Studying sequence effects of mRNA 5' cap juxtapositions on translation initiation rate using randomization strategy of the extreme 5' end of mRNA.

by Anjali Pai

B.Eng., M.Sc.

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Biochemistry

National University of Ireland, Cork. School of Biochemistry and Cell Biology December 2018

Head of School: Professor Rosemary O'Connor

Supervisor: Professor Pavel Baranov

Research supported by Health Research Board and Science Foundation Ireland

This thesis is dedicated to my wonderful grandparents, ajja and mamama whom I dearly miss.

Table of Contents

Declaration

I hereby declare that all work presented in this thesis is original and entirely my own unless otherwise stated. This thesis has not been submitted in whole or in part for a higher degree to this or any other university. Any assistance and contribution by others to this work are acknowledged within the text.

Signed: Date:

Anjali Pai

Statement of Contribution

In this thesis, the pGL3 vector was a kind gift from Dr. Dimitry Andreev from Lomonosov Moscow State University. Data analysis of 18 libraries produced in this work was carried out by Dr. Patrick O'Connor, UCC who generated Figures (3.14, 3.16-3.20), Table 3.3, Appendix Table 3 and Appendix Figures (3,5-7). The CAGEr dataset for HEK293 cells and RiboGalaxy datasets were processed by Stephen Kiniry, UCC used for this work.

All other work was performed by me.

This work was funded by the Health Research Board, Ireland and Science Foundation Ireland.

Acknowledgements

I want to firstly thank Dr. Kellie Dean for having accepted me in the PhD Scholars program and giving me an opportunity to pursue my PhD. My supervisor Dr Pasha Baranov has been a unique mentor in giving me the independence to complete my research the way I believe in it while still being there when his guidance mattered. I want to thank him for believing in me to complete this project.

I want to thank Dr. Dimitri Andreev for important inputs in my experiments and sharing his valuable expertise with me during my thesis. I want to specially acknowledge Dr. Anmol Kiran who helped me with learning python. I thank Patrick O'Connor and Stephen Kiniry for the bioinformatic data analysis and discussions. I want to thank Anirudh Jayasimha from the Dept. of Pharmacology for his guidance on the qRT thermal cycler. I would like to thank Dr. Audrey Michelle for giving important language inputs for my thesis.

I have completed this PhD only with the support of my dear family and friends. Firstly, I would thank my 'Fabulous XX chromosomes' gang for moral support and positivity during the most challenging part of this journey. You people made all the time in the WGB totally worth it. I want to thank all the members past and present from the recoding lab and LAPTI lab for the great cake episodes and parties we organised. I have been very lucky to get prompt replies from NEB, Epicentre, Promega and BGI for all the technical help they have given me and the amazing scientist crew in research gate who have answered and helped me solve most of my technical problems.

I have earned some lifelines in this journey and I wish there is no termination to these bonds, my closest friends who are like an extended family. I would like to bow down to my parents and thank them for shaping me into the person I have become today. I want to thank my family for giving me immense strength during my lows and finding umpteen quirky ways to cheer me up.

Abstract

Translation initiation is a complex process. The efficiency of translation initiation is determined not just by activity and availability of the translation initiation apparatus, but also the properties of mRNA 5'transcript leaders (5'TL). In most cases of cap-dependent translation, translation initiation begins with the formation of the preinitiation complex (PIC) loading and accommodation onto the m7G capped 5'end of mRNA, facilitated by m7G cap – eIF4F interactions. The PIC accommodation onto the 5'end of mRNA is a point of control in translation initiation where the role of 5' cap proximal mRNA sequence determinants are poorly understood.

To explore the effect of the nucleotides in the extreme 5' end of mRNA on translation initiation, a library of mRNA molecules was synthesized containing all possible permutations of the first 10 nucleotides, referred to as E5S (Early 5' Sequence). The library was transfected into HEK293T cells. The lysates obtained from transfected cells were separated on a sucrose density gradient to isolate mRNAs bound to polysomes. Based on the assumption that efficiently translated mRNAs are associated with polysomes, the effect of E5S on translation initiation was measured by comparing frequencies of nucleotides (and their combinations) at specific positions in E5S from mRNAs in polysome fractions to their frequencies in E5S of the original library using massively parallel sequencing. The second position of E5S was found to have a markedly higher influence on translation initiation than positions further downstream (for technical reasons it was not possible to estimate the influence of the first position of E5S). In this position G was the most enriched nucleotide, and U was the most depleted nucleotide. Analysis of available ribosome profiling datasets did not reveal a significant association between E5S and ribosome footprint densities at the coding regions. While this work clearly suggests the influence of nucleotide context on translation initiation, it is possible that such as uORFs and RNA secondary structures, have a higher influence on translation initiation than E5S. The E5S is a previously unappreciated determinant of translation initiation, and this work suggests that differences in mRNA 5' end accessibility defined by the cap proximal sequence may be an important determinant in modulating the rate of translation initiation.

List of Tables

List of Figures

List of Abbreviations

4EBP	translation initiation factor 4E-binding protein
ATF4	activating transcription factor 4
CAGE	cap analysis of gene expression
CBC	Cap-binding complex
CDS	coding sequence
CT	CBC dependant translation
E5S	Early 5’ Sequence
eIF	Eukaryotic initiation factor
ET	eIF4E-dependent translation
GEF	guanine exchange factor
HHR	Hammerhead ribozyme
IC	initiation complex
IVT	*In vitro* transcription
LAP	liver activating protein
LARP6	La ribonucleoprotein domain family member 6
LFQ	label-free quantification
LIP	liver inhibitory protein
lnO	Long oligo
LRS	leaky ribosomal scanning
MFC	multifactor complex
miRNA	microRNA
mORF	main open reading frame
mTORC1	mammalian target of rapamycin complex 1
NGS	Next-generation sequencing
NET-seq	native elongating transcript sequencing
PABP	Poly (A) binding protein
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate Buffer Saline
PIC	pre-initiation complex
PLB	Polysome lysis buffer
PR	Polysomal RNA
RACE	robust analysis of 5’ transcript ends
RBP	Ribosome binding protein
Ribo-seq	ribosome profiling

RNAPII	RNA polymerase II
RT-PCR	reverse transcription- polymerase chain reaction
sL	stem-loop
sO	Short oligo
sOBG	Short oligo Taq extended into double-stranded DNA
TC	Ternary complex
TE	Translation efficiency
TIRES	Translation Initiation Rate Enrichment Statistic
TISU	Translation Initiator of Short 5′ UTR
TL	Transcript leader
TOP	terminal oligopyrimidine tract
TR	Total RNA
TRAP	Translating ribosome affinity purification
TSS	Transcription start site
uAUG	Upstream AUG
UMI	Unique molecular identifier
uORF	Upstream open reading frame
UTR	Untranslated region
UTRdb	UTR database

1. Introduction

1.1 The role of translation in mammalian gene expression

The task of producing a protein molecule from its gene is a highly complex process. The regulation of protein production involves multiple ways which all act in a controlled, but stochastic and highly dynamic manner called 'gene expression regulation'. The regulation of gene expression involves the synthesis of mRNA and protein via transcription and translation [1] that are coordinated by various participating factors and pathways.

While a nucleotide sequence in the DNA determines the sequence of its mRNA during transcription, a mRNA sequence determines the amino acid sequence of the resulting peptide during translation. However, there is no trivial relationship between the transcript concentration vs that of its protein concentration at a particular genomic locus. Studies that have quantified transcripts and proteins revealed that the importance of establishing the expression level of a protein includes multiple processes. Some of them are mentioned below:

a) **Translation initiation rates** which are influenced by the sequence of the mRNAs containing upstream open reading frames (uORFs), alternative transcription start sites (TSS), and/ or upstream AUGs (uAUGs).

b) **Modulation of translation rates** can occur by protein binding elements to the regulatory elements on the transcript, e.g., microRNAs (miRNA), Ribosome binding proteins (RBPs) etc. or through relative availability of the transcript and/or (tRNA charged) ribosome.

c) **Modulating the half-life of a protein** that includes the complex ubiquitin-proteasome pathway or autophagy that can influence the concentration of the protein independent of its transcript concentration.

d) **Temporal delay in protein synthesis** based on changes in the transcript concentrations steered by mRNA export and the translation process.

e) **Transport of proteins** using mechanisms to export proteins includes the spatial disconnection of proteins from the transcripts that they were synthesised from.

Therefore, the direct comparison between protein and mRNA abundances from the same location or the same cell type may not be ideal. With the advent of high throughput sequencing technology, the genetic expression of a cell is defined by both its transcriptome and its translatome [2–5]. Our understanding of the relationship between mRNA and protein levels depends on significant recent advances to quantify transcripts and proteins to produce qualitative data using cutting edge technologies as shown in Figure 1.1.

In the mRNA transcript, the 5' transcript leader (TL) region of the mRNA carries various elements that can regulate the translational readout both quantitatively (amount of protein expressed) and qualitatively (sequence of proteins that are expressed). These regulatory elements include RNA secondary structures, protein binding sites, and uAUGs which dictates the translation of uORFs and produces proteins with N-terminal extensions [6–8]. In the coming sections, we will discuss cap-dependent translation initiation in mammals and narrow our focus on the effects of the 5'-cap proximal nucleotides of the mRNA on translation in mammals.

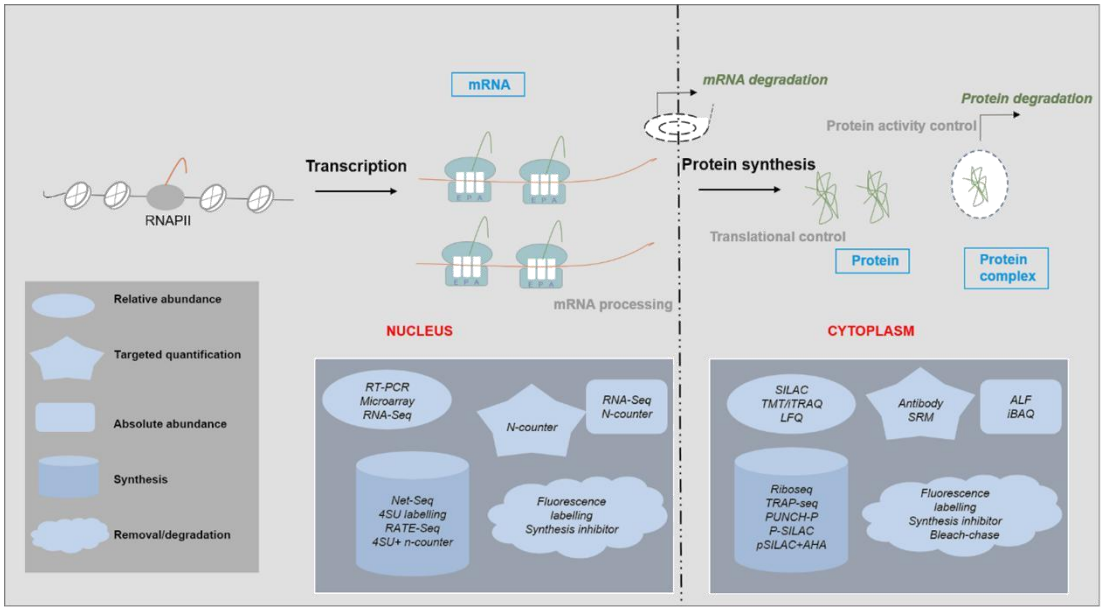


Figure 1.1: Mechanisms controlling gene expression thus illustrating the relationship between mRNA and proteins. Various Mechanisms, types of molecules involved, methods for their respective quantitative measurement and the properties measured by the respective methods are indicated.

Abbreviations: NET-seq, native elongating transcript sequencing; RATE-seq, RNA approach to equilibrium sequencing; Ribo-seq, ribosome profiling; SILAC, stable isotope labelling by amino acids in cell culture; TMT, tandem mass tag; iTRAQ, isobaric tag for relative and absolute quantification; LFQ, label-free quantification; SRM, selected reaction monitoring ; ALF, absolute label-free quantification ; iBAQ, intensity-based absolute quantification; TRAP-seq, Targeted purification of polysomal mRNA; pSILAC, pulsed stable isotope labelling by amino acids in cell culture; PUNCH-P, puromycin-associated nascent chain proteomics; AHA, azidohomoalanine labelling; 4SU labelling, 4-thiouridine labelling; RNAPII, RNA polymerase II and RT-PCR, reverse transcription- polymerase chain reaction.

This image is adapted and modified from [9].

1.2 Overview of the scanning model of translation

Upon reaching the cytoplasm, the 5' cap of the mRNA that governs the cap dependent translation in mammals proceeds in two distinct pathways: Cap Binding complex (CBC) dependent translation (CT) and eukaryotic translation initiation factor 4E (eIF4E)-dependent translation (ET). ET will be the primary focus of the coming sections. CT is believed to precede ET because CBC-bound mRNA is a precursor of eIF4E-bound mRNA[10]. While CT is largely involved in mRNA quality control, ET oversees the bulk of protein synthesis.

Cap dependant translation in mammals is a cyclic process that can be broadly divided into four stages: initiation, elongation, termination and ribosome recycling. Most regulation in translation occurs at the step of initiation. Translation initiation involves the recruitment of 43S pre-initiation complex (PIC) to the 5' end of the m7G-capped mRNA which is recognised and facilitated by eIF4F complex, through the multi-subunit eIF3. The 43S PIC scans the 5'TL for an AUG start codon based on its complementarity with the anticodon of Met-$tRNA_i$. AUG recognition triggers the hydrolysis of GTP in the ternary complex (TC) and release of eIF2-GDP from Met-$tRNA_i$ to produce a stable 48S initiation complex, followed by the joining of the large 60S ribosomal subunit, stimulated by eIF5B to form an 80S initiation complex. The elongation phase commences with decoding of the next triplet that is positioned in the ribosomal A-site [11–13]. The elongation phase incorporates amino acids into a growing polypeptide chain. The recognition of the stop codon triggers the termination of translation. Lastly, 80S initiation complex dissociates from mRNA and dissembles onto 40S and 60S subunits which are recycled to initiate the subsequent rounds of translation [11,14,15]. However, there are alternate mechanisms of translation that are cap independent and will not be discussed in this review [16–18].

1.3 eIF4E based translation initiation (ET)

Translation is a cyclical process, where ribosomal subunits that participate in translation initiation are derived from recycling of post-termination ribosomal complexes (post-TCs)[19–24]. Ribosomal recycling yields separate 40S and 60S ribosomal subunits [25].

Most mammalian mRNAs are translated by a scanning mechanism. The Met-tRNA$_i$ in a TC with GTP-bound eIF2 is loaded on the 40S ribosomal subunit, promoted by initiation factors eIF1,eIF1A, eIF5 and eIF3 to form 43S PIC (Figure 1.2) [8]. eIF4F comprises of m^{7}G cap-binding protein eIF4E, scaffolding subunit eIF4G, and DEAD box helicase eIF4A. In the next step, mRNA is activated by unwinding its 5'TL in an ATP-dependant manner by the eIF4F, eIF4B, and eIF4H along with the help of PABP [11,26–28]. The unwinding of long, highly structured 5'TL in some mRNAs require the presence of a DExH-box containing protein, DHX29 [11,27,29]. The mRNA with a circular 'closed-loop' configuration formed by the interaction of eIF4G-PABP, can improve translation efficiency by facilitating the utilization of recycled 40S ribosomes [30].

In the empty 40S subunit, the mRNA channel remains closed due to the interactions between its head and the body to form a latch [31]. The mRNA channel must be opened to allow initial loading of mRNA on the 40S subunit. .eIFs, 1 and 1A in unison are responsible for unlatching the 40S to cause 'open' confirmation of the 40S subunit i.e. conducive for the process of scanning [32,33]. The 43S PIC subsequently binds to the mRNA with the help of eIF4F, eIF4B, and eIF3. The 43S PIC, in an 'open conformation' with tRNA$_i$ not fully engaged in the P-site (P$_{OUT}$) (metastable state of TC), then scans base by base along the 5'TL in the 5'-3' direction of the mRNA in using complementarity with the anticodon of Met-tRNA$_i$ to identify the 'strength' of the AUG codon in the presence of eIF1, eIF1A and eIF5 [34]. GTP bound to eIF2 is hydrolysed by eIF5 in the scanning PIC, but the dissociated phosphate (P$_i$) is not released as it is blocked by eIF1 in the complex. The 43S PIC stops scanning when it encounters the first AUG codon (or near cognate codon, although with lower efficiency), if it is in a poor context, the scanning complex may pass AUG without translation initiation [35]. Recognition of the start codon causes the tRNA$_i$ to be accommodated in the P-site (P$_{IN}$) leading to a closed PIC, causing an arrest of the scanning process, switching the scanning complex to a 'closed' conformation. This

rearrangement triggers the release of eIF1, allowing eIF5 mediated eIF2-GTP hydrolysis and resultant dissociation of P_i [11,36–38]. A resulting stable 48S initiation complex is formed with an established codon-anticodon base pairing. In the next step, the 60S ribosomal subunit joins the 48S initiation complex causing a displacement of eIF2-GDP and initiation factors (eIF1, eIF3, eIF4F, eIF4B and eIF5) mediated by eIF4B. In the next step, the hydrolysis of eIF5B-GTP causes the displacement of eIF1A and eIF5B-GDP from the assembled elongation competent 80S ribosome.

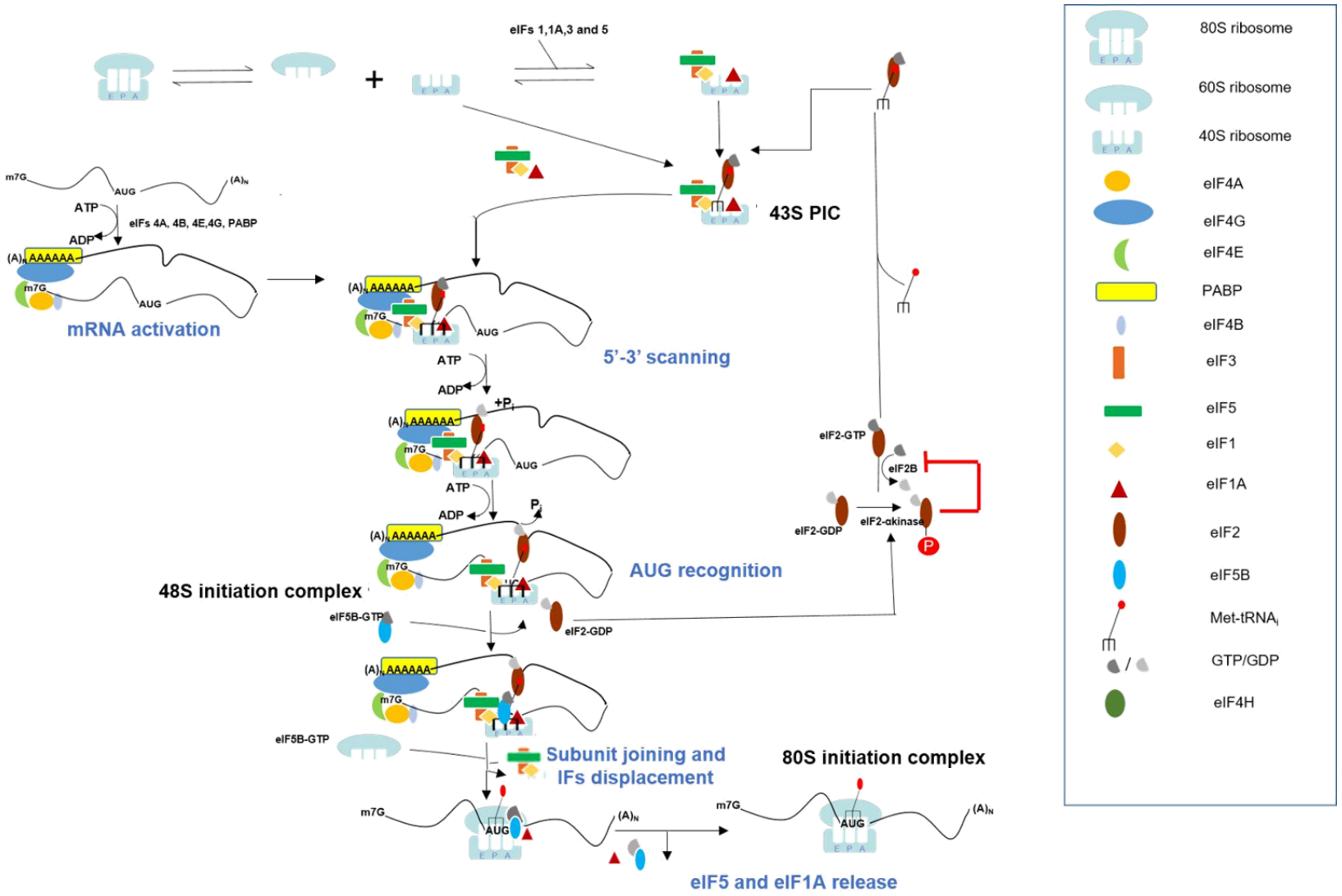

eIFs 1,1A,3 and 5
43S PIC
eIFs 4A, 4B, 4E,4G, PABP
ATP
ADP
mRNA activation
5'-3' scanning
AUG recognition
48S initiation complex
Subunit joining and
IFs displacement
80S initiation complex
eIF5 and eIF1A release
eIF2-GTP
eIF2-GDP
eIF2B
eIF2-akinase
eIF5B-GTP
80S ribosome
60S ribosome
40S ribosome
eIF4A
eIF4G
eIF4E
PABP
eIF4B
eIF3
eIF5
eIF1
eIF1A
eIF2
eIF5B
Met-tRNAi
GTP/GDP
eIF4H

Figure 1.2: **Mechanism of translation initiation.** The process of initiation is shown as a pathway of multiple reactions beginning with the dissociation of 80S ribosomes into free 40S and 60S subunits and the assembly of the 43S PIC on the small ribosomal subunit. 80S ribosomes and 40S subunits are represented with approximate locations of the aminoacyl-tRNA (A), peptidyl-tRNA (P), and exit (E) sites labelled in the 40S subunit. eIFs are labelled in the form of shapes as shown in the reference tab on the right. GTP and GDP are represented as dark/light grey structures respectively (shown in the reference tab). Translation is a cyclic process. Ribosome recycling yields separate 40S and 60S ribosomal subunits. eIF2, GTP and Met-tRNAi forms a TC called eIF2-GTP-Met-tRNA$_i$. The 43 PIC is formed that includes a 40S subunit, eIF1, eIF1A, eIF3, eIF2-GTP-Met-tRNA$_i$ and eIF5. mRNA is activated when the mRNA cap-proximal region is unwound in an ATP dependent manner by eIF4F and eIF4B. The 43S PIC attaches to the unwound mRNA and scans the 5'TL in the 5'-3' direction in search of an initiation codon. Upon initiation codon recognition, the scanning complex switches to a 'closed confirmation' and a 48S initiation complex is formed. This leads to the eIF5 mediated hydrolysis of eIF2-bound GTP and P_i is released leading to the displacement of eIF1. In the next step, the 60S subunit joins the 48S complex followed by the displacement of eIF2-GDP and other factors (eIF1, eIF3, eIF4B, eIF4F and eIF5) mediated by eIF5B. The hydrolysis of eIF5-GTP results in the displacement of eIF1A and eIF5-GDP from the elongation competent 80S ribosome. Following elongation, termination occurs (not shown in the figure) followed by recycling which generates separated ribosomal subunits and the process begins again. This figure is adapted and modified from [11,12].

Translation initiation is the most regulated step in mRNA translation. Several sequence elements present in the 5'TL of a mRNA molecule that influence translation initiation have been characterized that includes upstream ORFs initiated with AUG and near cognate start codons, specific secondary structures, as wells as specific sequence motifs e.g. Terminal oligo-pyrimidine tract (TOP). The elements of the 5'TL present in the cap proximal end of the mRNA that can influence the process of translation initiation will be discussed in the following section. However, there are no studies that exhaustively explore the effect of the context of nucleotides present in the cap proximal end of the mRNA on the rate of translation initiation.

1.4 Elements in 5'TL that modulate translation initiation

Secondary structures in the cap proximal end of mRNA

Scanning of the 40S ribosome along the 5'TL of the mRNA through structural barriers is an important regulatory step in translation initiation as discussed in the previous sections. Secondary structures located in the 5'TL of the mammalian transcripts can significantly alter the translation efficiency [39,40], initially reported by Kozak's experiments [41–43] . The presence of secondary structures in the proximity to the 5'end of mRNA decreases the efficiency of translation in α and β-globins, whereas having a minor effect on translation when placed downstream in the leader sequence – this was first demonstrated by Kozak [41]. The inhibition of translation in the presence of secondary structure was studied in the bovine growth hormone receptor where the translation was modulated up to 80-fold based on differences in the 5'TL splice variant. Insertion of various hairpins into 5'TL to study its effects of translation were first performed by Kozak. These initial studies in Cos 7 cells found that hairpins with the predicted thermal stability of −30 kcal/mol had no effect on translation, while hairpins of −50 kcal/mol reduced translation by 85%–95% [44]. Similar to *in vitro* studies [45], it was shown in live cells that mRNA structures are inhibitory when placed proximal to the 5′-mRNA cap between positions +1 - +9 [46].

RNA helicases are known to play a general role in translation initiation and have a role in unwinding RNA hairpin structures in an ATPase dependant fashion. Stable RNA secondary structures can resist the unwinding activity of the helicase elF4A, overcome partially by the overexpression of elF4A in partnership with elF4B [47]. Apart from eIF4A, other RNA helicases are also involved in translational control including DHX29, DHX9 (also referred to as RNA helicase A or RHA) and DDX3.

mRNAs containing moderate to strong 5'TL secondary structure (ΔG <-19 kcal/mol) use DHX29, a DEAH box protein to promote translation initiation. Rather than unwinding secondary structures in the RNA (due to poor helicase activity), DHX29 acts by altering the 40S conformation between the 'open' and 'close' state of mRNA entrance site by shuttling between its NTP and NDP bound states [29]. DHX29 is also known to associate with eIF1A to play a vital role in leaky scanning and start codon recognition [27].

DHX9 can impact translation initiation of specific mRNAs. DHX9 can bind to the 5'TL structural motif in *c-JUND* mRNA [48] and unwind the structural elements within its 5'TL to promote translation initiation. La ribonucleoprotein domain family member 6 (LARP6) binds to the 5' stem loop (sL) of type 1 collagen mRNA with high affinity. DHX9 forms a complex with LARP6 to promote translation of type 1 collagen mRNA [49]. The mechanism by which DHX9 probably unwinds the secondary structure of 5'sL, releasing LARP6 to promote translation remains unclear [50].

Translation of specific mRNAs containing an sL, the TAR RNA motif in their 5'cap proximal end, for example, HIV-1 gRNA, are enhanced in the presence of DDX3. sL can impede eIF4F binding and subsequent 43S PIC loading. DDX3 aids in unwinding the secondary structure of specific mRNAs populations with the help of eIF4G and eIF4F, facilitating the entry of the 43S to promote translation initiation [51].

Iron metabolism is regulated by the binding of RBPs namely iron regulatory protein (IRP) to hairpin structures called Iron Regulatory Elements (IRE) present in the 5'TL and 3'UTR of ferritin and transferrin transcripts respectively. An IRE hairpin is ~30 nts long and forms a 5′- CAGUGN- 3′ loop and a stem with moderate stability

(Δ*G*∼−7 kcal/mol), interrupted by an unpaired C residue [52,53]. mRNAs coding the H and L-ferritin, the iron storage protein contain a single IRE in their 5'TL. At low iron concentration, IRP binds to the 5′TL of the ferritin mRNA to prevent ferritin translation; at higher iron levels, the IRP is saturated with iron and falls off the ferritin mRNA. The release of the IRP from the ferritin mRNA leads to efficient translation of the mRNA [54,55]. Transferrin mRNA is responsible for iron uptake in cells. In the presence of excess iron, IRP binds to IREs in the 3'UTR of transferrin mRNA causing iron-dependant degradation [55].

Another form of structural impediment in the 5'TL for translation initiation can occur as G-quadruplexes. G-quadruplex is a non-canonical four stranded nucleic acid structure formed by guanine rich nucleotide sequences [56]. For example, the presence of secondary structures in the form of G-quadruplexes in the cap proximal ends of *NRAS* 5'TL can repress translation by acting as a roadblock to inhibit the progression of the ribosome during translation [57].

The role of uAUGs /uORFs in modulating translation

According to the scanning model, the 43S PIC enters the mRNA at its 5'cap and scans sequentially along the 5′TL until it positions the first AUG codon [45] in its P-site. The optimal context of an AUG start codon in mammals is GCCA/GCCAUGG (termed Kozak consensus) of which A at −3 and G at +4 (the A at the AUG codon being +1) [39,58] are critical in determining the strength of the start codon context. Sometimes 'near cognate' triplets that differ from AUG by a single base can be selected by the scanning PIC at lower frequencies, due to the mismatch with the anticodon of $tRNA_i$ and probable destabilisation of the 48S PIC. Near cognates rely on optimal context more heavily than AUG with NUG triplets functioning better than A(A/G)G triplets in start codon selection [12]. Various factors including initiation factors, structural elements in the $tRNA_i$ and the rRNA along with the protein components of the small ribosomal subunit 40S are involved in discriminating between AUGs and non-AUG triplets by the scanning PICs. In eukaryotes, while eIF1 promotes scanning and blocks the recognition of non-AUGs and poor context AUGs, eIF5 antagonises the function of eIF1 [59].

If an upstream AUG (uAUG) is in-frame with a downstream AUG uninterrupted by a stop codon, leaky scanning may occur to produce two protein isomers differing by an N-terminal extension, to produce a longer form usually targeted at a particular cellular component) [12,24,60]. Some uORFs have the ability to inhibit downstream ORFs; direct evidence for this is seen for a relatively small number of genes. Inhibitory uORFs are principally governed by two primary control mechanisms:

a) One class of regulatory uORF encodes a peptide that can stall the 80S ribosome engaged in synthesis at or near the uORF stop codon. Stalling by the uORF peptide prevents the scanning of 43S PICs that leaky scanned at the uORF-AUG codon by creating a 'roadblock' modulated by ligands (Figure 1.3b), for example, spermidine for *AMD1*.

b) The second class of regulatory ORFs can inhibit the downstream ORF start codon by hindering the 43S PICs, their encoded peptide being irrelevant to their inhibitory function (Figure 1.3a). Genome data warranted that the barrier created by such uORFs can be overcome by leaky scanning. uORFs whose AUG codons comply the rules of optimal Kozak context having a higher inhibitory effect to prevent downstream ORF synthesis [61,62].

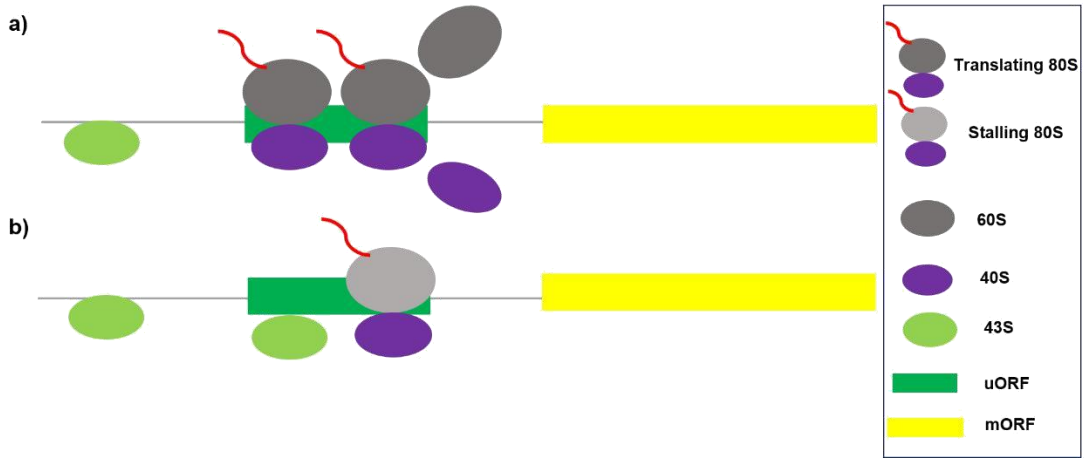


Figure 1.3 :Translational control mechanisms by uORFs a) When uORF is translated (shown as 80S ribosomes) by the scanning 43S PIC, upon termination free subunits dissociate from the mRNA to prevent the translation of the main ORF (mORF) b) The 80S ribosomes are stalled during elongation or termination by an uORF attenuator peptide generated by the leaky scanning of uORF-AUG codon blocks the scanning 43S PIC, preventing mORF translation.

