 1st-3rd Quartile TE distributions and the line inside the box represents the median TE of the dataset. The whiskers represent the standard deviation of the sample. The graph is grouped by TE of the CDS, 5'UTR and 3'UTR of standard annotated protein-coding transcripts. There is much higher TE of the CDS compared to the 3'UTR, but the range and values of 5'UTR TE versus the CDS TE is quite comparable. The lowest TE is in the *Em2* sample, with the least amount of variation. The TE range of 5'UTR of *Em2* is the highest of the three samples, which may be due to the quality of the RNA-Seq control datasets. The TE of 3'UTRs is negligible, showing the lack of any meaningful translation from these features.

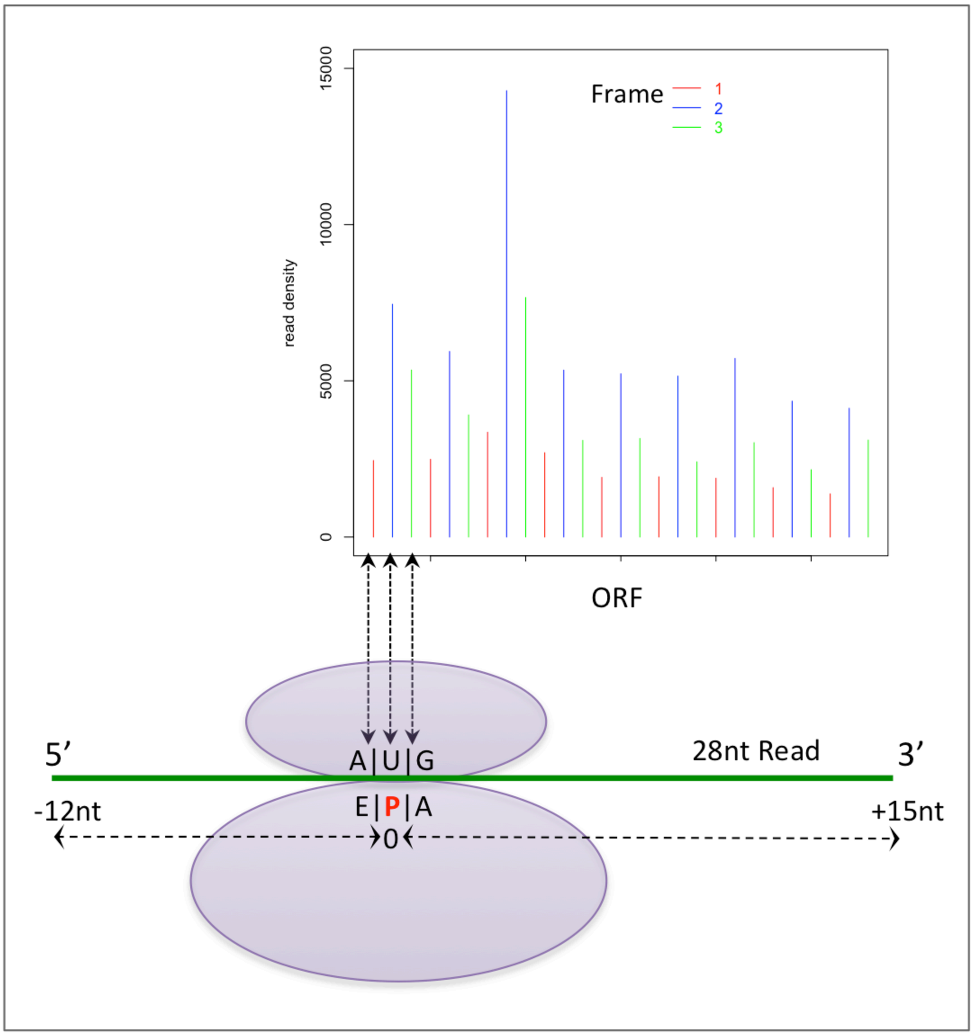


Figure 5.8 Framing and P-site mapping

The A-site (aminoacyl) in a ribosome is the first binding site of the tRNA-ribosome complex formed during translation, at which the codon is recognised by a ribosome. This is followed by the P-site (peptidyl) and E-site (exit). The P-site is where a tRNA is held during the addition of single amino acids to the translating polypeptide. The P-site is mapped by calculating the nucleotide offset of AUG reads from the 5' end of the read to the 'U' of the AUG, which is usually -12nt in a 28nt RBF. Calculating the P-site of RBFs can help determine codon coverage in-frame with the translation of an ORF. It can also help to calculate the number of reads that are in-frame with an ORF, dictating 'real translation' and help to define novel start and stop codons of uncharacterised ORFs.

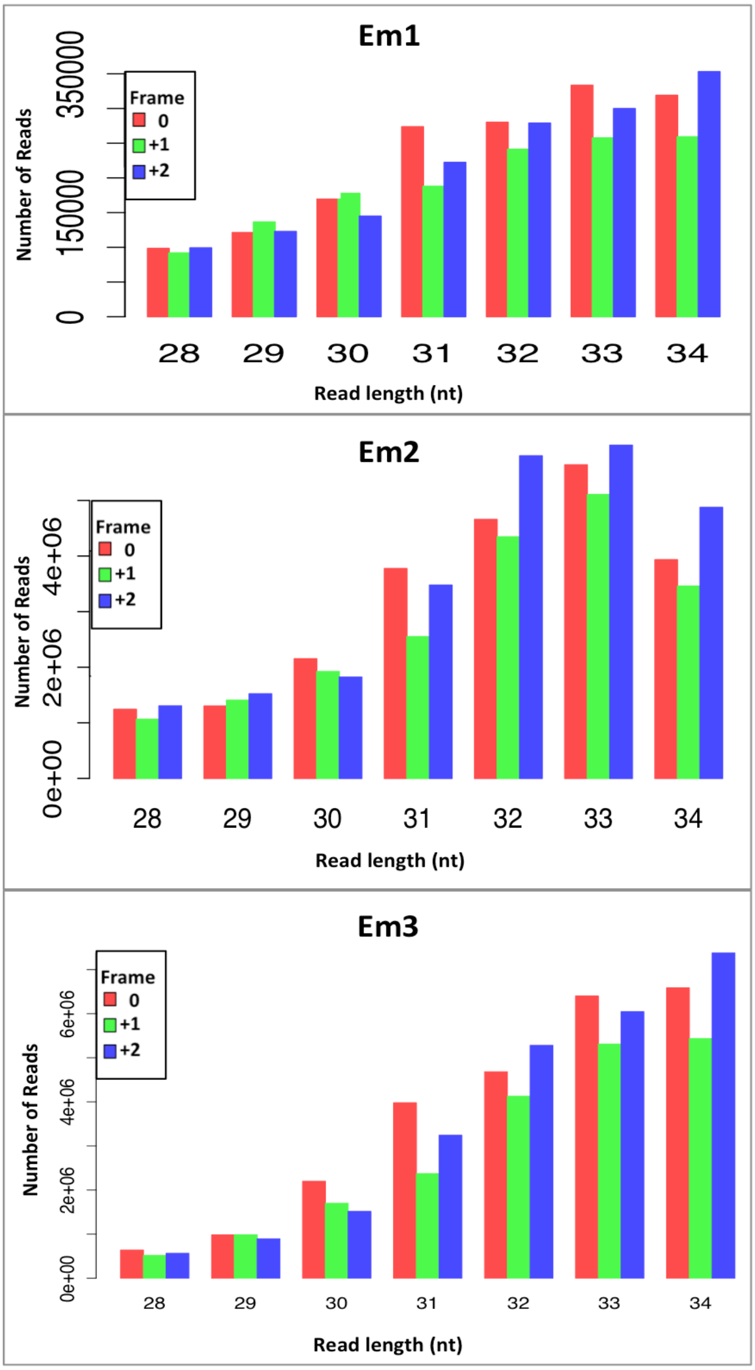
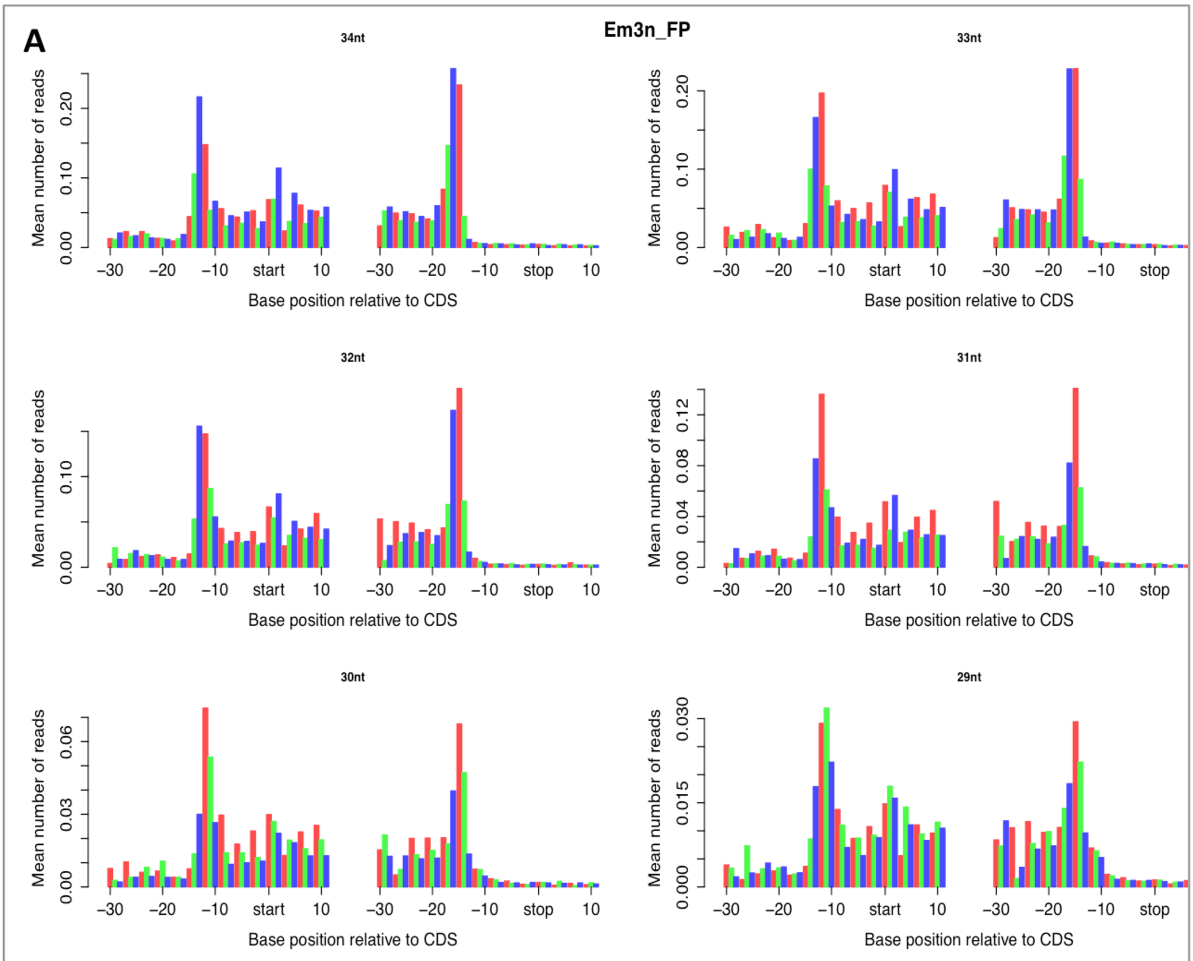


Figure 5.9 Framing of different read lengths using RiboSeqR

The Ribo-SeqR package in R (Chung *et al.* 2015) was used in order to visualise the framing of RBF reads in the *Em1, Em2 and Em3* datasets. The software is provided with a *FASTA* file containing all annotated mRNA transcripts and annotates all ORFs beginning with an ATG and an in-frame stop codon. It then maps the 5' end of all reads relative to the frame of translation (0, +1, +2) in these putative ORFs to calculate the framing for each RBF read length (28-34nt). As can be seen in the graphs, (the y-axis: number of reads) it can be observed that the RBF reads in our data do not show a clear preference or bias to any particular reading frame.



B

Read Length	**Em1 offset (nt)**	**Em2 offset (nt)**	**Em3 offset (nt)**
28	**12**/11	**11**	**12/11**
29	**12**/11	**12**/11	12/**11**
30	**12**/11	**12**	**12**
31	**13**	**12**/13	**12**
32	**12**/11	**13/12**	**13**=12
33	**13**/12	**13/12**	13/**12**
34	**13**=12	**13**	**13**

Figure 5.10 Estimation of 5' offset values for P-site mapping

A) Ribo-SeqR was used calculate the 5' nucleotide off-set from the proposed P-Site for each read length (29-34nt) using the *PlotTranscript* function. The results are visualised by plotting the 5' ends of all reads overlapping the putative 'start' and 'stop' codons detected by the FindCDS function, and mapped on to a collapsed ORF model. The graphs show that the 5' offset is not consistent between reads of the same length in the *Em3* dataset and therefore the P-site mapping is usually split between two nucleotide positions (and thus frames) for a given read length.

B) The results of this analysis carried out on all read lengths (28-34nt) in all three embryo datasets are summarised in the Table shown. The values represent the off-set value determined by visual scoring of the graphs shown in A) and the values highlighted in bold represent the off-set chosen for subsequent P-site mapping.