Leaky scanning of an inhibitory uORF is increased during stress conditions, eIF2(αP) at serine 51 decreases the levels of eIF2-GTP and acts as a competitive inhibitor for eIF2 Guanine exchange factor eIF2B causing decreased TC assembly. Decreased TC levels can lead to a delay in reinitiation that allows ribosomes to bypass inhibitory uORFs and translate the mORF. An example of this mechanism is described using the mammalian *ATF4*, the transcriptional regulating activating transcription factor (Figure 1.4 b). *ATF4* expression involves the differential contribution of two upstream ORFs (uORFs) in the 5′ TL of the mouse *ATF4* mRNA. The 5′ proximal uORF1 is a positive-acting element that facilitates ribosome scanning and reinitiation at downstream coding regions in the *ATF4* mRNA. When eIF2-GTP is abundant in non-stressed cells, ribosomes scanning downstream of uORF1 reinitiate at the next coding region, uORF2, an inhibitory element that blocks *ATF4* expression. During stress conditions, phosphorylation of eIF2 and the accompanying reduction in the levels of eIF2-GTP increase the time needed for the scanning ribosomes to become competent to reinitiate translation. This delayed reinitiation allows for ribosomes to scan through the inhibitory uORF2 and instead reinitiate at the ATF4-coding region [63,64].

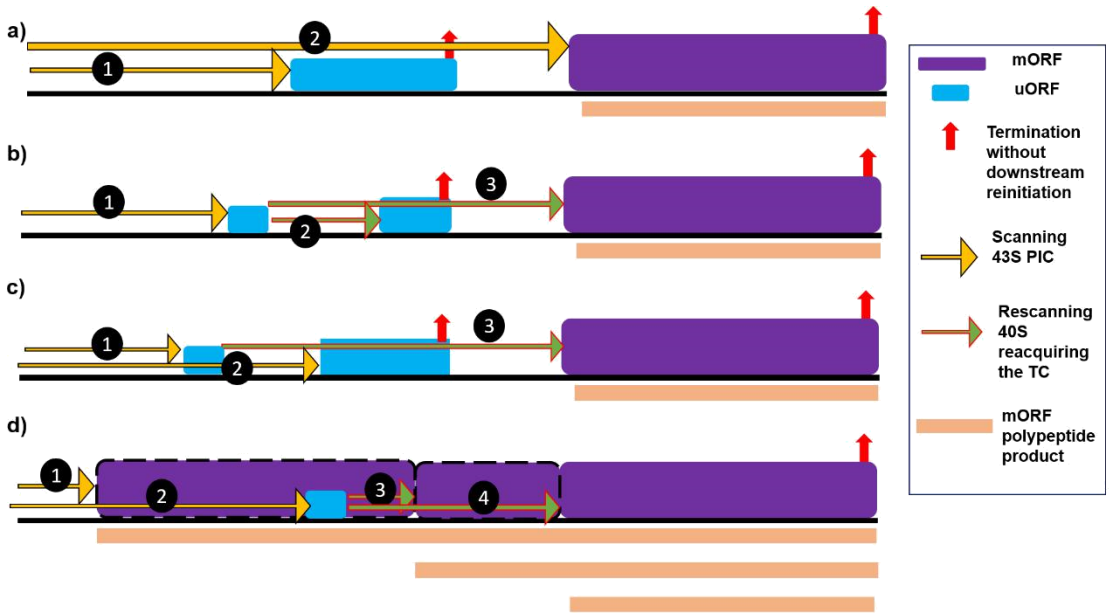


Figure 1.4: uORF translational control under different gene architectures a) 1. The scanning PICs translating uORF do not reinitiate at the mORF (Figure 1.3a) 2. Leaky scanning at the uORF with suboptimal start codon initiates at the mORF. In this case, leaky scanning can be inhibited depending upon elevated eIF5 levels [e.g.: lowering translation of eIF5 gene], by eIF2(αP) [e.g., IFRD1] and polyamines (e.g., AMD1 encoding SAM decarboxylase) b) 1. Scanning ribosomes translate a short uORF whose translation does not prevent reinitiation. 2. When scanning is resumed, TC reacquisition can lead to the translation of an inhibitory downstream uORF that can prevent further reinitiation 3. Slow acquisition of TC due to low TC concentration induced by eIF2(αP) allows for reinitiation at a downstream mORF. An example includes ATF4. c) 1. Scanning ribosomes can initiate translation at a uORF that permits reinitiation 2. Ribosomes that can leaky-scan at the first uORF translate a second inhibitory uORF that prevents reinitiation 3. Ribosomes can translate the first uORF resume scanning and upon bypassing the second inhibitory uORF (avoid its inhibitory effect), can reacquire TC and translate the mORF. d) When an upstream start codon is in frame with the mORF, the inhibitory uORF can be bypassed during elongation to produce protein isoform 'A' with specific properties 2. Scanning ribosomes can bypass a suboptimal in-frame start site and initiate at a downstream uORF 3. Rescanning and reacquisition of TC leads to reinitiation at a proximal start codon producing protein isoform 'B' 4. Slow reacquisition of TC can allow reinitiation at a farther downstream start codon to produce shortest protein isoform 'C' with activities opposing thos de C/EBP-α and C-EBP-β.

This image is adapted and modified from [24,65].

Small ORFs (smORFs)

smORFs are defined as small ORFs containing less than 100 codons that can be translated [66]. smORFs can exist within the 5'TL and encode functional proteins [67,68]. smORFs can either modulate downstream initiation events or have distinct biological functions. Detecting products from smORFs is technically challenging [69] and has been possible recently with the advent of ribosome profiling [70]. Genome wide analysis using ribosome profiling has identified the previously non-annotated smORFs with the potential to encode biologically active peptides [69]. For example, in mice, myoregulin (*MLN*) smORF expresses a 46aa peptide that plays a role in muscle contraction. Another example in human is humanin, a smORF (24 amino acids long) that has significant implications in apoptosis [69].

1.5 5' Terminal Oligopyrimidine Tract (TOP) motif

Transcripts containing a cysteine following the m7G cap and an uninterrupted stretch of 4-14 pyrimidines (TOP motif) are referred to as TOP mRNAs [71]. TOP mRNAs are known to encode components of the translational machinery including ribosomal proteins and elongation factors [72]. TOP mRNA translation is highly responsive to stress and growth conditions and is believed to be mediated via the mammalian target of rapamycin complex 1 (mTORC1) and its downstream effector eukaryotic translation initiation factor 4E-binding protein (4EBP) [73].

Translational control of 5'TOP mRNA relies on the regulation of eIF4E by 4EBP [73,74]. However, suppression of 4EBP-1/2 function is unlikely to be the only factors driving 5'TOP mRNA translation, eIF4E overexpression did not promote translation of 5'TOP mRNAs [75]. These findings suggest the presence of additional regulatory factors that can bind directly to the TOP sequence in regulating 5'TOP mRNA translation [76].

Early evidence suggested that the regulation of 5'TOP mRNA involved an unknown titratable repressor molecule [77]. Based on this observation, several candidates were proposed to have an association with 5'TOP elements and its immediate downstream region. Some of these candidates include the abundant La antigen or La-related protein

3 (LARP3) [78–80] , AUF1 [81], ZNF9 [80], TIA-1 [82], LARP1 [83] and LARP7 [84]. It seems likely that one, or several of these proteins can compete with eIF4F for binding to the 5'TOP and, hence, prevent 43S recruitment under conditions non-permissive for 5'TOP mRNA translation. However, there is a lack of definitive evidence for regulatory roles of these proteins in the translation of 5'TOP mRNAs.

In recent studies, ribosome-profiling suggested that mTOR almost exclusively stimulates TOP/ TOP-like mRNA translation [73,74]. In contrast, polysome profiling indicated that mTOR mediates the translation of non-TOP mRNA as well [85]. Gandin *et al* revealed that mTOR sensitivity is not based solely only on TOP motif but distinctive 5'TL features [86].The mechanisms that can control the specificity of TOP / TOP like mRNA regulation remains unclear and has been debated in recent studies [86].

1.6 RNA binding proteins and their role in translation initiation

An RBP can form ribonucleoprotein complexes (RNP) and associate with transcripts to influence their fate and function [87]. RBPs bind to specific sequence / structural motifs in the RNA via well-defined RNA binding domains (RBD) [88] such as the RNA recognition motif (RRM) [89], hnRNP K homology domain (KH) [90] or DEAD box helicase domain [91]. However, recent advances in structural biology have revealed the existence of complex protein-RNA interactions that do not require canonical RBDs [92].

RBPs can bind to the RNA to regulate mRNA stability, localisation and its translation [88,93]. RBP's can have both positive and negative effects on the translation of mRNAs depending on their interaction with specific RNA motifs. A few examples of RBPs rendering translation control are described below-

a) LARP1

LARP1 is an evolutionary conserved RBP containing a La motif, a 90 amino acid domain followed by an RRM-L5 and a highly conserved C-terminal region called the DM15 domain [94] /LARP1 motif [95]. LARP1 contains binding sites for PABP and RAPTOR (regulatory-associated protein of mTOR) and plays an important role in the

regulation of a subset of mRNAs containing the 5'TOP motif [96]. It can enhance or restrict translation based on cell type, RNA binding affinities and available protein-protein interactions. *In vitro* studies demonstrate the that LARP1 can bind to the m7G cap and the first cytidine of the TOP mRNA (higher affinity compared to eIF4E), thereby blocking the eIF4F complex on TOP mRNAs and repressing translation [97–99]. In contrast, LARP1 is observed to bind to a range of targets to activate proto-oncogenes and enhances total protein synthesis in various cancer cell lines [100–102].

b) PABP

PABP is a conserved family of eukaryotic RBP involved in various stages of post-transcriptional gene expression including pre-mRNA 3' end processing, translation initiation, termination, mRNA stability and turnover and mRNA-specific degradation mechanisms [24,103–106].

In the context of cap dependant translation, the interaction of eIF4G-PABP enhances the eIF4E-cap binding activity and eIF4A helicase activity [107–109]. During translation initiation, the interaction of PABP-eIF4G and poly(A), stabilises bound mRNAs to a 'closed-loop' formation that enhances the 43S PIC assembly and post-termination ribosome recycling [110,111].

PABP is regulated by PABP interacting proteins 1 and 2 (Paip1 and Paip2). Paip1 binds to PABP and enhances its affinity for eIF4G in the presence of eIF3 [112]. Paip2 is a competitive inhibitor of eIF4G-PABP interaction and can inhibit the interaction between PABP and poly (A) tail of the mRNA [113]. In mice, Paip2 knockout can silence transcription during spermiogenesis [114]. When Paip2 is knocked out, an increase in PABP can lead to non-productive eIF4G binding or competitive binding to 5'TL in a subset of mRNAs leading to infertility, decreased sperm count and abnormal spermatid structure in mice [28,115–119].

1.7 m7G cap and its role in translation initiation

The cap structure was first observed in several viral mRNAs before it was identified in cellular mRNA of HELA cells [120,121]. The cap structure is the first modification made to RNA polymerase II transcribed RNA. The cap structure is formed co-transcriptionally in the nucleus as soon as the first 25-30 nts are incorporated into the nascent transcript [122,123]. mRNA is capped by N7-methyl guanine (m7G) that are linked through an inverted 5′-5′ triphosphate bridge to the initiating nucleoside of a nascent transcript [124]. Three enzymatic activities namely RNA triphosphatase (TPase), RNA guanylyltransferase (GTase) and guanine N7 methyltransferase (guanine N7 MTase) are involved in the conversion of 5’ triphosphate of the nascent transcript into a cap 0 structure as shown in Figure 1.3 [125].

Additionally, the m7G-specific 2′O methyltransferase (2′O MTase) methylates the +1 and +2 ribonucleotides at the 2′O position of the ribose to generate the cap 1 and cap2 structures respectively. Although the cap 0 and cap 1 modification of a nascent mRNA occurs in the nucleus, cap 2 modification occurs in the cytoplasm. The human enzymes that methylate the 2′O position of the +1 and +2 ribose to form the cap 1 and cap 2 structures, respectively, have recently been identified [126,127]. Cap1 and cap2 methylations in U2 snRNA are required for spliceosome E complex formation and consequently for efficient pre-mRNA splicing [128].

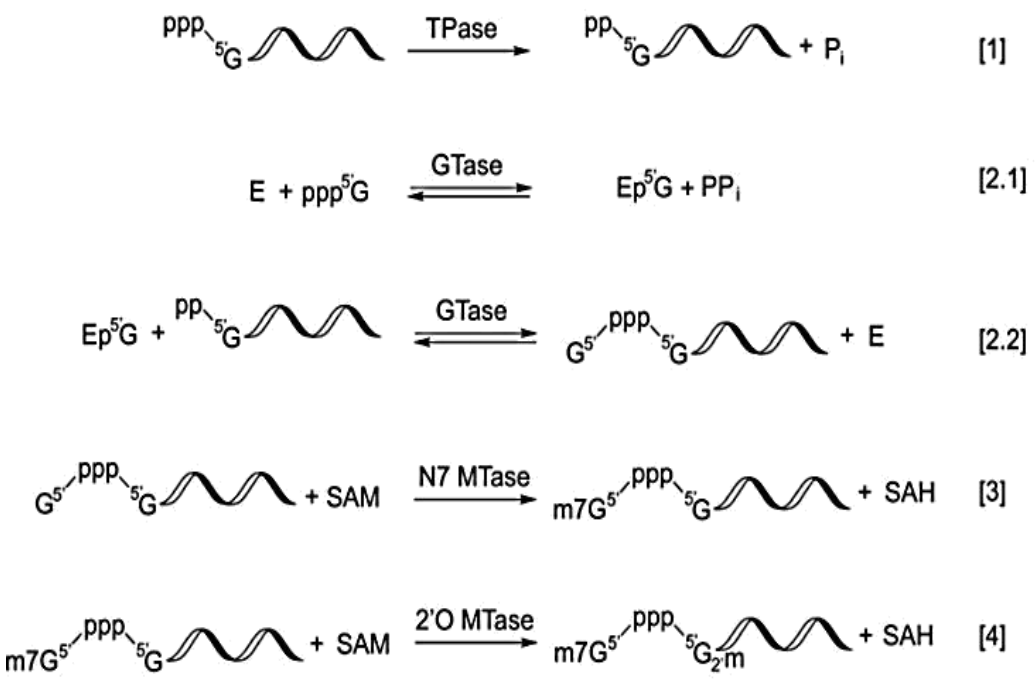


Figure 1.5: Enzymatic steps involved in RNA capping The RNA triphosphatase activity (TPase) removes the γ-phosphate from 5′ triphosphate, generating a diphosphate 5′ end and inorganic phosphate (reaction [1]). The guanylyltransferase (GTase) activity takes up a GTP molecule to form a covalent intermediate containing a lysyl-Nζ-5′-phosphoguanine (reaction [2.1]). In the presence of a 5′ diphosphate RNA, the GTase activity transfers the 5′-phosphoguanine (GMP) to the 5′ diphosphate, forming a 5′-5′ triphosphate linkage between the first base of the RNA and the capping base (reaction [2.1]). In the presence of S-adenosylmethionine (SAM), the guanine-N7 methyltransferase (MTase) activity adds a methyl group to N7 amine of the guanine cap to form the cap 0 structure (reaction [3]). Finally, the m7G cap-specific 2′O MTase modifies the 2′O of +1 ribose and generates the cap 1 structure (reaction [4]). This image is adapted from [129].

The cap structure is a critical part of the process of cap dependant translation. eIF4E is the cap binding protein that binds to m7GpppN (where N is any nucleotide). eIF4E recruit's mRNA transcripts onto the ribosome through its high affinity binding with eIF4G [11]. eIF4E is involved in two important processes: a) it binds to the m7G cap and recruits eIF4G/eIF4A to the 5' end of the mRNA transcript to form the eIF4F complex [6,13] and b) enables circularization of the mRNA [13]. Many mRNAs with highly structured TL are sensitive to eIF4E whereas housekeeping genes such as *GAPDH* and *Actin* containing short unstructured 5'TL are not eIF4E sensitive [130–133]. Overexpression of eIF4E is observed to increase the efficiency of transcripts

containing a highly structured TL in cap dependant translation initiation. eIF4G enhances the affinity of eIF4E to bind the m7G cap [134]. eIF4G interacts with eIF4E through the motif YX 4 LΦ (Y denotes tyrosine, X denotes any amino acid, L denotes Leucine and Φ denotes a hydrophobic residue), also conserved in 4E-BPs [135,136]. Competition between eIF4G and 4E-BP1 occurs due to the presence of a shared binding motif (YX 4 LΦ) available on the dorsal side of eIF4E. The interaction of 4E-BP1 to eIF4E is modulated by phosphorylation of multiple serine and threonine residues [137]. The translation levels are therefore lowered when 4E-BP1 is active and this activity is thought to be regulated by mTOR dependent phosphorylation [138].

mTOR is the mammalian target of rapamycin, a highly conserved serine/threonine kinase that plays a significant role in controlling cell growth and metabolism. The mTOR activity is regulated by growth factors and amino acid availability as well as the energy status of the cell [138]. When mTOR activity is low, 4E-BP1 is hypo-phosphorylated (i.e. phosphorylated in 2 out of its 4 phosphorylation sites) allowing efficient binding to eIF4E and blocks translation initiation. When mTOR activity is high, 4E-BP1 is phosphorylated in its 4 phosphorylation sites (S37, T46, T70, and S65) [139,140] causing it to release eIF4E, thus allowing initiation of cap dependent translation [141]. eIF4E does not use its lateral side to bind to eIF4G but comprises of non-canonical binding sites to accommodate 4E-BPs [142–146]. These non-canonical motifs can increase the affinity of 4E-BP to eIF4E up to threefold, contributing as an essential component in the 4E-BPs competition with eIF4G [145].

Despite major advances in our understanding of various regulatory elements within the mammalian 5′ TL modulating translational efficiency, we have recently started to appreciate the transcriptional heterogeneity of this process.

1.8 Alternative Transcription start site (TSS) and its role in translational regulation.

Regulation of gene expression at the transcriptional level leads to transcript diversity. Transcription begins from a TSS after the transcription initiation complex assembles on the corresponding promoter. However, many genes are known to have multiple transcript isoforms that contain alternative first exons corresponding to their alternative promoter, adding complexity at the level of transcription. In mammals, it is estimated that around 58% of the transcribed genes contained multiple promoters[147].

Variable 5'TLs can alter gene expression by producing different mRNA variants in a tissue specific manner [148–150] thereby influencing mRNA stability and translational efficiency. However, alternative first exons can differ in length and sequence but in extremely rare cases, they have similar length and nucleotide sequence, for example, gene clusters of *Pcdh* and *UGT1* [151].

In mammalian genes, most promoters are located within the CpG- rich regions and occur less frequently in TATA box regions. Whilst the TATA box enriched promoters are known to initiate in a well-defined site, CpG rich promoters are known to have a broad, plastic and evolvable initiation site for transcription [152]. A series of TSSs were observed over a very small 4-6 bp surrounding the principle TSS [153].

A true transcription site is identified with the presence of 7-methyl guanine cap structure to the 5' triphosphate of the first base of an RNA polymerase II transcribed mRNA. It is this unique feature of RNA that forms a basis for several methods aiming to enrich and identify capped messages to map the exact positions in the nucleotides to which the cap is added. The main methods extrapolating this mechanism is cap analysis of gene expression (CAGE) [154], oligo-capping [155], robust analysis of 5' transcript ends using 5' Rapid amplification of cDNA ends (5'RACE)[156] and NanoCAGE[157].

The 2'-3' diol structure of the cap nucleotide is also present in the extreme 3' end of an RNA molecule, exploited by the CAGE technology. The diol structure is oxidised chemically followed by biotinylation, selection of capped messages by immunoprecipitation with streptavidin. The enriched capped RNA is transcribed into cDNAs that span the entire lengths of the capped RNA molecules [154].

In the oligo-capping and 5'-RACE methods, the ability of the 5'cap to resist phosphatase treatment is exploited. The phosphatase treatment ensures the removal of tri, di, and mono-phosphates from the cleaved or degraded RNA. The cap is subsequently removed using the tobacco acid pyrophosphorylase leaving a 5'monophoshate that is then ligated to a linker molecule to mark the 5' extreme end of mRNA [155,156]. Full length c-DNA generated by one of the above-mentioned methods can be included with short DNA tags attached to the 5'end of the mRNA suitable for next generation sequencing [158]. The information on the exact position of cap addition sites for millions of RNA molecules can be generated by the amalgamation of cap-selection and next generation sequencing technologies [159–161], thus making digital information on the number of transcription initiation events at any genomic position easily available. The NanoCAGE method finds the TSS of mRNA molecules from low quantities of total RNA as input (~10ng) [157]. NanoCAGE combines a template switching method relying on the reverse transcription of the cap of the mRNA to enrich for 5'ends [162] as well as a semi-suppressive PCR to minimise PCR artefacts [163].

The potential of alternative TSS in altering 5'TL structure leading to enhanced or diminished levels of protein synthesis has been extensively studied. [164–166]. In a study combining polysome profiling with high throughput mRNA-5'end sequencing, the translational status of the mRNA isoform with distinct TSSs was studied. Among 9,951 genes expressed in mouse fibroblasts, 4,153 genes showed significant initiation at multiple sites, of which 745 genes exhibited significant isoform-divergent translation [167].TSS switches are of functional significance and have an association with a pathogenic phenotype such as BRCA1 in breast cancers [168,169]. In some genes such as the tumour protein *p53* and *GNAS*, alternative promoters were shown to be activated or silenced to modulate transcription levels [170]. A recent study exploring the mammalian genes for TSS switching events during cerebellar development revealed 9767 cross-over TSS switching events across 1511 genes suggesting that the dominant TSS shifts over time [171]. The alternative switching events have also been used to characterize the cellular phenotype based on its transcriptional landscape in human cell lines [172]. However, the magnitude of the effects of transcriptional switching on the ribosome associated mRNA population remains unclear.

1.9 5'cap proximal nucleotides and their possible role in translation control

During the conventional mechanism of scanning as explained earlier (Introduction, 1.2 and1.3), the 43S PIC moves from the 5'-3' direction of a mRNA in search of a start codon. The mechanism in which the mRNA positions on the 43S PIC or the factors that permit the release of the 43S PIC to permit scanning from the 5'end is not clear. Due to the advent of high resolution cross-linking studies, it is now possible to study the mRNA path on the 48S complexes [29,173] i.e. 5' distal when positioned over a start codon. RNase footprint studies have demonstrated that the 80S ribosome protects ~30 nts of the mRNA but the 48S complex binds to an additional 10-20 nts of RNA on the 5'end [174,175], in contact with other IFs particularly eIF3 [29]. This is consistent with previous observations from Kozak that efficient translation requires a minimum TL length of 20nt [39]. It is interesting to know the mechanism in which the start codon could be positioned in the P-site of the 43S PIC for transcripts containing short 5'TL <20nt.

There are various models that can be predicted to offer an explanation for this occurrence [176] as shown in figure 1.6 and listed as follows:

a) It is shown that the interaction between eIF4G-eIF4E is weakened upon eIF4E binding to the cap structure [177]. 3'-5' scanning has been suggested in recent studies [178]. One possibility is that upon 43S PIC loading, the contact between eIF4E-eIF4G breaks, allowing retrograde movement of the 43S PIC and positioning of the AUG in the P-site (Figure 1.4). Furthermore, displacement of a part of the cap binding complex, possibly eIF4G in association with a fraction of available eIF1, has been proposed to occur during initiation on Translation Initiator of Short 5′ UTR (TISU) elements [179], the details on how the process is driven remains unclear.

b) The interaction between eIF4E- 5'cap could be dampened (unknown mechanism) allowing the mRNA to slide over the surface of the 40S subunit until the AUG enters the P-site.

c) The presence of leaderless mRNA has been demonstrated in prokaryotes and *in vitro* systems in eukaryotes [180–182]. It is also known that eukaryotic cells

contain a large pool of "empty" 80S monosomes that are biologically inactive [183]. Direct loading of empty 80S particles that are devoid of eIFs can occur on the mRNA, i.e. then threaded through the ribosome until the AUG enters its P-site.

One group of cellular mRNAs possess a short 5'TL with the presence of a TISU motif located downstream juxtaposition to the TSS [184]. TISU element in mRNAs can control the initiation rates of both transcription and translation. The TISU motif comprises of the sequence SAASATGGCGGC in which S is C/G. mRNAs containing short 5'TL (< 12nt) in the presence of TISU elements was shown to facilitate translation initiation, rendered without the possibility of a scanning mechanism [185]. TISU mRNAs are insensitive to eIF1A induced leaky scanning and remain unaffected by the inhibition of eIF4A helicase action [185]. TISU translation was found to be strongly dependant on eIF1. eIF1/1B siRNA knockdown led to translation repression of TISU bearing reporter RNA [179]. Translation of TISU mRNA is maintained during stress when canonical cap dependant translation ceases [179]. However, the mechanism of TISU mRNA translation and its magnitude on regulating gene expression remains poorly understood.

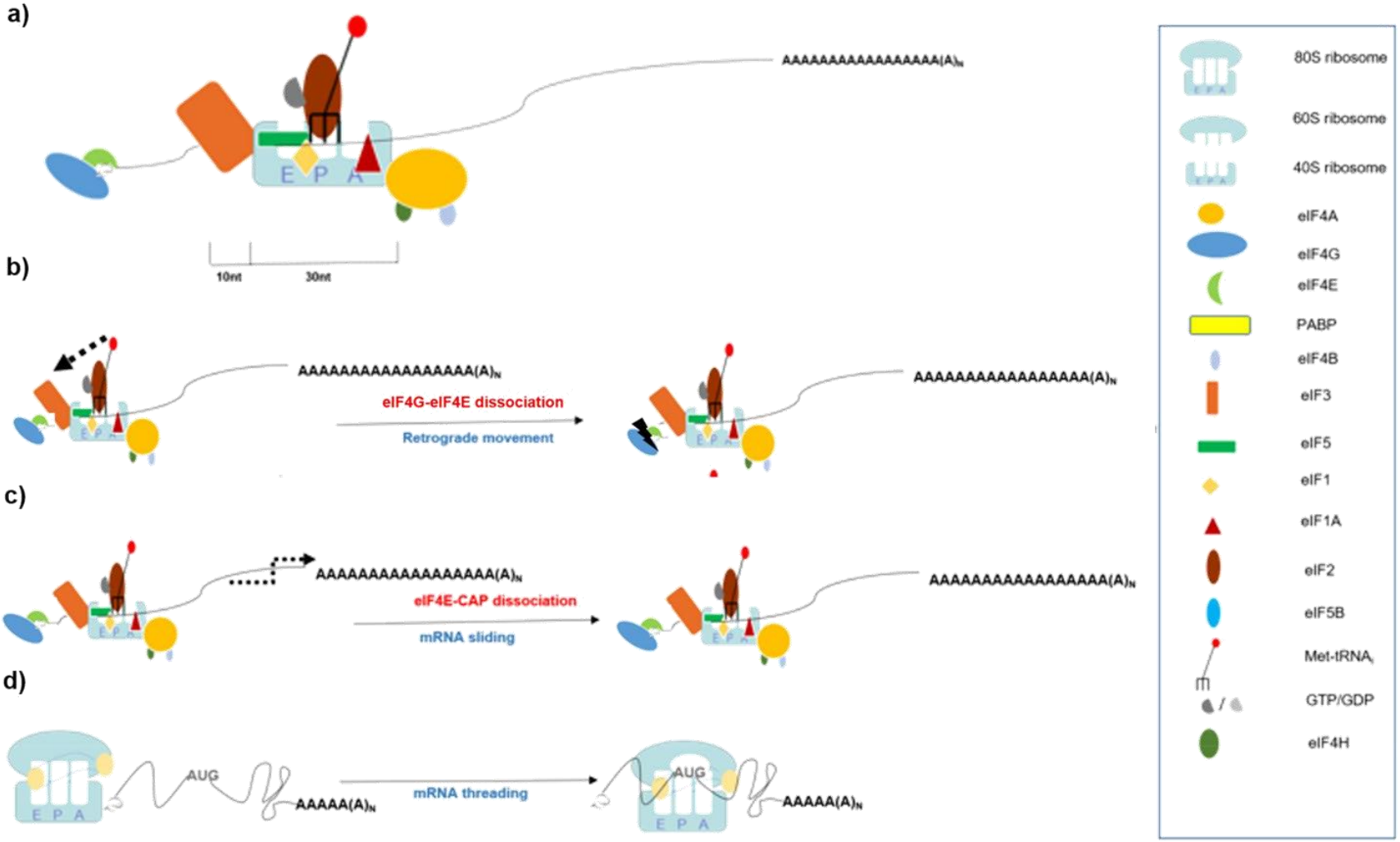

a)
AAAAAAAAAAAAAAAAA(A)N
E P A
10nt
30nt
b)
AAAAAAAAAAAAAAAAA(A)N
eIF4G-eIF4E dissociation
Retrograde movement
AAAAAAAAAAAAAAAAA(A)N
c)
AAAAAAAAAAAAAAAAA(A)N
eIF4E-CAP dissociation
mRNA sliding
AAAAAAAAAAAAAAAAA(A)N
d)
AUG
AAAAA(A)N
mRNA threading
AUG
AAAAA(A)N
80S ribosome
60S ribosome
40S ribosome
eIF4A
eIF4G
eIF4E
PABP
eIF4B
eIF3
eIF5
eIF1
eIF1A
eIF2
eIF5B
Met-tRNAi
GTP/GDP
eIF4H

Figure 1.6: Probable mechanisms in recognition of initiation codons proximal to 5'cap by the ribosome. a) During conventional scanning, the 43S PIC moves 5′–3′ on the mRNA until an initiation codon is positioned in the P-site. Various studies indicate that the 43S PIC pauses over an AUG start codon to establish contact with around 40nts of mRNA, 10 upstream nucleotides of which are in close contact with eIF3. Such a configuration is a probable explanation that AUG codons within the first 20nts of most mammalian transcripts are poorly recognized by the ribosome. However, a few models can be predicted to explain the initiation events observed on mRNA carrying TISU elements which have 5′ TLs shorter than 10nts as follows (B) Model 1: The eIF4E/4G contact is destroyed permitting retrograde movement (3′–5′) of the PIC. (C) Model 2: The eIF4E-5′ cap interaction is perturbed upon allowing the mRNA to slide over the surface of the 43S ribosome until the AUG enters the P-site. (D) Model 3: Transcripts carrying 5′ TISU elements are selectively recruited to empty 80S ribosomes and then treaded through the mRNA channel until the AUG enters the P-site. This image is adapted from [176].

Despite major advances in our understanding of various regulatory elements within the mammalian 5′ TL that modulate the translational efficiency, we are only beginning to appreciate the impact of transcriptional heterogeneity on this process.

Translation initiation is considered to be the most regulated phase of the translational cycle [8]. Translation initiation begins with the assembly of the 43S PIC and eIF4F complex loaded onto the 5' terminus of the mRNA. eIF4E performs the first critical step of eIF4F function via its interaction with the 5'-cap of the mRNA [186]. Biochemical and structural studies have elaborated the eIF4E-cap interaction [130,187–190].

It was observed that the flexibility of the C-terminal loop of eIF4E is reduced upon complex formation with m7GpppA *in-vitro* [191], indicating that eIF4E can establish contact with the first cap-proximal nucleotide. 5' cap proximal nucleotides that mediate RNA secondary structure does not inhibit the binding activity of eIF4E but influences the RNA exiting the eIF4E cap binding pocket in a translation inhibitory manner [192]. Cap proximal nucleotides can mediate translation repression via RBPs when bound to structures located within ~40 nucleotides of the cap supporting a steric mode of inhibition [193]. Cross-linking assays have demonstrated cap-dependent interactions of eIF4B, eIF4H, and eIF3a with mRNA up to 52 nts downstream from the 5′ terminus [194]. Although there have been various reports about the interactions of the various initiation factors with the 5'TL of mRNA molecules, the actual mechanism and magnitude by which cap-proximal nucleotides influence translation has not been investigated previously.

RBNS (RNA Bind-n-Seq) is a method for comprehensive, quantitative mapping of RNA binding specificity [195]. RBNS has been used to demonstrate RNA sequence preferences for a general initiation factor, which cells potentially exploit for translational control of specific mRNAs. This method has been modified for studying the binding affinities of yeast initiation factor eIF4G1 to yeast transcript leaders containing conserved oligo-uridine motifs [196]. The binding affinities of initiation factors that may have a role in identifying cap proximal nucleotides of the mRNA have not been studied extensively in human cell lines.