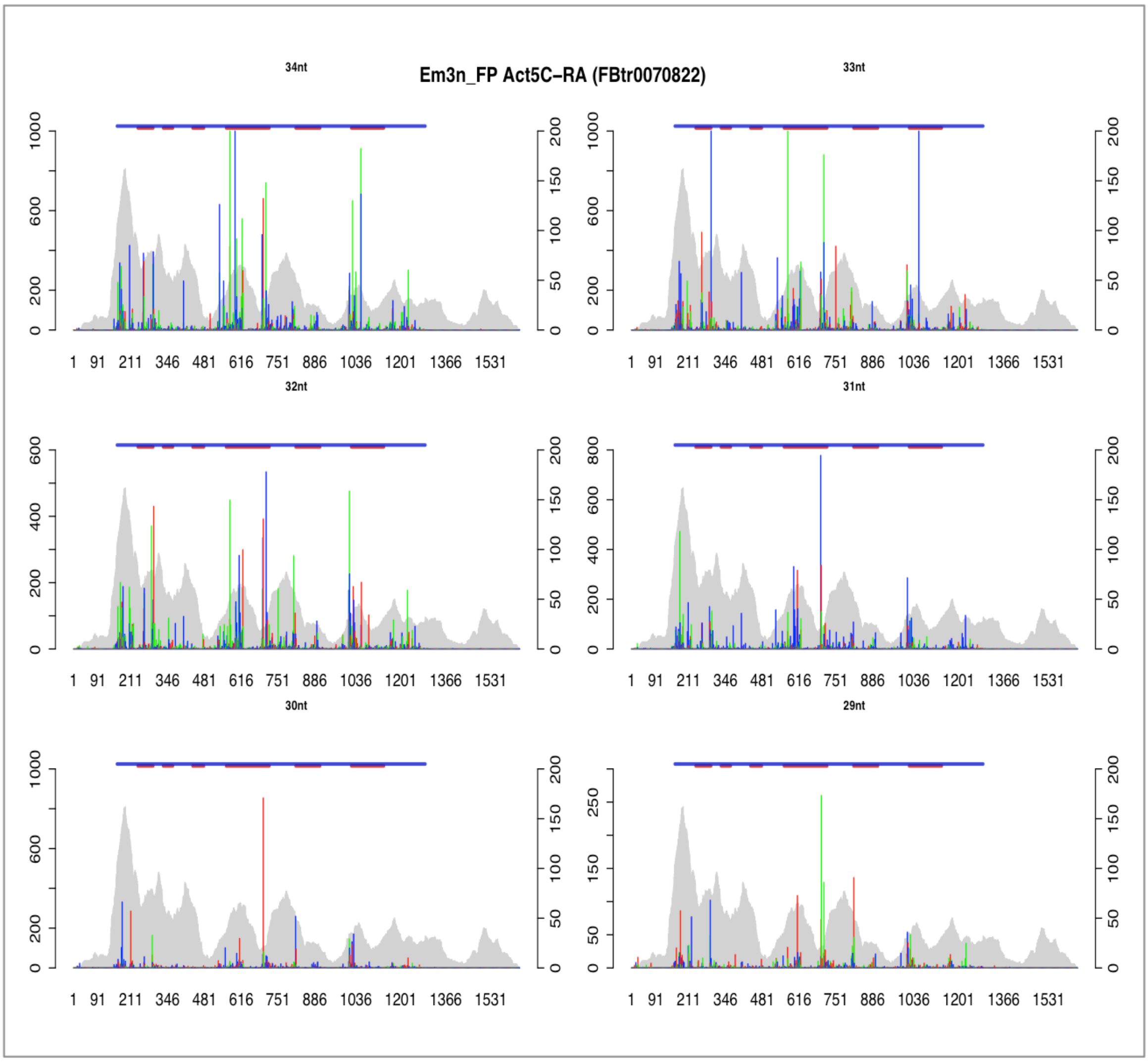


Figure 5.11 Visualisation of P-site mapped reads across the Actin5C CDS

This figure shows the mapped P-sites of RBF reads of each read length (29-34nt) using the values in bold in Figure 5.10B, that align to the highly transcribed and translated Actin-5C-RA transcript in the *Em3* dataset. The background in grey represents the coverage of RNA-Seq reads and the different coloured peaks represent the frame of translation of the mapped P-site position relative to the start codon (Red-Frame 0, Green-Frame 1, Blue-Frame 2). The y-axis shows the number of reads and it can be seen that framing can reveal the translation of the longer canonical ORF (Blue), but also identifies 6 small ORFs (Red), based on a minimum threshold of 50 in-frame reads, that overlap the annotated CDS. in the very highly transcribed and translated *Actin 5C* (*Act5C*) transcript. This data highlights a problem of P-site mapping in the absence of clear framing. Due to the ambiguity in mapping P-sites, it is difficult to discern whether these reads represent genuine translation of alternative reading frame ORFs or due to ambiguity of P- site mapping.

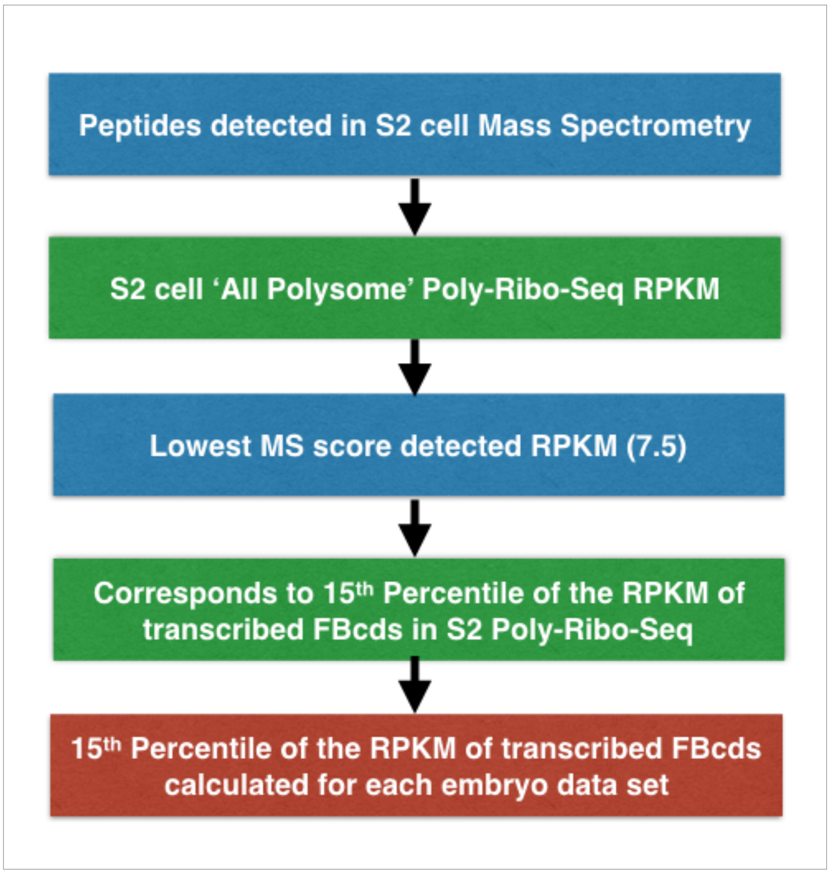

Peptides detected in S2 cell Mass Spectrometry
S2 cell ‘All Polysome’ Poly-Ribo-Seq RPKM
Lowest MS score detected RPKM (7.5)
Corresponds to 15th Percentile of the RPKM of transcribed FBcds in S2 Poly-Ribo-Seq
15th Percentile of the RPKM of transcribed FBcds calculated for each embryo data set

Figure 5.12 Adapting the *S2* MS-based translation cut-off to a 15th Percentile *RPKM* cut-off for Embryo data

This figure shows process by which I arrived at a translation cut-off based on the 15th Percentile of the *RPKM* values for the CDS of all annotated protein-coding genes transcribed at each embryo stage. Poly-Ribo-Seq abundance estimates (*RPKM*) of all of the FB annotated protein-coding genes that were detected by the small protein enriched MS experiments conducted in *S2* cells were analysed to select the lowest *RPKM* value of 7.54. I observed that this *RPKM* value corresponded to the 15th Percentile of the *RPKM* values for the CDS of all annotated protein-coding genes (FBcds) transcribed in *S2* cells. Since we had not performed MS on any embryo samples and the *RPKM* metric is not suitable for direct comparison across disparate samples, I used this 15th Percentile-based translation cut-off in each of the embryo datasets as it is more representative of the translational state of each sample.

A

Translated	*FBcds*	*FB smORF*	*ncrORF*	*uORF*
Em 1	*12,870 (83%)*	*376 (74%)*	*348 (15%)*	*2,863 (34%)*
Em 2	*14,387 (83%)*	*422 (76%)*	*125 (4%)*	*2,134 (21%)*
Em 3	*13,813 (83%)*	*430 (81%)*	*282 (13%)*	*3,926 (40%)*
Total Unique	*16,375*	*493*	*488*	*4,839*
New in Embryo	*5,477 (33%)*	*302 (61%)*	*334 (68%)*	*3,118 (64%)*

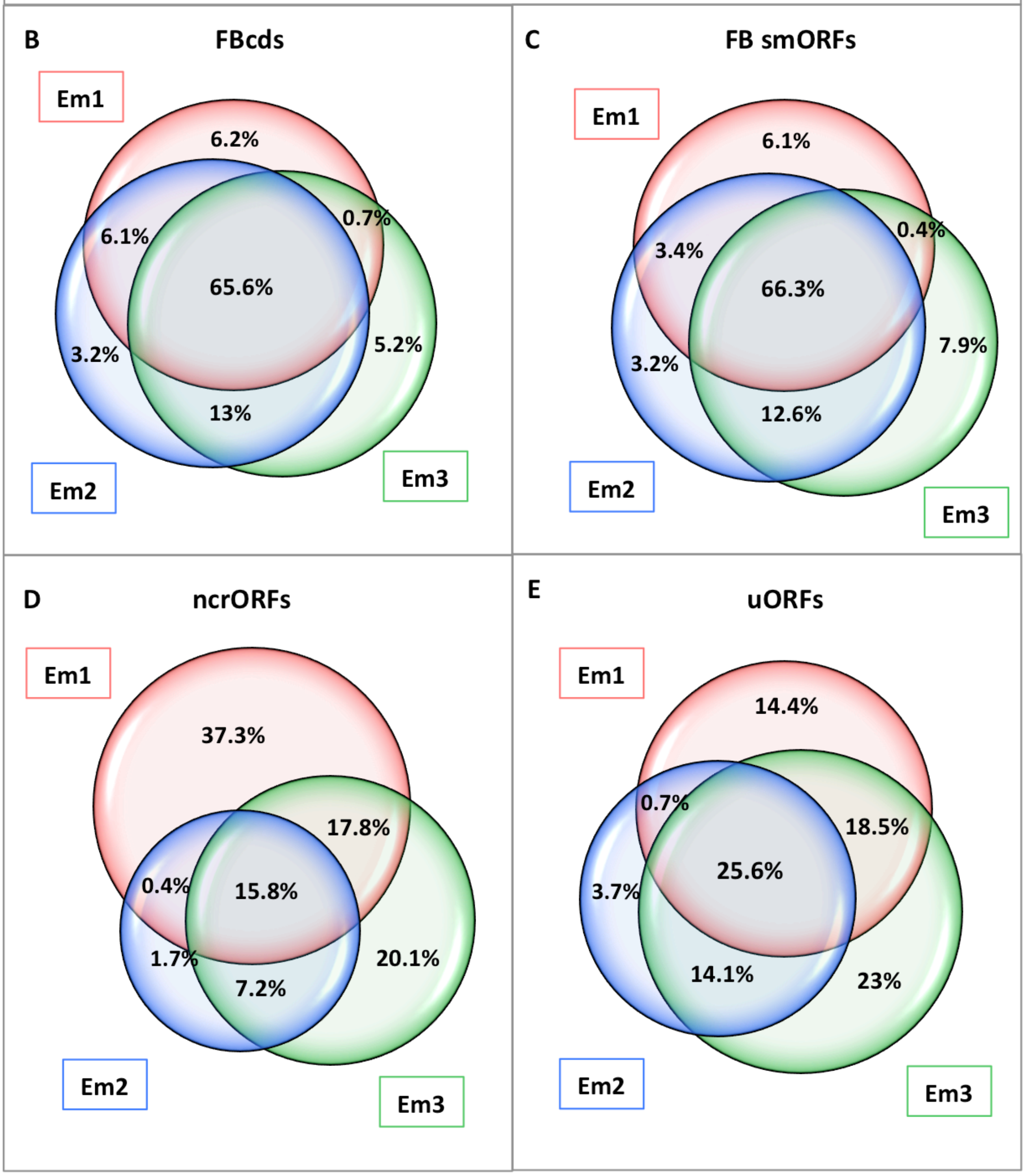


Figure 5.13 Numbers of translated ORFs using a 15th-percentile FBcds *RPKM* cut-off and patterns of translation across Embryogenesis

A) A summary table showing the final numbers of annotated longer protein-coding ORFs (FBcds), annotate smORFs (FB smORFs), long ncRNA smORFs (ncrORFs) and 5'UTR smORFs (uORFs) translated according to the translation cut-off based on the 15th Percentile of the *RPKM* values of transcribed FBcds at the *Em1* (Red), *Em2* (Blue) and *Em3* (Green) stage embryo datasets. The numbers in parenthesis show the proportion of transcribed ORFs that are translated using this cut-off. We can see translation of 493 FB smORFs, 488 ncrORFs and 4,839 uORFs and 16,375 FBcds across the whole of embryogenesis (Total Unique). About two-thirds of each type translated do not have previous evidence of translation (New in Embryo) in the published *S2* cell Poly-Ribo-Seq dataset (Aspden et al. 2014).

B) Venn diagrams of percentages in each data set of the total FBcds translated across embryogenesis. 65.6% of all Fbcds are translated throughout embryogenesis

C) Venn diagrams of percentages in each data set of the total FB smORFs translated across embryogenesis. Similar to FBcds, 66.3% of all annotated smORFs are translated in all three stages

D) Venn diagrams of percentages in each data set of the total ncrORFs translated across embryogenesis. Unlike FBcds and FB smORFs, only 15.8% of all ncrORFs are translated throughout embryogenesis. and only a further 9% are translated in *Em2* which has a very low level of ncrORF translation (4% of transcribed - Panel A). ncrORF translation appears to be mainly stage specific with 37.3% in *Em1* and 20.1% in *Em3*.

E) Venn diagrams of percentages in each data set of the total uORFs translated across embryogenesis. 25.6% of uORFs are translated throughout embryogenesis. The translation of uORFs is more prevalent at stage *Em3*, with 23% of uORFs specific to that stage and 81% of all uORFs being translated in *Em3* embryos.

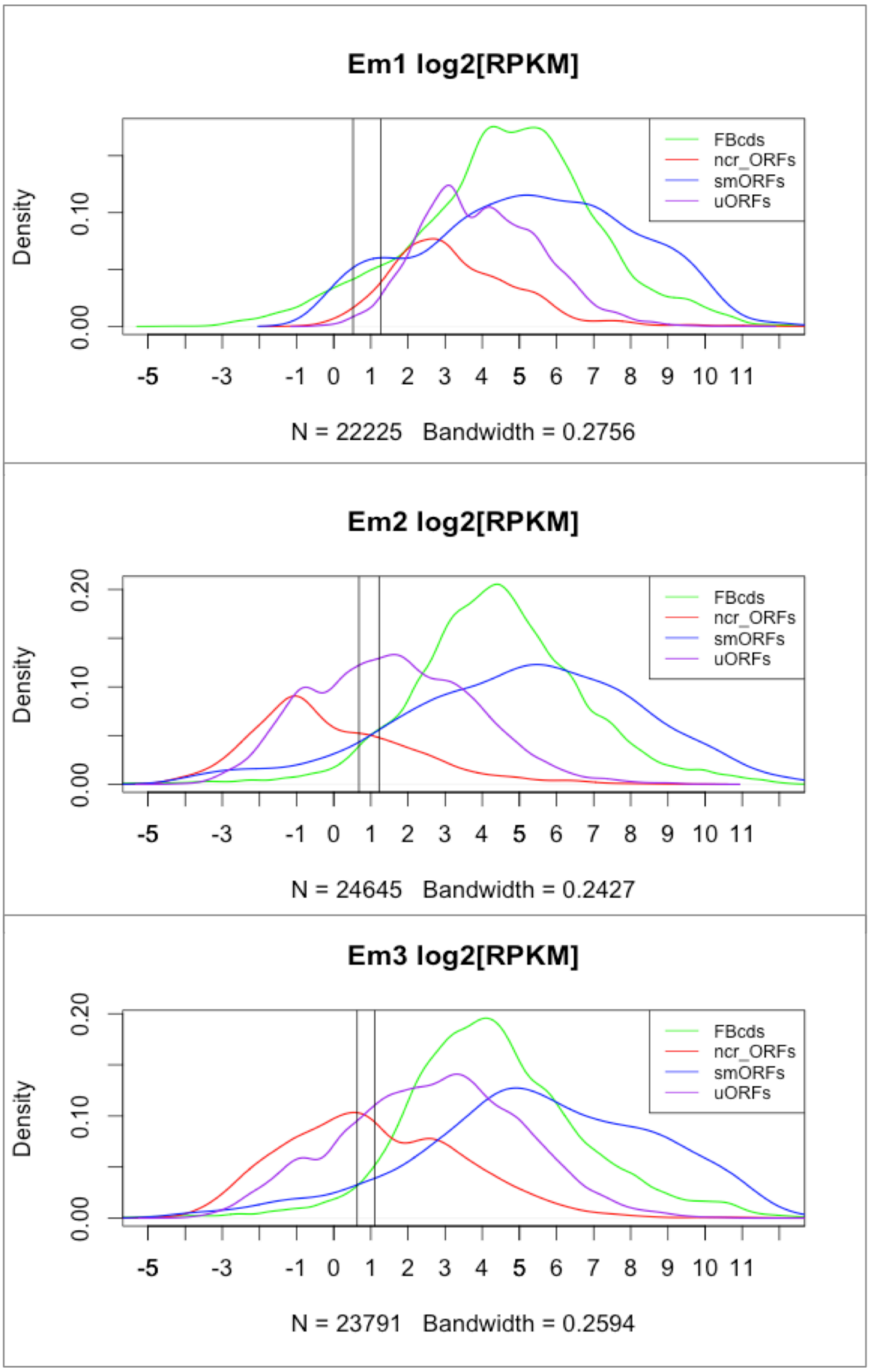
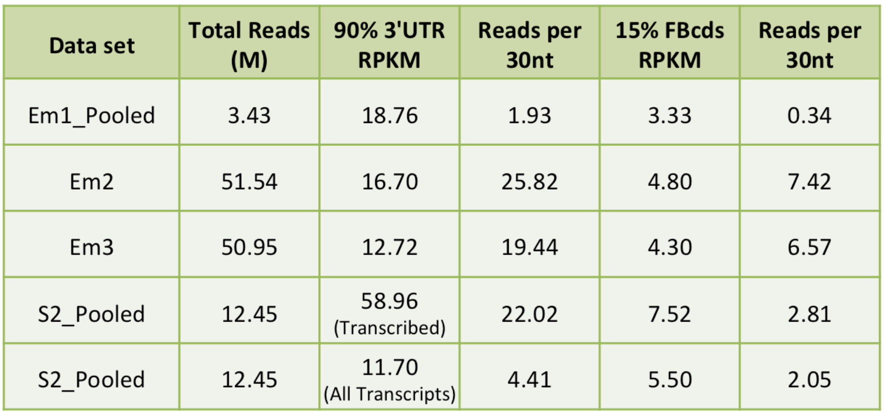


Figure 5.14 Density plot of log2 *RPKM* values of transcribed ORFs

This figure shows the density plot of the log2 *RPKM*s for FBcds (Green), ncrORFs(Red), smORFs (Blue), uORFs (Purple) transcribed in *Em1*, *Em2* and *Em3* datasets. The vertical lines represent the highest (90th percentile FBcds 3'UTR *RPKM*) and lowest cut-offs (15th Percentile FBcds *RPKM*) tested in this analysis. We can observe that there is a distinct distribution of *RPKM* values for each type of smORF plotted and the log2 of *RPKM* density distribution goes from wider to restricted of FBcds>FB smORFs > uORFs > ncrORFs. ncrORFs at the *Em2* show a distinct decrease in ribosome association in the *Em2* dataset.



Data set	Total Reads (M)	90% 3'UTR RPKM	Reads per 30nt	15% FBcds RPKM	Reads per 30nt
Em1_Pooled	3.43	18.76	1.93	3.33	0.34
Em2	51.54	16.70	25.82	4.80	7.42
Em3	50.95	12.72	19.44	4.30	6.57
S2_Pooled	12.45	58.96 (Transcribed)	22.02	7.52	2.81
S2_Pooled	12.45	11.70 (All Transcripts)	4.41	5.50	2.05

Table 5.5 Comparison of translation cut-offs in terms of number of reads

This table compares the 90th percentile transcribed annotated protein coding transcript 3'UTR *RPKM* cut-off with the 15% FBcds CDS *RPKM* cut-offs with regard to the total number of reads in the BAM file (in Millions) and number of reads per 30nt, which is minimum ORF size used in the identification of dwarf smORFs. The Poly-Ribo-Seq datasets compared are *Em1*, *Em2* and *Em3*. For comparison, the pooled *S2* data is included, which shows the numbers for cut-offs based on transcribed transcripts and 'All transcripts' used in the published data (Aspden *et al.* 2014). The final cut-off used for the Embryo analysis (15% FBcds *RPKM*) gives some 6.5-7.5 reads per every 30bp for the highly sequenced *Em2* and *Em3* dataset.

Chapter 6: General Discussion

Putative small open reading frames (smORFs), potentially coding for smORF-encoding peptides (SEPs) less than 100 amino acids, have been shown to outnumber annotated protein-coding genes by several orders of magnitude in all genomes through computational examination of the sequence (Mumtaz, and Couso 2015). However, smORFs have generally been excluded from genome annotation efforts due to their sheer numbers, lack of homology to known protein sequences and lack of functionally characterised examples, problems that originate from difficulty in detecting SEPs by traditional biochemical methods. Conventional bioinformatics strategies used to score homology to known protein coding sequences, or the ratios of synonymous to non-synonymous substitutions in the nucleotide sequences are generally unsuccessful in identifying functional smORFs. Due to their small size, smORFs usually fail to pass the significance threshold used for canonical gene annotation because short smORF sequences do not provide enough sample for rigorous sequence analysis (Ladoukakis *et al.* 2011). Although, a few genome-wide studies had detected SEPs using mass spectrometry, (MS) (Svensson *et al.* 2003; Oyama *et al.* 2004; Frith *et al.* 2006) and functional smORF genes had been identified in yeast using functional genomics (Kastenmayer *et al.* 2006), these were generally regarded as the exception rather than the rule, and did not provide a large enough pool to compare SEPs to a wider pool of small peptides.

RNA-Seq has enabled the high-throughput examination of transcriptomes and vastly expanded our knowledge of the transcriptome; showing that up to 85% of the genome is transcribed in many different species (Hangauer *et al.* 2013). These findings have culminated in the discovery and attempted characterisation of widespread, novel genomic elements such as non-coding RNA genes, which have led to the revision of the definition of a gene that was based on the central dogma of biology. The most interesting of these non-coding RNAs are the mRNA-like ‘long’ non-coding RNAs, many of which are not really non-coding, as they contain ‘Dwarf’ smORFs (Mumtaz, and Couso 2015). Due to the disparity between peptide detection of these smORFs, the default course of action upon encountering such small, putatively coding sequences has been to annotate these transcripts as putative long non-coding RNAs. This self-feedback loop of mis-annotation has resulted in smORFs being largely ignored.

These genome annotation conventions have hampered the studies that would provide a basis for in-depth functional characterisation of smORF genes and their corresponding peptides, as quite simply, they are ‘flying under the radar’. The usually reliable strategy of detecting peptides through Immuno-staining by specific antibodies has proven largely unsuccessful with SEPs, therefore these studies rely upon alternative methods to prove the existence of these peptides, such as epitope-tagging for detection, and then using indirect genetic methods, to subsequently show that the functional unit of the gene is indeed the translated peptide and not the RNA (Galindo *et al.* 2007; Magny *et al.* 2013; Pueyo *et al.* 2016). The findings of these studies have revealed the roles of tiny peptides in development, physiology and cell biology. Therefore, there had arisen a significant need for a more high-throughput strategy to assess the translation of smORF sequences, as to establish them as a significant class of genes, to which the evidence has been mounting, upon the emergence of single-smORF gene studies.

The recently developed technique of Ribo-Seq, which is the deep sequencing of ribosome footprints (Ingolia *et al.* 2009), has presented a comprehensive biochemical approach towards addressing the issue of detecting smORF translation by setting a precedent of showing pervasive translation outside of annotated protein-coding regions (reviewed in Ingolia (2014) and Jackson and Standart (2015)). Our lab applied an adaptation of this technique, called Poly-Ribo-Seq, in *Drosophila S2* cells to show the translation of longer annotated ‘Flybase’ smORFs (median length 80aa) at a similar proportion to that of canonical protein coding genes (81% of total transcribed) and lower levels of translation (34% of transcribed) of ‘Dwarf’ smORFs (median length: 20aa) in long ncRNAs and 5’UTRs (Aspden *et al.* 2014). In a similar controversy to RNA-Seq (Kapranov, and St Laurent 2012), the results of these early Ribo-Seq studies generated a significant amount of skepticism regarding smORF translation, particularly of those in long ncRNAs (Bánfai *et al.* 2012; Guttman *et al.* 2013; Chew *et al.* 2013).

The work presented in this thesis includes the further exploration of the translation of smORFs in *Drosophila melanogaster,* firstly by the independent corroboration of SEP translation, with complementary techniques such as transfection-tagging and peptidomic methods, as has been described in Chapter 3. To further expand the catalogue of smORFs expressed in the fruit fly I then performed a high-throughput assessment of smORF translation of three discrete stages that cover the entirety of

Drosophila embryonic development. While adapting the Poly-Ribo-Seq protocol to *Drosophila* embryos, I implemented significant improvements to the yield and quality of the data (Chapter 4), which help to reduce some of the technical challenges of the application of Ribo-Seq to samples that are difficult to collect. Finally, Chapter 5 addressed the immense debate in the field with regards to data analysis of Ribo-Seq experiments. Various computational metrics have been developed, aimed at discerning 'real' translation events compared to background noise (reviewed in Brar and Weissmann (2015)). Chapter 5 explores the suitability of using these metrics for scoring the translation of smORFs and led to the establishment of a translation cut-off suitable for the high confidence translational assessment of smORFs.

In vitro Translation methods are unsuitable for small peptide detection

Our initial attempt at more high-throughput methods of SEP detection and corroboration of smORF translation was to try and corroborate the results of Poly-Ribo-Seq defined translation of smORFs by using *in vitro* translation (IVT). The IVT method is a widely accepted proof of translation that can be achieved rapidly in a cell-free system. IVT detection is effected through the incorporation of fluorescently labeled amino acids; therefore it does not require the use of a specific antibody, which is a method that we have been moving away from due to the largely unsuccessful application of these towards the detection of SEPs in previous instances (Galindo *et al.* 2007; Magny *et al.* 2013; Pueyo *et al.* 2016). In this attempt, a variety of different IVT systems were tested, in order to produce detectable peptides, however these attempts proved unsuccessful at detecting the endogenous SEP, as I was not able to observe any fluorescent signal directly from labeled SEPs using IVT of native smORF constructs. This could be postulated to have happened due to the low numbers of Lysines that naturally occur in short SEP sequences, although the more likely explanation for not being able to detect SEPs using this method was due to the masking of any fluorescent band in the SEP size range on the gel due to the high level of background signal from Globin. As previously discussed, Globin is required for the IVT reaction and there was no option to exclude it from the final samples (Pelham, and Jackson 1976).

In order to test the hypothesis of insufficient labeling, I used a template containing the small FLAG-tag (28aa) in smORF constructs to detect translation products by Western blotting. This approach was again not particularly successful as

SEPs could not be detected and Globin background was still a hindrance using gel-based, Western blot detection. However, when this experiment was repeated using the much larger Venus tag (238aa), Venus-fusion SEPs made by IVT could be detected. despite the poor resolution between the Venus tag-only band and SEP-Venus fusion protein. This supported the observation of band masking by Globin and explained that the lack of SEP detection is due to their small size. In addition, there is evidence in the literature that peptides smaller than 100 amino acids are actively degraded in bacterial IVT systems (Loose 2007) and the same may be true for eukaryotic systems; therefore the large Venus tag, which is more than double the size of the largest SEPs, may have enabled SEPs to escape degradation whilst the endogenous product was being degraded too quickly to be detected. As the use of tagged constructs reduces the throughput of IVT, the *in vitro* translation method was therefore deemed unsuitable as a high-throughput assessment of smORF translation.

The Tagging transfection assay is an effective approach for the reliable detection of FLAG tagged SEPs

FLAG and Venus tagged smORF constructs were made in our lab to study the expression and sub-cellular localisation of SEPs through fluorescent microscopy performed by another member of the lab (Unum Amin). However, the high levels of translation of the Venus tag through downstream start codons, as detected by the Western blots presented in this thesis, informed that we could not use the Venus tagged constructs for IVT (as discussed above) and in the microscopy study due to false positives and Venus signal presenting as background in the cell and the nucleus. As the Venus tag is widely used as an epitope tag, these results should inform future studies that use a similar imaging approach to visualise the subcellular localisation and expression of proteins.

Accordingly, we restricted the use of the tagging-transfection assay to express only 3x FLAG-tagged SEPs in *S2* cells. The tagging-transfection assay was carried out in *S2* cells using C-terminal tagged smORF constructs with an SV40 3'UTR sequence and a constitutive pAct promoter, using the translational machinery of the cells to translate the fusion peptide, and then subject the lysate of these transfections to Western blotting for detection of the expressed FLAG, which would only occur if the smORF sequence (and 5'UTR) initiated the translation. Although, this method is more labour intensive than IVT due to cloning of recombinant smORF constructs (performed by R.J. Phillips and designed by J.L. Aspden) and subsequent transfection of *S2* cells; this approach was far better suited to SEP detection than IVT, as it allowed us to take a multi-pronged approach to their detection, using both microscopy and Western blotting.

The results from my tagging-transfection assay experiments offers an excellent proof-of-principle for the translation of the *S2* cell Poly-Ribo-Seq smORF sequences that were tested, as 17 of 18 FB smORFs tested were detected using Western blotting as well. In addition, 7 smORFs occurring in 2 long ncRNA genes were tested in this assay and the translation of 4 ncrORFs across both genes could also be observed. The microscopy approach detected all the smORFs tested, as the expression could be searched for by eye using fluorescent labelling of FLAG in the cells. However, the Western-blot method was valuable in helping to determine biochemical properties of the smORFs. The other advantage of this approach was the capability to corroborate the size of peptide on the gel and compare that to the proposed transcript to better inform the gene model, and to see which smORFs in polycistronic transcripts have translation potential. For example, pncr009:3L ORF4 is detected at just below the 15KDa marker on the gel as opposed to the predicted size of 8.2KDa, which led us to re-examine the cloned smORF sequence. This led to the discovery of an upstream in-frame start codon as the most likely source of translation initiation, which would not have been determined by microscopy.

Some of the SEPs detected in this assay included those with metrics below the translation cut-offs used in the *S2* cell data (Aspden *et al.* 2014). However, as the amount of signal in the Western blot of the SEPs did not correlate well overall with the Poly-Ribo-Seq data *RPKM*, coverage and translation efficiency (TE) metrics for these smORFs, it was difficult to determine whether this was due to the cut-offs. Their detection might be due to the fact that smORF-FLAG transcripts are transcribed at artificially high levels using the Actin5C promoter. In addition, the smORF sequences are not in an endogenous context since the construct does not include 3'UTR sequences. Finally, and perhaps most importantly, this approach does not take into account the sampling conditions of Poly-Ribo-Seq, the stability of the endogenously translated peptide or its detectability by the Western blotting technique, the latter two of which are dependent on the size of the peptide. This is highlighted by the fact that the smORF-FLAG constructs that were detected in the transfection-tagging assay coupled with

Western blotting generally encoded for SEPs that are larger in size (median: 13KDa) than those that could not be detected (median: 7KDa).