Selection of transcription start sites and alternate promoter usage is an important element of regulation of gene expression. However, little is known about the effects

of TSS and AP on translation. Tamarkin-Ben-Harush *et al*. showed that the ability of eIF4E to bind to capped mRNA with different +1 nucleotides that modulated with stress, with the lowest binding affinity observed for 5' cytidine [197]. However, in the control condition, there was no significant effect of initiating nucleotides on the translational response.

Kozak postulated the scanning model for translation initiation where particular sequences immediately surrounding the AUG, especially those including a purine at position −3, enhance AUG selection by the scanning PIC for an optimum context, which in mammals is 5′-(A/G)NNAUGG-3′[45,58,198]. The influence of nucleotide context of start codons has various implications on translation efficiency warranted in many recent studies [59,199,200]. Considering that nucleotide context in the start of a coding sequence can play an important role in determining the efficiency of translation, there is a possibility of sequence context in the cap proximal nucleotides of mRNA having a potential role in determining the rate of translation initiation.

2. Materials and Methods

The methods section has been divided as follows:

a) Section 2.1- 2.3 contains the methods involved in designing the experimental protocol for this work.
b) Section 2.4-2.20 includes methods in the optimization and development of the protocol used for this work.
c) Section 2.21- 2.28 contains the methods used for bioinformatic analysis of this work.

All chemicals used are of molecular biology grade and were purchased from Sigma-Aldrich unless described otherwise. Oligonucleotides were purchased from IDT (Integrated DNA Technology) unless described otherwise.

2.1 RNA ligation of small molecules

DNA oligos A and B used for RNA ligation were transcribed using T7 Ribomax^{TM} express large-scale RNA production system (P1320, Promega). Oligo A and B were ligated in the ratio of 1:1 using T4 RNA ligase 1 using manufacturer's instructions (NEB, M0437M). Bacterial RppH was used to dephosphorylate the 5' ends of the RNA according to the manufacturer's instructions (NEB, M03565). The secondary structure of RNA oligos was analysed using RNAfold software [201]. The sequence of Oligo A and B are as follows:

Oligo name	Sequence (5'-3')
Oligo A	attgggacaactgtgttcactagcaacc
Oligo B	attgggagtcagttcaacactagcaata
Oligo C	attgggacaactgtgaacactagcaata

2.2 RNA ligation of large molecules

T4 RNA ligase was used as described previously [202] and reverse transcription (RT) primer was used to reverse transcribe the ligated product that was visualized in an 8% denaturing PAGE UREA gel.

2.3 Cleavage of RNA using RNase H

RNase H was used according to manufacturer's instructions (NEB, M0297S) in varying ratios of RNA: RNase H. Splint DNA was generated complimentary to the RNA molecule under consideration.

2.4 Two step Polymerase Chain reaction for amplification of lnO

Plasmid pGL3 (a kind donation from Dr. Dimitri Andreev) was used as a template for PCR amplification of the firefly luciferase gene. Primer sO was obtained from *Trilink Technologies* in triplicates (sO1, sO2, and sO3).

The primers with the following sequences were used:

Primer name	**Sequence (5'-3')**
sO	cgccgtaatacgactcactatagnnnnnnnnnnacaactgtgttcactagcaa
a131	ttttttttttttttttttttttttttttttttttttttttttttttttttttaacttgtttattgcagcttataatgg
aFla50	acaactgtgttcactagcaacctcaaacagacaccatggcctgcagggaagacgccaaaaacataaa

A two-step PCR method was used to generate the lnO DNA template using primers a131, sO, and aFla50 and Phusion polymerase (NEB, M0530L) (Figure 2.1).

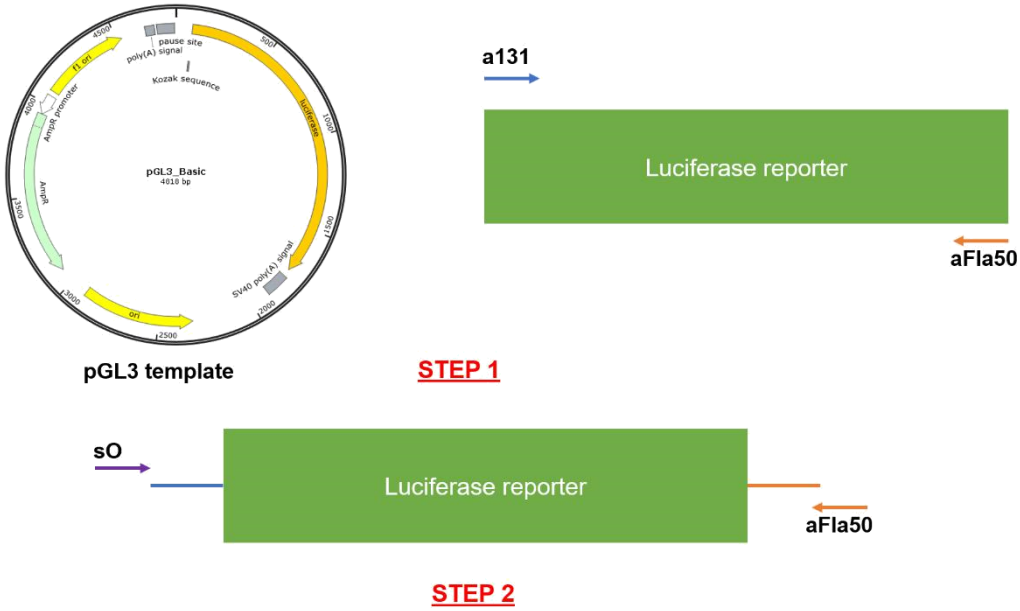


Figure 2.1: Two-step PCR method for generation of the lnO template. A part of the pgl3 vector was amplified using an a131 primer and aFLA50 in PCR-1; sO, and aFLA50 in PCR-2 to incorporate E5S into the lnO template.

Table 2.1: PCR conditions for Phusion PCR

Component	50 µl Reaction	Final Concentration	
Nuclease-free water	32.5 µl		
5X Phusion HF Buffer		10 µl	1X
10 mM dNTPs		1 µl	200 µM
10µM Forward Primer (a132)	2.5 µl	0.5 µM	
10µM Reverse Primer (aFla50)	2.5 µl	0.5 µM	
Template DNA (pGL3 vector)	1 µl	20ng	
Phusion DNA	0.5 µl	1.0 units/50 µl PCR	

Table 2.2: PCR conditions for producing lnO DNA

STEP	TEMP	TIME
Initial Denaturation	98°C	30 seconds
30 Cycles	98°C 65°C 72°C	10seconds 30seconds 30 seconds per kb
Final Extension	72°C	10 minutes
Hold	4°C	

PCR-2 incorporates sO into the lnO DNA template to include a T7 promoter and the random region (N=10) sequence. The reaction components of the PCR reaction were set up as in table 2.3. The thermo-cycling conditions were as follows:

Table 2.3: Thermo cycling conditions for PCR2

STEP	TEMP	TIME	λ (°C/s)
Initial Denaturation	98°C	30 seconds	3
15 Cycles	98°C 65°C 72°C	10 seconds 30 seconds 30 seconds	2.2 2.2 2.2
Final Extension	72°C	10 minutes	3
Hold	4°C		

2.5 Taq Polymerase extension

sO was annealed and extended to its complement (RVG_long) to form sOBG using an annealing mix of 10 µM forward primer (sO), 10 µM reverse primer (RVG_long) and 4 µl of 0.2 M Tris HCl at pH=7.5.

Primer name	Sequence (5'-3')
sO	cgccgtaatacgactcactatagnnnnnnnnnnacaactgtgttcactagcaa
RVG_long	tatgtttttggcgtcttccctgcaggccatggtgtctgtttgaggttgctagtgaacacagttgt

Table 2.4: Optimal set-up condition for taq polymerase reaction

TEMPERATURE	TIME
95°C	5 minutes
70°C	11 minutes
37°C	5 minutes

40 µl of the above annealing mix was added to the extension mix comprising of 40 ul of 5x superscript 3 buffer, 10 µl of 0.1 M DTT,8 µl dNTPs (10mM each), 3 µl of SuperScript™ III Reverse Transcriptase and 99 µl of water. The extension and annealing mixes were resuspended thoroughly and left at 37°C for 45 minutes and purified using isopropanol extraction.

2.6 *In vitro* Transcription (IVT)

RNA was transcribed using the manufacturer's protocol from AmpliScribe™ T7 High Yield Transcription Kit (Epicentre, AS3107) using 1 µg of DNA as a template.

2.7 RNA purification

RNA was purified according to the manufacturer's protocol using the Zymo research RNA clean and concentrator kit (Zymo research, R1015). The quality of RNA was verified on a 7.5 % PAGE-urea gel with recipe described in Table 2.6.

Table 2.5: Recipe for 7.5% PAGE urea gel

40% acrylamide/bis-acrylamide (19:1)	2.82 ml
Urea	7.2 g
10X TBE	1.5 ml
Water	4.72 ml
37°C to dissolve then filter before adding	
10% APS (ammonium persulfate)	37.5 µl
TEMED	7.5 µl
final volume	15 ml

2.8 Capping of RNA

RNA was capped using the manufacturer's protocol from ScriptCap™ Capping Enzyme (Cellscript, C-SCCE0610).

2.9 RNA transfections

HEK293T cells were plated at 40-45% confluence in a 15cm dish. After 12 hours, the media was changed. A transfection mix containing 30-40 µg RNA/ 15 cm dish, 100 µl of Lipofectamine 2000 and 3.8 ml of DMEM was mixed in an RNase free tube and incubated at room temperature for 20 minutes. The transfection mix was added to the cells. The cells were kept at 37◦C in a HEPA filter CO_2 incubator for 2 hours. Cells were then lysed, and transfection efficiency was measured using a reporter assay.

2.10 Cell lysis

Polysome lysis buffer (PLB) was made with the following components: 20mM Tris HCl, 250mM NaCl, 1.5 mM MgCl2, 1mM DTT and 0.5% Triton-X 100. To 1 ml of PLB, 1µl of cyclohexamide (100mg/ml) and 10 µl of TURBO DNase was added. HEK293T cells were taken from the incubator. The media was aspirated from the 15 cm plate and washed with cold Phosphate Buffer Saline (PBS) containing cyclohexamide (100mg/ml) followed by the addition of 400 µl of PLB after thorough scraping of the cells from the surface of the dish. The collected cell lysate was added to a 1.5ml RNase free tube and left on ice for 10 minutes followed by centrifugation at 18000 g for 10 minutes. The supernatant was collected and used for the consequent steps. 30 µl of the lysate was used to measure the transfection efficiency (in triplicates) using the reporter luciferase assay [59].

2.11 Luciferase assay

The firefly luciferase activity was determined using the Luciferase Stop & Glo® Reporter Assay System (Promega). Relative light units were measured on a Veritas Microplate Luminometer (Turner Biosystems). The light units in triplicates were measured and standard error bars were plotted using the Graphpad Prism software.

2.12 Sucrose gradients preparation

Sucrose gradients of 60% and 10 % density were prepared respectively comprising of 20 mM Tris-HCl (pH 7.5), 250 mM NaCl, 15 mM $MgCl_2$, 1mM DTT and 100mg/mL cyclohexamide. Approximately 5.5 ml of 10% sucrose was slowly layered onto the same volume of 60% sucrose in a Beckman centrifuge tube (Beckman Coulter, 331372). The gradient tubes were sealed using a parafilm and slowly placed horizontally for 4 hours to allow spontaneous gradient formation. This was followed by carefully and slowly inverting the tube to a vertical position without disturbing the gradient. The cell lysate was loaded onto the gradient and centrifuged at 35,000 g for 3 hours at 4 ° C using an ultracentrifuge (Beckman, Optima XE-100k).

2.13 Polysome fractionation

The fraction collector was carefully washed with RNase free water. A UV lamp was switched on and allowed to warm up. The tubes were removed from the rotor and placed on ice. The pump was set to 6ml/min and the tube was filled with chasing solution (60% (w/v) caesium chloride containing 0.02% bromophenol (w/v). It was ensured that there were no bubbles introduced into the pump syringe or tubing. A Tracer DAQ analysis program was launched. Settings used by Gandin *et al* [203] were used to obtain the digital polysome profile. The pump was set to collect fractions at 1.5ml/min. The pump was put to a remote position, then the pump and fraction collector was started. The fractions were collected in a 96 well UV plate every 11 seconds (~250 µl). At the same time, the DAQ tracer was switched on and an upward displacement of the gradients was started along with simultaneous detection of UV absorbance at 254nm. The settings on the DAQ tracer were as per the Gandin *et al*. protocol [203]. When the first drop of chasing solution came out, the fraction collection was stopped. The polysome trace was saved in .csv format. The values corresponding to channel 0 (corresponding to 254nm absorbance) were selected and plotted as a scatter plot.

2.14 RNA extraction

Total RNA was isolated from 15 cm plates using manufacturer's protocol for TRIZOL LS (Invitrogen™, 10296028). The Trizol method was the preferred method for RNA extraction [204] from sucrose fractions. The fractions were separated based on their absorbance values into monosome fraction, light fraction and heavy fraction respectively where, Light (2-5 ribosomes) and heavy (>5 ribosomes) respectively. Equal amounts of polysome fractions were flash frozen with equal amounts of Trizol LS. RNA was extracted as per the manufacturer's protocol and precipitated using isopropanol precipitation.

2.15 Isopropanol precipitation

The sample to be isolated was precipitated with 10% volume of sodium acetate, 1.5 µl of glycoblue (Ambion, AM9515) and 1-1.5 volumes of isopropanol and left at -80°C

for 1 hour. This mixture was centrifuged at 12000 g for 20 minutes followed by an ethanol wash of 500 µl of 80% ethanol centrifuged at 12000 g for 20 minutes. The resulting precipitate was eluted in water to the desired volume.

2.16 Poly A purification

mRNA was obtained from the total RNA fraction using Purist poly A Mag kit (Ambion™, AM1922) following the manufacturer's protocol. The percentage of mRNA obtained varied from 0.7-1% of the total RNA extracted.

2.17 Reverse Transcription (RT)

RNA was purified by phenol-chloroform extraction, re-suspended in 10 µl of water and 2 µl of reverse transcription primer (RT_primer or RT_primer_modified) was added. This premix was denatured at 80°C for 2 minutes and then placed on ice for 2 minutes. The RT reaction was set up as tabulated in table 2.7 below and incubated for 30 min at 48 °C in a thermal cycler:

Table 2.6: Reaction set up for the reverse transcription reaction

Component	Amount per reaction (µl)	Final
Ligation and primer	12.0	
First-strand buffer (5×)	4.0	1×
dNTPs (10 mM)	1.0	0.5 mM
DTT (0.1 M)	1.0	5 mM
SUPERase·In (20 U μl^{-1})	1.0	20 U
SuperScript III (200 U μl^{-1})	1.0	200 U

The RT reaction mixture was incubated at 48°C for 30 minutes and RNA was hydrolysed at 95 °C for 20 minutes. The reaction was transferred to a 1.5 mL tube with 156 µl water, 20 µl sodium acetate (3M, pH 5.5), 2 µl glycoblue and 300 µl isopropanol was added to precipitate the cDNA.

RT_primer:

5’Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTG GTCGC/iSp18/CACTCA/iSp18/TTCAGACGTGTGCTCTTCCGATCTAGTTTGA GGTTGCTAGTGAAC3’

RT_primer_modified:

5’/5Phos/AGANNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGA TCTCGGTGGTCGC/iSp18/CACTCA/iSp18/TTCAGACGTGTGCTCTTCCGATC TAGTTTGAGGTTGCTAGTGAAC3’

Post precipitation, the product was dissolved in 10 µl of water and mixed with 3X RNA loading dye. The purified cDNA was visualized on a 7.5% PAGE urea gel prepared according to the following recipe:

Table 2.7:7.5% urea TBE gel

40% acrylamide/bis-acrylamide (19:1)	2.82 ml
Urea	7.2 g
10X TBE	1.5 ml
Water	4.72 ml
37°C to dissolve then filter before adding	
10% APS (ammonium persulfate)	37.5 µl
TEMED	7.5 µl

Once the gel was set, the gel was pre-run at 15mA for 30 minutes. The samples were mixed in 3X loading dye and heated at 80°C for 2 minutes followed by 2 minutes on ice and sample loading on the PAGE-UREA gel. The control RT sample contained 2.0 µl of reverse-transcription primer (1.25 µM), 8 µl of water and 3X RNA loading

dye. The gel was run at 15 mA (~ 300V) for 60-70 min to separate the non-extended primer from the RT product. The gel was visualized using SYBR gold in 10X TBE buffer under blue light.

The RT product band was purified from the denaturing gel using 600 µl of DNA extraction buffer (300 mM NaCl, 10 mM Tris (pH 8) and 1 mM EDTA) and left at room temperature in a nutator overnight. On the following day, cDNA was precipitated from the extraction buffer using isopropanol extraction and DNA was eluted in 16.5 µl of water as described previously.

2.18 Circularization

RT sample was taken and circularized using the reaction components that were added as follows: 2 µl of CircLigase buffer (10x), 1 µl of MnCl2 and 1 µl of CircLigase II. The reaction was set at 60°C for 2 hours followed by a denaturation step at 80°C for 10 minutes. The circularized DNA was purified using isopropanol precipitation as described previously and re-suspended in 10 µl of water.

2.19 Library Preparation

DNA libraries were amplified by Phusion polymerase PCR suitable for HiSeq3000. Circularized DNA was used as a template along with standard forward and reverse Illumina sequencing primers (Table 2.7). A trial PCR was performed to find the optimal number of PCR cycles required for each sample for a 20 µl reaction using the setup shown in table 2.9.

Forward primer: 5′-AATGATACGGCGACCACCGAGATCTACAC-3′
Indexed reverse library PCR primers:

5′-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGA CG TGTGCTCTTCCG-3′ (The underlined NNNNNN indicates the reverse complement of the index sequence used during Illumina sequencing shown in table 2.9).
PCR cycles were optimised for 6-18 cycles to obtain the ideal amount of PCR product required for Illumina sequencing [205].

Table 2.8: Reverse NGS primers used for library preparation

Forward index (5′→3′)	Indexed reverse library PCR primer (5′→3′)
ACGACT	CAAGCAGAAGACGGCATACGAGATAGTCGTGTGACTGGAG TTCAGACGTGTGCTCTTCCG
GCAGCT	CAAGCAGAAGACGGCATACGAGATAGCTGCGTGACTGGAG TTCAGACGTGTGCTCTTCCG
TACGAT	CAAGCAGAAGACGGCATACGAGATATCGTAGTGACTGGAG TTCAGACGTGTGCTCTTCCG
GCTACG	CAAGCAGAAGACGGCATACGAGATCGTAGCGTGACTGGAG TTCAGACGTGTGCTCTTCCG
ATCACG	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAG TTCAGACGTGTGCTCTTCCG
CGATGT	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAG TTCAGACGTGTGCTCTTCCG
TTAGGC	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAG TTCAGACGTGTGCTCTTCCG
TGACCA	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAG TTCAGACGTGTGCTCTTCCG
ACAGTG	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAG TTCAGACGTGTGCTCTTCCG
GCCAAT	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAG TTCAGACGTGTGCTCTTCCG
CAGATC	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAG TTCAGACGTGTGCTCTTCCG
ACTTGA	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAG TTCAGACGTGTGCTCTTCCG
GATCAG	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAG TTCAGACGTGTGCTCTTCCG
TAGCTT	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAG TTCAGACGTGTGCTCTTCCG
GGCTAC	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAG TTCAGACGTGTGCTCTTCCG
CTTGTA	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAG TTCAGACGTGTGCTCTTCCG

Table 2.9: PCR for library amplification of circularized DNA

Initial denaturation	98°C	30 seconds
Cycle 2-24	98°C 65°C 72°C	10 seconds 10 seconds 5 seconds
Final elongation	72°C	5 seconds

Ramping ΔCt 2.2°C/sec for each step

The ramping temperature is crucial in avoiding biases in the PCR reactions [206]. The PCR products were visualized on an 8 % PAGE gel using SYBR gold in 10XTBE buffer and viewed under blue light. The recipe for 8% PAGE gel is as follows:

Table 2.10 : Recipe to make 8% PAGE gel

10X TBE solution	500µl
19:1 acrylamide: bisacrylamide solution (40%)	1 ml
water	3.5 ml
TEMED	10 µl
Ammonium Persulfate (10% w/v) (APS)	35 µl
final volume	5ml

The amplified DNA products were separated from the control library on an 8% PAGE gel. The DNA libraries were extracted from the gel using a DNA extraction buffer overnight and precipitated using isopropanol extraction with 2 µl of glycoblue and dissolved in 10 µl of water. The quality of the PCR products was verified on an 8% PAGE gel and the quantity was measured on The Agilent 2100 Bioanalyzer as per manufacturer's instructions using the Agilent DNA 1000 Kit. The samples were sent for next generation sequencing (NGS) to BGI, Hong Kong.

2.20 Candidate confirmation

Candidates were selected based on their TIRES (Methods, 2.24). The candidate oligos were incorporated on the lnO DNA template replacing the (NNNNNNNNNN) sequence and amplified using Phusion polymerase. The lnO DNA template was transcribed, capped and transfected into HEK293T cells for 2 hours. The firefly luciferase activities were measured on a Veritas Microplate Luminometer (Turner Biosystems). Standard error and mean were calculated for triplicates of individual candidates.

Bioinformatics analyses

2.21 Clipping of identifier sequence

The raw data was available in a FASTQ format (The FASTQ format). We began the bioinformatics analyses by using the Cutadapt tool 208 in order to remove the identifier sequence ACAACTGTGTTCACTAGCAACCTCA from all reads. This was adequate for most of the duplicates and it produced sequences that were either of the expected read length or one nucleotide longer. Cutadapt failed to successfully clip three datasets owing to poor sequence quality at the 3' ends of the read. Therefore, a custom script was written to count the number of occurrences of every permutation. These permutations were included in our analysis if they passed the following:

All nucleotides in the random region were annotated as a base (A, C, T or G). Those denoted with an N were discarded.

The random region (length=10nt) was in the expected location of either positions 9 to 18 or 10 to 19 in the FastQ read.

For reads that were not successfully trimmed with Cutadapt, the nucleotides following the random region must be ACA, representing the start of the HBB TL sequence.

2.22 Aggregation of datasets

The total number of reads obtained in each library was calculated. As the number of reads obtained from datasets PR2-2 and TR2-2 was particularly low, they were aggregated with PR2-1 and TR2-1 respectively.

2.23 UMI correction

Unique molecular identifiers (UMIs) are short sequences or "barcodes" added to each reading in NGS protocols. They serve to reduce the quantitative bias introduced by additional PCR cycles. UMIs were incorporated in the RT_primer_modified to remove sequences that are likely to represent duplicates generated by the library construction PCR. When multiple sequences with the same nucleotide identity share the same extended UMI, one sequence is selected arbitrarily, and the others discarded using a custom script.

2.24 Calculation of TIRES and $TIRES_G$ values

Owing to differences in the level of sequencing depth, the polysomal reads selected to total RNA reads (PR/TR) ratios of every motif present in the libraries were produced from rescaled read counts, as described below:

a) The TIRES ratio of every motif present in the five samples (1-1, 1-2, 2, 3-1 and 3-2 respectively) was calculated using the formula:

$$I_{jk} = \frac{P_{jk}\sum_{i \in J} T_{ik}}{T_{jk}\sum_{i \in J} P_{ik}} \quad \{1\}$$

Where $\boldsymbol{I_{jk}}$ *is* TIRES of an N-nucleotide long variant *j* from the set of 4^n random variants ***J*** calculated for the data obtained in the sample ***k*** (1-1, 1-2, 2, 3-1 or 3-2). ***P*** and ***T*** are the number of reads from PR and TR libraries, respectively.

The maximum effect of $TIRES_G$ on E5S was seen at N=8nt and is maintained consistently in this work unless mentioned otherwise.

b) $TIRES_G$ was computed as the geometric mean of TIR ratio of every motif present in the five libraries 1-1, 1-2, 2, 3-1 and 3-2 respectively.

2.25 Analysis of the NanoCAGE dataset

The NanoCAGE dataset was downloaded from86 and the sequences of transcripts isolated from polysomes was extracted. The data were processed using custom python scripts and plotted using tools from the Microsoft Office suite.

2.26 Sequence logo

The WebLogo3 software suite using custom settings was used to create sequence logos (http://WebLogo.threeplusone.com/create.cgi).

2.27 CAGE data analysis

The CAGEr package was used to extract clusters. Clusters were linked to gene names by finding the closest annotated coding gene downstream. If there was no gene downstream within 51,884 nucleotides the cluster was discarded, this limit was chosen as 95% of annotated TSS's are less than this distance from the annotated coding start. In the case of a gene with multiple clusters, the cluster with the highest read count was chosen.

2.28 Ribosome occupancy in HEK293T cells

Data for the studies (control condition) from Andreev et al and Sidrauski et al was processed using the RiboGalaxy platform 68,209,210. RiboGalaxy uses the Galaxy8 framework for the pre-processing, alignment and analysis pipelines. Custom python scripts were used to calculate the TE values as the ratio of the ribo-seq counts against the RNA counts. The TE values are indicative of the ribosome occupancy on a

transcript. The sequence for each transcript containing a TE value was obtained using the CAGE data (HEK293T cells). The first 11 nucleotides were isolated for each transcript. The $TIRES_G$ for each of these motifs was obtained.

3. Results

Aim

The impact of 5'TL on translation efficiency is sanguine, the influence of cap proximal mRNA nucleotides on translation at a systems level has not been addressed previously. This question becomes important in the context of TSS heterogeneity, thus making possible that mRNA leaders that differ only in a few nucleotides may possibly have different TEs.

To understand the effects of cap proximal nucleotides in the 5'TL, it is important to consider the first few nucleotides in the 5'TL and examine their effects on translation. To understand the effect of each nucleotide on a certain position in the cap proximal nucleotides of 5'TL, it was essential to randomize the region of 5'TL under consideration. In this work, cap proximal nucleotides were randomised upto a length of 10 nts referred to as Early 5' Sequence (E5S) and variations in this sequence context could potentially influence the rate of translation initiation. Although the information available would not directly give us information about the translation initiation rate, it would help us understand a significant statistic influencing the rate of translation initiation called TIRES (Translation Initiation Rate Enrichment Statistic). The aim of this work was to explore the effect of E5S on the TIRES by:

a) Creating a library of molecules including all possible E5S variations and

b) Monitoring the changes in translation initiation rate enrichment statistic (TIRES) across all possible E5Ss.

Experimental considerations

I. Library size and string length preferences for E5S

The library of molecules generated to study the effects of E5S on TIRES was achieved by a randomization strategy. Each position along E5C has one of four nucleotides, making the number of possible variants 4n where n is the nucleotide string length. As the desired library comprised of a string length of 10nt, it contains 410 variants i.e. 1048576. The string length of 10 as E5S was chosen due to the following considerations:

1) The optimal start codon context preferences for translation initiation postulated by Kozak spans a 7nt sequence from -3 to +4 positions surrounding the AUG start codon. A string length of 7 nts was considered as the minimal size and 3 additional nts were added for the extra scope.

2) The longer the sequence length, the higher the possibility of a secondary structure. However, the probability of having a stable secondary structure with 10 nucleotides is low, therefore, any structural impediments occurring in the translational efficiency can be ruled out. The E5S downstream sequence was also kept devoid of high GC content to avoid the possibility of secondary structures.

3) The cap proximal nucleotides of RNA could potentially include a motif for RBPs which might influence TIRES. The RNA binding motifs for different RBPs in 24 diverse eukaryotes have been described by 211. In 102 cases identified in humans, the RNA binding motifs had an average length of ~5-8nt. If the change in TIRES mediated by E5S was due to an RBP, a minimum string length of 10nt should be enough to cover the majority of the RBP consensus binding motifs.

II. Design of control and reporter constructs

Two DNA constructs were generated. The first construct called short oligo (sO) was chemically synthesized by Trilink technologies. The second construct, Long oligo (lnO) was generated by a two-step PCR approach.

III. Architecture of sO

sO, consisted of a T7 promoter followed by 10 random nucleotides (N) and the 5' transcript leader of the human β globin gene (HBB) (NCBI Gene ID: 3043). Certain design considerations outlined below were kept in mind while designing the short oligo.

a) T7 promoter

T7 and SP6 are DNA dependent RNA polymerases that produce RNA transcripts from a DNA template, exhibiting high specificity for their respective promoters 212. T7 RNA polymerase is highly specific in recognising the T7 promoter sequence (TAATACGACTCACTATA) 213. It also requires a double stranded DNA template and Mg2+ ion as a cofactor for the synthesis of RNA 214. The T7 promoter was chosen as: a) the tools available for in vitro transcription (IVT) from the T7 promoter are well developed and robust and b) a wide range of high quality T7 RNA polymerase based in vitro transcription (IVT) kits are available having the capacity to produce large amounts of RNA in a short duration of time (30-60 minutes).

b) Synthesis of random nucleotide (N=10) string in sO

sO was generated in triplicate (sO1, sO2, and sO3) by Trilink technologies using chemical synthesis techniques. Essentially, a DNA string was synthesized containing 10nts distributed equally were (N=A, C, G, and T) inserted at each position. Quality control of the sO was analysed using mass spectrometric analysis by Trilink and found to be within the acceptable wobble range (range not shown in the QC sheets) ("TriLink | Long RNA Synthesis, Longmer RNA,").

Quality control details of the sOs generated by Trilink technologies are outlined in Appendix Table 1.

c) 5' transcript leader (TL) of HBB

The E5S (randomized) was synthesised on to the 5' end of a pre-existing 20nt sequence from the TL of the HBB. This was extended on an overlapping complementary 35nt oligonucleotide region from the HBB using SuperScript™ III Reverse Transcriptase to

generate the final 45bp double stranded sOBG (sO-beta globin TL) containing the E5S and 35 bases of the TL immediately upstream of the AUG start codon of the HBB gene. The HBB 5'TL was chosen due to the following reasons:

1) HBB is highly expressed and translated in many human cell lines.

2) The 5'TL of HBB was analysed using Oligo Calculator (NEB) tool and stable secondary structures (ΔG (kcal. mole-1> -9) were absent.

IV. The architecture of long oligo (lnO)

To facilitate the use of sOBG in downstream experiments, the addition of a reporter gene was necessary. Reporters can be i) readily assayed after transfection, ii) used as markers for screening successfully transfected cells, iii) used for studying the regulation of gene expression and iv) can serve as controls for standardizing transfection. Luciferase assay is the preferred reporter assay system due to its broad dynamic range, high sensitivity, and easy use.

For the generation of the lnO, the coding sequence (CDS) and 3'UTR of firefly luciferase (Fluc) were amplified from the modified pGL3 vector and a 50 nts polyA tail was added using an appropriately designed primer. The start codon of the firefly luciferase was designed in an optimal Kozak context for a) optimal protein production and b) to avoid products of leaky scanning as a result of a poor start codon context 216,217. The poly A tail protects the mRNA molecule from enzymatic degradation in the cytoplasm, and aids in transcription termination, export of the mRNA from the nucleus, and translation 218. sOBG was then amplified with the reporter gene as outlined in the methods to generate the lnO shown in Figure 2.1. LnO was used as a standard construct for IVT of RNA used in all downstream experiments described in this thesis unless otherwise mentioned.

V. Transfecting mRNA into cells for protein production

HEK293T cells were the preferred choice of cell line due to reliable growth rates, propensity for transfection and general robustness. DNA transfection is considered a robust way to initiate protein production in cells and is less technically demanding with respect to creating a target mRNA library. When DNA is transfected into cells, both transcription and translation biases can occur. However, with mRNA transfection,

biases resulting from transcription are eliminated as the mRNA is directly provided to the cells. In addition, mRNA transfections are reliable and easily quantifiable. Hence, mRNA transfections were chosen as the method of introducing the target lnO library in HEK cells.

VI. Choice of a method to study mRNA libraries that are highly translated

Various techniques can be used to study mRNAs that are highly-translated:

1) Polysome profiling uses sucrose gradients to separate highly translating mRNA population from untranslated ones 20,203.
2) Ribosome profiling measures the translation of ribosome protected fragments by deep sequencing 8. Using this technique, the position of the ribosome at codon resolution can be determined to allow discoveries of new coding transcripts and protein isoforms as well as accurate measurement of translation rates 219.
3) Translating ribosome affinity purification (TRAP) is a technique used to analyse cell specific translation responses. TRAP involves the generation of engineered cells that express a tagged ribosomal protein in-vivo under a tissue specific promoter. These tagged ribosomes are then purified, and associated mRNAs are identified by microarray or deep sequencing [5].