Small protein enrichment with Mass Spectrometry enables effective detection of longer, more abundant SEPs

Recent improvements in mass spectrometric (MS) technology, such as nano HPLC coupled with LC-MS/MS, have greatly improved the sensitivity of this peptide detection technique. MS combined with small-protein fractionation has given rise to the field of Peptidomics, which has in recent years shown some significant success in the detection of SEPs (Slavoff *et al.* 2013; Ma *et al.* 2014; Vanderperre *et al.* 2013; Bazzini *et al.* 2014). This thesis described a number of different methods that were tested in *Drosophila S2* cells in order to enrich for small proteins, such as differential solubilisation, ultrafiltration and SDS-PAGE. The SDS-PAGE enriched sample was the most successful in the mass spectrometric detection of 75 Flybase annotated SEPs, of which 60 SEPs were matched with high confidence according to generic MS metrics. Over half of the total SEPs detected did not have any previous peptidomic evidence, which highlights the value of using small protein enrichment in a proteomics workflow geared towards detecting SEPs, as mass spectrometry technique tends to detect only the most abundant peptides (Aspden *et al.* 2014).

Dwarf smORF encoded peptides cannot be easily detected using conventional Mass Spectrometry methods

The MS spectra from the SDS-PAGE enriched protein samples were also matched against a custom database designed to contain peptide sequences of all the ncrORFs and uORFs annotated in our lab. This failed to detect any of these ‘Dwarf’ smORFs using the conventional probability-based confidence score (Perkins *et al.* 1999). In this regard, the small size of these particular SEPs yet again poses a significant challenge to their biochemical detection as the majority of the peptides generated by trypsin digestion can arise from larger proteins. This is due to the fact that very small peptides do not have the capacity to generate a large number of trypsin-generated fragments, while larger proteins can generate numerous trypsin-generated fragments depending on their size. High confidence scoring in MS is based on fragment number and abundance and therefore by default, SEPs are bound to obtain low scores in proteomics analysis. The average size of a peptide generated by trypsin digestion is 14 amino acids (Burkhart *et al.* 2012) therefore a dwarf SEP of ~20 amino acids would generate 2 unique tryptic peptides at the most. This highlights the main reason why scoring SEPs by conventional methods developed for canonical protein coding genes, which give rise to an average of 8 tryptic peptides per protein (Brunner *et al.* 2007), have had difficulty in the detection of small peptides using MS.

Our proteomics study was able to detect peptide fragments that mapped to 33 uORFs and 13 ncrORFs, but these did not pass the MASCOT confidence-based score, therefore a manual curation of matched peptides was performed; all peptide matches less than 8 amino acids in length were discarded as well as any remaining peptides that matched to multiple regions in the genome using tBLASTn. Using this approach, we were left with 18 uORFs and 8 ncrORFs, which were then compared to the *S2* cell RNA-seq data to check for transcript expression and eliminate any false-positives, leaving behind 16 uORFs and 1 ncrORF. The low level of overlap between this much smaller pool of MS detected dwarf SEPs and the *S2* cell Poly-Ribo-Seq (3 uORFs) may be a sampling issue of the cell line and technique.

This low rate of detection for Dwarf smORF peptides could be some what expected, given the similarly low number of Dwarf smORF peptides detected by MS in other studies (Slavoff *et al.* 2013; Bazzini *et al.* 2014) and the limitations of the technique discussed above and reviewed in Chu *et. al.* (2015) and Andrews and Rothnagel (2014). Consequently it was decided that Poly-Ribo-Seq is a better-suited approach for a comprehensive, high throughput assessment of smORF translation as MS measurements are not strictly a measure of the translational status of the cell, because they are dependent on the stability of the protein. This is a similar case to where RNA-Seq measurements give an estimate of the abundance of a transcript in a sample but do not factor in the extensive post-transcriptional regulation and stability of the transcripts. Therefore it is generally accepted that RNA levels are not a suitable read out of the actual transcriptional state of the cell (Hayles *et al.* 2010) and thus techniques that measure RNA polymerase activity (such as ChIP-seq, NET-seq, GRO-seq) have been developed to directly assess the transcriptional state of a sample (reviewed in Ferrari *et al.* (2014)). To address these issues, I aimed to improve the data collection and analysis of Poly-Ribo-Seq in *Drosophila* embryos.

Application of Poly-Ribo-Seq to *Drosophila* Embryogenesis

The application of Poly-Ribo-Seq in *Drosophila S2* cells revealed the translation of a large number of smORFs (Aspden *et al.* 2014), however, only one-third of the total FB smORFs and uORFs and only 13% of ncrORFs found in *Drosophila melanogaster* are actually transcribed in the *S2* cell line. Therefore, we decided to implement this technique in whole *Drosophila* embryos to expand the catalogue of translated smORFs as well as to provide a more endogenous and organismal context for their translation. Three 8-hour embryo windows were used as individual samples, in order to cover the 24-hours of *Drosophila* embryogenesis. These stages helped to facilitate in the collection of large amounts of material required for the Poly-Ribo-Seq protocol and provide a developmental context. The datasets were divided into Early-embryogenesis 0-8h After Egg Laying (AEL) (*Em1*), Mid-embryogenesis 8-16h AEL (*Em2*) and Late-embryogenesis 16-24h AEL (*Em3*). As could be imagined, adapting the technique to embryos required some optimisation from the original protocol performed on a simple, cultured cell line.

In *S2* cells, we used Poly-Ribo-Seq to try and enrich for smORFs by selecting for RNAs in the small polysome fraction (≤6 Ribosomes). This technique also allowed the reduction of biochemical noise from non-productive ribosomal binding by selecting for actively translating RNAs (2≥ Ribosomes). The incorporation of polysome fractionation in to the Ribo-Seq technique made an already demanding protocol even more challenging, therefore in embryos, the ‘All Polysomes’ fraction (2≥ Ribosomes) was used rather than just the small (2-6) ribosome fraction to allow for more starting material for the experiment. Furthermore, since uORFs are located in the 5’UTRs of longer protein-coding genes, we may have been losing out on signal from the discarded transcripts where downstream ORFs have more than 6 ribosomes attached. In addition, many lncRNAs are polycistronic and results from the *S2* cell data showed that multiple ORFs may be translated at the same time and so they may be associated with many ribosomes. Finally, as we would be performing the first of this kind of experiment in *Drosophila* embryos across the entirety of embryogenesis, it would be ideal to capture the whole sample, so that we can compare smORFs relative to canonical coding sequences.

As described in Chapter 4, we address the recently highlighted issue of degradation of *Drosophila* ribosomes by RNase I by comparison of the quality of digestion with MNase, which is the preferred enzyme used by others who have performed Ribo-Seq in *Drosophila* (Dunn *et al.* 2013; Miettinen, and Björklund 2015). MNase did not work well in our hands, probably due to the huge difference in digest conditions between Poly-Ribo-Seq and traditional Ribo-Seq. The Poly-Ribo-Seq digest is carried out in an approximately 500-fold greater reaction volume, with 10% Sucrose. Furthermore, it has recently been reported that the use of RNase I results in a larger proportion of reads in the library mapping to small ORFs as well as that MNase digestion yields a larger proportion of 3'UTR mapped reads (Miettinen, and Björklund 2015). This increase in 3'UTR reads may have affected the abundance measurements of the stop codon read-through in *Drosophila* reported by Dunn *et al.* (2013) and therefore it would be interesting to compare this in RNase I generated embryo data such as those generated in this study. My temperature and concentration optimisation experiments with RNase I showed that ribosome degradation is lower at 4°C, which may be due to the increased viscosity of the sucrose solution that could affect the physical conformation and accessibility of Ribosomes to RNase I (Reboud *et al.* 1984). Therefore, as our lab (Aspden *et al.* 2014) and others (Kronja *et al.* 2014) have successfully used RNase I for ribosome footprint generation, I decided to continue using RNase I.

Despite the optimisation of earlier steps, which included embryo harvesting, lysis, polysome fractionation and digestion of mRNA, the main limitation that occurred was that after completing the Aspden *et al.* (2014) protocol, we were not left with enough material for even a small scale (MiSeq) sequencing run (Em JA sample) and this led to the testing of the NEB library preparation kit (Em NEB). The NEB kit offers significantly greater yield (10-fold), due to fewer intermediate gel-purification and RNA precipitation steps. The improvement in yield was surprisingly efficient and the protocol can now be used for other material that is difficult to gather, including specific tissues such as *Drosophila* testes and heads, which based on tissue specific RNA-Seq data, are expected to be a rich source of novel genes such as smORFs (Findlay *et al.* 2009; Young *et al.* 2012; Reinhardt *et al.* 2013). I observed a good general correlation of data between the two library preparations (Em JA and Em NEB) in terms of both abundance measurements (R^2=0.79) and the scoring of smORF translation (80% overlap of FB smORFs). Finally, any lingering concerns regarding nuclease choice and library preparation bias were assuaged by the excellent TE correlation (R^2=0.89) between Kronja *et al.* (RNase I – dual adapter library preparation) and Dunn *et al.* (2013)

(MNase – ssDNA circularisation library preparation) datasets produced from 0-2 hour AEL *Drosophila* embryos (Kronja *et al.* 2014). This led to the decision to switch permanently to the NEB kit for library preparation. Combined with the recent development of using immuno-purification to isolate epitope-tagged Ribsome complexes (Ingolia *et al.* 2014; Williams *et al.* 2014) that can be expressed in specific tissues, this improved protocol proffers exciting avenues in to further discovery of novel protein coding sequences.

Transfer RNAs Are A Major Contaminant In Poly-Ribo-Seq Libraries And Improvements to the protocol Using Sucrose Cushion Purification

In the original Ribo-Seq protocol, Ribosome-RBF complexes are purified from nuclease-treated lysate using a sucrose cushion purification step, which removes most contaminants and leaves behind contaminating sequences that are represented by a handful of short, specific rRNA sequences (Ingolia *et al.* 2012; Dunn *et al.* 2013). Due to our addition of polysome fractionation prior to nuclease digest, the Poly-Ribo-Seq protocol did not use a sucrose cushion purification step and therefore the rRNA sequences are broadly distributed in our samples. Interestingly, my analysis of contaminating sequences showed that transfer RNAs (tRNAs) make up a large proportion of the contaminant sequences in Poly-Ribo-Seq libraries. These tRNA reads could be mapped to the first 32nt of tRNA sequences (5'tRFs) of two abundant tRNAs (tRNA^{Glu} and tRNA^{Asp}) and their proportion changed depending on the library preparation method used; suggesting these highly structured RNA fragments may be preferentially ligated by T4 RNA ligase 1 and therefore over-represented in the sequencing data. 5' and 3'tRFs have recently been emerged as a new class of regulatory small non-coding RNAs with wide ranging regulatory functions (reviewed in Kirchner and Ignatova (2015) and in Wilusz (2015)). Therefore it is important for future studies to be aware of this distortion to abundance measurements.

Although I managed to achieve excellent depletion of these 5'tRFs using specific targeting of tRNA^{Glu} and tRNA^{Asp} for depletion, which allowed a yield up to 46% mRNA reads, their presence informed of an opportunity to further improve the biochemical purification of RBFs through the removal of other polysome-associated background, such as fragments of RNA protected from nuclease digestion by RNA binding proteins. RBFs are size selected (28-34nt) by gel-purification prior to library preparation, however, a significant proportion of RNA fragments outside this range are carried over into the library and sequencing reads outside this size range are then bioinformatically removed during the data analysis step. The incorporation of a sucrose cushion purification step into our protocol resulted in not only removal of tRNA sequences, but also by a significant decrease in the proportion of sequencing reads that were outside the 28-34nt RBF size range, thereby improving the overall sequencing efficiency.

In addition to providing better purification, the sucrose cushion step also facilitated improved rRNA depletion by limiting 70% of rRNA reads to six specific sequences, that could be efficiently removed with biotinylated oligonucleotides. These were therefore also used in supplement to the rRNA depletion beads developed in Aspden *et al.* (2014). This combined depletion strategy allowed me to bring the proportion of rRNA reads down to as low as 33% of the total sequencing reads which is a immense improvement on the 70-90% rRNA reads observed in published datasets.

Overall improvements to the Poly-Ribo-Seq Protocol and the need to Improve Sequencing Efficiency

Just as in the case of mass spectrometry, the small size and lower abundance of smORFs posed a significant challenge to their detection using Ribo-Seq. Dwarf smORFs such as ncrORFs are, on average, 35 times smaller than canonical protein-coding sequences and hence generate far fewer reads, as they cannot accommodate as many ribosomes. Combined with the generally low expression of lncRNAs (Cabili *et al.* 2011), the proportion of reads mapping to ncrORFs in Ribo-Seq data is estimated to be only 0.1% of the total reads (Housman, and Ulitsky 2015). Problematically, the Ribo-Seq libraries had a very low sequencing efficiency due to the high level of rRNA contamination and difficulties in specifically mapping such short sequencing reads. This is exemplified by the fact that only 3.5% of the initial sequencing reads in the *S2* cell Poly-Ribo-Seq dataset could be mapped to ORFs. Therefore Ribo-Seq libraries need to be sequenced at a very high depth in order to accrue a tractable number of reads in order to score the translation of Dwarf smORFs. This high depth of sequencing is especially relevant in the embryo datasets as the 8-hour windows are highly complex samples as compared to *S2* cells, due to the fact that the total number of mapped reads are split between larger number of transcripts. Furthermore there is much higher variation in the level of transcription and translation, such as between a gene that is ubiquitously expressed throughout the 8-hour window and a gene with highly restricted spatio-temporal expression.

The *Em3* dataset achieved 45% of the total reads uniquely mapping to ORFs which is an approximately 15-fold improvement compared to the 3.5% yielded by the *S2* cell Small Polysome dataset published in Aspden *et al.* (2014). This improvement was accomplished by 1) significantly improving rRNA depletion, 2) increasing the proportion of 28-34nt reads by and 3) lowering 3'UTR signal (thus increasing the proportion of reads mapping to ORFs), as discussed above. Using this improved protocol, I was able to obtain ~46 million reads mapping to ORFs for the *Em2* and *Em3* datasets which is about 8 times as many reads as in the *S2* cell dataset (5.9 million reads). This means that the *Em2* and *Em3* datasets have 35x sequencing coverage of coding sequences, which is a significant improvement over the 4.5x sequencing coverage of the *S2* cell dataset and thus allowed a much more reliable scoring of smORF translation as discussed in Chapter 5.

3'UTR signal is an unsuitable measure for defining a translation cut-off

Despite the large number of metrics already developed for trying to score translation in Ribo-Seq data, scoring the translation of smORFs remains significant challenge (reviewed in Mumtaz and Couso (2015) and in Brar and Weissman (2015)). The data analysis pipeline based on Aspden *et al.* (2014) had significant room for improvement as this approach relied on using 3'UTR signal (background) in annotated protein coding transcripts as a measure of defining translation cut-offs. However I questioned the suitability of this approach as the 3'UTR based cut-off as the proportion of 3'UTR reads can change depending on the Library preparation method used. Furthermore there is increasing MS (Slavoff *et al.* 2013; Vanderperre *et al.* 2013; Ma *et al.* 2014) and Ribo-Seq (Bazzini *et al.* 2014; Ingolia *et al.* 2014; Ji *et al.* 2015) evidence of translation of putative downstream ORFs in 3'UTRs (dORF) as well as stop codon read-through in *Drosophila* (Dunn *et al.* 2013); thus challenging the assumption that signal in 3'UTR regions is indeed 'Background' or noise.