Identification of elements in the 5'TL corresponding to translated mRNA isoforms is essential for the analysis of gene expression regulation 220. Polysome profiling provides access to the full-length translated mRNAs including the untranslated regions (UTRs). In contrast, ribosome profiling can map ribosome protected fragments only to coding sequences. Hence, polysome profiling was chosen to study highly-translating mRNA populations in E5S libraries.

3.1 Library requirement to study the effect of E5S on TIRES

A number of next generation sequencing libraries were used in this work (n=18). These libraries account for differences caused by technical variation across samples and account for biases that may be introduced by the experimental protocol adopted to study the effect of E5C on TIRES. Below is a general summary of the various indexes and library names that is used constantly throughout this thesis:

1) **The sOBG library** is a control library obtained from the taq extension of the synthetic oligo (sO) generated from the commercial provider, Trilink. sO was obtained in triplicate namely sO (1-3) to account for any biases in the chemical synthesis of the random sO oligo. sO was extended to form sOBG (1-3) to form three libraries.

2) **The lnO library** is obtained upon the addition of the 5'TL of HBB and CDS+3'UTR of the luciferase reporter to sOBG. sOBG (1-3) generated lnO (1-3). lnO control library was generated to account for any PCR biases generated in the PCR approach used for lnO generation. There are three lnO libraries in total.

3) **Total RNA (TR) libraries** were generated when lnO (1-3) were transfected into HEK293T cells in duplicates. Cell lysates were collected 2 hours post transfection to isolate total RNA to represent the original control library. There are six libraries in total from lnO (1-3) in duplicates leading to TR1-1, TR1-2, TR2-1, TR2-2, TR3-1, and TR3-2.

4) **Polysomal RNA (PR)** libraries were generated when lnO (1-3) was transfected into HEK293T cells in duplicates. Cell lysates were collected 2 hours post transfection and loaded on a sucrose gradient to isolate (actively translating) polysomal RNA. There are six libraries in total from lnO (1-3) in duplicates leading to PR1-1, PR1-2, PR2-1, PR2-2, PR3-1, and PR3-2.

5) **Translation Initiation Rate Enrichment Statistic (TIRES)** of each library was calculated as shown in the methods (Methods, 2.26) representing the ratios of TR/PR, to form six libraries in total as 1-1, 1-2, 2-1, 2-2, 3-1 and 3-2 respectively.

Overview of possible experimental designs

3.2 Production of the target E5S RNA library

The primary aim of this work was to create a library of molecules including all possible E5S variations (410). Different approaches were used to produce the target RNA library. One of the simplest methods considered was the use of chemically synthesized RNA. However, chemical synthesis of RNA is restricted to sizes of 50-100 nts 221, considered too small for the experimental approach used here.

Bacteriophage polymerases such as T7, SP6 and T3 available commercially were used for the transcription of long RNA molecules. These RNA polymerases preferentially initiate RNA synthesis with 1-3 guanine (G) residues (Trilink | Long RNA Synthesis, Longmer RNA). Robust kits are also available commercially for the same purpose and were used.

The T7 RNA polymerase consensus sequence in the T7 promoter contains a G, GG or GGG at its 3' end. The 5' end of the T7 transcribed mRNA always contains a G at the +1 position which is a critical part of the T7 promoter 213. Thus, all T7 mRNAs in the RNA library start with a G at +1 position.

We aimed to create an RNA library starting with a random nucleotide containing the E5S, following the 5'cap. It was, therefore, essential to remove the +1G generated from T7 IVT. Modified T7 promoters without guanines at positions +1, +2, and +3 were tested for in vitro synthesis of mRNA from the DNA template. T7 polymerase did not support the IVT of RNA in the absence of G at +1. The presence of G at the +2 and +3 positions of the 5' mRNA terminus improved efficiencies of in vitro RNA production, but their absence was permissive to T7 polymerase mediated RNA production (data not shown). Therefore, the removal of +1G from IVT RNA was considered using the following approaches: a) Ligation of a chemically synthesised RNA molecule containing random nucleotides (N=10) onto the luciferase reporter RNA, b) Cleavage of the +1G using RNase H and c) Cleavage of the +1G using a self-cleaving enzyme such as hammerhead ribozyme.

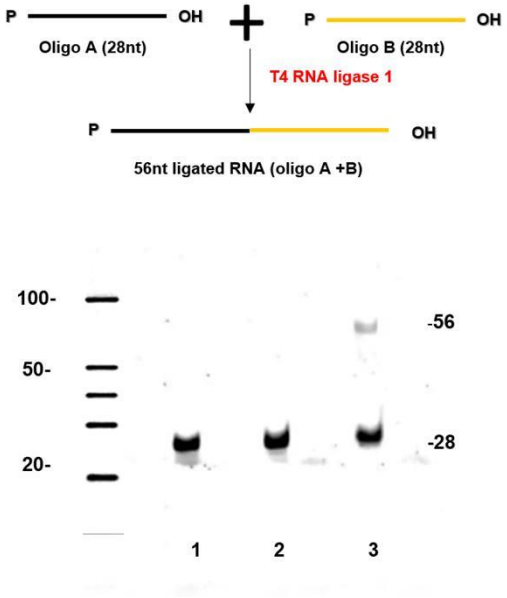
a) Using an RNA ligation method to generate the target E5S library

T4 ssRNA ligase 1 catalyses the ligation of 5' phosphoryl terminated nucleic acid to a 3' hydroxyl terminated nucleic acid using single stranded RNA molecules as substrates[222–224]. For an *in vitro* transcript that contains a 5' triphosphate, dephosphorylation of RNA molecules using RNA 5´ pyrophosphohydrolase will generate a 5´ monophosphate [225]. To produce the target mRNA E5S library containing +1N (Nϵ(A,C,T,G)) juxtaposition to the cap, the first approach considered was the ligation of a chemically synthesised RNA molecule containing E5S onto an in vitro synthesised luciferase reporter RNA.

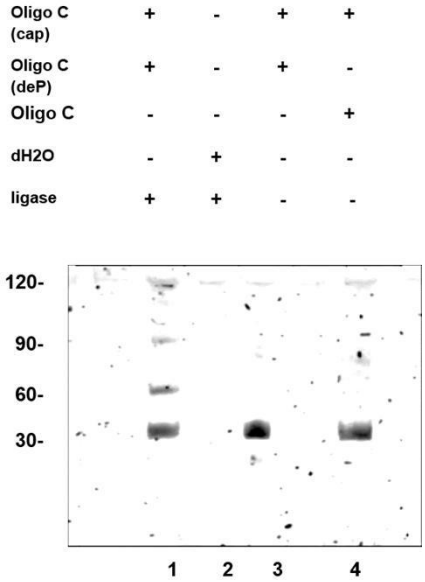
Preliminary experiments were performed to estimate the efficiency of ligation in different conditions (Figure 3.1a and b). When two RNA oligos of different nucleotide compositions were ligated under standard conditions, the ligation reaction was successful but with low product yield (Figure 3.1 a). A similar result was seen when homologous (same nucleotide composition and size) RNA oligos were ligated using a standard ligation protocol in Figure 3.1b. It was observed that in the absence of a dephosphorylated *in vitro* transcript, the ligation reaction did not occur (Figure 3.1b).

To test the efficiency of ligation of a small molecule (oligo A) and a long reporter molecule, the long molecule was dephosphorylated, and standard T4 ssRNA ligase ligation conditions were used as recommended by the manufacturer. As the difference between the short and long RNA molecule was minimal (28nt), it was not possible to visualize the products of the ligation reaction using a denaturing PAGE urea gel. Therefore, the ligation products were reverse transcribed and amplified for 20 cycles using Phusion polymerase. A successful PCR product indicated successful ligation of long and short RNA molecules. Figure 3.1c shows that the ligation of short and long RNA molecules was unsuccessful. This is likely due to inefficient dephosphorylation of the long reporter molecule.

a)


P
OH
Oligo A (28nt)
P
OH
Oligo B (28nt)
T4 RNA ligase 1
P
OH
56nt ligated RNA (oligo A +B)
100-
50-
20-
-56
-28
1
2
3

b)


Oligo C (cap) + - + +
Oligo C (deP) + - + -
Oligo C - - - +
dH2O - + - -
ligase + + - -
120-
90-
60-
30-
1
2
3
4

c)

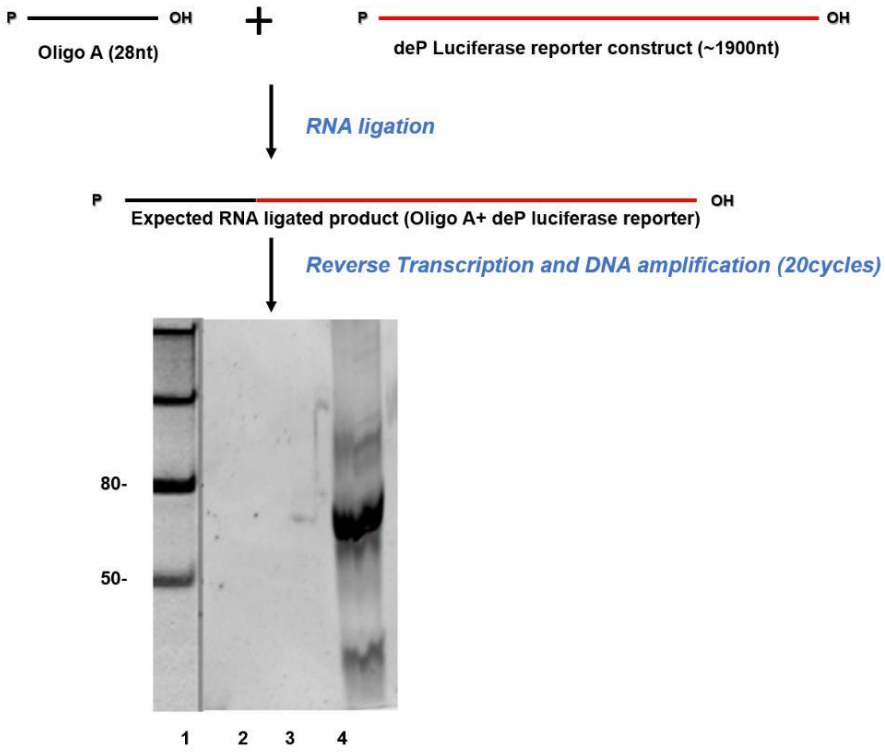

P
OH
Oligo A (28nt)
P
OH
deP Luciferase reporter construct (~1900nt)
RNA ligation
P
OH
Expected RNA ligated product (Oligo A+ deP luciferase reporter)
Reverse Transcription and DNA amplification (20cycles)
80-
50-
1
2
3
4

Figure 3.1: Analysis of ligation efficiency of in vitro transcribed RNA fragments using T4 ssRNA ligase 1.RNA ligations were performed on in vitro transcribed RNA molecules as described in the methods (2.3) and visualized using 15% PAGE urea gel stained using 2% SYBR gold. a) Left lane, RNA marker (sizes 20-100 bp); lane 1, RNA Oligo A; lane 2, dephosphorylated RNA Oligo B; lane 3, RNA Oligo A and dephosphorylated RNA oligo B ligated using T4 ssRNA ligase 1, b) Incubation of oligo C (30nt) with T4 ssRNA ligase1.Lane1, oligo C capped and oligo C (dephosphorylated with 5'pyrophosphohydrolyase) ligated using T4 ssRNA ligase 1; Lane 2, control (ligase only, no oligos); lane 3, control (oligos only, no ligase); lane 4, capped oligo C and oligo C incubated with T4 ssRNA ligase1. c) ligation of short (oligo A) and dephosphorylated long reporter RNA molecules (luciferase) that were reverse transcribed, amplified at 20 cycles using Phusion polymerase and visualized on a 8% PAGE gel stained using 2% SYBR gold where Lane1, low ssRNA marker; Lane 2: short+ long RNA without T4 ssRNA ligase 1; lane 3: short+ long RNA with T4 ssRNA ligase 1; Lane 4: Amplified control oligo (oligo A plus a 20nt long 5'TL of luciferase oligo).

The small molecules were ligated (Figure 3.1a and b) under standard conditions recommended as per manufacturer's instructions, but the efficiency of ligation was observed ~ <25%. Dephosphorylation of the RNA oligo 5' end is required for successful RNA ligation. The ligation of a short oligo to a luciferase reporter RNA was not detectable by PCR (Figure 3.1c). Unsuccessful dephosphorylation of the luciferase reporter mRNA (~1950nt long) may have contributed to the inefficiency of ligation. Consistent with this, Dr. Stephen Rader 202 (personal communication) observed similar limitations with long oligos.

Splint ligation of RNA was considered. Using this method, specific RNA molecules are ligated together using T4 DNA ligase and a bridging DNA oligo complementary to the RNAs226. Stark et al employed a splint ligation strategy for RNA molecules of lengths 100-120nt [227]. Various attempts were made to optimise this method including changing the lengths of splint DNA molecules from 18nt to 25nt, however, this did not improve the ligation of long RNA molecules, and this strategy was unsuccessful (data not shown). Although the ligation of small RNA molecules is well studied, there are few studies addressing the ligation of long RNA molecules. It is not clear why the ligation reaction in this study was problematic, and unsuccessful ligation of small and long RNA molecules created a considerable challenge. As a result, various alternative strategies were tested to cleave the +1G produced during in vitro transcription of the lnO RNA.

b) Removal of +1G from in vitro transcribed mRNA using a site-specific cleavage reaction

To obtain the lnO RNA library comprising of homogenous +1N ends in the 5' terminus, the cleavage of +1G produced from IVT using T7 RNA polymerase was essential. Two approaches were considered for +1G cleavage: a) using an endoribonuclease like RNase H and b) using a self-cleavage ribozyme such as the hammerhead ribozyme.

Endoribonuclease RNase H cleavage of +1G from in vitro transcribed RNA

RNase s H are a family of widely expressed non-sequence-specific endonucleases that hydrolyse RNA from an RNA/DNA hybrid [228]. Figure 3.2a illustrates the mechanism of using RNase H and a complementary DNA oligonucleotide to cleave an RNA molecule. RNase H cuts RNA at the 3' end of the DNA in a partial RNA: DNA hybrid. The design of the RNase H experiment included certain modifications of the lnO oligo as shown in Figure 3.2a. The extended lnO oligo RNA contains X nucleotides preceding the E5S region (N=10). A complementary DNA oligo with sequence Y, where Y is complementary to X is added to form a partial DNA: RNA hybrid cleavage site as shown in Figure 3.2a. Post cleavage, it is critical to separate the cleaved RNA oligos based on their size differences. As the length of the lnO template was around ~2.5 Kb in size, it was essential to have a considerable difference in size between cleaved products for visualisation on a denaturing PAGE urea gel. The size difference between the cleaved products was around 500nt as shown. A complementary DNA oligo of 50nt (Y) is required to cut specifically at the 3' site of the extended lnO oligo before the random nucleotide N. The lengths of the extended DNA: RNA hybrid oligo containing X nucleotides (where length of Xmer =20nt) and complementary oligo containing Y nucleotides (where length of Ymer =20nt) have been modified in the figure (Figure3.2a) for clarity, the actual lengths vary as explained above.

The RNase H cleavage efficiency on long RNA molecules was investigated. Bacterial rRNAs isolated from BL21 *E. coli* cells were used. The length of the bacterial 16s rRNA is 1542 nt. Two DNA oligos complementary to the 16S rRNA region 1100-1117 and 1491-1506 were designed and hybridised to the bacterial rRNA. This was subsequently incubated with RNase H and the appropriate buffer (Methods, 2.32). The expected lengths of products post cleavage are 1117 + 425 nts (oligo probe in the region 1100-1117) and 1506 + 36nt (oligo probe in the region 1491-1506) respectively. However, if the cleavage of the RNA hybridised to DNA probe 1100-1117 occurs in the absence of the cleavage of RNA hybridised to the other DNA probe (1491-1506), a 390nt cleavage product is expected. The efficiency of cleavage of the bacterial 16S rRNA as outlined above was investigated using different concentrations of RNase H relative to RNA (Figure 3.2b). Initially, RNase H (1 unit/20 pmol of RNA).

was incubated with the RNA: DNA hybrid. However, only a small amount of cleavage product was observed. The efficiency of RNA cleavage increased with increasing

amounts of RNase H relative to RNA (Figure 3.2b). When the RNase H concentration was increased 25-fold, cleavage of approximately 50-60% of the 16S rRNA was observed.

Based on these findings, the RNase H approach was considered for cleavage of small RNA molecules using a 37 nts IVT RNA oligo. A lower percentage of RNA cleavage was observed at low RNase H concentrations (Figure 3.2c) and at high concentrations of RNase H, the RNA cleavage efficiency was observed to be ~50%. RNase H cleavage reactions contained ~350 fmol of purified bacterial rRNA and 50 nmol of RNA oligos as the target RNA (Figure 3.2b and c). However, the downstream transfection experiments require large amounts of purified lnO RNA (40-50 pmol (10-30 µg)) depending on the cell density and size of the culture plates used. The results (Figure 3.2 a and b), indicate that to accomplish complete cleavage of the target lnO RNA library, large amounts of RNase H would be required (the RNase H activity is significantly lower than expected). While the approach using RNase H to prepare lnO is somewhat feasible on a small scale, it was considered unfeasible for the large scale required for the downstream experiments.

Various methods including chimeric (RNA/DNA chimeras) molecules, LNA (locked nucleic acid) 229–232 and optimisation of reaction conditions for cleavage of small RNA molecules were investigated (data not shown). Both the chimeric and LNA modifications increased the cleavage efficiency marginally to 60-65% (vs 50% without modification).

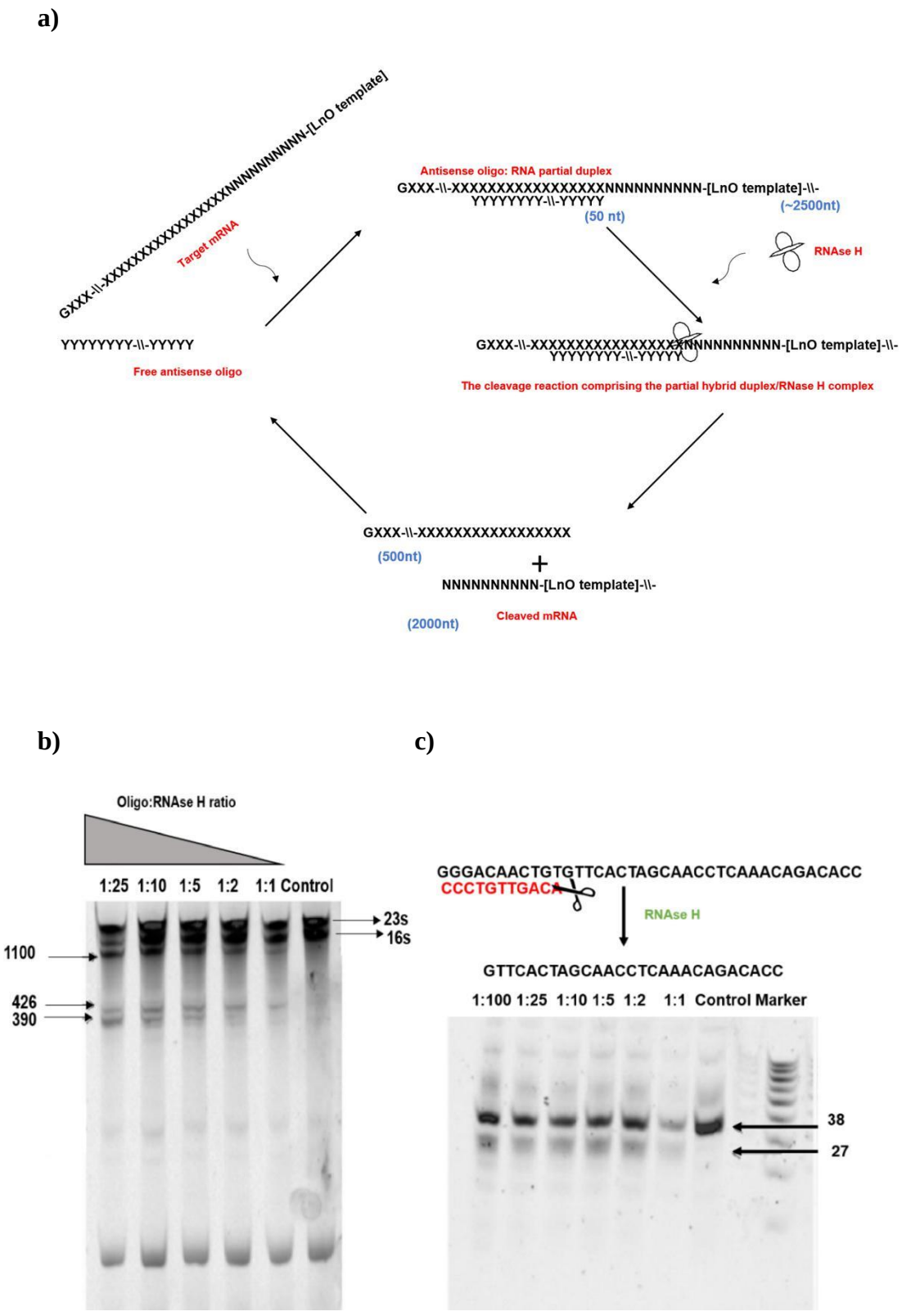


Figure 3.2: Cleavage of RNA/DNA hybrids using RNase H. a) Schematic diagram of the mechanism of catalytic RNase H promoted cleavage of the target lnO RNA b) Cleavage of 16S bacterial rRNA / DNA hybrids (generated using DNA oligonucleotide probes 1100-1117 and 1491-1506) with increasing concentrations (right to left) of RNase H c) RNase H cleavage of a 38 nts RNA (in vitro transcribed) / 10nt DNA hybrid using increasing (right to left) concentrations of RNase H.

Using Hammerhead Ribozyme for endolytic cleavage of lnO RNA

Next, consideration was given to the usage of a self-cleaving ribozyme, like the hammerhead ribozyme, to cleave the +1G at the start of the lnO library. The hammerhead ribozyme (HHR) catalyses the site-specific attack of an activated 2'OH nucleophile and its adjacent 3' phosphate causing cleavage of the P'-O5' phosphodiester linkage to form a 2',3' cyclic phosphate and a 5' alcohol 233. Figure 3.3 illustrates the minimum sequence required by the HHR to cleave target mRNA.
In addition, HHR sequence can be altered for complete cleavage of small RNA molecules 234. The previous methods of preparing lnO RNA using RNA ligation and RNase H were rejected due to their low product yield. Although the use of HHR was considered for cleavage of lnO RNA at the +1G site, it was not used in the final experimental protocol due to predicted issues with the extraction and recovery of cleaved RNA from a denaturing gel, again resulting in limited RNA availability for downstream experiments.

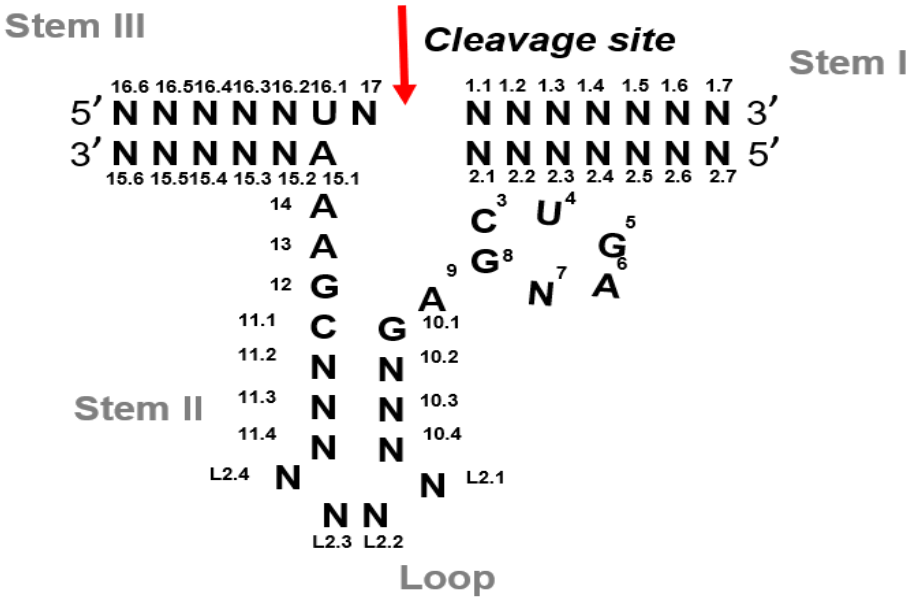


Figure 3.3: The minimal consensus structure required in the hammerhead ribozyme used for efficient autocatalytic self-cleavage.

This figure is adapted from [233].

Optimisation of experimental protocol

3.3 Amplification of template DNA for *in vitro* transcription using a novel PCR strategy

For large scale *in vitro* transcription by a T7 polymerase, a high-quality DNA template (lnO) is critical. The DNA template for the lnO library was generated using a two-step PCR approach (Figure 3.4) (methods, 2.6).

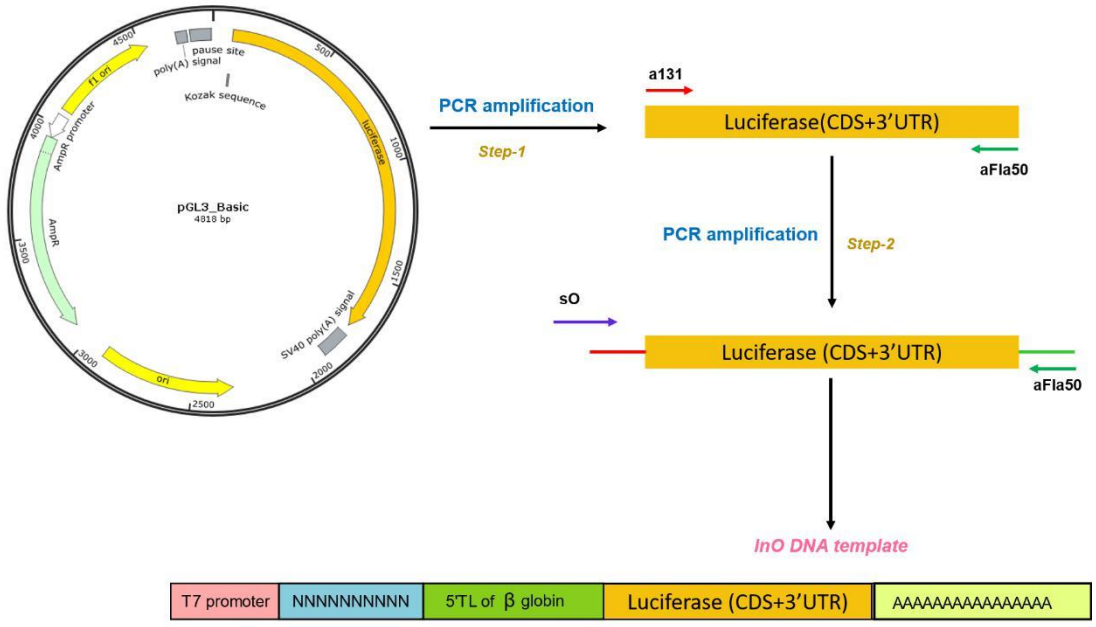


Figure 3.4:The two-step PCR approach used to generate lnO DNA template

The minimum quantity of PCR template used for IVT RNA production with a >99% probability of including 4^{10} variants was calculated using the GLUE web interface program. GLUE [235] is a program for estimating completeness and diversity in randomised libraries. The main aim of the two-step PCR was to incorporate all unique variants of the library exclusive of biases in the amplified DNA template. Using GLUE, it was estimated that a minimum of 3.2 attomoles of DNA was required as starting material to acquire 99% library completeness amounting to 2.1ng of the ~2 Kb lnO (in triplicate).

Three short control oligos were generated, sOBG1, sOBG2 and sOBG3 from each chemically synthesised oligo to be amplified to obtain long oligos lnO1, lnO2 and

lnO3 respectively. The short control oligos were duplexed using Superscript III (Methods, 2.7). The control lnO and sOBG (1, 2 and 3), visualised on a 2% agarose gel, are shown in Figure 3.5. The ~2.1 Kb band represents the lnO and 98bp band represents sOBG control samples.

Phusion polymerase was used to generate the desired DNA template using a two- step PCR with a minimum number of cycles (n=15), and Phusion HF buffer. The approximate error rate of Phusion HF buffer is ~4.4 x 10^{-7}. Specific quantities of DNA i.e. 125ng (100 fmol) were used as a template in each step of the PCR reaction.

PCR amplification and instrument biases are known to occur. Aird *et al* showed that using a ramp rate (λ) of 2.2 C/sec significantly reduced biases in comparison to the standard λ=3℃/sec[206]. As a result, a ramp rate of 2.2 C was adopted to minimise amplification biases. Low number of amplification cycles (n= 15) are also important in reducing biases and were incorporated in the two-step PCR approach [236,237].

Template DNA was purified and used downstream for large scale *in vitro* transcription reactions. 1µg of template DNA was used to generate 100-120 µg of purified RNA using AmpliScribe™ T7-Flash™ Transcription Kit. The quality of RNA was verified on an 8% denaturing PAGE UREA gel. A sharp band was present at ~2 Kb in size (Appendix Figure 1) without any denatured by-products.

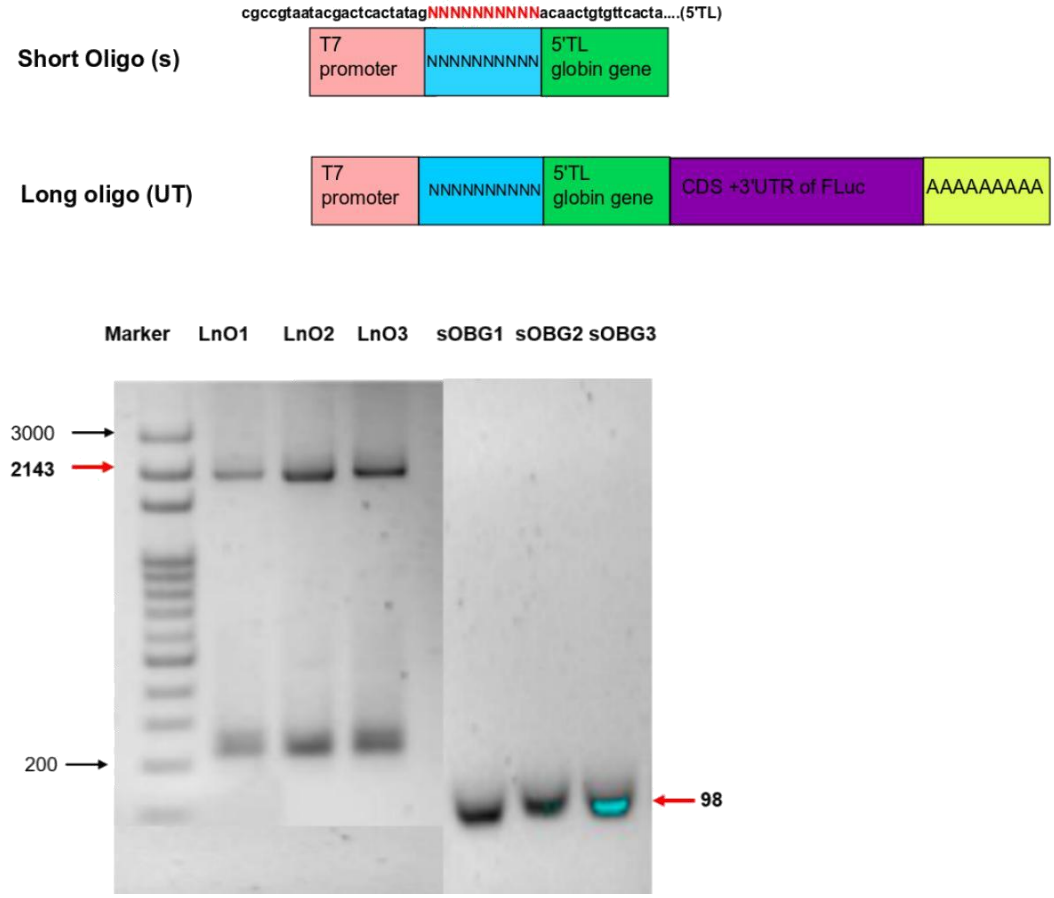


Figure 3.5 : Generation of sOBG and lnO. Single stranded triplicates of sO were extended using a primer complementary to its 3' end and superscript III polymerase. This extension generated the 98bp dsDNA templates sOBG1, sOBG2, and sOBG3, respectively. lnOs were generated in triplicates using a two-step PCR amplification approach using λ= 2.2℃/sec, size 2143 bp. Samples were run on a 2% agarose gel and visualised by staining with safe view and photographed under UV. The lower band above 200bp in the lanes corresponding to lnO triplicates 1, 2 and 3 respectively is the unused ultramer primer (a131) used in the PCR reaction.