Finally, the Aspden *et al.* (2014) data analysis pipeline could not be used due to the cut-off used for the published *S2* cell dataset being calculated as the 90th percentile value of the 3’UTR *RPKM* of all protein-coding transcripts regardless of their transcription in *S2* cells. I employed the use of a more appropriate cut-off, based on the genes transcribed in each dataset, which resulted in much higher cut-off for *S2* cells (*RPKM:* 58.96); thus significantly reducing the proportion of transcribed ORFs that are translated in *S2* cells, to only 25% of FBcds, 64% of FB smORFs, 4% of ncrORFs and 9% of uORFs that are transcribed. Except for the very low proportion of FBcds in *S2* cells, these proportions were similar to what was observed in the embryo datasets using these new, dataset specific 3’UTR cut-offs (Table 5.4). This re-analysis of the *S2* cell data led me to conclude that the 3’UTR signal is not a logically sound property on which to base translation cut-offs.

Translation efficiency is inherently unsuitable for defining the translation of Dwarf smORFs

Translation Efficiency (TE) is a useful metric for highlighting translational regulation and comparing the translation of genes across different datasets through the normalisation of ribosome association by the abundance of the transcript in the sample. Due to this normalisation, TE calculations are dependant on the quality and depth of the control RNA-Seq datasets and as such was not suitable in the case of this analysis; where the sequencing of the RNA-seq libraries was insufficient, and did not show a reliable number of reads across the ORFs. The main issue with TE when it comes to trying to define translation is that on its own, TE is only a ratio and therefore does not differentiate between high and very low expression of genes and therefore usually a threshold of a minimum number of reads is applied (10-50). However this minimum number of reads threshold works against very small ORFs that generally accrue fewer reads. Furthermore, Ribo-seq reads predominantly map to ORFs but RNA-Seq reads do not show such a predisposition; therefore even though a significant number RNA-Seq reads may map to a transcript, very short ORFs such as ncrORFs, which on average make-up only 4% of the transcript, struggle to accrue sufficient RNA-Seq reads. This observation also led me to re-evaluate the transcription cut-off to an *RPKM*>1 across the whole transcript as opposed to across the ORF. TE works well for longer ORFs such as FBcds, which, on average, constitute 70% of the transcript but is unsuitable for dwarf smORFs.

Using Mass Spectrometric detection to inform an *RPKM* cut-off for discerning confidently translated ORFs

Mass Spectrometry (MS) is the most widely accepted proof of peptide translation as this method directly detects protein fragments. From the results of the small protein enriched MS experiment in *S2* cells, it could be observed that MS detection of SEPs can be roughly correlated with Ribo-Seq *RPKM*. As discussed earlier in this chapter, Ribo-Seq is an accurate representation of the translational state of a sample compared to MS, as it is not influenced by the stability of the translated protein, but as peptide detection is the benchmark for Ribo-Seq, it was reasonable to base a cut-off for translation on the lowest *RPKM* value in the *S2* cell Poly-Ribo-Seq data, of a protein detected in my *S2* cell MS data, as it represents the limit of protein detection by MS.

The use of this approach was further justified by further correlation analysis carried out by another member of the lab (P. Patraquim) using previously published and publicly available MS data from *S2* cells (Brunner *et al.* 2007). Figure 6.1 shows the results of this analysis, where a nearly linear correlation (Pearson's r = 0.64, p < 0.0001) can be observed between the *RPKM* of an ORF to the number of tryptic peptides detected by MS (normalised to the length of the protein) that correspond to the protein translated by that ORF. This linear correlation remains statistically significant for an *RPKM* value of 4.2 in *S2* cells, and correlates to the 90th percentile of *S2* cells FBcds *RPKM* across the ORF (proposed cut-off). This statistically viable cut-off is even lower than the value calculated by my analysis of the in-house *S2* cell MS data (*RPKM*: 7.5). Recent proteo-genomic approaches have focussed on using Ribo-Seq and RNA-Seq to improve the MS discovery of novel translated ORFs (Krug *et al.* 2011; Menschaert *et al.* 2013; Koch *et al.* 2014), however, there is no reason why this approach cannot be applied the other way around, by using quantitative MS to confer a reliability of the cut-offs placed upon the Ribo-Seq experiment being conducted. If the two experiments are conducted in parallel, the power of the *RPKM* cut-off being dictated by MS would make the Ribo-Seq assessment indisputable.

RPKM, sample complexity and adapting the MS based cut-off to embryos

RPKM is a measure of the abundance of a transcript relative to all the other transcripts within that sample (Mortazavi *et al.* 2008), therefore it is unsuitable to compare highly disparate samples (Wagner *et al.* 2012) such as the relatively simple *S2* cell line and the data generated from the embryo datasets. The developing embryo can be expected to have much higher variation of transcript abundance than a cell line for example, between a gene that has ubiquitous expression throughout embryogenesis and gene that is expressed in a small subset of embryonic cells, for a fraction of the 8-hour window. This is supported by the calculations which show that a larger proportion of annotated genes are transcribed in embryonic stages (70-79%) when compared to *S2* cells (59%); therefore same number of sequencing reads from each sample would get split amongst many more transcripts in the embryo data. In regards to this, a *RPKM* cut-off of 7.5 may not represent the same absolute abundance between *S2* cell and embryo samples.

This concept can be illustrated by the genes *Distal-less* (*Dll*), *Serotonin transporter* (*sertT*) and *BarH1,* which have *RPKM* values between 2 and 6 across all three embryo Poly-Ribo-Seq datasets, even though endogenous protein expression of all three has been detected in embryos by antibody staining. *Dll* is a homeobox gene that is detected in imaginal disc primordium as well as in a small group of cells that define the anterior/posterior compartment boundary of the head and thoracic hemisegments at embryonic stage 11 (Grenier, and Carroll 2000; Gebelein *et al.* 2004). *sertT* has restricted expression in the ventral nerve cord in stage 15-16 embryos and sertT protein is detected by antibody staining in 2 neurons per hemisegment (Couch *et al.* 2004). BarH1 antibodies stain sub-populations of PNS neurons and dorsal intersegmental muscles in mid and late stage embryos (Higashijima *et al.* 1992). This difference in abundance of transcripts that are temporally regulated could also be interpreted in my analysis using the 80th (medium confidence) and 90th Percentile (high confidence) 3'UTR *RPKM*s cut-offs, which showed that transcripts with a higher level of *RPKM* based abundance are expressed throughout embryogenesis, while those that are in the lower category show stage specific translation. In light of these observations it was important that I developed a translation cut-off that reflects the transcriptional state of the sample.

When the *S2* cell MS-derived *RPKM* cut-off was applied to the *S2* cell Poly-Ribo-Seq data, the proportion of transcribed ORFs that are translated was 84% of FBcds, 77% of FB smORFs, 18% of ncrORFs and 45% of uORFs. These proportions were roughly similar to those published in Aspden *et al.* (2014) and the stringency of these had already been validated using the independent approaches discussed in Chapter 3 of this thesis. The MS-based cut-off conveniently corresponded almost exactly to the 15th percentile of the *RPKM* of the CDS in all transcribed annotated protein-coding transcripts (FBcds) in the *S2* cell dataset. This proportion of translation (15%) has been independently corroborated by the findings of another study that use an framing-based ORF score to assess translation and found that 85% of transcribed, annotated genes in zebrafish are translated across various stages of embryonic development (Bazzini *et al.* 2014). Therefore we decided to use a cut-off based on the 15th percentile of FBcds *RPKM* to define translation in the embryo datasets. The most ideal approach would be to conduct quantitative proteomics in parallel, preferably on an aliquot of the embryonic lysate used for generating the Poly-Ribo-Seq data.

RPKM is the most suitable measure of translation in the absence of framing in *Drosophila* datasets

Despite best attempts, the lack of framing observed in *Drosophila* Ribo-Seq datasets is an unfortunate occurrence, since framing provides the most unambiguous evidence of translation. As was shown in Chapter 5, it is difficult to definitively map the P-site of RBF reads in *Drosophila*, which may be a consequence of the unique structure of *Drosophila* ribosomes and/or their sensitivity to RNase I. This poses a significant disadvantage to the Ribo-Seq studies in *Drosophila*, especially since the majority of the recently developed computational programs specifically developed for advanced analysis of Ribo-Seq data, rely on the tri-nucleotide periodicity of RBFs across the ORF being assessed (Ji *et al.* 2015; Calviello *et al.* 2016). The only exception to these framing-bases programs is the recently developed Fragment Length Organisation Similarity Score (FLOSS), which relies on scoring the similarity of the distribution of different RBF read lengths across an ORF and comparing these distributions to those of annotated protein coding sequences (Ingolia *et al.* 2014). FLOSS and its suitability for scoring smORF translation will be evaluated in further analysis of my datasets in future work.

The examination of different metrics revealed that in the absence of framing in *Drosophila*, similarly to Aspden *et al.* (2014)*,* an *RPKM* based cut-off worked best for scoring smORF translation in embryos. The additional filter for *Coverage* was not used in this analysis since it is dependent on the depth of sequencing and our datasets were sequenced to a very high depth, and showed the highest Coverage of 1 across the ORF for FBcds, FB smORFs and uORFs. Furthermore, it was not necessary to apply the minimum 5 read threshold across the ORF, used in Aspden *et al.* (2014)*,* to my *Em2* and *Em3* datasets due to the significant improvement to the protocol and the consequent increase in the number of reads. Even the smallest ORF size used in our putative smORF annotation (30 nt) would accrue more than 5 reads at the *RPKM* cut-off used to define translation. Accordingly, the mapping of RBF reads and translation of these ORFs is a robust phenomenon, corroborated by a large number of studies (Ruiz-Orera *et al.* 2014), using a variety of approaches (Popa *et al.* 2016).

An in-depth analysis of previously published datasets by Ruiz-Orera *et al.* (2014) shows that more than 95% annotated protein coding transcripts and around 35% of lncRNAs transcripts are shown to be bound by ribosomes, above an abundance threshold of *RPKM* 2. Therefore depending on the tissue or cell line tested and the metrics used to define a translation cut-off, the main difference in findings between the different studies is only the proportion of ORFs translated and there is no longer a question remaining on whether or not translation is occuring (Ji *et al.* 2015). Overall, my study of smORF translation across the whole of embryogenesis shows the translation of approximately 500 Flybase smORFs, 500 ncrORFs (smORFs found in lncRNAs) and 5,000 uORFs, with high confidence. Two-thirds of the numbers of each type of ORF are novel and have not previously defined by other Ribo-Seq experiments.

Translation of ncrORFs is highly regulated during Mid-Embryogenesis

Interestingly, there is an overall decreased ribosomal association of the population of transcribed lncRNAs in *Em2* when compared to the other classes of ORFs at this stage, and between stages. Therefore, the *Em2* dataset shows translation of a significantly lower proportion of ncrORFs (4%) as compared to the *Em1* (15%) and *Em3* (13%) datasets even though a large proportion (20%) of embryo transcribed lncRNAs are exclusively transcribed at this stage. This observation might be explained by the overlap of the *Em2* stage with the highly conserved embryonic phylotypic stage, which overlaps key early developmental processes such as germ band elongation, retraction and head involution and the expression of highly conserved body patterning genes is tightly regulated. This phylotypic stage occurs at 7-13 hours After Egg Laying (AEL) (Domazet-Lošo, and Tautz 2010). It is likely that the translation machinery of the embryo may be recruited to the expression of these conserved developmental genes at this time and thus may not be available for association with long non-coding RNAs. Consequently if 'Dwarf' smORFs in long ncRNAs represent new genes and hence evolutionary novelties as previously suggested (Ruiz-Orera *et al.* 2014), their translation may not be tolerated at this highly conserved stage or as advantageous as at the divergent earlier and later stages or perhaps both.

Concluding Remarks

After expanding and cataloguing the translation of smORFs, the real question that remains now is what proportion of these translated smORFs encode functional peptides. Our lab's previous work with *tarsal-less*, *Sarcolamban* and Hemotin, as well as other SEP studies, have shown that peptides as small as 11 amino acids are functional (reviewed in Saghatelian and Couso (2015)). Furthermore, an RNAi screen conducted in our lab to assess specifically the cellular function of smORFs, showed that when dsRNA is used to knock down the expression of mitochondrial-localised SEPs, there is a direct and observable mitochondrial phenotype for a large proportion of the smORFs tested. Improved bioinformatic tools are being developed to address the issue of function by looking at conservation and motif analysis of small peptides, but so far all this data analysis is trained on the characteristics of canonical protein coding genes (Mackowiak *et al.* 2015). Until there is a tractable number of functionally characterised smORFs to train these programs, bioinformatics will lag behind their discovery.

smORFs also may not be sharing the same functional and bioinformatics characteristics, especially 'Dwarf' smORFs, which are of a unique size category to those which often get detected in various biochemical screens and almost never annotated as protein-coding. Therefore until a significant proportion of these unique smORFs are characterised, the field is in a catch-22 situation where not enough characterised examples exist to provide a training set for computational analysis and the existance of an overwhelming amount of putative smORFs that need further characterisation. This is where the value of the data shown in this thesis is realised, by providing high-quality, reliable data, which can be used for further analysis, not only for identification of translated smORF candidates for further characterisation in *Drosophila* embryos, but also indication towards further bioinformatic analysis of smORF charateristics and the regulation of smORF translation, across embryogenesis.

Chapter 6 Figure

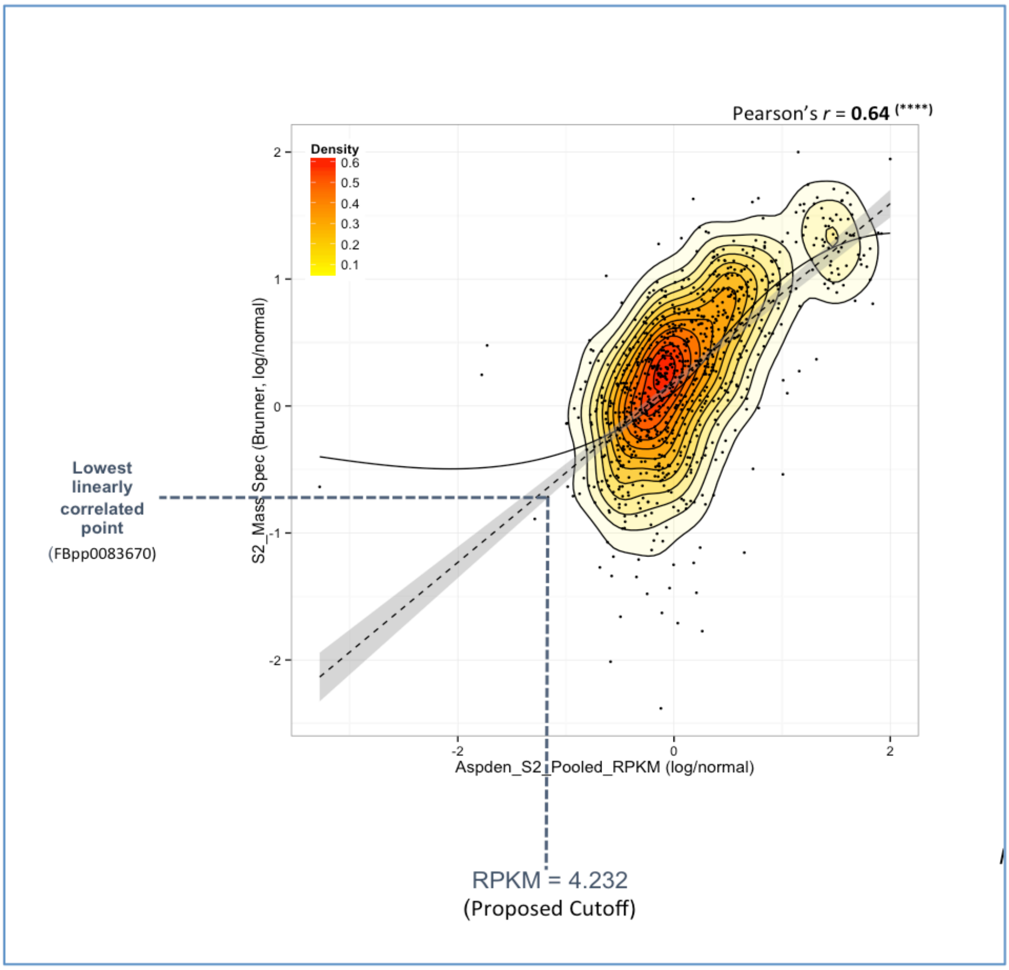


Figure 6.1 Correlation analysis between *S2* Poly-Ribo-Seq data using the correlation between *RPKM* and published mass spectrometry datasets

The correlation between Ribo-Seq *RPKM* (Aspden *et al.* (2014)) and unique MS detected peptides divided by protein length in amino acids using a previously published MS data (Brunner *et al.* (2007)) is shown for all ORFs that feature in both datasets in *S2* cells. The coordinate of each point in the analysis is determined by its *RPKM* value (x axis) and number of Mass Spectrometry hits (y axis) in *S2* Cells. The black dotted line indicates linear regression, while the surrounding grey area shows its 95% confidence interval; the solid black line denotes the results of the lowess regression. A local lowess regression yields identical results to a linear regression analysis for most *RPKM*/Mass spec hits values (see high-density regions orange and red). This indicates that most of the *S2* cell Ribo-Seq *RPKM* range is linearly and positively correlated with mass-spec Hits/a.a. (Pearson's r = 0.64, p<0.0001). Based on this, we determined a translation confidence *RPKM* cutoff by inspecting the linearly correlated point with the lowest *RPKM*. In *S2* cells, this corresponds to the only annotated isoform of the CG31156 gene [all data is log10/normalized].