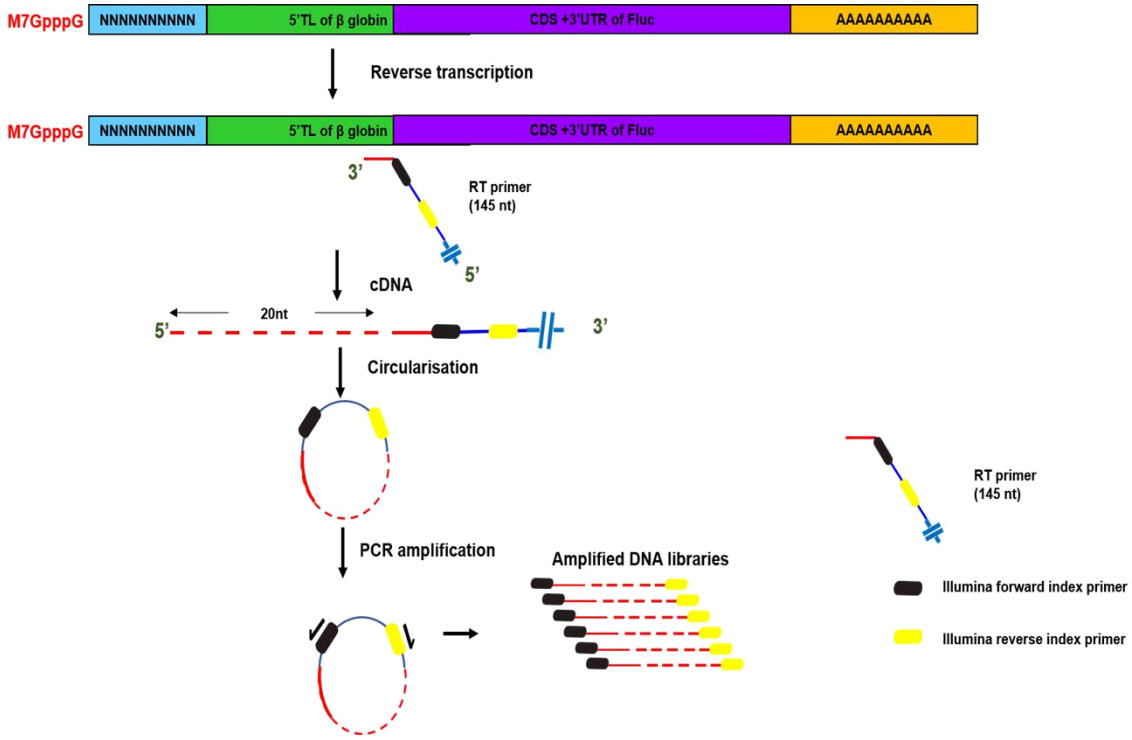


Figure 3.6: Schematic outline of the steps involved in the DNA library preparation from an RNA template. RNA was isolated from HEK293T cells post RNA transfection (2 hours), purified and reverse transcribed using a custom reverse transcription primer (RT_primer/ RT_modified_primer) containing Illumina forward and reverse primers as shown in the figure. The cDNA was then circularized using CircLigase™ ssDNA Ligase. The circularized cDNA was amplified to generate the NGS libraries. The DNA libraries were sequenced using the MiSeq platform.

It was important to verify the success of the two-step PCR approach by verifying the number of biases generated from sO generated lnO DNA. To test this, control libraries were prepared from short oligo controls (sOBG1 and 2). The lnO DNA template was *in vitro* transcribed, capped, reverse transcribed into cDNA using a specific RT primer, circularised and amplified by PCR to produce NGS libraries (Figure 3.6). All *in vitro* transcribed RNA molecules started with a +1G from the T7 promoter and included E5S (N=10) in positions 2-11. When lnO1 RNA was transfected in cells followed by cell lysis and total RNA extraction, duplicate control samples denoted as TR were generated (TR1-1 and TR1-2 are duplicates created from lnO1). These libraries were sequenced using the high throughput sequencing technology, MiSeq.

3.4 Confirming the incorporation of E5S in lnO based control libraries

Each DNA library was sequenced using MiSeq and analysed for the total number of unique variants representing E5S. T7 polymerase used in IVT ensured the presence of G in the +1 position. Under ideal circumstances, it is expected that due to the presence of four nucleotides A, C, G and T, the frequency of occurrence of each nucleotide in positions 2 to 11 should be 0.25 (shown as the red dashed line against the relative frequency of 0.25 in Figure 3.7). The aim of the experiment was to ensure the random distribution of nucleotides in the intended positions of 2 to 11. With minor differences of ± 10%, the nucleotides incorporated at each position were found to be close to the expected frequency of 0.25 apart from nucleotide C at position 4 which was at a frequency of ~0.10 in all samples. This could be due to the way the chemical synthesis of the sO oligos proceeded. The preference for nucleotides in all other positions (2 to 11)was close to the expected frequency. There was no single nucleotide in any specific position over-represented or under-represented in the control samples of sOBG1 and TR1 duplicates as shown in Figure 3.7 indicating the success of the two-step PCR approach.

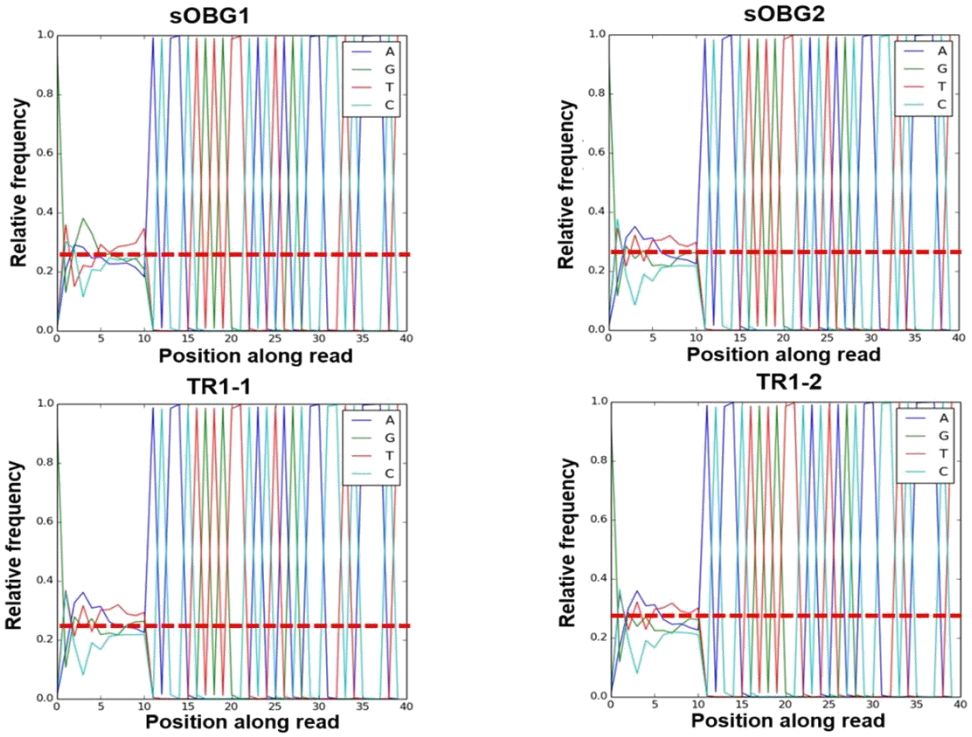


Figure 3.7: NGS data from MiSeq (Triniseq) confirms the occurrence of random oligonucleotides along E5S with the presence of G at the +1 position. sOBG stands for short oligo control in duplicates 1 and 2, TR stands for total RNA control in duplicates 1 and 2.

However, there was a difference in sequencing depth between samples owing to differences in the expression of their E5S unique variants. Variation in the NGS DNA library sample quality (Table 3.1) and a lack of sample equalisation at the MiSeq facility led to a significant variation in the total number of reads i.e. 4790610 to 11164, between samples. It was observed that the highest number of unique variants were associated directly with the highest number of total reads obtained from the MiSeq data. Despite these technical variations, no biases were observed in the representation of a single nucleotide in a specific position of 2 to 11, excluding the aforementioned bias at position 4. The nucleotides were observed at a frequency of 1 at all positions downstream of position 11 marking the end of E5S region.

Table 3.1 Data obtained from MiSeq for samples sOBG1, sOBG2, TR1-1, and TR1-2, respectively.

Sample number	Number of unique reads	Total number of reads
sOBG1	385389	665774
sOBG2	911995	4790610
TR1-1	231795	320914
TR1-2	10662	11164

After the successful generation of lnO DNA template, the next steps included the use of polysome profiling with massively parallel sequencing to study the effects of E5S on TIRES.

3.5 +1G occurs at high frequency in the TSS of annotated human transcripts

The TSS marks the start of transcription of mRNA at the 5'-end of a gene sequence. The determination of the exact TSS position is crucial for the identification of various regulatory elements present in the 5’TL region immediately flanking the TSS 238.

The 5’TL sequence of numerous transcripts are available across several databases. Given that the first position in lnO RNA is guanine, it was of interest to examine the TSS sequences from annotated transcripts (human) extracted from different databases to determine the frequency of guanine at the TSS position. The data was extracted from the following databases, a) GENCODE (v 27) b) UTRdb and c) FANTOM for HEK cells. The distribution (in percentage) of TSS corresponding to individual bases (G, A, T, and C) is shown in Figure 3.8.

The GENCODE Consortium curates gene features in the human genome using a combination of computational analysis, manual annotation and experimental

validation [239]. The current release of GENCODE 27 data (human), containing 79,911 annotated transcripts, was downloaded and the 5'TL including TSS information was extracted for these transcripts. Guanine is the most frequent +1 base (37.12%) at the TSS of all annotated transcripts from the GENCODE 27 database, followed by A (31.6%) (Figure 3.8).

Sequences collated in UTRdb were recovered from the National Centre for Biotechnology Information (NCBI) RefSeq transcripts using custom software [240]. For human genes, a comprehensive collection of UTRs [derived from the full set of over 300 000 alternative full-length transcripts collected in ASPicDB [241]] was generated by a thorough analysis of all available EST/mRNA. The human 5'UTR sequences were downloaded from the UTRdb in the FASTA format and TSSs were investigated for individual base distribution. A (36.7%) was the most frequently occurring +1 base at the TSS, closely followed by G (35.5%) in 124215 annotated transcripts (human) obtained from UTRdb.

CAGE was performed previously across a large collection of primary cell types and revealed that many mammalian promoters are composite entities containing multiple closely-separated TSSs with independent cell-type-specific expression profiles. The FANTOM5 promoter centric expression atlas provides expression profiles for most coding and non-coding transcripts in the human and mouse genomes [242]. The FANTOM5 consortium, containing HEK cells TSS data, was downloaded. G (36.2%) was found to be the most frequently occurring +1 base at the TSS, followed by C (30.6%), in 22213 annotated transcripts in the FANTOM database for HEK cells.

It was expected that the lnO RNA generates the E5S with equal probabilities of all four bases A, C, T, and G in each position along the E5S. Interestingly, the frequency of TSS composition varied in all investigated annotated transcripts (human) from three different databases. However, G was found to be the most frequently occurring base in position 1 of most annotated transcripts in the databases investigated as shown in Figure 3.8. The lnO investigated in this work has guanine at position 1 and thus represents ~35% of available annotated transcripts (human).

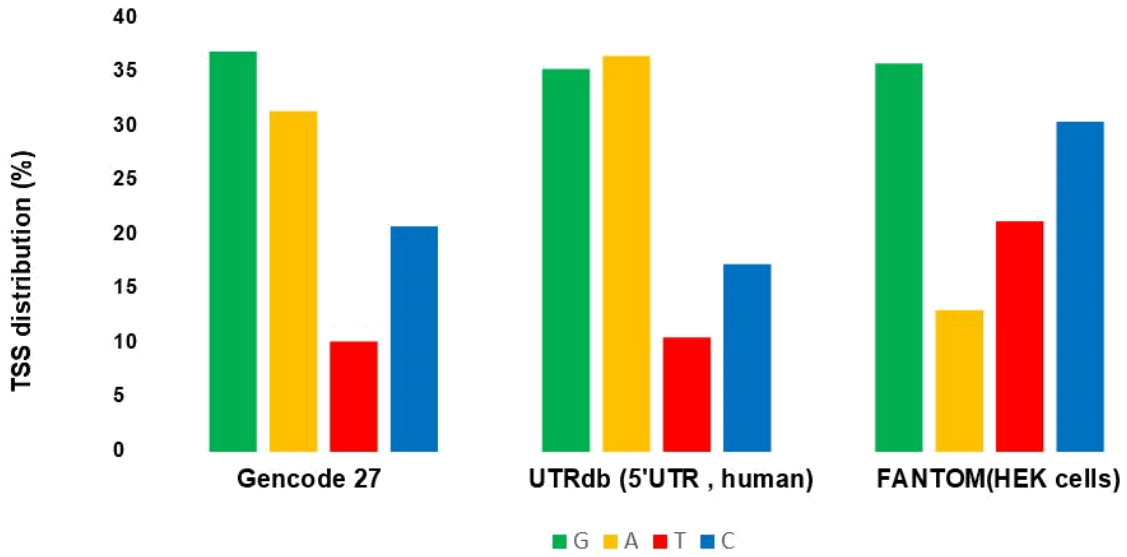


Figure 3.8: **Frequency distribution of individual bases in the TSS of human annotated transcripts from different databases,** GENCODE 27, UTRdb and FANTOM (HEK cell)**.**

3.6 Optimising the polysome profiling protocol for generation of a high quality NGS library

To understand the effects of E5S on TIRES, it was important to develop and optimise a strategic experimental approach. The approach that was used in this work is as follows:

a) mRNA transfection aspects

To understand the in vitro effects of E5S on TIRES, lnO RNA was transfected into HEK293T cells. mRNA transfection of lnO RNA was optimised according to standard protocols. Firefly luciferase activity was the readout method used to determine RNA transfection efficiency. For each lnO RNA (1, 2 and 3), four transfections were performed, and cell lysates were prepared 2hrs after transfection. Out of the four cell lysates, the two samples with the highest luciferase activity were used as duplicates for polysome RNA isolation (PR) and the rest were used as duplicates for total RNA control (TR) for each lnO respectively.

b) Polysome isolation

During mRNA translation, the ribosomal subunits (the 40S and 60S) bind to target mRNA forming an 80S complex i.e. a monosome. Ribosomes move along the mRNA during translation elongation in association with tRNAs. When there is an active translation, many monosomes can associate on the same mRNA molecule forming polysomes. Specific mRNAs bound to polysomes indicates the active translational status of the mRNAs. To investigate the RNA molecules with high TIRES, polysome profiling was the chosen readout for actively translating mRNAs. There are various factors which influence polysome formation such as translation initiation factors, the context of the start codon, elongation factors and the presence of secondary structures in the 5'TL 243,244. In this experiment, the molecular architecture in the lnO library were identical except for differences in their E5S (positions 2 to 11). Therefore, any changes in the TIRES can be considered to be an effect of the E5S on mRNA translation initiation.

c) Library preparation of mRNA controls and polysome RNA

This work aimed to produce deep sequencing libraries, upon analysis would ideally generate sequence information for all unique variants from the E5S library isolated from the polysomes. After transfection of lnO RNA into HEK293T cells, total RNA and polysomal associated RNA were isolated and subsequently used in library preparation as described in Figure 3.6. Based on the assumption that efficiently translated mRNAs are associated with heavy polysomes, the effect of E5S on translation initiation was measured by comparing frequencies of nucleotides (and their combinations) at specific positions in E5S from mRNAs in polysome fractions to their frequencies in E5S of the original library (total RNA) using massively parallel sequencing. While both ribosome and polysome profiling have previously been used for various applications, the use of polysomes in the context of studying the E5S of TIRES has not been reported. This novel experiment, therefore, included various optimisation and development protocols, which are described in the following sections.

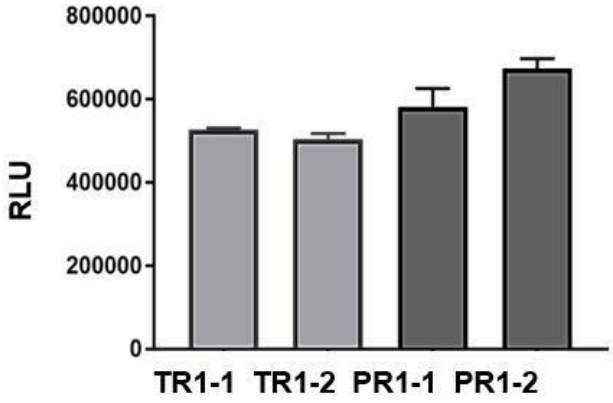
d) mRNA transfections

Lipofectamine mRNA transfection is based on the principle of liposome formation between the cationic lipofectamine and negatively charged nucleic acid molecule. These liposomes can fuse with the negatively charged plasma membrane of living cells, allowing nucleic acid to cross into the cytoplasm and contents to be available to the cell for expression [245,246].

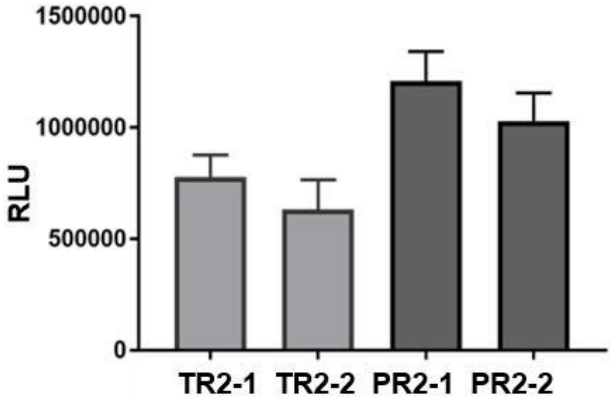
mRNA lnO libraries (1, 2 and 3) were transfected in duplicate into HEK293T cells with the transfection reagent Lipofectamine 2000. Transfection was optimised with respect to cell density. Different capping reagents were evaluated, and the best performing reagent was used. The quantity and quality of mRNA were evaluated prior to transfection. The following conditions were observed for optimal transfection and translation of firefly luciferase from the lnO RNA in HEK293T cells: a) cell confluence was maintained at 70-75% during transfection and b) the duration of transfection was two hours. It was observed that uncapped mRNA showed 500 times less luciferase activity in comparison to capped mRNAs (appendix Figure 2).

Luciferase reporter assays showed expression of firefly luciferase protein, a measure of successful mRNA transfection. Figure 3.9 shows high luciferase reporter values upon transfection of lnO RNA (1, 2 and 3 in duplicate) in HEK293T cells. The lysates with the highest luciferase expression were used for polysome isolation, the others were used as total RNA controls.

a)



b)



c)

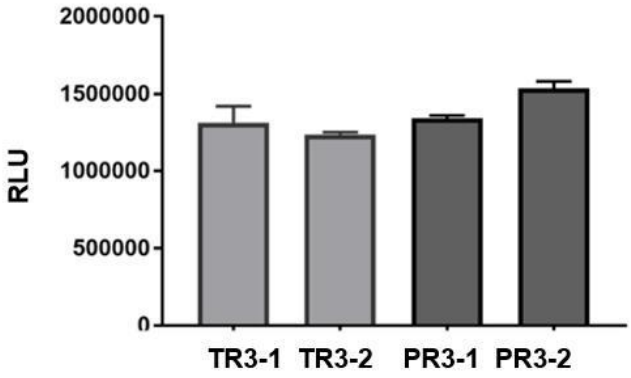


Figure 3.9: Transfection efficiency of samples indexed PR and TR (in triplicate) determined using the luciferase reporter assay. HEK293T cells were transfected with lnO RNA at 30-40 µg/15cm plate. Cell lysates were prepared 2 hours after transfection and firefly luciferase light units were measured to evaluate transfection efficiencies. Transfection efficiencies are tabulated of each lnO RNA used for TR and PR libraries a) 1 b) 2 and c) 3. Data shown as the mean + standard deviation.

3.7 Isolating specific mRNA populations from polysome fraction

Polysomes can be size-fractionated using sucrose density gradient centrifugation. Specific mRNA bound to the polysomes indicates the active translational status of the mRNA. Highly translating mRNAs were isolated from polysomes using a continuous sucrose density gradient (10-60%) (Methods,2.12) (Figure 3.10 (a and d)).

To understand the effects of E5S on TIRES, it was essential to obtain various fractions from the sucrose density gradient (Methods,2.12) as monosomes, light polysomal fraction and heavy polysomal fraction (Figure 3.10a) to indicate their respective translational status in the cell. Various fractions in the sucrose gradient are representative of different stages of active mRNA translation and will yield valuable information about the effects of E5S on TIRES. Total RNA and RNA isolated from polysomes libraries were used to compare the TIRES changes in the unique variants of E5S as explained previously.

Total RNA is ideally comprised of ~5% mRNA. Using polyA mRNA isolation kits, the fraction of mRNA obtained from a pool of total RNA ranged from 0.5-2% (data not shown). The mRNA pool consists of all mRNAs in the cell along with the transfected lnO RNA library. The transfected lnO RNA is a very minor fraction of the total mRNA purified from HEK293T cells. Extracting transfected lnO RNA from the polysomal fraction (PR) for the downstream processing steps and next generation library preparation was the most challenging part of the experiment, and the following optimisation strategy was employed: Initially, a 10cm dish was used to transfect HEK293T cells with lnO mRNA for two hours (four 10 cm dishes). The RNA transfection protocol was maximised for high efficiency as discussed previously (Results, 3.6 d). Lysates with the highest luciferase activities indicative of successful transfection were loaded on a sucrose gradient and processed for polysome isolation in duplicate (two 10 cm dishes). The remaining lysates were used as control to isolate total RNA in duplicate (two 10 cm dishes). In the first attempt, the monosome, light polysome, and heavy polysome fractions were isolated and processed for library preparation (Figure 3.10 (a-c)). Figure 3.10a shows a polysome profile obtained from a 10-60% sucrose gradient of HEK293T cell lysate obtained from a 10cm dish

transfected with 10 µg RNA /dish. High transfection efficiency was a considered as a readout for efficient lnO translation. Total RNA was isolated from different fractions as monosomes, light and heavy polysome outlined in green boxes. Following RNA isolation, reverse transcription of the RNA was performed using a custom reverse primer (Figure 3.10b).

The RT products were then purified and visualised using a 7.5% denaturing PAGE UREA gel. Figure 3.10b shows an unextended reverse transcription (RT) primer used as a control. RT products from the monosome, light, and heavy fractions are expected to form an extended product higher (~20nt) than the unextended primer outlined in the red box (Figure 3.10b). Faint bands indicative of extended products were seen in this area. These were then gel excised, purified, circularised, and used as a template for DNA library preparation using Illumina index primers.

The PCR products were visualised on an 8% PAGE gel stained with 2% SYBR gold containing RT products from the monosome, light polysome fraction and heavy polysome fraction respectively along with the control template (produced from circularisation of the RT primer). The PCR products were found to be the same size as the control template. This could reflect a) unsuitable conditions for PCR amplification of the input templates and/or b) insufficient amounts of RT product in the reaction for downstream processing steps. However, the control template is produced in large quantities indicates that the conditions used for the PCR reaction were favourable. Thus, it is probable that low amounts of RT product was insufficient for downstream processing steps. However, in this case, the presence of two clear bands (above the control unextended primer band) could indicate the presence of non-hydrolysed RNA left in the reaction. To avoid the masking of the non-hydrolysed RNA with the RT product formation on 8% PAGE urea gel, the RNA hydrolysis step was included after the RT reaction. After the inclusion of the hydrolysis step, the RT reaction did not show any bands above the unextended control (data not shown). Although the inclusion of this step was not useful in producing the desired PCR products (data not shown), it helped in further optimisation of the protocol by avoiding non-specific product formation. Overall, it was concluded that the amount of RT product produced by these reactions was insufficient for preparation of successful NGS DNA libraries.

To increase the amount of RT product, it was essential to have a viable quantity of RNA extracted and purified from polysomes. However, it was critical to identify the minimum amount of RNA required for downstream processing and production of high-quality DNA libraries. A q-RT PCR (data not shown) was employed in parallel to confirm the minimum amount of RNA required to produce a successful DNA library against the standard library preparation steps including reverse transcription, circularization and amplification shown in Appendix Figure 10.

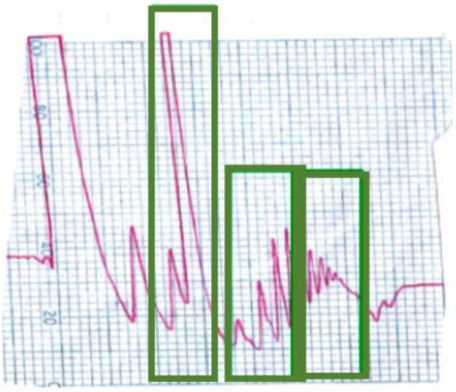
To improve the efficiency of the RT reaction, the mRNA fraction was isolated from the total RNA pool using polyA fractionation. Following this, all polysomal fractions were pooled from the 10cm dish and used for DNA library preparation. However, no RT product was seen in the 8% PAGE urea gel and upon further processing, no PCR products were observed (data not shown). Thus, it was predicted that total RNA from the polysomal fraction of a 10cm dish as an input was not sufficient for DNA NGS library preparation.

Consequently, the yield of polysomal RNA was increased using the following modifications a) increasing the size of the dish (15cm) with 70-80% confluence for transfection, b) increasing the amount of RNA used for transfection to 30µg/15cm dish, c) using a modified RT primer containing a unique molecular identifier (UMI)

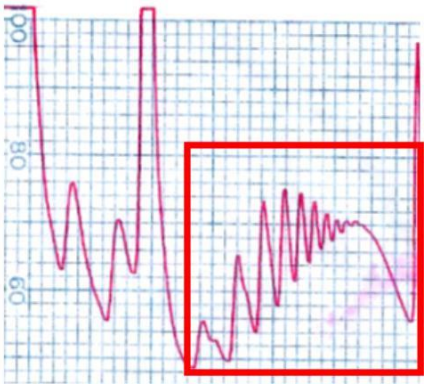
d) polyA purification of total RNA was employed to purify the pool of mRNA fraction from total RNA which helped reduce the background for visualisation of the RT product. The successful results are shown in (figure 3.10 (d-f)). The polysome profile obtained from the cell lysate of HEK293T cells transfected with lnO RNA 1 (in duplicate) as shown in figure 3.10d. The polysome fractions were extracted and pooled together. Subsequently, total RNA was extracted from the polysome fraction, polyA purified to isolate mRNA, reverse transcribed using a custom RT_primer_modified and an RNA hydrolysis step was incorporated. The RT product produced from the polysome fraction visualised on an 8% PAGE UREA gel is shown in Figure 3.10e. A clear band (outlined in the red box) was visible above the control band containing the unextended RT_primer. This band was excised from the gel, purified, and processed downstream to produce DNA NGS libraries. The successful library amplification PCR reaction was visualised on an 8% PAGE gel as shown in figure 3.10f. A clear band was seen above the control band upon PCR amplification of 14 cycles. This was an indication of a successful library suitable for

processing on a Hi-Seq3000. Previously, a similar approach to the successful experiments described above was assayed using a smaller dish (a 10 cm dish with 70-80% confluence) for transfection and incorporating all the other modifications. This approach was unsuccessful (data not shown) and it was concluded that the minimum cell number required to generate successful target DNA libraries suitable for HiSeq3000 from polysomes, was approximately 16 x 106 cells i.e. 15 cm dish at 70-75% confluence.

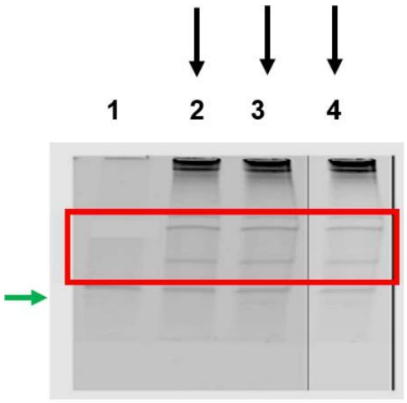
a)



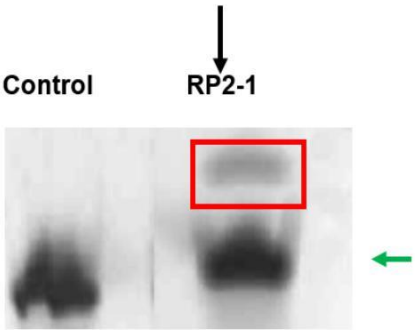
d)



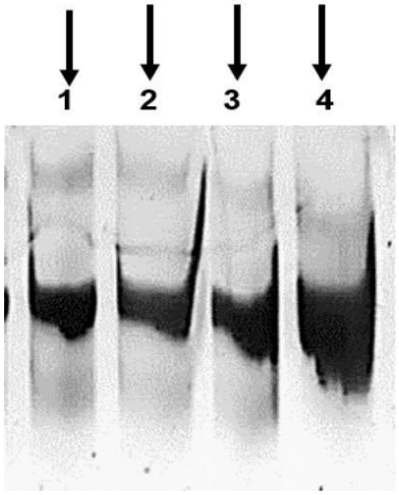
b)



e)



c)



f)

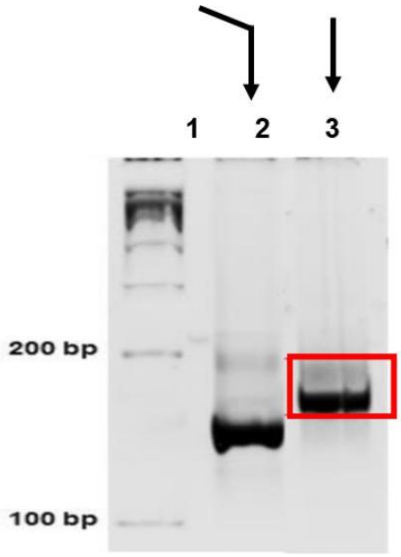


Figure 3.10: Isolation of polysomal fractions for successful library preparation a) A polysome profile obtained from 10-60% sucrose gradient of HEK293T cell lysate from a 10cm dish transfected with 10 µg lnO RNA per dish. OD was monitored at 254nm using a spectrophotometer connected to the fractionator setup (Y axis) across increasing sucrose density (X axis). Total RNA was isolated from different fractions as monosomes, light and heavy polysomes outlined in green boxes. b) the total RNA from various fractions of the polysome was reverse transcribed purified and visualized using a 7.5% denaturing PAGE urea gel; lane 1, unextended reverse transcription (RT) primer; lanes 2, 3 and 4, RT products from monosome, light and heavy polysome fraction respectively. The green arrow indicates the unextended RT primer to be avoided and the RT products expected are outlined in the red box selected for purification by gel excision (the expected RT product should be about 20nt longer than the primer. Two main bands were visible above the control band, the RT product was unclear, hence both were recovered by gel excision). These products were then circularized and used as a template for PCR amplification. c) DNA library products following amplification of the circularised RT products analysed on 8% PAGE gel stained with 2% SYBR gold. Libraries prepared from RT products obtained from the monosome, light and heavy polysomal fraction are shown. lane1: control library ~151 bp from the unextended rt primer, lanes 2-4: libraries from monosome, light, and heavy polysomal fractions respectively. d) the polysome profile showing all polysomes obtained from a HEK293T cell lysate transfected with 30 µg lnO RNA per 15cm dish which was used for total RNA isolation and polyA purification. e) a 7.5% denaturing PAGE urea gel stained with 2% SYBR gold facilitating visualization of the RT product from RNA obtained from the polysomal fraction. The green arrow indicates the unextended RT primer to be avoided and the RT products are outlined in the red box (about 20bp larger in size) for gel excision and purification. f) 8% PAGE gel stained with 2% SYBR gold used to visualize libraries generated from circularised RT product obtained from polysomal fraction alone. Lane 1: DNA marker used to indicate the 100bp and 200 bp, lane 2: control library ~151 bp from the unextended RT primer lane 3: libraries obtained upon 8-18 cycles depending upon the sample. of the PCR reaction (~171 bp in size) were excised for deep sequencing.

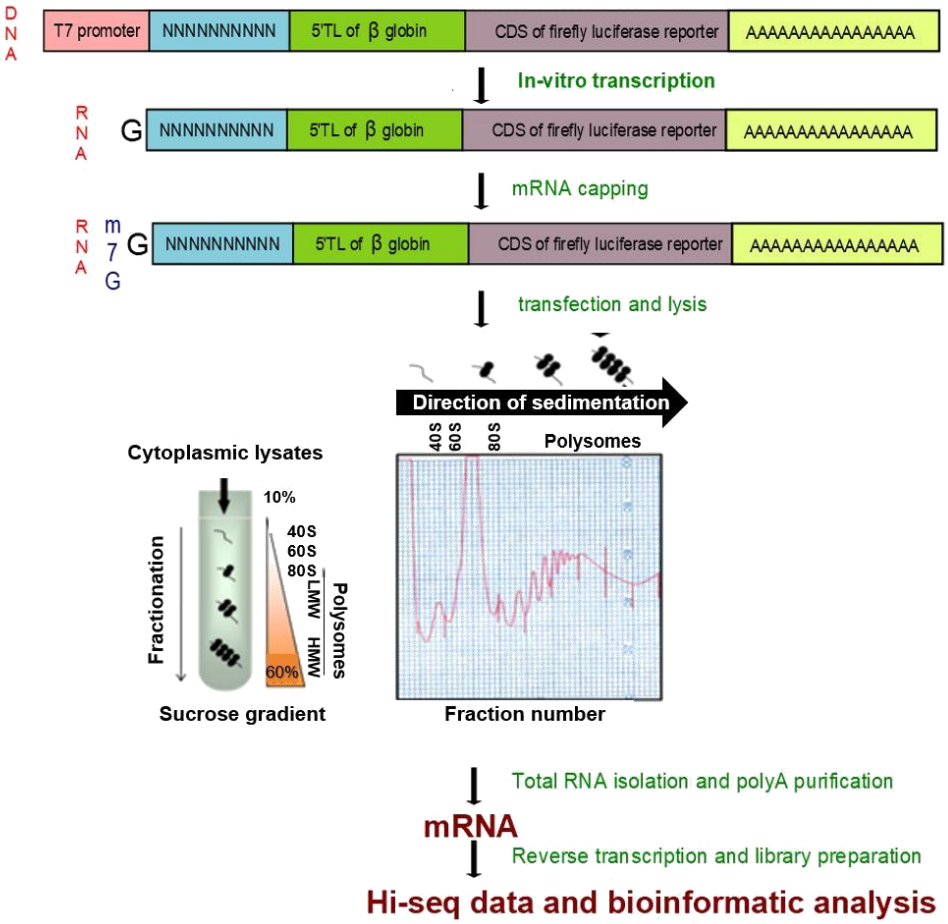


Figure 3.11: Schematic of the experimental protocol used to study the effect of E5S on TIRES

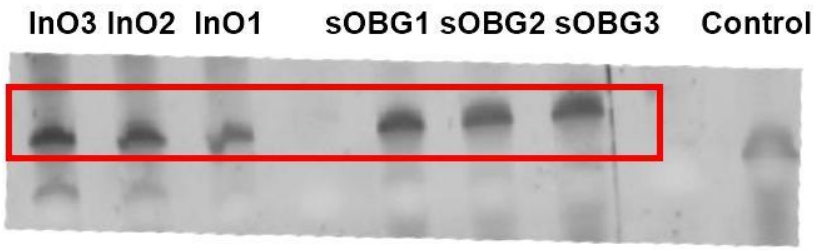
A random oligo is chemically synthesized and incorporated with a firefly luciferase reporter followed by *in vitro* transcription, capping, transfection, polysome isolation, total RNA extraction, polyA mRNA purification, library preparation, deep sequencing and computational analyses as shown in the figure.

After careful consideration and review of the optimisation strategy, an experimental protocol was designed to study the effects of E5S on TIRES (Figure 3.11). The final protocol was as follows – a DNA template was created using a two-step PCR approach, in vitro transcribed and capped enzymatically. RNA quality was assessed at each step using 8% PAGE urea gel. mRNAs were transfected into HEK293T cells in 15cm dishes, lysed, loaded onto a 10-16% sucrose gradient and the polysomal fractions isolated. Total RNA and polyA purification from polysomal fraction resulted in a pool of pure mRNAs. Reverse transcription using a custom UMI

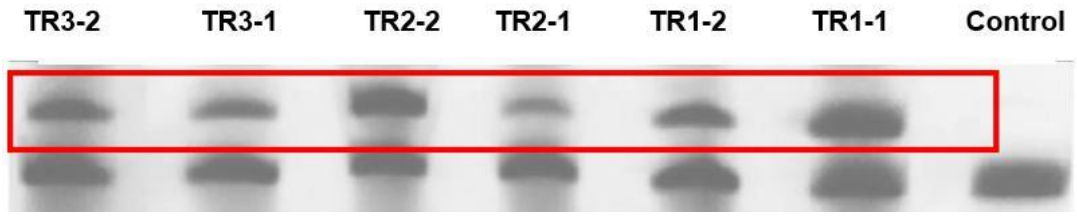
primer, circularisation and amplification steps resulted in a library representative of the actively translating transcript population differing in their E5S.

Following the protocol from Figure 3.11, 18 libraries were prepared as discussed previously (Results,3.1). The RT products used in generating the 18 libraries are shown in Figure 3.12. Figure 3.12a shows the RT samples of controls lnO1,2 and 3 and sOBG1,2 and 3 above the control template (RT primer unextended) (~145 bp) outlined in a red box that was excised and processed further for the creation of lnO and sOBG NGS libraries. In cases where the RT product produced was low in quantity, additional PCR cycles were necessary to increase the yield. However, the increase in cycles could potentially introduce PCR biases. To control for such biases an RT modified primer containing a UMI (unique molecular identifier) sequence was used for TR and PR samples. Figure 3.12b shows the RT products obtained from TR (1,2 and 3 respectively) in duplicates. A clear band was seen boxed in red above the control band comprising the RT modified primer used as a control (unextended) (~150bp). The bands were excised and processed for library preparation. Figure 3.12c shows the RT products obtained from PR (1,2 and 3 respectively) in duplicates. A clear band was seen represented as a red box above the control band comprising the RT modified primer used as a control (unextended). The bands on PR1-1 and PR1-2 were faint but were excised, processed, and successfully used to obtain the desired DNA libraries. The Illumina library preparation step includes the PCR amplification of a DNA template comprising of 8-16 cycles depending on the quantity of cDNA extracted from different libraries shown in Figure 3.12 247.

a)



b)



c)

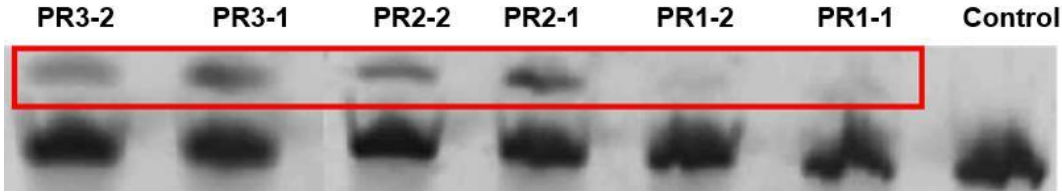


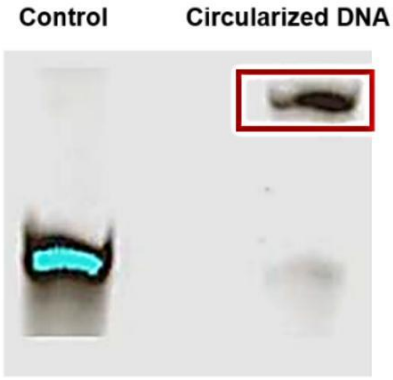
Figure 3.12: Purification of reverse transcription products (RT) of all library samples including short control (sOBG), long control (lnO), total RNA control (TR) and polysome fractions (PR) in duplicates respectively. RNA retrieved from TR and PR samples was reverse transcribed using a UMI primer to generate cDNA. The RT products were visualized in a 7.5% denaturing PAGE urea gel using 2% SYBR gold staining. The unextended RT primer serves as the control and the samples outlined in the red box represent the RT product to be excised for purification.

As the experiments in this thesis relied on low quantities of product, the aim was to optimise each reaction to obtain the maximum amount of final product. Upon successful isolation, each RT product was purified and circularised for library preparation shown in Figure 3.13.

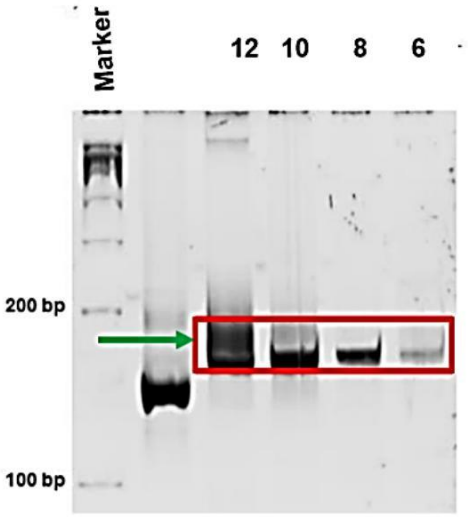
CircLigaseTM was used in the process of circularisation. This enzyme is a thermostable ligase that catalyses the intramolecular ligation (i.e. circularization) of ssDNA templates. In this case, the cDNA produced from the experiment was circularised as seen in Figure 3.13a. Using optimal conditions (Methods, 2.18), it is seen that the amount of circularised DNA was ~>95% of the control DNA.

The PCR amplification step was the final step of the library preparation. It was important to optimise the number of cycles in the PCR for two main reasons a) to obtain optimal DNA quality suitable for sequencing using the HiSeq platform and b) to limit biases by avoiding over amplification of the product. Figure 3.13b shows amplification over different number of PCR cycles. The PCR product visualised on an 8% PAGE gel begins to appear after 6 amplification cycles seen as a faint band. Further, upon reaching 8 amplification cycles, an intense band is seen. However, a smear appears when the PCR amplification is continued to 10 cycles that is intensified upon further amplification at 12 cycles. The smear is indicative of over amplification during the PCR reaction. As shown in Figure 3.13b, 8 amplification cycles were considered optimal and used for next-gen library generation.

a)



b)



c)

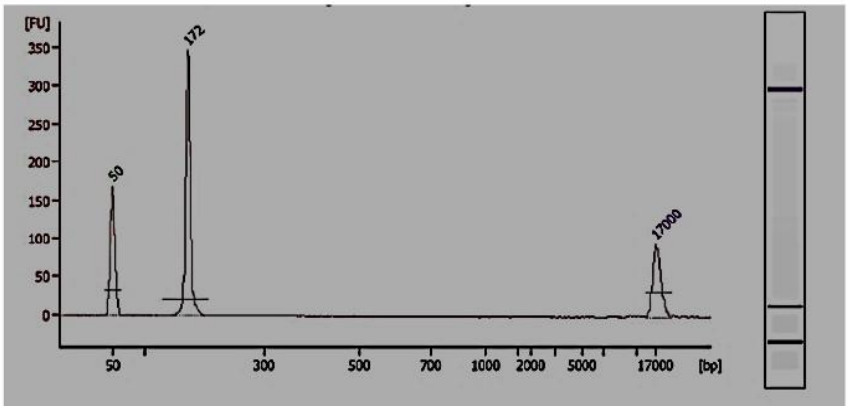


Figure 3.13: Library preparation and quality analysis a) Circularization of purified cDNA visualized on a 15% denaturing PAGE urea gel stained with 2% SYBR gold. The control has the unextended RT primer and the circularized product is shown in a red box. b) Purification of PCR products. The red box indicates the ~171-nt band that was gel excised and purified for deep sequencing. The control indicates the ~151-nt background band derived from unextended RT primer was avoided during gel excision. The green arrow indicates the partial duplexes resulting from reannealing as the PCR amplification approaches saturation. c) BioAnalyzer profile of a high-quality sequencing library of sample TR1-1. A single 172-nt peak is present (the peak at 35 and 17000 are the vendor's internal standard, present in all profiles).

Analysis using an Agilent Bioanalyzer 2000 was used to confirm the quality of the DNA libraries. The Bioanalyzer is used to quantify yield and detect artefacts post-PCR amplification. Bioanalyzer profiles are used regularly in the assessment of library quantity and quality. Figure 3.13c shows a bioanalyzer profile of a library of TR1-1 produced from lnO1 RNA. A single 172-nt peak is visible (the peak at 35bp and 17000bp are the vendor's internal DNA standard, present in all profiles). This profile denotes a high-quality DNA sequencing library comprising of a population of 172bp DNA fragments.

Table 3.2 : Bioanalyzer analyses of libraries qualified for deep sequencing.

Library Name	Fragment size	Concentration (ng/µl)	Molar concentration (nmol/l)
sOBG1	172	16.74	152.32
sOBG2	177	10.51	87.88
sOBG3	163	1.58	18.09
lnO1	174	12.64	97.31
lnO2	171	5.60	54.66
lnO3	171	7.92	93.12
TR1-1	172	1.53	6.7
TR1-2	181	8.87	80.95
TR2-1	177	8.67	68.4
TR2-2	171	4.63	23.91
TR3-1	179	11.05	113.96
TR3-2	178	13.07	82.84
PR1-1	154	5.63	75.90
PR1-2	176	20.81	49.68
PR2-1	177	6.75	18.09
PR2-2	174	7.45	20.74
PR3-1	173	13.09	38.86
PR3-2	175	20.32	16.88

Note: One of the libraries, sOBG2, was misplaced by the vendor and library generation had to be repeated. While processing this library, the RT reaction was performed with an incorrect primer, RT_primer_modified instead of an RT_primer. This was a technical error which was only observed later. Although the results are not affected by this, it added an additional step of removing PCR duplicates from the data available from the library.

Table 3.2 shows the molar concentrations of DNA in the final libraries. Most of the samples have the desired fragment size of >165nt. All samples produced a single peak indicating a high-quality sequencing library. The desired molar concentrations recommended by the vendor states that each library must have a minimum concentration of 2-4nm/l. All the libraries generated met the desired criteria for the BGI HiSeq platform.

3.8 Massively parallel sequencing shows a high percentage of inclusion of all possible variants in the libraries

Eighteen libraries were sequenced with Illumina HiSeq3000 instrument by a commercial provider (BGI, China). Following successful library generation and quality assessment, the libraries were sequenced in two lanes containing nine libraries in each lane. The typical throughput of HiSeq3000 lane is ~300 million reads. The number of reads expected per library is shown below:

Expected number of reads per lane using Hi-Seq-3000 platform = 300 million
Number of libraries multiplexed per lane =9
Number of reads expected per library = 33.33 million
Number of unique variants expected in a single library = 410 = 1.04 million
The number of times each unique read occurs in a library is calculated as:
The number of reads expected per library / Number of unique variants expected in a single library = 33.33 million/1.04 million = 32

Upon parallel sequencing, data containing the total number of clean reads with their quality scores from the two different sequencing lanes respectively is shown in Appendix Table 2 (a and b). The total number of reads obtained from the two sequencing lanes was 298.2 million and 264.7 million respectively. These numbers were slightly lower than the expected 300 million reads/lane.

The highest number of unique variants possible in each library is 410 i.e. 1048576 representing a complete library (100%). Library completeness is calculated as the

number of unique reads obtained divided by the maximum number of reads possible in a single library (1048576). The library completeness of each library is shown in Table 3.3. All TR and PR libraries comprised of >95% of possible unique variants except for libraries PR2-2 and PR1-1.

The FASTQC tool was used to assess the quality of reads obtained from all libraries. Base calling is the process by which raw data from the sequencing instrument is converted to nucleotide sequences 248. Base calling accuracy is typically measured by a Q score (Phred quality score) and is a common metric to assess the accuracy of a sequencing run. Q scores are defined as logarithms of base calling error probability[248].

Table 3.3: Evaluation of the completeness of libraries

RNA library	No. reads (millions)	No. permutations (millions)	% completeness observed
sOBG1	26.4	1001470	95.5
sOBG2	28.5	1012639	96.5
sOBG3	20.6	1000039	95.3
lnO1	31.4	1021488	97.4
lnO2	35.7	1022746	97.4
lnO3	27.6	1019764	97.25
TR1-1	45	1041306	99.30668
TR1-2	8	1005306	95.87345
TR2-1	23	1031308	98.3532
TR2-2	31	1033711	98.58236
TR3-1	29	1032177	98.43607
TR3-2	30	1032967	98.51141
PR1-1	7.8	982937	93.74018
PR1-2	23	1028695	98.104
PR2-1	19	1027989	98.03667
PR2-2	1.47	683174	65.15255
PR3-1	23	1034968	98.70224
PR3-2	29	1034968	98.70224

The sequences produced with Illumina include the adapter sequence/identifier sequence. To remove the adapter sequence (ACAACTGTGTTCACTAGCAACCTCA) from all datasets, the Cutadapt tool 208 was used. The read lengths post Cutadapt clipping is shown for 18 libraries in Figure 3.16. The libraries sOBG (1 and 3) and lnO (1-3) had high concentration of RNA available for RT, therefore an RT_primer without a UMI was used in the generation of these libraries. The expected lengths of the processed reads upon adapter trimming are 11nts. The samples indexed TR and PR were processed for library generation using a UMI sequence (RT_primer _modified) due to their low concentrations after

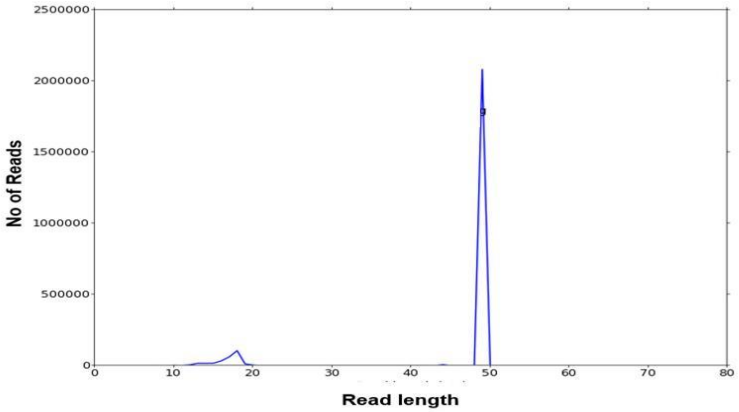
downstream processing. The expected lengths of the processed reads upon adapter trimming are 18nts. In addition, sOBG2 was reverse transcribed using RT_primer_modified as mentioned previously.

Cutadapt was successful in clipping most of the libraries which produced reads with expected length or with one additional nucleotide. However, Cutadapt failed to successful clip three datasets (Figure 3.16 (b and c)) owing to poor sequence quality at the 3' ends of the read. To overcome these issues, a set of criteria was set for adapter trimming of these libraries outlined in methods (Methods, 2.23).

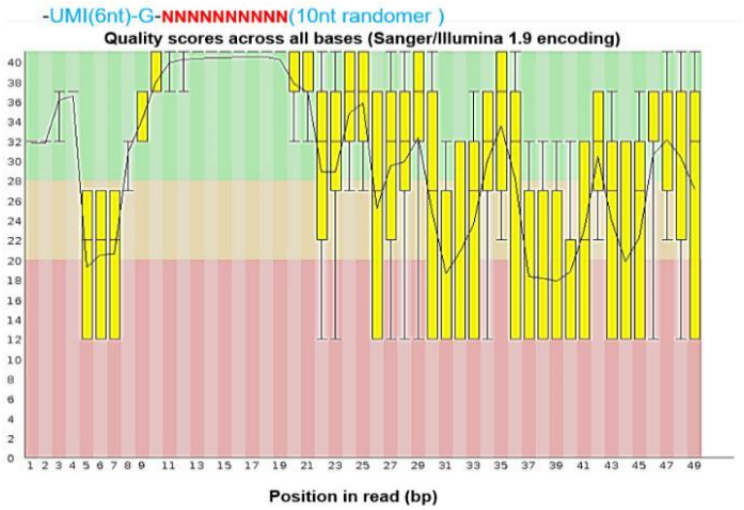
However, in a few cases, the clipped reads also had a significant number of reads that were one nucleotide longer than expected i.e. 12 or 19 nucleotides in length. By examining the nucleotide frequency at each position of read length being 18 or 19nt, the extra nucleotide was found to be added immediately 5' of the +1G position. The major fraction of this extra nucleotide was G and a smaller fraction was T. The extra nucleotide is likely caused by a non-templated nucleotide addition during library generation 249, thus in a subsequent analysis, this additional nucleotide was excluded since it did not reflect the actual sequence of mRNA at the 5’ end. It was observed that the median number of occurrences of each unique variant in the libraries indexed TR and PR varied between 6 and 37.

Unsuccessful adapter trimming of PR2-2 reads can be seen in Figure 3.14 (a). However, it was found that the quality of the reads was lower in the region flanking E5S comprising of sequence GNNNNNNNNNN. Figure3.14 b shows that the desired E5S region had a quality score of >40. As a result, PR2-2 reads were then filtered using an alternative method (Methods, 2.23) to obtain the desired, higher quality reads along the E5S for the PR2-2 library. A plot of read position against the number of reads containing a specific nucleotide in a position shows that all four nucleotides are represented at similar frequencies in the E5S region for library PR2-2 (shown as NNNNNNNNNN in Figure 3.14c). Successful adapter clipping was observed in 15 out of 18 libraries sequenced.

a)


No of Reads
Read length

b)


-UMI(6nt)-G-NNNNNNNNNN(10nt randomer)
Quality scores across all bases (Sanger/Illumina 1.9 encoding)
Position in read (bp)

c)

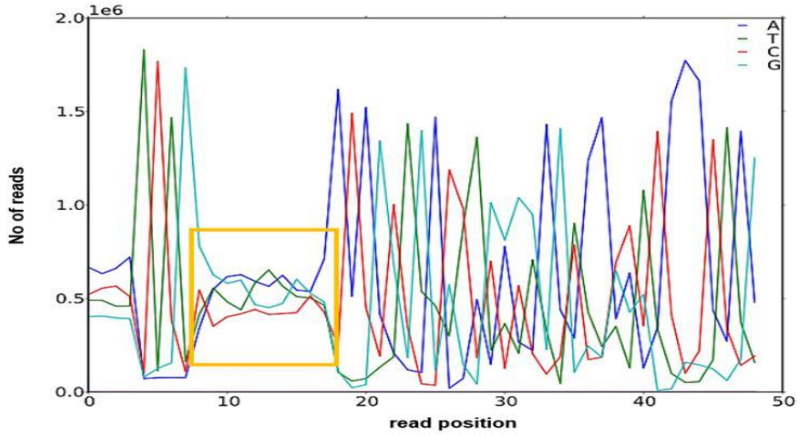

1e6
A
T
C
G
No of reads
read position

Figure 3.14: High quality of base calling in sample PR2-2 observed along the E5S a) Adapter clipping using Cutadapt was unsucessful in sample PR2-2. The read length of the majority of reads was 49nt instead of the expected 18nt after adapter clipping. b) FastQC quality along the E5S was >40 c) Frequency of each nucleotide at different coordinates of sequencing reads. The orange square correspinds to the region of E5C where frequencies of individual nucleotides where Nϵ(A,C,T,G) are shown.

The reads from libraries indexed sOBG/lnO vs TR/PR were processed differently based on the UMI index and adapter clipping as shown in Figure 3.15.

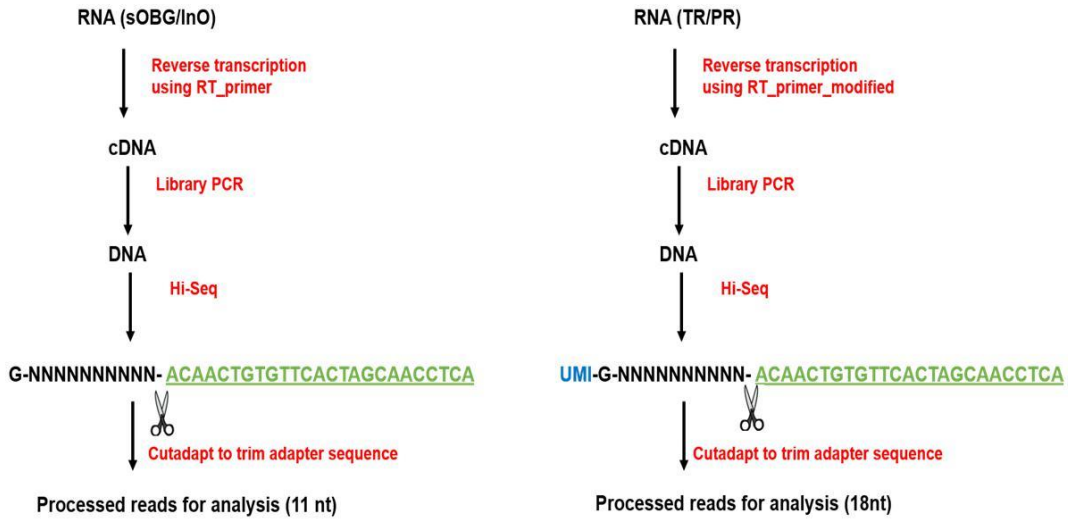
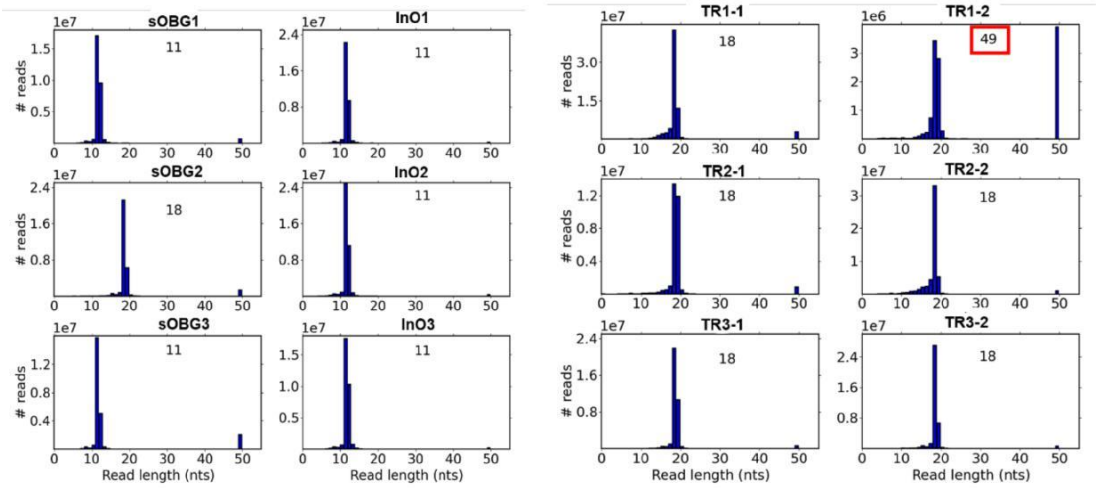


Figure 3.15: Processing reads in different E5S libraries. The adapter was trimmed from raw reads and processed for further analyses. The reads that were not clipped using Cutadapt were clipped using a set criterion (Methods, 2.23).

a) **b)**



c)

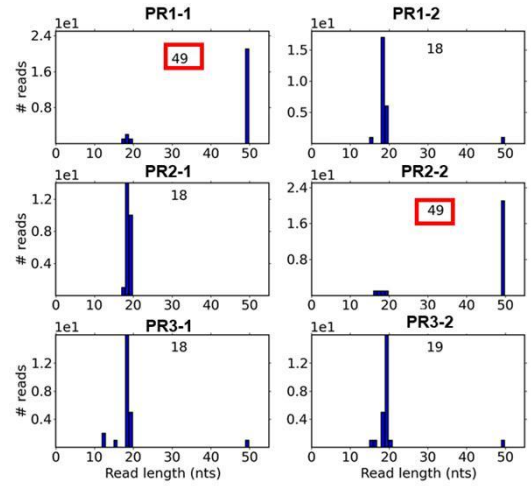


Figure 3.16: Read lengths of the libraries post adapter trimming. The identifier sequence was trimmed from the raw sequences of the libraries and the read lengths were calculated and plotted for samples indexed a) sOBG and lnO, b) TR and c) PR. The 49nt represented by red squares for libraries TR1-2, PR1-1 and PR2-2. indicates unsuccessful adapter clipping which was processed using an alternative method (Methods,2.23).

Preliminary data analysis of library samples

3.9 Unique Molecular Identifier (UMI) correction removes PCR duplicates from the library

UMI are random sequences of bases used to tag each molecule (fragment) prior to library amplification thereby aiding in the identification of PCR duplicates 250 (Appendix Figure 4). UMIs were used in libraries indexed TR and PR (12 libraries in all) due to the low amounts of starting material in them. Such samples typically require additional cycles of PCR making them prone to PCR duplications. Figure 3.17 shows a plot of the number of unique variants in the libraries plotted against the number of occurrences of each read in the dataset. The comparison of UMI corrected samples vs. the total number of reads in all samples showed no significant differences.

Calculation of percentage error following UMI correction is shown in Table 3.4. The PCR duplication biases contributed an average error of 11% across 12 libraries under consideration. Following UMI correction, the highest percentage of PCR duplicates were found to be 20.6% in the library TR1-1 and the lowest percentage of 1.5% in the library PR2-2. This corresponds to the total number of reads that occurred in each library before UMI correction (Table 3.4), where TR1-1 had the highest (56 million) and PR2-2 had the least (1.5 million) number of reads.

The quality of libraries indexed PR2-1 and PR2-2 using Agilent Bioanalyzer showed that they were high-quality, single-fragment DNA libraries without any DNA contamination. On written communication with BGI, it was confirmed that the sequencing depth for sample PR2-2 was compromised due to unknown reasons. However, on close inspection of the total number of reads, it is observed that for samples generated from lnO3 (TR3-1, TR3-2, PR3-1, and PR3-2) the total number of reads remained consistent as expected even upon UMI correction.

Table 3.4: Calculation of percentage error following UMI correction

Sample name	Total number of reads	Number of reads post UMI correction	% of reads removed after UMI correction
TR1-1	56749003	45012166	20.68201445
TR1-2	8834780	8305027	5.996221751
TR2-1	25618339	23336172	8.908333206
TR2-2	38663009	31967374	17.3179356
TR3-1	32926495	29771654	9.581466233
TR3-2	34115498	30234230	11.37684697
PR1-1	8589484	7800876	9.181087013
PR1-2	25655168	22829261	11.01496198
PR2-1	21928544	19006669	13.32452807
PR2-2	1501788	1479325	1.495750399
PR3-1	26028323	23411900	10.05221504
PR3-2	34688810	29131045	16.02178051

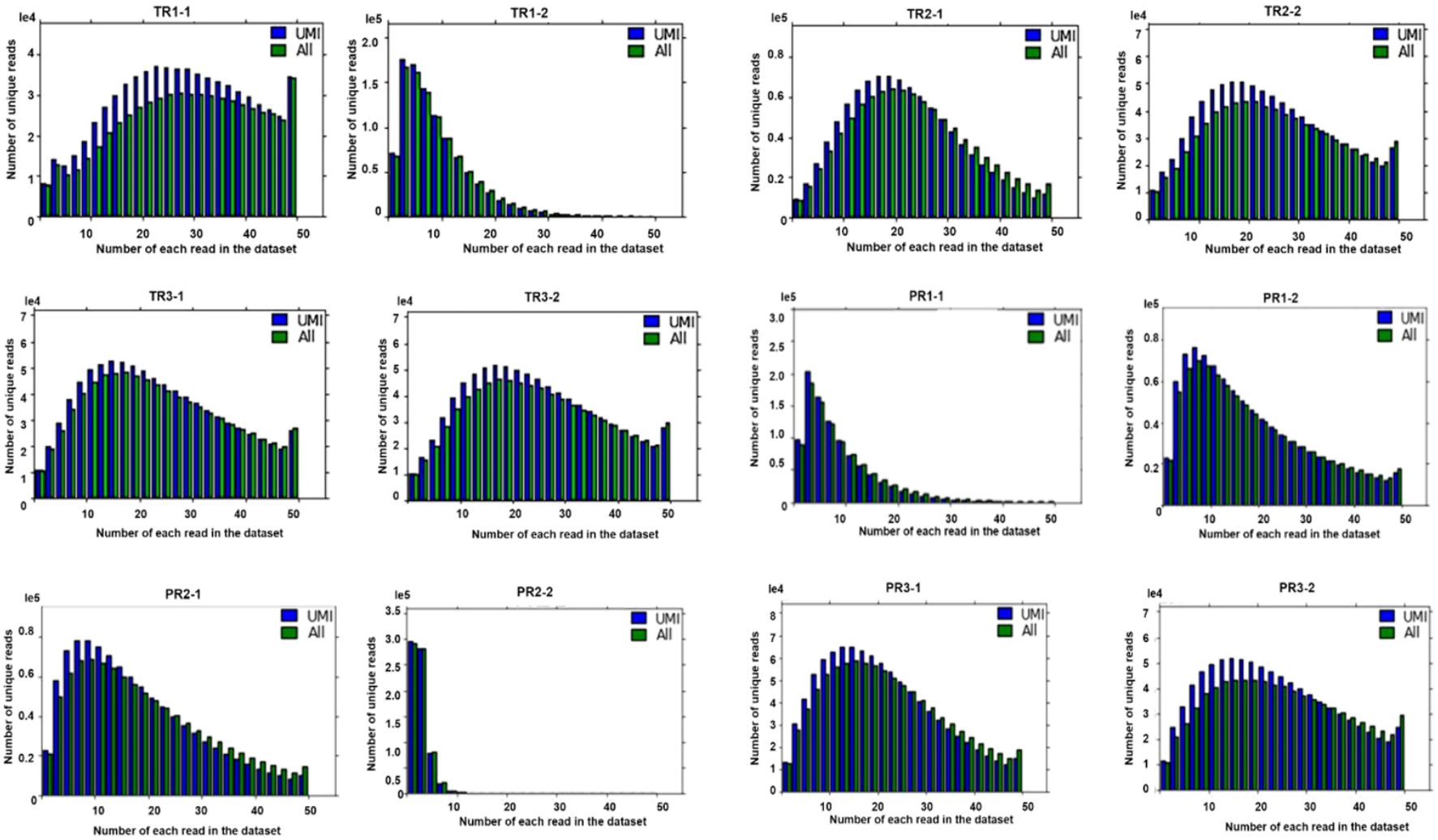

TR1-1
TR1-2
TR2-1
TR2-2
TR3-1
TR3-2
PR1-1
PR1-2
PR2-1
PR2-2
PR3-1
PR3-2
UMI
All
Number of unique reads
Number of each read in the dataset

Figure 3.17: The effect of UMI correction on samples indexed TR and PR. The plot contains the number of unique reads plotted against the number of occurrences of each read in the dataset before (green) and after (blue) UMI correction.