Chapter 7: References

Ahmad, Q.R., Nguyen, D.H., Wingerd, M.A., Church, G.M. and Steffen, M.A., 2005, Molecular weight assessment of proteins in total proteome profiles using 1D-PAGE and LC/MS/MS, *Proteome Sci*, 3(1), p. 6.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A., 2011, Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, *Genome biology*, 12(2), p. R18.

Allen, J.E., Pertea, M. and Salzberg, S.L., 2004, Computational gene prediction using multiple sources of evidence, *Genome research*, 14(1), pp. 142-8.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990, Basic local alignment search tool, *Journal of Molecular Biology*, 215(3), pp. 403-10.

Andrade, M.A., Daruvar, A., Casari, G., Schneider, R., Termier, M. and Sander, C., 1997, Characterization of new proteins found by analysis of short open reading frames from the full yeast genome, *Yeast (Chichester, England)*, 13(14), pp. 1363-74.

Andrews, S.J. and Rothnagel, J.A., 2014, Emerging evidence for functional peptides encoded by short open reading frames, *Nat Rev Genet*, 15, pp. 193-204.

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), pp. 3889-94.

Aspden, J.L., Eyre-Walker, Y.C., Phillips, R.J., Amin, U., Mumtaz, M.A., Brocard, M. and Couso, J.P., 2014, Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq, *eLife*, 3, p. e03528.

Barbosa, C., Peixeiro, I. and Romão, L., 2013, Gene expression regulation by upstream open reading frames and human disease, *PLoS genetics*, 9(8), p. e1003529.

Barreau, C., Paillard, L. and Osborne, H.B., 2005, AU-rich elements and associated factors: are there unifying principles? *Nucleic acids research*, 33(22), pp. 7138-50.

Basrai, M.A., Hieter, P. and Boeke, J.D., 1997, Small Open Reading Frames: Beautiful Needles in the Haystack, *Genome research*, 7(8), pp. 768-71.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C. and Giraldez, A.J., 2014, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation, *The EMBO journal*, 33(9), pp. 981-93.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., Krajewski, K., Strahl, B.D., Chen, X., Bickel, P., Giddings, M.C., Brown, J.B. and Lipovich, L., 2012, Long noncoding RNAs are rarely translated in two human cell lines, *Genome research*, 22(9), pp. 1646-57.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M., 2004, Global identification of human transcribed sequences with genome tiling arrays, *Science (New York, N.Y.)*, 306(5705), pp. 2242-6.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.K., Ooi, H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N., Yu, Y., Ruan, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Armengol, L., Bird, C.P., de Bakker, P.I., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I.B., Huppert, J., Zody, M.C., Abecasis, G.R., Estivill, X., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. and de Jong, P.J., 2007, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, 447(7146), pp. 799-816.

Blanco, S., Dietmann, S., Flores, J.V., Hussain, S., Kutter, C., Humphreys, P., Lukk, M., Lombard, P., Treps, L., Popis, M., Kellner, S., Hölter, S.M., Garrett, L., Wurst, W., Becker, L., Klopstock, T., Fuchs, H., Gailus-Durner, V., Hrabě de Angelis, M., Káradóttir, R.T., Helm, M., Ule, J., Gleeson, J.G., Odom, D.T. and Frye, M., 2014, Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders, *The EMBO journal*, 33(18), pp. 2020-39.

Brar, G.A. and Weissman, J.S., 2015, Ribosome profiling reveals the what, when, where and how of protein synthesis, *Nature Reviews Molecular Cell Biology*.

Broadie, K., Skaer, H. and Bate, M., 1992, Whole-embryo culture of *Drosophila*: development of embryonic tissues *in vitro*, *Roux's archives of developmental biology*, 201(6), pp. 364-75.

Brunner, E., Ahrens, C.H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E.W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P.G., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A.J., Hafen, E., Schlapbach, R. and Aebersold, R., 2007, A high-quality catalog of the *Drosophila* melanogaster proteome, *Nat Biotechnol*, 25(5), pp. 576-83.

Burkhart, J.M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. and Zahedi, R.P., 2012, Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics, *Journal of proteomics*, 75(4), pp. 1454-62.

Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Biaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J.L. and Raj, A., 2015, Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution, *Genome biology*, 16, p. 20.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L., 2011, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, *Genes and development*, 25(18), pp. 1915-27.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B. and Ohler, U., 2016, Detecting actively translated open reading frames in ribosome profiling data, *Nature methods*, 13(2), pp. 165-70.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C.A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y. and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), 2005, The transcriptional landscape of the mammalian genome, *Science (New York, N.Y.)*, 309(5740), pp. 1559-63.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., Forrest, A.R., Carninci, P., Biffo, S., Stupka, E. and Gustincich, S., 2012, Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat, *Nature*, 491(7424), pp. 454-7.

Cech, T.R. and Steitz, J.A., 2014, The noncoding RNA revolutiontrashing old rules to forge new ones, *Cell*, 157(1), pp. 77-94.

Chew, G.L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F. and Valen, E., 2013, Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs, *Development*, 140(13), pp. 2828-34.

Chu, Q., Ma, J. and Saghatelian, A., 2015, Identification and characterization of sORF-encoded polypeptides, *Critical reviews in biochemistry and molecular biology*, 50(2), pp. 134-41.

Chung, B.Y., Hardcastle, T.J., Jones, J.D., Irigoyen, N., Firth, A.E., Baulcombe, D.C. and Brierley, I., 2015, The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis, *RNA (New York, N.Y.)*, 21(10), pp. 1731-45.

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), pp. 19428-33.

Clark, I.E., Wyckoff, D. and Gavis, E.R., 2000, Synthesis of the posterior determinant Nanos is spatially restricted by a novel cotranslational regulatory mechanism, *Current Biology*, 10(20), pp. 1311-4.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E. and Mattick, J.S., 2012, Genome-wide analysis of long noncoding RNA stability, *Genome research*, 22(5), pp. 885-98.

Clynen, E., Baggerman, G., Husson, S., Landuyt, B. and Schoofs, L., 2008, Peptidomics in drug research, *Expert Opin Drug Discov*, 3(4), pp. 425-40.

Cohen, S.M., 2014, Everything old is new again: (linc)RNAs make proteins!, *The EMBO journal*, 33(9), pp. 937-8.

Costa, E.P., Menschaert, G., Luyten, W., De Grave, K. and Ramon, J., 2013, PIUS: peptide identification by unbiased search, *Bioinformatics (Oxford, England)*, 29(15), pp. 1913-4.

Couch, J.A., Chen, J., Rieff, H.I., Uri, E.M. and Condron, B.G., 2004, robo2 and robo3 interact with eagle to regulate serotonergic neuron differentiation, *Development*, 131(5), pp. 997-1006.

Craig, D., Howell, M.T., Gibbs, C.L., Hunt, T. and Jackson, R.J., 1992, Plasmid cDNA-directed protein synthesis in a coupled eukaryotic *in vitro* transcription-translation system, *Nucleic acids research*, 20(19), pp. 4987-95.

Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W., Van Damme, P. and Menschaert, G., 2015, PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration, *Nucleic acids research*, 43(5), p. e29.

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G. and Menschaert, G., 2013, Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs, *BMC Genomics*, 14, p. 648.

Crick, F., 1970, Central dogma of molecular biology, *Nature*, 227(5258), pp. 561-3.

Crowley, K.S., Reinhart, G.D. and Johnson, A.E., 1993, The signal sequence moves through a ribosomal tunnel into a noncytoplasmic aqueous environment at the ER membrane early in translocation, *Cell*, 73(6), pp. 1101-15.

Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Bienkowska, J., Adams, R.M., Smith, T.F. and Lindelien, J., 1997, Biology's new Rosetta stone, *Nature*, 385(6611), pp. 29-30.

del Prete, M.J., Vernal, R., Dolznig, H., Müllner, E.W. and Garcia-Sanz, J.A., 2007, Isolation of polysome-bound mRNA from solid tissues amenable for RT-PCR and profiling experiments, *RNA (New York, N.Y.)*, 13(3), pp. 414-21.

Dhahbi, J.M., Spindler, S.R., Atamna, H., Yamakawa, A., Boffelli, D., Mote, P. and Martin, D.I., 2013, 5' tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction, *BMC Genomics*, 14(1), p. 298.

Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E., Ru, K., Soldà, G. and Simons, C., 2008a, Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation, *Genome research*, 18(9), pp. 1433-45.

Dinger, M.E., Pang, K.C., Mercer, T.R. and Mattick, J.S., 2008b, Differentiating protein-coding and noncoding RNA: challenges and ambiguities, *PLoS computational biology*, 4(11), p. e1000176.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. and Gingeras, T.R., 2012, Landscape of transcription in human cells, *Nature*, 489(7414), pp. 101-8.

Domazet-Lošo, T. and Tautz, D., 2010, A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns, *Nature*, 468(7325), pp. 815-8.

Duncan, C.D. and Mata, J., 2014, The translational landscape of fission-yeast meiosis and sporulation, *Nature Structural and Molecular Biology*, 21(7), pp. 641-7.

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R. and Weissman, J.S., 2013, Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila* melanogaster, *eLife*, 2, p. e01179.

Falth, M., Skold, K., Norrman, M., Svensson, M., Fenyo, D. and Andren, P.E., 2006, SwePep, a database designed for endogenous peptides and mass spectrometry, *Molecular and cellular proteomics: MCP*, 5(6), pp. 998-1005.

Fatica, A. and Bozzoni, I., 2013, Long non-coding RNAs: new players in cell differentiation and development, *Nat Rev Genet*, 15(1), pp. 7-21.

Ferrari, F., Alekseyenko, A.A., Park, P.J. and Kuroda, M.I., 2014, Transcriptional control of a whole chromosome: emerging models for dosage compensation, *Nature Structural and Molecular Biology*, 21(2), pp. 118-25.

Fickett, J.W., 1995, ORFs and genes: how strong a connection? *J Comput Biol*, 2(1), pp. 117-23.

Findlay, G.D., MacCoss, M.J. and Swanson, W.J., 2009, Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*, *Genome research*, 19(5), pp. 886-96.

Finoulst, I., Pinkse, M., Van Dongen, W. and Verhaert, P., 2011, Sample preparation techniques for the untargeted LC-MS-based discovery of peptides in complex biological matrices, *Journal of biomedicine and biotechnology*, 2011, p. 245291.

Firth, A.E. and Brierley, I., 2012, Non-canonical translation in RNA viruses, *The Journal of general virology*, 93(Pt 7), pp. 1385-409.

Fonslow, B.R., Carvalho, P.C., Academia, K., Freeby, S., Xu, T., Nakorchevsky, A., Paulus, A. and Yates III, J.R., 2011, Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT, *Journal of proteome research*, 10(8), pp. 3690-700.

Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T.L. and Grimmond, S.M., 2006, The abundance of short proteins in the mammalian proteome, *Plos Genetics*, 2(4), pp. 515-28.

Fuchs, R.T., Sun, Z., Zhuang, F. and Robb, G.B., 2015, Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure, *PLoS One*, 10(5), p. e0126049.

Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A. and Couso, J.P., 2007, Peptides encoded by short ORFs control development and define a new eukaryotic gene family, *Plos Biology*, 5(5), pp. 1052-62.

Gebelein, B., McKay, D.J. and Mann, R.S., 2004, Direct integration of Hox and segmentation gene inputs during *Drosophila* development, *Nature*, 431(7009), pp. 653-9.

Gerashchenko, M.V. and Gladyshev, V.N., 2014, Translation inhibitors cause abnormalities in ribosome profiling experiments, *Nucleic acids research*, 42(17), p. e134.

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), pp. 669-81.

Gilbert, S.F., 1997, *Dev Biol,* Sinauer Associates,.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G., 1996, Life with 6000 genes, *Science*, 274(5287), pp. 546, 563-7.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N. and Booth, B.W., 2011, The developmental transcriptome of *Drosophila* melanogaster, *Nature*, 471(7339), pp. 473-9.

Grenier, J.K. and Carroll, S.B., 2000, Functional evolution of the Ultrabithorax protein, *Proceedings of the National Academy of Sciences of the United States of America*, 97(2), pp. 704-9.

Guttman, M. and Rinn, J.L., 2012, Modular regulatory principles of large non-coding RNAs, *Nature*, 482(7385), pp. 339-46.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W. and Cassady, J.P., 2009, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, *Nature*, 458(7235), pp. 223-7.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E. and Lander, E.S., 2011, lincRNAs act in the circuitry controlling pluripotency and differentiation, *Nature*, 477(7364), pp. 295-300.

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. and Lander, E.S., 2013, Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins, *Cell*, 154(1), pp. 240-51.

Guydosh, N.R. and Green, R., 2014, Dom34 rescues ribosomes in 3' untranslated regions, *Cell*, 156(5), pp. 950-62.

Han, Y., Gao, X., Liu, B., Wan, J., Zhang, X. and Qian, S.B., 2014, Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation, *Cell research*, 24(7), pp. 842-51.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H. and Shiu, S.H., 2007, A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection, *Genome research*, 17(5), pp. 632-40.

Hangauer, M.J., Vaughn, I.W. and McManus, M.T., 2013, Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs, *PLoS genetics*, 9(6), p. e1003569.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M. and Gerstein, M., 2002, A question of size: the eukaryotic proteome and the problems in defining it, *Nucleic acids research*, 30(5), pp. 1083-90.