3.10 Verifying the technical reproducibility between samples

The use of massively parallel sequencing enables quantification of many variants in a single experiment, in this case, 410 variants (assuming 100% completeness of original randomer library). Measuring reproducibility between duplicates is a critical component in assessing the quality of data obtained from these experiments.

To study the reproducibility between duplicates, the ratio of the number of reads sequenced from the polysomal selected to total RNA (PR/TR) was determined along the E5S inclusive of shorter permutations. The shorter unique variants were calculated upon aggregating read counts containing the same sequence motif beginning at the 5' end of the E5C (Methods,2,24).

It was observed that the reproducibility of PR/TR values along the length of E5C was weak. Spearman's ranking correlation was used to rank duplicates based on their similarity in PR/TR values. Spearman's correlation coefficient, (ρ) measures the strength and direction of the association between two ranked variables. It may be observed that the pairwise Spearman's correlation increases as the length of the E5S are reduced. However, the average Spearman's correlation of the pairwise comparison for all library combinations was 0.31. These pairwise correlations improved significantly upon consideration of shorter permutations (the average correlation of nucleotide stretches of 7nt in length was 0.61) seen in Figure 3.18.

The grey lines in Figure 3.18 represent duplicates generated from the same lnO. It is observed that (1-1 Vs 1-2) and (3-1 Vs 3-2) had a high Spearman correlation value of >0.80 for E5C of 8nt length. Interestingly, the libraries generated from lnO2 and those lnO1 (1-1 Vs 2 and 1-2 Vs 2 respectively) had a Spearman's correlation of >0.70 for E5S of 8nt length represented by blue lines in Figure 3.18.

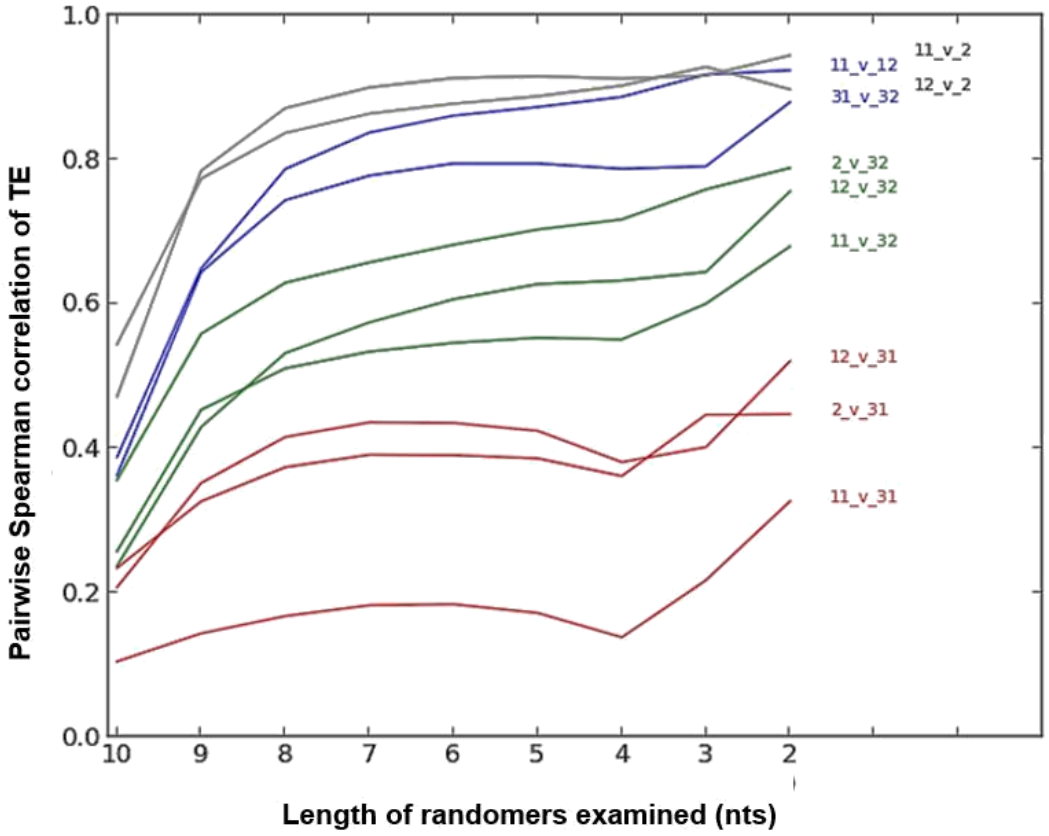


Figure 3.18: Technical replicates produced from lnO1(1-1, 1-2) and 3 (3-1,3-2) are highly reproducible. Spearman's ranking correlation was used to rank duplicates to calculate their similarities. It was seen that the duplicates obtained from the same lnO (1-1 vs 1-2) and (3-1 vs 3-2) had a high-ranking correlation of >0.8.

3.11 Influence of E5S on TIRES

The most important aim of this work was to examine the effect of E5S on TIRES. The three lnOs experiments were performed in duplicates. Due to the low coverage observed in PR and TR libraries generated from lnO2, the libraries were combined to form a single library referred to as 2 hereon.

The TIRES ratio of every motif present in the five samples (1-1, 1-2, 2, 3-1 and 3-2 respectively) was calculated using the formula:

$$I_{jk} = \frac{P_{jk}\sum_{i \in J} T_{ik}}{T_{jk}\sum_{i \in J} P_{ik}} \quad \{1\}$$

Where Ijk is TIRES of an N-nucleotide long variant j from the set of 4n random variants. J calculated for the data obtained in the sample k (1-1, 1-2, 2, 3-1 or 3-2). P and T are the number of reads from PR and TR libraries, respectively. The maximum effect of $TIRES_G$ on E5S was seen at N=8nt and is maintained consistently in this work unless mentioned otherwise.

Figure 3.19 shows the TIRES effect of individual nucleotides along the E5S of the five libraries: 1-1,1-2,2,3-1 and 3-2 respectively. An equiprobable frequency model was used to predict the influence of each nucleotide along E5S based on its effects on TIRES. There was a lack of consistency in the preference of a specific nucleotide at any specific position across E5C to influence TIRES across the 5 libraries under consideration (Figure 3.19(a-e)). The nucleotide at position 2 had a strong influence on TIRES. In samples 1-1, 1-2 and 2 at position 2, G was found to have the highest enrichment and U was depleted. However, in samples 3-1 and 3-2, A was found to have the highest enrichment while U remained depleted at position 2. These differences in patterns could have arisen due to multiple factors including differences in the chemical synthesis of sO, differences in lnO used in each library and differences in technical reproducibility across samples including sequencing reads obtained from the libraries.

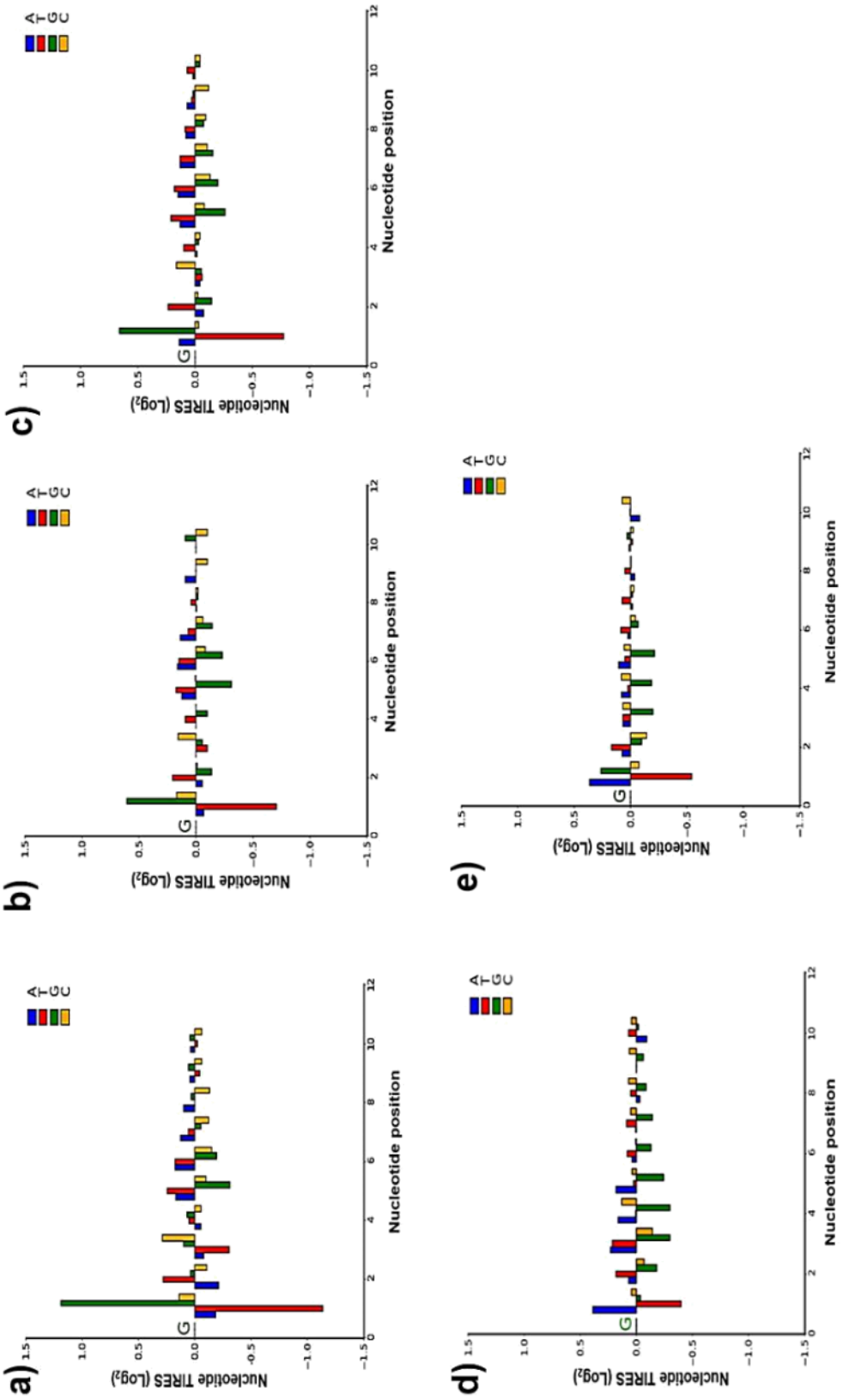


Figure 3.19: Analysis of sequence context preference for TIRES based on the E5S observed in individual samples a) 1-1 b) 1-2 c) 2 d) 3-1 and e) 3-2 respectively.

$TIRES_G$ was computed as the geometric mean of TIRES ratio of every motif present in the five libraries 1-1, 1-2, 2, 3-1 and 3-2 respectively. The geometric mean was used instead of the arithmetic mean to reduce the large differences in values between samples that could otherwise cause a disproportionate influence on the result. The measurement of $TIRES_G$ included certain considerations: a) $TIRES_G$ was only measured for unique variants which were observed in all libraries (1-1,1-2,3-1 and 3-2 respectively), b) if a unique variant was not present in any library and c) if TIRES value was computed to be 0 in any of the libraries, the reads were discarded.

Due to the presence of random nucleotides along E5S, we expect the nucleotide in each position to have an equal probability of occurrence. It is expected that the frequency of each of the four possible outcomes (A, C, T and G) to be 0.25. For example, if it was found that the real frequency of a nucleotide in position 2 is 0.05 The observed to expected ratio would be 0.05/0.25= 0.2 (observed value is five times lesser than expected value). If the observed frequency is the same as the expected frequency of 0.25, the observed/expected ratio will be 1 (baseline of plot 3.20a). Based on the model above, y= log2($TIRES_G$). In Figure 3.20(a) U, has a nucleotide $TIRES_G$ log2(-0.7) which is ~0.61. Therefore, U occurs almost 60% less frequent than what would be expected at position 2. The sequences in the y axis with a value less than ±0.05 had less than ~4% deviation from the expected value and this was considered an arbitrary threshold for significance of the effects of E5S on $TIRES_G$. Based on this consideration, the effect of E5S on $TIRES_G$ was insignificant along positions 9-11 (Figure 3,20a).

Sequence logos were generated (Methods, 2.28) using WebLogo 251 for the top (10%) and bottom (10%) candidates based on their $TIRES_G$ values. Each logo consists of stacks of symbols, one stack for each position in the sequence. The height of symbols within the stack indicates the relative frequency of each nucleotide at that position. Figure 3.20b shows that unique variants containing the highest $TIRES_G$ values (top 10% of 16386 unique variants). Based on the information theory, A/U in position 3 and G in position 2 was enriched. Similarly, unique variants containing the lowest $TIRES_G$ values (bottom 10% of 16386 unique variants) had an enrichment for U in position 2-5 and G in position 6.

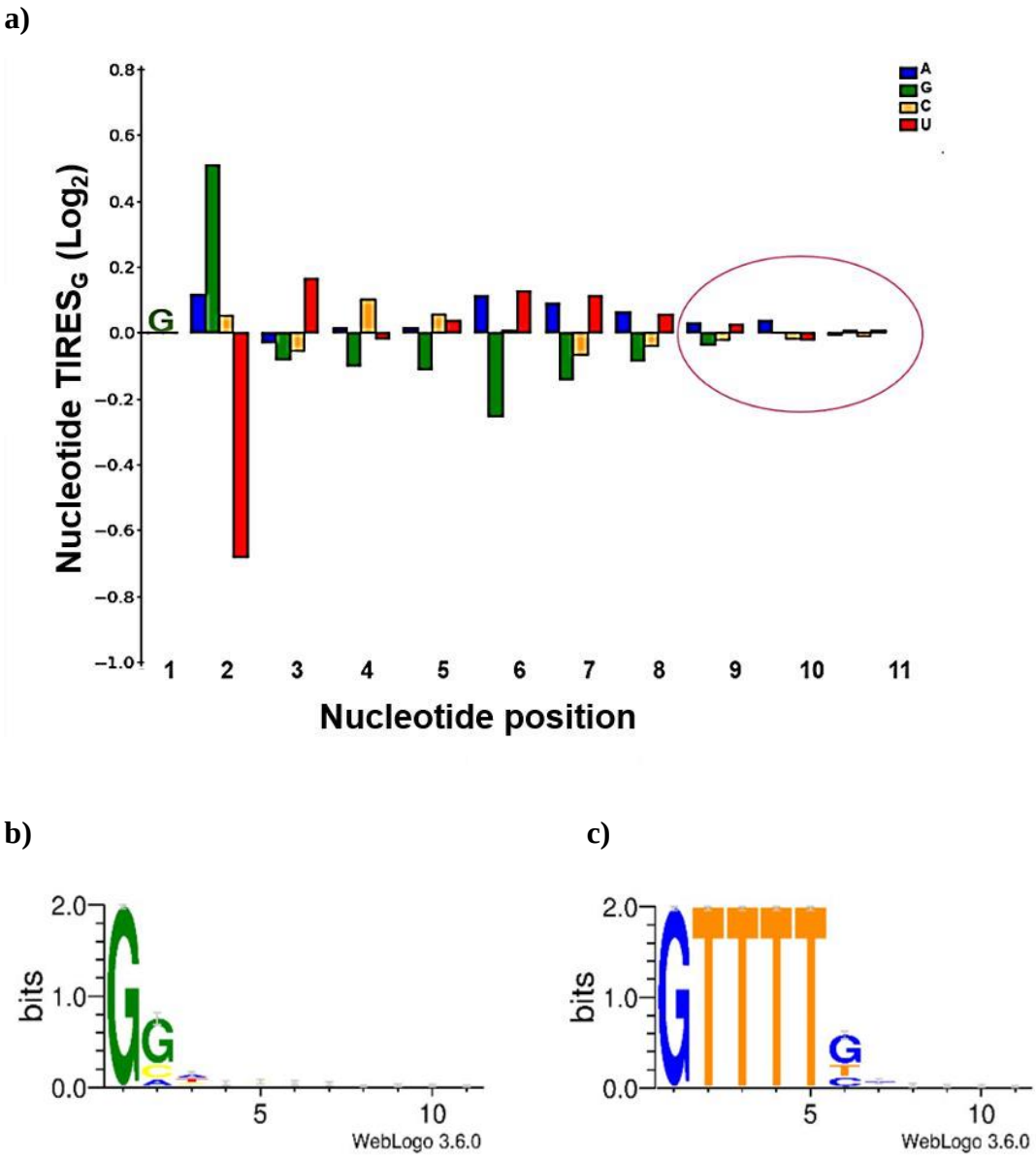


Figure 3.20: Sequence context preference for translation initiation in the E5S. a) TIRES was calculated (Methods, 2.24) per nucleotide for libraries 1-1,1-2,2,3-1 and 3-2 respectively along positions 2-11 of E5S along with +1G b) WebLogo of 1000 out of 1048286 unique variants having highest TIRES values along the E5S and c) WebLogo of 1000 sequences out of 1048286 unique variants having lowest TIRES values along E5S. WebLogo was created using the online WebLogo3 software.

3.12 E5S position 2 has an influence on TIRES values

As seen in Figure 3.20a, the second nucleotide position of the E5S appears to have a greater influence on $TIRES_G$ in comparison to nucleotides at positions 3-11. To measure the strength of the nucleotide context of nucleotide position 2 in E5S influencing $TIRES_G$, scatter plots were used. Scatter plots are used to identify potential associations between any two datasets 252. An upward trend of the plot is indicative of positive association and a downward trend represents the negative association. A correlation coefficient evaluates the existence of a linear relationship between two samples under consideration 253.

Appendix Figure 6 compares the relationship between technical replicates using lnO1 and 3 respectively (samples 1-1 vs 1-2 and 3-1 vs 3-2). A strong linear relationship was observed between the technical duplicates 1-1 vs 1-2 and 3-1 vs 3-2 with a correlation coefficient of 0.837 and 0.774 respectively indicating the similarity in measures at position 2 of E5S on TIRES value in the respective library.

The scatter plots along other positions were measured showed similar measures between duplicates (data not shown), the effect of the second nucleotide is shown as it had the most significant effect along the E5S. The cooperative effects of adjacent nucleotides along the E5S were calculated but did not have a significant influence on the TIRES. However, the nucleotides with positive cooperative effect on E5S is listed in Appendix Table 3.

3.13 Validation of the effect of E5S on translation in HEK293T cells using reporter assays

Figure 3.20 (a) indicates that specific nucleotide positions in the E5S can influence the $TIRES_G$. To validate these findings, candidates (n=4) with the highest and lowest TIRES values were selected. $TIRES_G$ values are representative of the cumulative effect of TIRES on all five libraries. Upon candidate selection, the NNNNNNNNN of the lnO template was replaced by the E5S context of the candidate listed in Figure 3.21 (High denoted as H and low as L). However, the candidate constructs were generated using a similar approach as that of the lnO template (Figure 3.6). These constructs were transcribed in-vitro, capped, and transfected into HEK293T cells for 2 hours after which their luciferase values were measured (Methods,2.20). It was expected that if E5S had a large influence on $TIRES_G$, the reporter assay would show substantial differences between the selected candidates depending on their E5S context. The box plots in Figure 3.21 illustrate there were no obvious differences in the luciferase activities of high and low TE candidates. These experiments were repeated using HEK cell free lysates using the same candidates, and no significant differences were observed in the luciferase activities of high and low $TIRES_G$ candidates (data not shown).

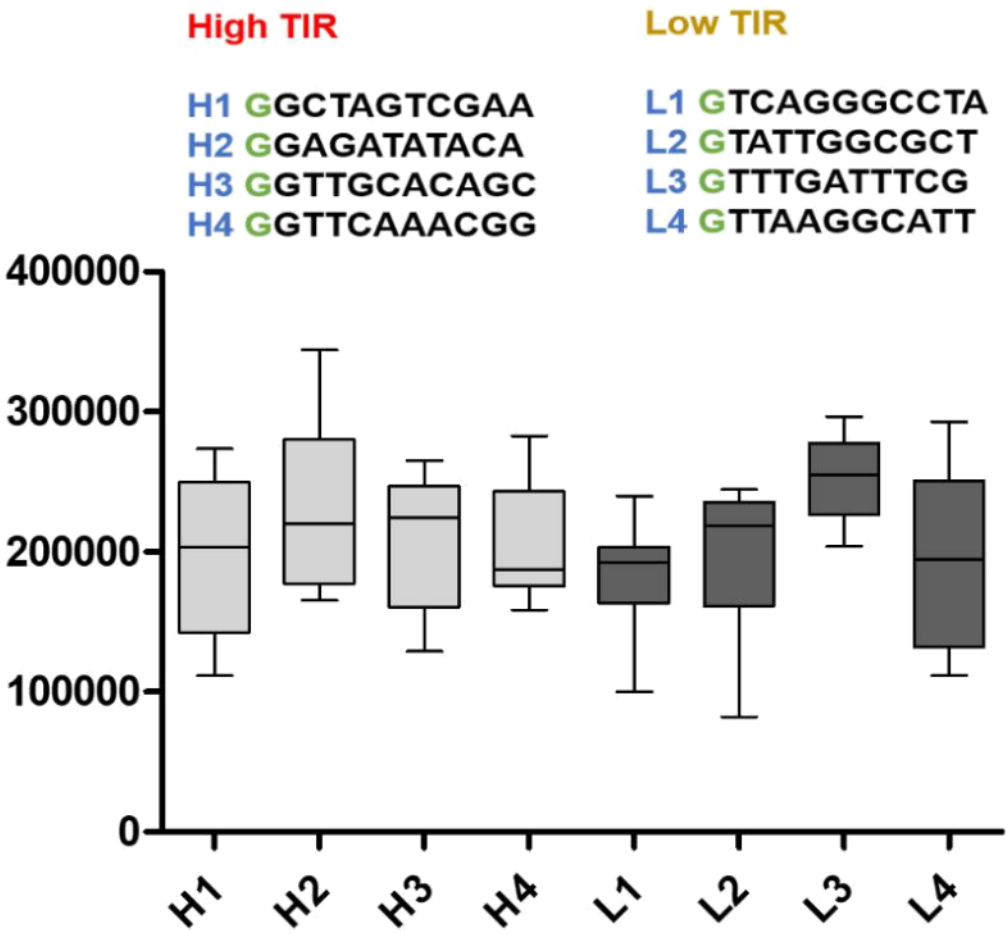


Figure 3.21: Validation of the effects of specific E5S context on TIRES. High and low TIRES candidates were selected based on their E5S context. These candidates were inserted into the lnO DNA template replacing the NNNNNNNNNN region with the desired E5S context (high and low). The candidate constructs were transcribed in-vitro, capped, and transfected in HEK293T cells for two hours after which firefly luciferase reporter values were measured and plotted for all candidates.

3.14 Percentage of GC in E5S influences translational efficiency

In mammalian cells, the dependency of translational efficiency on thermal stability, location and GC content of mRNA hairpin structures in the 5'TL have been previously evaluated 254 (Bebendure, 2006). It was found that the presence of high GC content and a stable secondary structure in the 5'TL had a negative impact on the translation efficiency in mammalian cells. In this work, experiments were performed in HEK293T cells, thus the potential variation of $TIRES_G$ values according to the GC content along the E5S is of interest. The GC percentage of each individual variant in the E5S was plotted against its respective $TIRES_G$ value. It was observed that the increase in GC content in the E5S causes a decrease in the observed $TIRES_G$ values (Figure 3.22). This is consistent with previous reports that high GC content observed in 5'TL can cause translation repression 255. An example of secondary structure prediction in model RNAs containing different distributions of GC% (E5S+ 5'TL of β-globin as in lnO) is shown in Appendix Figure 9.

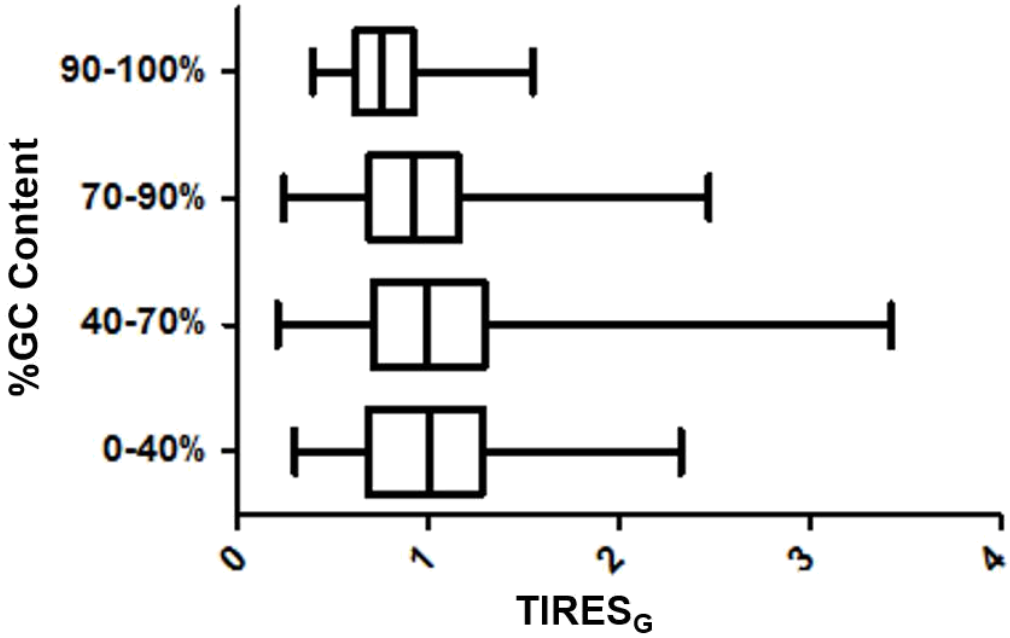


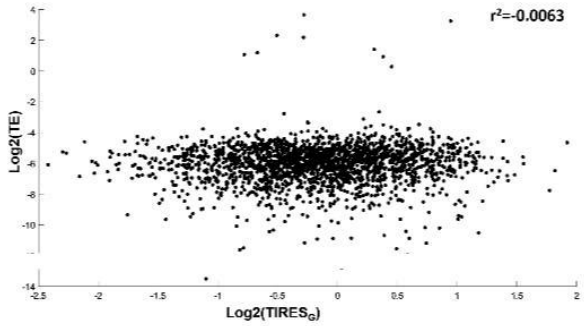
Figure 3.22: Percentage of GC content in the E5S affects TIRES$_G$ in HEK293T cells. The percent GC content increases from top to bottom in the figure. The average mean TIRES$_G$ value for each range of %GC content represented in the figure was calculated. Error bars represent the standard error of the mean of the fields indicated across different ranges of GC% calculated against their TIRES$_G$ values.

3.15 Comparison of translation efficiencies between artificially designed and naturally occurring mRNAs

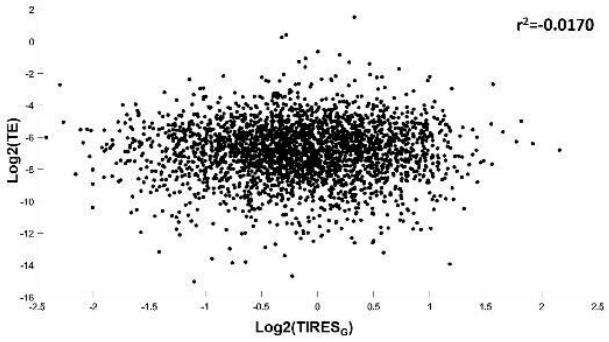
If there is a certain binding preference of a specific initiation factor for a specific E5S context selected for high $TIRES_G$, it should be maintained across all endogenous mRNAs containing the specific E5S sequence. To investigate this, the published transcriptional start site 256 and ribosome footprinting 68,210 datasets obtained from HEK cells were compared to E5S candidates with high and low $TIRES_G$ values. Importantly, only endogenous 5'TLs beginning with a guanine were examined, as this feature was a constant in all E5S libraries used in this work.

No overall correlation between the ribosome occupancy in the footprinting datasets and the $TIRES_G$ values obtained from the E5S library was observed as seen in Figure 3.23 (a and b). Information theory was used to measure the information content of 10% of the top and bottom candidates ranked by their TE values using the Andreev and Sidrauski ribosome footprinting datasets using Sequence Logos 68,210. The sequence preference was observed in the E5S region for the high and low TE candidates obtained from ribosome footprint studies differed between the ribosome profiling datasets. When the data available for all candidates ranked by their TE was analysed using information theory, sequence logos indicated similar patterns as seen in Figure 3.23 (b, c, and e, f). While the Andreev et al footprint study had top candidates with a significance of G and bottom candidates with U in position 2 which is coherent with our observations in this work. These observations were not consistent with the Sidrauski study that is expected.

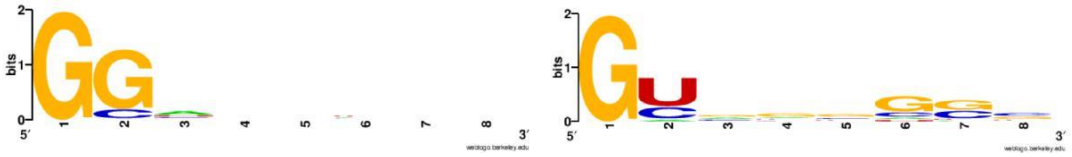
a)



b)



c) **d)**



e) **f)**

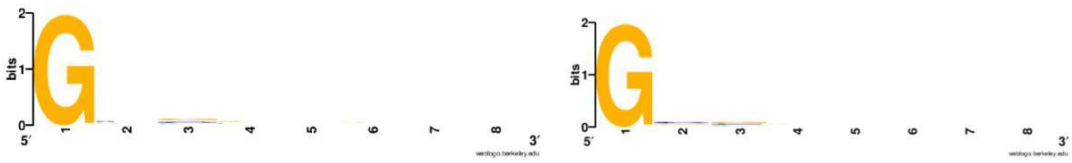


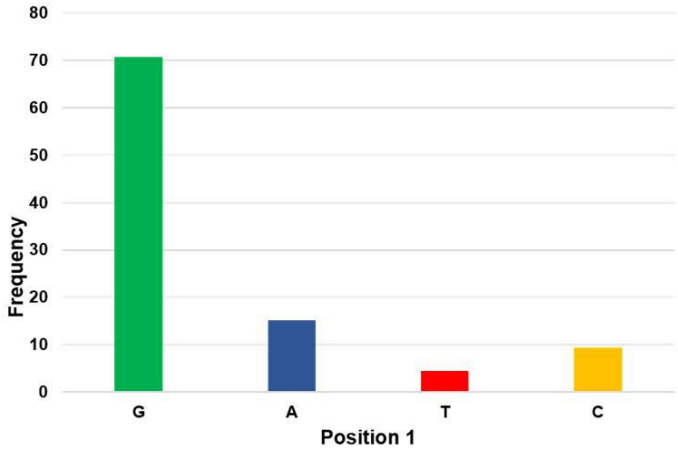
Figure 3.23: E5S identity does not correlate with mRNA translation in HEK293T cells. Ribosome occupancy of endogenous mRNAs that begin with a +1 guanine does not correlate with the $TIRES_G$ values obtained for E5S. The log2 $TIRES_G$ for each E5SC was plotted against the log2 ribosome occupancy of endogenous mRNAs from two different datasets. a) Andreev et al, b) Sidrauski et al. The TE values were calculated for all mRNAs from the two ribosome footprinting studies. Top and bottom 10% TE value candidates were chosen to study their motif preferences along the region of E5S using WebLogo. The high and low TE value nucleotide preferences for top 10% of endogenous mRNAs from two datasets c and d) Top candidates from Andreev et al and Sidrauski et al, e, and f) bottom candidates from Andreev et al and Sidrauski et al.