Hartenstein, V., 1993, *Atlas of Drosophila development,* Cold Spring Harbor Laboratory Press,.

Hayden, C.A. and Bosco, G., 2008, Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila* melanogaster and other dipteran species, *BMC Genomics*, 9, p. 61.

Hayles, B., Yellaboina, S. and Wang, D., 2010, Comparing transcription rate and mRNA abundance as parameters for biochemical pathway and network analysis, *PLoS One*, 5(3), p. e9908.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R. and Ordoukhanian, P., 2014, Library construction for next-generation sequencing: overviews and challenges, *BioTechniques*, 56(2), pp. 61-4, 66, 68, passim.

Heyer, E.E. & Moore, M.J., 2016, Redefining the Translational Status of 80S Monosomes, Cell, 164(4), pp. 757-69

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P. and Ulitsky, I., 2015, Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species, *Cell reports*, 11(7), pp. 1110-22.

Higashijima, S., Michiue, T., Emori, Y. and Saigo, K., 1992, Subtype determination of *Drosophila* embryonic external sensory organs by redundant homeo box genes BarH1 and BarH2, *Genes and development*, 6(6), pp. 1005-18.

Housman, G. and Ulitsky, I., 2015, Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs, *Biochimica et biophysica acta*.

Inagaki, S., Numata, K., Kondo, T., Tomita, M., Yasuda, K., Kanai, A. and Kageyama, Y., 2005, Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*, *Genes to Cells*, 10(12), pp. 1163-73.

Ingolia, N.T., 2014, Ribosome profiling: new views of translation, from single codons to genome scale, *Nature Reviews Genetics*, 15(3), pp. 205-13.

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), pp. 1534-50.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J., Jackson, S.E., Wills, M.R. and Weissman, J.S., 2014, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes, *Cell reports*, 8(5), pp. 1365-79.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. 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), pp. 218-23.

Ingolia, N.T., Lareau, L.F. and Weissman, J.S., 2011, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes, *Cell*, 147(4), pp. 789-802.

International Human Genome Sequencing Consortium, 2004, Finishing the euchromatic sequence of the human genome, *Nature*, 431(7011), pp. 931-45.

Ivanov, P., Emara, M.M., Villen, J., Gygi, S.P. and Anderson, P., 2011, Angiogenin-induced tRNA fragments inhibit translation initiation, *Molecular cell*, 43(4), pp. 613-23.

Jackson, R. and Standart, N., 2015, The awesome power of ribosome profiling, *RNA (New York, N.Y.)*, 21(4), pp. 652-4.

Jackson, T.J., Spriggs, R.V., Burgoyne, N.J., Jones, C. and Willis, A.E., 2014, Evaluating bias-reducing protocols for RNA sequencing library preparation, *BMC Genomics*, 15, p. 569.

Jagus, R. and Beckler, G.S., 2003, Overview of eukaryotic *in vitro* translation and expression systems, *Current protocols in cell biology / editorial board, Juan S. Bonifacino*, Chapter 11, pp. Unit 11.1.

Jahn, D., Verkamp, E. and Söll, D., 1992, Glutamyl-transfer RNA: a precursor of heme and chlorophyll biosynthesis, *Trends in Biochemical Sciences*, 17(6), pp. 215-8.

Ji, Z., Song, R., Regev, A. and Struhl, K., 2015, Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins, *eLife*, 4.

Johannes, G., Carter, M.S., Eisen, M.B., Brown, P.O. and Sarnow, P., 1999, Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray, *Proceedings of the National Academy of Sciences of the United States of America*, 96(23), pp. 13118-23.

Kapranov, P. and St Laurent, G., 2012, Dark Matter RNA: Existence, Function, and Controversy, *Frontiers in genetics*, 3, p. 60.

Karaiskos, S., Naqvi, A.S., Swanson, K.E. and Grigoriev, A., 2015, Age-driven modulation of tRNA-derived fragments in *Drosophila* and their potential targets, *Biol Direct*, 10, p. 51.

Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.-C., Yang, H., Carter, C.D., Wheeler, D., Davis, R.W., Boeke, J.D., Snyder, M.A. and Basrai, M.A., 2006, Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae, *Genome Res.*, 16(3), pp. 365-73.

Kawashima, Y., Fukutomi, T., Tomonaga, T., Takahashi, H., Nomura, F., Maeda, T. and Kodera, Y., 2010, High-yield peptide-extraction method for the discovery of subnanomolar biomarkers from small serum samples, *Journal of proteome research*, 9(4), pp. 1694-705.

Kirchner, S. and Ignatova, Z., 2015, Emerging roles of tRNA in adaptive translation, signalling dynamics and disease, *Nat Rev Genet*, 16(2), pp. 98-112.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J., 2012, Counting absolute numbers of molecules using unique molecular identifiers, *Nature methods*, 9(1), pp. 72-4.

Koch, A., Gawron, D., Steyaert, S., Ndah, E., Crappé, J., De Keulenaer, S., De Meester, E., Ma, M., Shen, B., Gevaert, K., Van Criekinge, W., Van Damme, P. and Menschaert, G., 2014, A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites, *Proteomics*, 14(23-24), pp. 2688-98.

Kozak, M., 1978, How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell*, 15(4), pp. 1109-23.

Kronja, I., Yuan, B., Eichhorn, S.W., Dzeyk, K., Krijgsveld, J., Bartel, D.P. and Orr-Weaver, T.L., 2014, Widespread changes in the posttranscriptional landscape at the *Drosophila* oocyte-to-embryo transition, *Cell reports*, 7(5), pp. 1495-508.

Krug, K., Nahnsen, S. and Macek, B., 2011, Mass spectrometry at the interface of proteomics and genomics, *Molecular bioSystems*, 7(2), pp. 284-91.

Kuersten, S., Radek, A., Vogel, C. and Penalva, L.O., 2013, Translation regulation gets its 'omics' moment, *Wiley Interdiscip Rev RNA*, 4(6), pp. 617-30.

Kung, J.T., Colognori, D. and Lee, J.T., 2013, Long noncoding RNAs: past, present, and future, *Genetics*, 193(3), pp. 651-69.

Ladoukakis, E., Pereira, V., Magny, E.G., Eyre-Walker, A. and Couso, J.P., 2011, Hundreds of putatively functional small open reading frames in *Drosophila*, *Genome Biol*, 12(11), p. R118.

Lander, E.S. and Waterman, M.S., 1988, Genomic mapping by fingerprinting random clones: a mathematical analysis, *Genomics*, 2(3), pp. 231-9.

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 Biol*, 10(3), p. R25.

Lareau, L.F., Hite, D.H., Hogan, G.J. and Brown, P.O., 2014, Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments, *eLife*, 3, p. e01257.

Larsson, O., Tian, B. and Sonenberg, N., 2013, Toward a genome-wide landscape of translational control, *Cold Spring Harb Perspect Biol*, 5(1), p. a012302.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B. and Qian, S.-B., 2012, Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution, *Proceedings of the National Academy of Sciences*, 109(37), pp. E2424-32.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. and Dewey, C.N., 2010, RNA-Seq gene expression estimation with read mapping uncertainty, *Bioinformatics (Oxford, England)*, 26(4), pp. 493-500.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup, 2009, The Sequence Alignment/Map format and SAMtools, *Bioinformatics (Oxford, England)*, 25(16), pp. 2078-9.

Lin, M.F., Carlson, J.W., Crosby, M.A., Matthews, B.B., Yu, C., Park, S., Wan, K.H., Schroeder, A.J., Gramates, L.S., St Pierre, S.E., Roark, M., Wiley, K.L.J., Kulathinal, R.J., Zhang, P., Myrick, K.V., Antone, J.V., Celniker, S.E., Gelbart, W.M. and Kellis, M., 2007, Revisiting the protein-coding gene catalog of *Drosophila* melanogaster using 12 fly genomes, *Genome Res*, 17(12), pp. 1823-36.

Lin, X., Cook, T.J., Zabetian, C.P., Leverenz, J.B., Peskind, E.R., Hu, S.C., Cain, K.C., Pan, C., Edgar, J.S., Goodlett, D.R., Racette, B.A., Checkoway, H., Montine, T.J., Shi, M. and Zhang, J., 2012, DJ-1 isoforms in whole blood as potential biomarkers of Parkinson disease, *Scientific reports*, 2, p. 954.

Lipman, D.J., Souvorov, A., Koonin, E.V., Panchenko, A.R. and Tatusova, T.A., 2002, The relationship of protein conservation and sequence length, *Bmc Evolutionary Biology*, 2(1), p. 20.

Loevenich, S.N., Brunner, E., King, N.L., Deutsch, E.W., Stein, S.E., Aebersold, R. and Hafen, E., 2009, The *Drosophila* melanogaster PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation, *BMC bioinformatics*, 10, p. 59.

Loose, C.R., Langer, R.S. and Stephanopoulos, G.N., 2007, Optimization of protein fusion partner length for maximizing *in vitro* translation of peptides, *Biotechnol Prog*, 23(2), pp. 444-51.

Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M. and Saghatelian, A., 2014, Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue, *Journal of proteome research*, 13(3), pp. 1757-65.

Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M. and Obermayer, B., 2015, Extensive identification and analysis of conserved small ORFs in animals, *Genome biology*, 16, p. 179.

MacPhee, D.J., 2010, Methodological considerations for improving Western blot analysis, *Journal of pharmacological and toxicological methods*, 61(2), pp. 171-7.

Magny, E.G., Pueyo, J.I., Pearl, F.M., Cespedes, M.A., Niven, J.E., Bishop, S.A. and Couso, J.P., 2013, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames, *Science*, 341(6150), pp. 1116-20.

Maier, T., Güell, M. and Serrano, L., 2009, Correlation of mRNA and protein in complex biological samples, *FEBS Lett*, 583(24), pp. 3966-73.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M., 2005, Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, 437(7057), pp. 376-80.

McGowan, S.J., Terrett, J., Brown, C.G., Adam, P.J., Aldridge, L., Allen, J.C., Amess, B., Andrews, K.A., Barnes, M. and Barnwell, D.E., 2004, Annotation of the human genome by high-throughput sequence analysis of naturally occurring proteins, *Current Proteomics*, 1(1), pp. 41-8.

Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K. and Van Damme, P., 2013, Deep Proteome Coverage Based on Ribosome Profiling Aids Mass Spectrometry-based Protein and Peptide Discovery and Provides Evidence of Alternative Translation Products and Near-cognate Translation Initiation Events*, *Molecular and cellular proteomics*, 12(7), pp. 1780-90.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F. and Mattick, J.S., 2008, Specific expression of long noncoding RNAs in the mouse brain, *Proceedings of the National Academy of Sciences*, 105(2), pp. 716-21.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S. and Rinn, J.L., 2012, Targeted RNA sequencing reveals the deep complexity of the human transcriptome, *Nature biotechnology*, 30(1), pp. 99-104.

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 research*, 22(11), pp. 2219-29.

Miettinen, T.P. and Björklund, M., 2015, Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions, *Nucleic acids research*, 43(2), pp. 1019-34.

Molecular Cell Biology, L.H.B.A.Z.S.e.a. 2000a, Section 11.6, Processing of rRNA and tRNA, in *Molecular Cell Biology,* W. H. Freeman.

Molecular Cell Biology, L.H.B.A.Z.S.e.a. 2000b, Section 9.1, Molecular Definition of a Gene, in *Molecular Cell Biology,* W. H. Freeman.

Morris, K.V. and Mattick, J.S., 2014, The rise of regulatory RNA, *Nat Rev Genet*, 15(6), pp. 423-37.

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), pp. 621-8.

Mullis, K.B. and Faloona, F.A., 1987, Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction, *Methods in enzymology*, 155, pp. 335-50.

Mumtaz, M.A. and Couso, J.P., 2015, Ribosomal profiling adds new coding sequences to the proteome, *Biochemical Society transactions*, 43(6), pp. 1271-6.

Norman, C., Runswick, M., Pollock, R. and Treisman, R., 1988, Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element, *Cell*, 55(6), pp. 989-1003.

Ohno, S., 1972, Brookhaven Symp Biol, *So much junk DNA in our genome.* pp. 366-70.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schönbach, C., Gojobori, T., Baldarelli, R., Hill, D.P., Bult, C., Hume, D.A., Quackenbush, J., Schriml, L.M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K.W., Blake, J.A., Bradt, D., Brusic, V., Chothia, C., Corbani, L.E., Cousins, S., Dalla, E., Dragani, T.A., Fletcher, C.F., Forrest, A., Frazer, K.S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I.J., Jarvis, E.D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R.M., King, B.L., Konagaya, A., Kurochkin, I.V., Lee, Y., Lenhard, B., Lyons, P.A., Maglott, D.R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W.J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J.U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J.C., Reed, D.J., Reid, J., Ring, B.Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C.A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M.S., Teasdale, R.D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L.G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E.S., Rogers, J., Birney, E., Hayashizaki, Y. and RIKEN Genome Exploration Research Group Phase I and II Team, 2002, Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, *Nature*, 420(6915), pp. 563-73.

Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L. and Menschaert, G., 2016, sORFs.org: a repository of small ORFs identified by ribosome profiling, *Nucleic acids research*, 44(D1), pp. D324-9.

O'Neil, S.T. and Emrich, S.J., 2013, Assessing De Novo transcriptome assembly metrics for consistency and utility, *BMC Genomics*, 14, p. 465.

Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T. and Sugano, S., 2004, Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs, *Genome research*, 14(10B), pp. 2048-52.

Pauli, A., Norris, M.L., Valen, E., Chew, G.L., Gagnon, J.A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., Tsai, S.Q., Joung, J.K., Saghatelian, A. and Schier, A.F., 2014, Toddler: an embryonic signal that promotes cell movement via Apelin receptors, *Science (New York, N.Y.)*, 343(6172), p. 1248636.

Pavlakis, G.N., Jordan, B.R., Wurst, R.M. and Vournakis, J.N., 1979, Sequence and secondary structure of *Drosophila* melanogaster 5.8S and 2S rRNAs and of the processing site between them, *Nucleic Acids Res*, 7(8), pp. 2213-38.

Pelham, H.R. and Jackson, R.J., 1976, An efficient mRNA-dependent translation system from reticulocyte lysates, *Eur J Biochem*, 67(1), pp. 247-56.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S., 1999, Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, 20(18), pp. 3551-67.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A.D., Dehay, C., Igel, H., Ares, Manuel, Vanderhaeghen, P. and Haussler, D., 2006, An RNA gene expressed during cortical development evolved rapidly in humans, *Nature*, 443(7108), pp. 167-72.

Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S. and Koller, D., 2014, Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation, *Molecular systems biology*, 10, p. 770.

Popa, A., Lebrigand, K., Barbry, P. and Waldmann, R., 2016, Pateamine A-sensitive ribosome profiling reveals the scope of translation in mouse embryonic stem cells, *BMC Genomics*, 17(1), p. 52.

Poston, C.N., Krishnan, S.C. and Bazemore-Walker, C.R., 2013, In-depth proteomic analysis of mammalian mitochondria-associated membranes (MAM), *Journal of proteomics*, 79, pp. 219-30.

Pueyo, J.I. and Couso, J.P., 2008, The 11-aminoacid long *Tarsal-less* peptides trigger a cell signal in *Drosophila* leg development, *Developmental biology*, 324(2), pp. 192-201.

Pueyo, J.I., Magny, E.G., Sampson, C.J., Amin, U., Evans, I.R., Bishop, S.A. and Couso, J.P., 2016, Hemotin, a Regulator of Phagocytosis Encoded by a Small ORF and Conserved across Metazoans, *PLoS biology*, 14(3), p. e1002395.

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), p. e1002603.

Qin, X., Ahn, S., Speed, T.P. and Rubin, G.M., 2007, Global analyses of mRNA translational control during early *Drosophila* embryogenesis, *Genome Biol*, 8(4), p. R63.

Quinlan, A.R. and Hall, I.M., 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics (Oxford, England)*, 26(6), pp. 841-2.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Raabe, C.A., Tang, T.H., Brosius, J. and Rozhdestvensky, T.S., 2014, Biases in small RNA deep sequencing data, *Nucleic acids research*, 42(3), pp. 1414-26.

Reboud, A.M., Dubost, S. and Reboud, J.P., 1984, Sucrose modifies ribosomal stability and conformation, *Biochimie*, 66(3), pp. 251-5.

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J. and Jones, C.D., 2013, De Novo ORFs in *Drosophila* Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences, *PLoS genetics*, 9(10), p. e1003860.

Rinn, J.,.L., kertesz, M., Wang, J.K., Squazzo, S.,.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E. and Chang, H.Y., 2007, Functional Demarcation of Active and Silent Chromatin Domains in Human *HOX* Loci by Noncoding RNAs, *Cell*, 129, pp. 1311-23.

Rinn, J.L. and Chang, H.Y., 2012, Genome regulation by long noncoding RNAs, *Annual review of biochemistry*, 81, pp. 145-66.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M. and Harvey, D.A., 2000, A *Drosophila* complementary DNA resource, *Science*, 287(5461), pp. 2222-4.

Ruiz-Orera, J., Messeguer, X., Subirana, J.A. and Alba, M.M., 2014, Long non-coding RNAs as a source of new peptides, *eLife*, 3.

Saghatelian, A. and Couso, J.P., 2015, Discovery and characterization of smORF-encoded bioactive polypeptides, *Nature chemical biology*, 11(12), pp. 909-16.

Sanger, F., Nicklen, S. and Coulson, A.R., 1977, DNA sequencing with chain-terminating inhibitors, *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463-7.

Schneider, I., 1972, Cell lines derived from late embryonic stages of *Drosophila* melanogaster, *J Embryol Exp Morphol*, 27(2), pp. 353-65.

Schrader, M., Schulz-Knappe, P. and Fricker, L.D., 2014, Historical perspective of peptidomics, *EuPA Open Proteomics*, 3, pp. 171-82.

Shirai, A., Matsuyama, A., Yashiroda, Y., Hashimoto, A., Kawamura, Y., Arai, R., Komatsu, Y., Horinouchi, S. and Yoshida, M., 2008, Global analysis of gel mobility of proteins and its use in target identification, *The Journal of biological chemistry*, 283(16), pp. 10745-52.

Slavoff, S.A., Heo, J., Budnik, B.A., Hanakahi, L.A. and Saghatelian, A., 2014, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining, *The Journal of biological chemistry*, 289(16), pp. 10950-7.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L. and Saghatelian, A., 2013, Peptidomic discovery of short open reading frame-encoded peptides in human cells, *Nature chemical biology*, 9(1), pp. 59-64.

Smith, J.E., Alvarez-Dominguez, J.R., Kline, N., Huynh, N.J., Geisler, S., Hu, W., Coller, J. and Baker, K.E., 2014, Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae, *Cell reports*, 7(6), pp. 1858-66.

Sobala, A. and Hutvagner, G., 2013, Small RNAs derived from the 5' end of tRNA can inhibit protein translation in human cells, *RNA biology*, 10(4), pp. 553-63.

Steitz, J.A., 1969, Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA, *Nature*, 224(5223), pp. 957-64.

Svensson, M., Sköld, K., Svenningsson, P. and Andren, P.E., 2003, Peptidomics-based discovery of novel neuropeptides, *Journal of proteome research*, 2(2), pp. 213-9.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S.P. and Bafna, V., 2007, Improving gene annotation using peptide mass spectrometry, *Genome research*, 17(2), pp. 231-9.

Tautz, D., 2008, Polycistronic peptide coding genes in eukaryotes - how widespread are they? *Briefings in functional Genomics and Proteomics*, submitted.

Tinoco, A.D. and Saghatelian, A., 2011, Investigating endogenous peptides and peptidases using peptidomics, *Biochemistry*, 50(35), pp. 7447-61.

Tinoco, A.D., Tagore, D.M. and Saghatelian, A., 2010, Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform, *J Am Chem Soc*, 132(11), pp. 3819-30.

Towbin, H., Staehelin, T. and Gordon, J., 1992, Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. 1979, *Biotechnology (Reading, Mass.)*, 24, pp. 145-9.

Trapnell, C., Pachter, L. and Salzberg, S.L., 2009, *TopHat*: discovering splice junctions with RNA-Seq, *Bioinformatics (Oxford, England)*, 25(9), pp. 1105-11.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L., 2012, Differential gene and transcript expression analysis of RNA-seq experiments with *TopHat* and *Cufflinks*, *Nature protocols*, 7(3), pp. 562-78.

Application Note, Sartorius AG, Treatment of Vivaspin concentrators for improved recovery of low-concentrated protein samples, doi: SLL4007-e06111

Illumina Technical Note, Estimating Sequencing Coverage, Accessed on 20th Feb 2016 at: http://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf

Ulitsky, I. and Bartel, D.P., 2013, lincRNAs: genomics, evolution, and mechanisms, *Cell*, 154(1), pp. 26-46.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H. and Bartel, D.P., 2011, Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution, *Cell*, 147(7), pp. 1537-50.

van Bakel, H., Nislow, C., Blencowe, B.J. and Hughes, T.R., 2010, Most "dark matter" transcripts are associated with known genes, *PLoS biology*, 8(5), p. e1000371.

Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.M. and Roucou, X., 2013, Direct detection of alternative open reading frames translation products in human significantly expands the proteome, *PLoS One*, 8(8), p. e70698.

van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., MacInnes, A.W., Cuppen, E. and Simonis, M., 2014, Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes, *Genome biology*, 15(1), p. R6.

Wagner, G.P., Kin, K. and Lynch, V.J., 2012, Measurement of mRNA abundance using RNA-seq data: *RPKM* measure is inconsistent among samples, *Theory in biosciences = Theorie in den Biowissenschaften*, 131(4), pp. 281-5.

Walsh, C.T., 2006, *Posttranslational Modification of Proteins: Expanding Natures Inventory,* Roberts and Company Publishers, Englewood, CO USA.

Wang, Z., Gerstein, M. and Snyder, M., 2009, RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev Genet*, 10(1), pp. 57-63.

Wapinski, O. and Chang, H.Y., 2011, Long noncoding RNAs and human disease, *Trends Cell Biol*, 21(6), pp. 354-61.

Washburn, M.P., Wolters, D. and Yates, J.R., 2001, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nature biotechnology*, 19(3), pp. 242-7.

Washietl, S., Kellis, M. and Garber, M., 2014, Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals, *Genome research*, 24(4), pp. 616-28.

Watson, J.D. and Crick, F.H., 1953, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, *Nature*, 171(4356), pp. 737-8.

White, K., Grether, M.E., Abrams, J.M., Young, L., Farrell, K. and Steller, H., 1994, Genetic control of programmed cell death in *Drosophila*, *Science*, 264(5159), pp. 677-83.

Wilkins, M.R., Appel, R.D., Van Eyk, J.E., Chung, M.C., Görg, A., Hecker, M., Huber, L.A., Langen, H., Link, A.J., Paik, Y.K., Patterson, S.D., Pennington, S.R., Rabilloud, T., Simpson, R.J., Weiss, W. and Dunn, M.J., 2006, Guidelines for the next 10 years of proteomics, *Proteomics*, 6(1), pp. 4-8.

Williams, C.C., Jan, C.H. and Weissman, J.S., 2014, Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling, *Science (New York, N.Y.)*, 346(6210), pp. 748-51.

Wilson, B.A. and Masel, J., 2011, Putatively noncoding transcripts show extensive association with ribosomes, *Genome biology and evolution*, 3, pp. 1245-52.

Wilusz, J.E., 2015, Controlling translation via modulation of tRNA levels, *Wiley Interdiscip Rev RNA*, 6(4), pp. 453-70.

Wilusz, J.E., Sunwoo, H. and Spector, D.L., 2009, Long noncoding RNAs: functional surprises from the RNA world, *Genes Dev*, 23(13), pp. 1494-504.

Wolin, S.L. and Walter, P., 1988, Ribosome pausing and stacking during translation of a eukaryotic mRNA, *The EMBO journal*, 7(11), pp. 3559-69.

Yang, X., Tschaplinski, T.J., Hurst, G.B., Jawdy, S., Abraham, P.E., Lankford, P.K., Adams, R.M., Shah, M.B., Hettich, R.L., Lindquist, E., Kalluri, U.C., Gunter, L.E., Pennacchio, C. and Tuskan, G.A., 2011, Discovery and annotation of small proteins using genomics, proteomics, and computational approaches, *Genome Res*, 21(4), pp. 634-41.

Yates, J.R., Ruse, C.I. and Nakorchevsky, A., 2009, Proteomics by mass spectrometry: approaches, advances, and applications, *Annual review of biomedical engineering*, 11, pp. 49-79.

Yoon, J.H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G. and Gorospe, M., 2012, LincRNA-p21 suppresses target mRNA translation, *Molecular cell*, 47(4), pp. 648-55.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L. and Ponting, C.P., 2012, Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila* melanogaster genome, *Genome biology and evolution*, 4(4), pp. 427-42.

Zalokar, M., 1976, Autoradiographic study of protein and RNA formation during early development of *Drosophila* eggs, *Developmental biology*, 49(2), pp. 425-37.

Zhuang, F., Fuchs, R.T., Sun, Z., Zheng, Y. and Robb, G.B., 2012, Structural bias in T4 RNA ligase-mediated 3'-adapter ligation, *Nucleic acids research*, 40(7), p. e54.

Appendix I - Primers

Primers for rRNA depletion beads

2S Rev: biotin-TACAACCCTCAACCATATGTAGTCCAAGCA

5S Fw: GCCAACGACCATACCACGCT

5S Rev: biotin-AAAAAGTTGTGGACGAGGCC

5.8S Fw: AACTCTAAGCGGTGGATCAC

5.8S Rev: biotin-CAGCATGGACTGCGATATGCG

Set 1: Used to generate 1Kb fragments

18S Fw A: ATTCTGGTTGATCCTGCCAG

18S Rev A: biotin-CAAGAATTTCACCTCTCGCGT

18S Fw B: GACCGTCGTAAGACTAACTT

18S Rev B: biotin-TAATGATCCTTCCGCAGGTTC

28S Fw A: TTATATACAACCTCAACTCAT

28S Rev A: biotin-AAGTATAGTTCACCATCTTTC

28S Fw B: GATCAGGTTGAAGTCAGGGG

28S Rev B: biotin-CATGCTCTTCTAGCCCATCTA

28S Fw C: ACATATACTGTTGTGTCGATA

28S Rev C: biotin-AAATACATAAATGCATCGTTT

28S Fw D: TTGATTTGAAAATTTGGTATA

28S Rev D: biotin-TCGAATCATCAAGCAAAGGAT

Set 2 : Used in combination with Set 1 to generate 0.5Kb fragments

18S Fw A: CCGAGGCCCTGTAATTGGAAT

18S Rev A: biotin-ATATGAGTCCTGTATTGTTATTTT

18S Fw B: ATTGTGTTTGAATGTGTTTATGTAAG

18S Rev B: biotin-AAGCATTTTACTGCCAACATGAAT

28S Fw A: ATATAAGGACATTGTAATCTATTAGC

28S Rev A: biotin-GGAAAAAATGCACACTATTCTCAT

28S Fw B: GCGCTTAAGTTGTATACCTATAC

28S Rev B: biotin-CATCCATTTTAAGGGCTAGTTG

28S Fw C: GCGGGTGTTGACACAATGTGA

28S Rev C: biotin- TAGGGCCATCACAATGCTTTGT

28S Fw D: CAAAACGTTGTTGCGACAGCA

28S Rev D: biotin-TCATTAGTAGGGTAAAACTAACC

rRNA and tRNA Depeletion Oligos

28S

rRNA depletion 1:biotin-GGGTAGTCCCATATGAGTTGAGGTTG

rRNA depletion 2: biotin-ATTGTGGAACTTTCTTGCTAAAATTTTTAAGA

rRNA depletion 3: biotin-TATAAACTTTAAATGGTTTAGAAGCCATACAATGC

18S

rRNA depletion 4: biotin-CGCTTGGTTTTAGCCTAATAAAAGCACAC

rRNA depletion 5: biotin-ATACGATCTGCATGTTATCTAGAGTTCAACCAATA

rRNA depletion 6: biotin-GGGACAAACCAACAGGTACGGCTCCACTTAC

tRNA

tRNA-Glu: biotin-CCGGATATCCTAACCACTAGACAATATGGGA

tRNA-Asp: biotin-AGGCGGGGATACTAACCACTATACTATCGAGGA

Overlapping PCR Primers for dual tagged dicistronic smORF constructs

CG42497 HA Rev:
AGTCCGGGACGTCATAGGGATAGCCCGCATAGTCAGGAACATCGTATGGGTACGCGGTGTAATCGAATCGCCTGTC

CG42497 HA Fw:
CTATGACGTCCCGGACTATGCAGGATCCTATCCATATGACGTTCCAGATTACGCTGCTTAAGCGGCAGCATCCA

CG32736 HA Fw: CTATGACGTCCCGGACTATGCAGGATCCTATCCATATGACGTTCCAGATTACGCTGCTTAAGCAAAATGCCCGCC

CG32736 HA Rev: AGTCCGGGACGTCATAGGGATAGCCCGCATAGTCAGGAACATCGTATGGGTACGCGGTGTTGGTGTTATTCGCGG

CG43194 HA Rev: AGTCCGGGACGTCATAGGGATAGCCCGCATAGTCAGGAACATCGTATGGGTAATAGATATCAACGCTATTCC

CG43194 HA Fw: CTATGACGTCCCGGACTATGCAGGATCCTATCCATATGACGTTCCAGATTACGCTGCTTAAAATCCTAGCTTTGTAGAT

CG42371 HA Rev: AGTCCGGGACGTCATAGGGATAGCCCGCATAGTCAGGAACATCGTATGGGTAGTCGCTTTTCTTCAAAGC

CG42371 HA Fw: CTATGACGTCCCGGACTATGCAGGATCCTATCCATATGACGTTCCAGATTACGCTGCTTAGAGGAGACCGCTTCCAA

Primers to generate PCR templates for IVT

CG32230 T7 Fw:

TAATACGACTCACTATAGGGTAATATATAGTTCCATTCTGTTTTATTGGA

CG32230 Rev: CAAATCCACTATGTTTATTTATAATTTGAA

CG44242 T7 Fw:

TAATACGACTCACTATAGGGGCGCTTGTATGAGTTTACAGTCA

CG44242 Rev: ATTTTCATCTTTTTAATTCGAACTTG

CecB-RA T7 Fw: TAATACGACTCACTATAGGGCATCAGTCGCACAGTTCTCA

CecB-RA Rev: GTTGTATATAGTGTCTTAATTTGTTTTTATT

pAct Fw: GAGCATTGCGGCTGATAAGG

SV40 Rev: AACGGGATCCAGACATGATAAGATAC