3.16 Sequence preference in E5S of mRNA isolated from polysomes in MCF7 cells

It is well known that certain features (e.g. uORFs, secondary structures, G quadruplexes etc) in the 5'TL can influence translation 8. To understand the effects of the 5'TL on TIRES, it is important to define a 5'TL of a mRNA containing accurate TSS information.

There are no resources which provide TSS information for commonly used cell lines. To understand the precise relationship between the 5'TL features and translational control, it would be ideal to have both TSS and translation efficiency determined in the same cell line. Polysome profiling is a standard technique to study translatomes. Gandin et al studied the mTOR sensitive mRNAs (>5000 mRNAs) in MCF7 cells for their TSS and translational efficiency. While the TSS information in this study was assessed using NanoCAGE, the translatome of the desired cell line was studied using polysome profiling. Data extracted from the mRNA using NanoCAGE analysis and polysome fractions in the mock condition (control) are illustrated in figure 3.24.Figure 3.24 shows the frequency of the TSS in the RNA isolated from the control conditions of MCF7 cells in highly translating polysomes (ribosome >=3) for 6551 genes. It is observed that most efficiently translated mRNA prefer G in their first and second positions. This converges with the observations made in this work where there is a preference for G in position 2 in actively translating cells containing a +1G along the E5S. The accurate information of TSS and the translatome on MCF7 cells was evaluated using NanoCAGE and polysome profiling. The data obtained from polysomes was tabulated for its sequence preference along various positions.

a)



b)

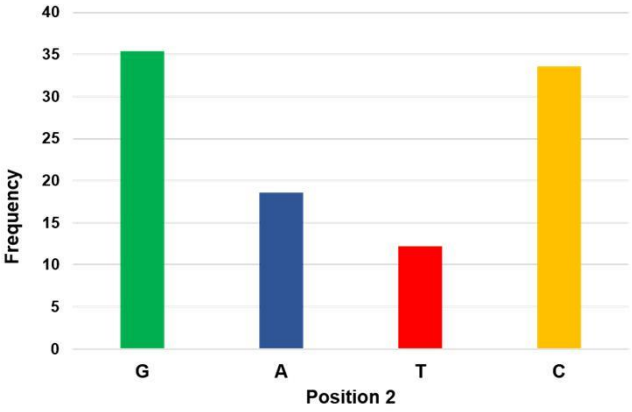


Figure 3.24: Sequence preference for mRNAs isolated from polysomes in MCF7 cells using NanoCAGE data [86] a) Nucleotide frequency at position 1 and b) Nucleotide frequency at position 2.

4. Discussion and future perspectives

Delineating the effects that can influence the translation initiation of cap proximal nucleotides in the 5'TL can help us in understanding how different regulatory context inputs can either restrict or promote the translation of specific messages. The central approach taken to address this question was manipulating the cap proximal sequence on mRNA namely a stretch of 10nt in length called E5S and study its effect on a translation initiation statistic measured as TIRES.

To study the contribution of the E5S on TIRES values, the effect of randomising its sequence on translation initiation was investigated. A similar approach previously was used to study the effect of start codon context on translation [200]. To produce the large amounts of RNA required for this investigation, an IVT method using T7 RNA polymerase was employed. The T7 promoter region used for IVT is conserved in positions -17 to +6 positions [257]. However, use of the T7 promoter introduced a minimum of a single guanosine at the +1 position that was critical for transcription. Milligan et al observed that the strength of transcription using a T7 polymerase was the highest in the presence of +GGG in the T7 promoter (+1 to +3 positions). While +GG (+1 and 2 positions) was essential for efficient transcription, the presence of +G (+1 position) was the minimal requirement critical for transcription. Indeed, replacing the +1 positions with C or A led to a 10-fold decrease in transcription efficiency. A similar observation was made in this work (data not shown) where transcription efficiency was determined for DNA templates varying in the presence of G/GG and GGG in positions +1 - +3 of the T7 promoter region. Despite efforts as outlined in the results, it proved impossible to bypass the +1G requirement for transcription. Consequently, to produce RNA in large quantities for downstream experiments, the presence of guanosine was included in the T7 promoter of the DNA template used for all IVT reactions discussed in this work. In-vitro systems using other promoters like Sp6 were not considered as they also added a +1G in their transcripts upon IVT.

Various strategies were considered to remove the +1G post transcription. The aim was to generate RNA including all nucleotides in the +1 position to study the effects of E5S on TIRES. RNA ligase and RNase H methods were successful but inefficient in

producing a large amount of RNA as explained in the results. Recently, Nelissen et al used a novel recombinant strategy employing tRNA scaffolds and combining T7 promoter IVT along with hammerhead ribozyme. It allowed them to excise the RNA of interest containing the desired +1 nucleotide position and limit insert length to 200nt [258]. Their recombination overexpression strategies included a large number of cloning steps and extensive downstream purification steps [259.] Due to insert length restrictions and technically challenging downstream processing steps, this method was considered but not used for IVT of the desired lnO transcript. However, keeping these challenges in mind, a novel two-step PCR approach was successfully used to generate large amounts of lnO template required for downstream experiments.

In library generation for massively parallel sequencing, Aird et al aimed at minimising biases caused during Illumina library preparation [206]. It was seen that low temperature ramp rates (2.2°C/s) were a critical factor involved in producing PCR products with minimal biases. A lower number of PCR cycles was preferred allowing minimisation of errors introduced in the PCR 260. Phusion polymerase is considered one of the best enzymes for PCR aimed at producing the least amount of PCR biases and the highest sequence integrity [260.] The novel two step PCR approach employed in this thesis used Phusion polymerase in combination with a lower number of PCR cycles and a lower ramp temperature resulting in minimum biases in the lnO template subsequently leading to minimal biases in the NGS library. Previous studies creating a randomized pool of mRNA transcripts for library preparation [261] used a selection strategy based on bacterial plasmids. The novel approach used here proved to be a fast and effective way of producing a randomized library containing all expected variants as shown in the results devoid of a selection pressure from a bacterial system.

Based on equal probabilities, all nucleotides were expected to occur at a frequency of 0.25 at each position of the E5S. Due to technical challenges in T7 promoter based IVT, the inevitable presence of G at +1 position restricted the study to only 25% of the possible E5S variants. However, in the case of annotated human transcripts obtained from three different databases, it was observed that G occurs in the +1 position at a higher frequency (~35%) by comparison with other nucleotides. Thus, the +1G transcripts used reflect approximately one third of the annotated human transcripts.

Polysome profiling separates translated mRNAs on a sucrose gradient based on the number of bound ribosomes [244]. In this work, a 10-60% continuous sucrose gradient was used for polysomal isolation. Polysome-profiling is used to study translatomes and extraction of efficiently translated mRNA (associated with >3 ribosomes) from a large volume across many fractions is challenging[262]. This property makes polysome-profiling inconvenient for larger experimental designs or for use with samples with low RNA yields. However, isolating a single RNA species of interest from the polysomal fraction for library preparation was technically challenging. Although most of the steps were optimized to obtain the highest yield in the experimental protocol used in this work, the RNA obtained from polysomes resulted in limiting quantities that differed between samples. Recently, Liang et al optimized a non-linear sucrose gradient (three sucrose fractions -5%, 34%, and 55%) which enriches for efficiently translated mRNA in only one or two fractions, thereby reducing sample handling by 5–10-fold [262]. This step if incorporated into the experimental protocol used here could potentially minimise losses during total RNA isolation from polysomes and thereby increase the library yield.

Various studies that characterized the 5' end of mRNAs revealed that transcription start sites in most mRNAs are not constricted to a single, well defined position but can often occur at multiple sites or be distributed around a specific site[152,153].'Alternative TSS in the 5'TL can modulate the translation efficiency [263]. The importance of TSS in determining the translation of a transcript has been well studied. It is known from previous work that some transcripts that can contain multiple TSS reflect the selection of particular sites by transcription factors[152.] When there is a narrow distribution of TSS around a particular site for a transcript, it is unlikely that its transcription reflects the presence of different transcription factors. While it is not understood, it is possible that these TSSs may also be regulated. The findings presented here suggest that small differences around the TSS can influence the enrichment preferences for translation initiation.

There has been a gap of knowledge in determining whether the nucleotides immediately downstream of the cap of mRNA can influence its translation initiation. Our findings suggest that cap proximal nucleotides namely E5S can influence translation initiation based on their TIRES values as described in this work. A

frequency-based model is proposed here which illustrates that minor changes in the nucleotides along the E5S can influence its preference for translation initiation. Tamarkin-Ben-Harush et al performed TSS mapping of the translatome and observed a significant change in translation due to cap proximal nucleotides during stress conditions [197]. However, in the control conditions, the length of the 5'TL had a pronounced effect of translation while the cap proximal nucleotides had no observed role in modulating translation. The exact mechanism by which the E5S proximal to the 5' cap can influence translation is not known. Our findings suggest that the E5S may influence the accessibility of the 5' end of a mRNA, and it is probable that minor changes in this sequence can modulate translation initiation, but the magnitude of this change remains unclear. It is likely that TSS selection for most transcripts is controlled by a regulatory mechanism. Therefore, comparison of the effect of E5S on TIRES during conditions of stress against the control condition as demonstrated in this work can potentially reveal factors that could control TSS selection.

Although the usage of geometric mean $TIRES_G$ averaged the effects of TIRES in various libraries, it is important to note that the context preference for each library was different. The individual pattern of sequence context preference for E5S based on TIRES was slightly different in position 2 for samples generated from lnO3 (libraries 3-1 and 3-2) where C was preferred against samples generated from lnO2 and lnO1 (libraries 1-1,1-2 and 2) where G was the preferred. These preferences could have potentially been the same across the libraries if the steps towards library preparation and the sequencing depth of all libraries could be equalized across the libraries.

The second position of E5S was found to have a markedly higher influence on translation initiation than positions further downstream (for technical reasons it was not possible to estimate the influence of the first position of E5S). In this position G was the most enriched nucleotide, and U was the most depleted nucleotide. Similar observations were made by another study (unpublished) where the effects of cap proximal nucleotides influencing eIF4E binding were studied[264.] Although the methodology used in this study is specific and cannot be directly compared to several databases containing information on HEK293T cells, it can be used to gain insight on the translation initiation rate attributed to the E5S. The TSS annotations of various transcripts are simplified against various complexities. A previous study comprising

of two arms: one of NanoCAGE to study TSS information and the other of polysome profiling to study genes active in the process of translation is a robust way to study the regulation of translation in the context of sequence information in a given cell line. Gandin et al used a similar approach to study mTOR responses in MCF7 cells [86,] TSS information of 6551 transcripts isolated from polysomes in the control condition was obtained. These genes analysed for their sequence in the E5S showed that G was preferred in position 1 and 2. This is coherent with our observations for E5S preferences in HEK293T cells. The influence of +2 position in the mRNA 5'TL has not been well studied in the context of translation. However, in mammals it is observed that the ribose of the +1 and +2 nucleotides of mRNA is methylated at the 2' positions.

However, the functions of this modification remains unclear for most mRNAs. A recent study in HELA cells showed that when the enzyme responsible for 2'O methylation at +1 position was knocked down, the global translation remained unaffected[127]. However, the 2' methylation in the +1 and +2 positions of mRNA influenced the ribosome binding and translational efficiency in specific mRNAs[124,265,266]. It is possible that 2'O methylation can influence the translation initiation by an unknown mechanism or influence the binding of the cap binding protein eIF4E; given that the modified +1 nucleotide in mRNA can contact eIF4E directly[189].

The intensity of the effects of $TIRES_G$ observed across E5S in this work was not reflected in reporter assays using a luciferase reporter gene constructed with high and low $TIRES_G$ candidates in HEK293T cells and in HEK cell free systems. Over forty years ago, Lodish proposed the model that translation of mRNAs that initiate protein synthesis at lower rates will be preferentially inhibited when initiation is globally reduced, and this was experimentally demonstrated by the comparison of translation of alpha and beta globin [267]. A difference in the affinity of mRNAs for the general translation initiation factor eIF2 was observed to arbitrate selective translation of a particular viral mRNA over globin mRNA in a cell-free system based on competition between mRNA molecules [268.] However, it is important to note that in the experiment where lnO RNA molecules were transfected, 4^{10} RNA molecules were competing for the components of the translational machinery. The TIRES effects are likely to be influenced by competition between mRNA molecules that differ only in their E5S.

The experimental design used in this work reflects the translation initiation enrichment in a randomized artificial mRNA library. Translation efficiency from artificial mRNA observed in the experiment did not correlate with endogenous RNA data obtained from ribosome profiling. Nonetheless, this work provides new findings on the sequence influence along E5S affecting the rate of translation initiation and provides significant insight into factors that may influence translation initiation on endogenous mRNA.

Secondary structure in the RNA can affect the binding capacities of 43S PIC to the mRNA and thereby inhibit their translation efficiencies. This work shows that increased GC% across E5S has a negative influence on its $TIRES_G$ values. The effect of secondary structure based on distance, position, and difference in thermal stability of RNA hairpins has been reported previously. The higher GC% contribution to lower translation efficiency was examined across different mammalian cell lines 46. It was seen that hairpin structures placed at various positions between +1 to +9 had maximal effect on modulating translational efficiency [46]. Future mutational studies on modulating secondary structures present in E5S variants containing high and low $TIRES_G$ values will be necessary to verify if mRNA translation is defined by the accessibility of the 5' end of the mRNA.

Analysis of available ribosome profiling datasets did not reveal a significant association between E5S and ribosome footprint densities at the coding regions. This work clearly shows the influence of nucleotide context on translation initiation, it is possible that other factors such as uORFs and RNA secondary structures have a higher influence on translation initiation than E5S. This work increases our understanding of how mRNAs are chosen for translation based on their E5S.

In the future, to fully assess the role of the E5S in translation, it will be important to develop methods to readily synthesize capped mRNAs encoding different +1 nucleotides including A, U, and C. mRNAs with different +1 nucleotides can be generated by improving existing chemical synthesis methods [269] or by identifying an RNA polymerase that can produce mRNAs with various +1 nucleotides that is adaptable to robust IVT. A third possibility would be to identify enzymes that can phosphorylate RNA 5' ends to produce 5'- triphosphorylated RNA which could be used as a substrate for existing in vitro capping systems [270].

Cap dependant translation initiation begins with the scanning mechanism in most cases. The 40S ribosomal subunit binds to the capped mRNA and scans along the 5'TL in search of an optimal start codon. Continuous measurement of protein synthesis in-situ revealed that ribosome migration occurs in a unidirectional motion. The rate of migration of the ribosome is virtually independent of mRNA sequence and secondary structure [271]. If the candidates with high and low TIRES values have a preference due to mRNA competition, measuring their respective protein synthesis rates can elucidate the context preference of E5S in modulating translation initiation. The time required for scanning along 5'TL was calculated using precise translation kinetic studies in the case of Vassilenko et al study including differences in the lengths of their 5'TL [271]. To implement a similar methodology in this work would be challenging due to similar lengths of 5'TL between all variants of E5S. In- vitro translation is dependent on various factors. To implement the calculation of the differences in time taken to scan along the 5'TL it is important to comprehend elongation rate, termination rate, luciferase maturation rate and its effects based on a change in salt concentration, however, some of these aspects are not possible using currents methods or are technically challenging.

It is well known that RBPs can influence the translation of a subset of mRNAs [92]. The candidates containing lower $TIRES_G$ could potentially be scanned for a motif preference for known RBP binding preference giving insight into mechanisms governing the relationship between translation and the E5S of mRNAs. In the future, RBP studies could probe for novel RBPs that bind candidates having a lower $TIRES_G$ value that can be validated using overexpression/knockdown studies of the desired RBP. A similar approach could be used for certain E5Ss that are preferred for high $TIRES_G$ value (a consequence of an RBP leading to enhanced translation).

The 5'TOP motif is the most well-known motif influencing translation efficiency that occurs in the 5' cap proximal region. The 5'TOP motif is known to modulate translation during stress conditions [272]. In the future, with the possibility of successful incorporation of a random nucleotide in the +1 position of the mRNA construct, it would be interesting to study the influence of 5'TOP motif using the experimental approach used in this work in basal and stress conditions.

The identification of the mechanism in which the cap structure binds the cap binding protein, eIF4E is crucial to our understanding of cap dependant translation initiation. It is possible that eIF4E has a binding preference to nucleotides that are proximal to the cap structure. Various structural and biophysical studies have demonstrated the binding of eIF4E to different analogues of the 5' cap [189–191]. The nucleotide in the +1 position is known to form different contacts with different initiation factors. For example, eIF4Es binding to the nucleotide in the +1 nucleotide position and the 5' cap can vary based on the nucleotide identity [273,274]. The advent of RBNS (RNA bind and Seq) could help us to understand the binding affinities of different mRNAs to specific initiation factors [195]. eIF4E is a limiting factor in cells, it remains bound to eIF4G/4EBP [275]. Zinshteyn et al showed that the translation initiation factor eIF4G1 preferentially binds yeast transcript leaders containing conserved oligo-uridine motifs 276. In yeast, it was also seen that conformational coupling between eIF4G and eIF4E is important to trigger ribosome loading onto mRNA [273,277]. Although these coupling preferences have not been studied in humans, it is highly probable that the binding affinities of eIF4E alone vary from that of eIF4E bound to eIF4G/4EBP. In the future, it will be interesting to study the effects of E5S on TIRES based on overexpression or knockdown of the initiation factors eIF4E and eIF4G.

The E5S is a previously unappreciated determinant of translation initiation, and this work suggests that differences in mRNA 5' end accessibility defined by the cap proximal sequence maybe an important determinant in modulating the rate of translation initiation.

Bibliography

1. McManus, J., Cheng, Z. & Vogel, C. Next-generation analysis of gene expression regulation--comparing the roles of synthesis and degradation. *Mol. Biosyst.* **11,** 2680–9 (2015).
2. Sultan, M. *et al.* A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science.* **321,** 956–960 (2008).
3. Eswaran, J. *et al.* Transcriptomic landscape of breast cancers through mRNA sequencing. *Sci. Rep.* **2,** 264 (2012).
4. King, H. A. & Gerber, A. P. Translatome profiling: methods for genome-scale analysis of mRNA translation. *Brief. Funct. Genomics* **15,** 22–31 (2014).
5. Thomson, S. R. *et al.* Cell-Type-Specific Translation Profiling Reveals a Novel Strategy for Treating Fragile X Syndrome. *Neuron* **95,** 550–563.e5 (2017).
6. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science.* **352,** 1413–1416 (2016).
7. Pickering, B. M. & Willis, A. E. The implications of structured 5′ untranslated regions on translation and disease. *Semin. Cell Dev. Biol.* **16,** 39–47 (2005).
8. Sonenberg, N. *et al.* Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136,** 731–45 (2009).
9. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. (2016). doi:10.1016/j.cell.2016.03.014
10. Lejeune, F., Ishigaki, Y., Li, X. & Maquat, L. E. The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: dynamics of mRNP remodeling. *EMBO J.* **21,** 3536–3545 (2002).
11. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11,** 113–27 (2010).
12. Hinnebusch, A. G. Molecular Mechanism of Scanning and Start Codon Selection in Eukaryotes. *Microbiol. Mol. Biol. Rev.* **75,** 434–467 (2011).
13. Hinnebusch, A. G. The scanning mechanism of eukaryotic translation initiation.*Annu. Rev. Biochem.* **83,** 779–812 (2014).
14. Livingstone, M., Atas, E., Meller, A. & Sonenberg, N. Mechanisms governing the control of mRNA translation. *Phys. Biol.* **7,** 021001 (2010).

15. Silvera, D., Formenti, S. C. & Schneider, R. J. Translational control in cancer.*Nat. Rev. Cancer* **10,** 254–66 (2010).

16. Seton-Rogers, S. Translation: Switching to cap-independence. *Nat. Rev. Cancer 2008 81* (2008).

17. Wellensiek, B. P. *et al.* Genome-wide profiling of human cap-independent translation-enhancing elements. *Nat. Methods* **10,** 747–50 (2013).

18. Lacerda, R., Menezes, J. & Romão, L. More than just scanning: the importance of cap-independent mRNA translation initiation for cellular stress response and cancer. *Cell. Mol. Life Sci.* **74,** 1659–1680 (2017).

19. Cheung, Y.-N. *et al.* Dissociation of eIF1 from the 40S ribosomal subunit is a key step in start codon selection in vivo. *Genes Dev.* **21,** 1217–30 (2007).

20. Valásek, L., Szamecz, B., Hinnebusch, A. G. & Nielsen, K. H. In vivo stabilization of preinitiation complexes by formaldehyde cross-linking. *Methods Enzymol.* **429,** 163–83 (2007).

21. Lorsch, J. R. & Dever, T. E. Molecular view of 43 S complex formation and start site selection in eukaryotic translation initiation. *J. Biol. Chem.* **285,** 21203–7 (2010).

22. Pisarev, A. V. *et al.* The Role of ABCE1 in Eukaryotic Posttermination Ribosomal Recycling. *Mol. Cell* **37,** 196–210 (2010).

23. Vassilenko, K. S., Alekhina, O. M., Dmitriev, S. E., Shatsky, I. N. & Spirin, A.S. Unidirectional constant rate motion of the ribosomal scanning particle during eukaryotic translation initiation. *Nucleic Acids Res.* **39,** 5555–67 (2011).

24. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352,** 1413–6 (2016).

25. Becker, T. *et al.* Structural basis of highly conserved ribosome recycling in eukaryotes and archaea. *Nature* **482,** 501–6 (2012).

26. Rogers, G. W., Richter, N. J., Lima, W. F. & Merrick, W. C. Modulation of the Helicase Activity of eIF4A by eIF4B, eIF4H, and eIF4F. *J. Biol. Chem.* **276,** 30914–30922 (2001).

27. Pisareva, V. P. & Pisarev, A. V. DHX29 and eIF3 cooperate in ribosomal scanning on structured mRNAs during translation initiation. *RNA* **22,** 1859– 1870 (2016).

28. Eliseeva, I. A., Lyabin, D. N. & Ovchinnikov, L. P. Poly(A)-binding proteins: Structure, domain organization, and activity regulation. *Biochem.* **78,** 1377– 1391 (2013).

29. Pisareva, V. P., Pisarev, A. V., Komar, A. A., Hellen, C. U. T. & Pestova, T. V. Translation Initiation on Mammalian mRNAs with Structured 5′UTRs Requires DExH-Box Protein DHX29. *Cell* **135,** 1237–1250 (2008).

30. Kahvejian, A., Roy, G. & Sonenberg, N. The mRNA closed-loop model: the function of PABP and PABP-interacting proteins in mRNA translation. *Cold Spring Harb. Symp. Quant. Biol.* **66,** 293–300 (2001).

31. Passmore, L. A. *et al.* The Eukaryotic Translation Initiation Factors eIF1 and eIF1A Induce an Open Conformation of the 40S Ribosome. *Mol. Cell* **26,** 41– 50 (2007).

32. Pestova, T. V., Borukhov, S. I. & Hellen, C. U. T. Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons. *Nature* **394,** 854–859 (1998).

33. Lee, J. H. *et al.* Initiation factor eIF5B catalyzes second GTP-dependent step in eukaryotic translation initiation. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16689–94 (2002).

34. Hinnebusch, A. G. Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. *Trends Biochem. Sci.* **42,** 589–611 (2017).

35. Kozak, M. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234,**187–208 (1999).

36. Luna, R. E. *et al.* The C-terminal domain of eukaryotic initiation factor 5 promotes start codon recognition by its dynamic interplay with eIF1 and eIF2β. *Cell Rep.* **1,** 689–702 (2012).

37. Weisser, M., Voigts-Hoffmann, F., Rabl, J., Leibundgut, M. & Ban, N. The crystal structure of the eukaryotic 40S ribosomal subunit in complex with eIF1 and eIF1A. *Nat. Struct. Mol. Biol.* **20,** 1015–7 (2013).

38. Saini, A. K. *et al.* Eukaryotic translation initiation factor eIF5 promotes the accuracy of start codon recognition by regulating Pi release and conformational transitions of the preinitiation complex. *Nucleic Acids Res.* **42,** 9623–9640 (2014).

39. Kozak, M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *Journal of Biological Chemistry* **266,** 19867–19870 (1991).

40. Gebauer, F. & Hentze, M. W. Molecular mechanisms of translational control.

Nat. Rev. Mol. Cell Biol. **5,** 827–835 (2004).

41. Kozak, M. Features in the 5’ non-coding sequences of rabbit alpha and beta-globin mRNAs that affect translational efficiency. *J. Mol. Biol.* **235,** 95–110 (1994).

42. Kozak, M. Leader length and secondary structure modulate mRNA function under conditions of stress. *Mol. Cell. Biol.* **8,** 2737–44 (1988).

43. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361,** 13–37 (2005).

44. Kozak, M. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 2850–4 (1986).

45. Kozak, M. The scanning model for translation: An update. *Journal of Cell Biology* **108,** 229–241 (1989).

46. Babendure, J. R., Babendure, J. L., Ding, J.-H. & Tsien, R. Y. Control of mammalian translation by mRNA structure near caps. *RNA* **12,** 851–61 (2006).

47. Rozen, F. *et al.* Bidirectional RNA helicase activity of eucaryotic translation initiation factors 4A and 4F. *Mol. Cell. Biol.* **10,** 1134–44 (1990).

48. Hartman, T. R. *et al.* RNA helicase A is necessary for translation of selected messenger RNAs. *Nat. Struct. Mol. Biol.* **13,** 509–516 (2006).

49. Manojlovic, Z. & Stefanovic, B. A novel role of RNA helicase A in regulation of translation of type I collagen mRNAs. *RNA* **18,** 321–334 (2012).

50. Zhang, Y. & Stefanovic, B. LARP6 Meets Collagen mRNA: Specific Regulation of Type I Collagen Expression. *Int. J. Mol. Sci.* **17,** 419 (2016).

51. Soto-Rifo, R. *et al.* DEAD-box protein DDX3 associates with eIF4F to promote translation of selected mRNAs. *EMBO J.* **31,** 3745–3756 (2012).

52. Muckenthaler, M., Gray, N. K. & Hentze, M. W. IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap- binding complex eIF4F. *Mol. Cell* **2,** 383–8 (1998).

53. Hentze, M. W. & Kühn, L. C. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 8175–82 (1996).

54. Aziz, N. & Munro, H. N. Iron regulates ferritin mRNA translation through a segment of its 5’ untranslated region. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 8478– 82 (1987).

55. Wang, Y. H., Sczekan, S. R. & Theil, E. C. Structure of the 5' untranslated regulatory region of ferritin mRNA studied in solution. *Nucleic Acids Res.* **18,** 4463–8 (1990).

56. Bugaut, A. & Balasubramanian, S. SURVEY AND SUMMARY 5 0 -UTR RNA G-quadruplexes: translation regulation and targeting. doi:10.1093/nar/gks068

57. Kumari, S., Bugaut, A., Huppert, J. L. & Balasubramanian, S. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.* **3,** 218–221 (2007).

58. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44,** 283–292 (1986).

59. Loughran, G., Sachs, M. S., Atkins, J. F. & Ivanov, I. P. Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. *Nucleic Acids Res.* **40,** 2898–2906 (2012).

60. Hershey, J. W. B., Sonenberg, N. & Mathews, M. B. Principles of translational control: an overview. *Cold Spring Harb. Perspect. Biol.* **4,**(2012).

61. Raney, A., Law, G. L., Mize, G. J. & Morris, D. R. Regulated Translation Termination at the Upstream Open Reading Frame in *S* -Adenosylmethionine Decarboxylase mRNA. *J. Biol. Chem.* **277,** 5988–5994 (2002).

62. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci.* **109,** E2424–E2432 (2012).

63. Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11269–74 (2004).

64. Harding, H. P. *et al.* Regulated translation initiation controls stress-induced gene expression in mammalian cells. *Mol. Cell* **6,** 1099–108 (2000).

65. Barbosa, C. *et al.* Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genet.* **9,** e1003529 (2013).

66. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7,** 768–71 (1997).

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

68. Andreev, D. E. *et al.* Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4,** e03971 (2015).

69. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* **11,** 909–916 (2015).

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

71. Avni, D., Biberman, Y. & Meyuhas, O. The 5' Terminal Oligopyrimidine Tract Confers Translational Control on Top Mrnas in a Cell Type-and Sequence Context-Dependent Manner. *Nucleic Acids Res.* **25,** 995–1001 (1997).

72. Meyuhas, O. Synthesis of the translational apparatus is regulated at the translational level. *Eur. J. Biochem.* **267,** 6321–30 (2000).

73. Thoreen, C. C. *et al.* A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* **485,** 109–13 (2012).

74. Hsieh, A. C. *et al.* The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* **485,** 55–61 (2012).

75. Shama, S., Avni, D., Frederickson, R. M., Sonenberg, N. & Meyuhas, O. Overexpression of initiation factor eIF-4E does not relieve the translational repression of ribosomal protein mRNAs in quiescent cells. *Gene Expr.* **4,** 241– 52 (1995).

76. Gentilella, A. & Thomas, G. The director's cut. *Nature* **485,** 50–51 (2012).

77. Biberman, Y. & Meyuhas, O. TOP mRNAs are translationally inhibited by a titratable repressor in both wheat germ extract and reticulocyte lysate. *FEBS Lett.* **456,** 357–60 (1999).

78. Pellizzoni, L., Cardinali, B., Lin-Marq, N., Mercanti, D. & Pierandrei-Amaldi,P. A Xenopus laevis Homologue of the La Autoantigen Binds the Pyrimidine Tract of the 5′ UTR of Ribosomal Protein mRNAs in Vitro: Implication of a Protein Factor in Complex Formation. *J. Mol. Biol.* **259,** 904–915 (1996).

79. Crosio, C., Boyl, P. P., Loreni, F., Pierandrei-Amaldi, P. & Amaldi, F. La protein has a positive effect on the translation of TOP mRNAs in vivo. *Nucleic Acids Res.* **28,** 2927–34 (2000).

80. Cardinali, B., Carissimi, C., Gravina, P. & Pierandrei-Amaldi, P. La protein is associated with terminal oligopyrimidine mRNAs in actively translating polysomes. *J. Biol. Chem.* **278,** 35145–51 (2003).

81. Kakegawa, T. *et al.* Identification of AUF1 as a rapamycin-responsive binding protein to the 5′-terminal oligopyrimidine element of mRNAs. *Arch. Biochem. Biophys.* **465,** 274–281 (2007).

82. Damgaard, C. K. & Lykke-Andersen, J. Translational coregulation of 5'TOP mRNAs by TIA-1 and TIAR. *Genes Dev.* **25,** 2057–68 (2011).

83. Tcherkezian, J. *et al.* Proteomic analysis of cap-dependent translation identifies LARP1 as a key regulator of 5'TOP mRNA translation. *Genes Dev.* **28,** 357– 71 (2014).

84. Markert, A. *et al.* The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. *EMBO Rep.* **9,** 569–75 (2008).

85. Larsson, O. *et al.* Distinct perturbation of the translatome by the antidiabetic drug metformin. *Proc. Natl. Acad. Sci.* **109,** 8977–8982 (2012).

86. Gandin, V. *et al.* NanoCAGE reveals 5' UTR features that define specific modes of translation of functionally related MTOR-sensitive mRNAs. *Genome Res.* **26,** 636–648 (2016).

87. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.* **3,** 195–205 (2002).

88. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8,** 479–490 (2007).

89. Cléry, A., Blatter, M. & Allain, F. H.-T. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18,** 290–298 (2008).

90. Valverde, R., Edwards, L. & Regan, L. Structure and function of KH domains. *FEBS J.* **275,** 2712–2726 (2008).

91. Linder, P. & Jankowsky, E. From unwinding to clamping — the DEAD box RNA helicase family. *Nat. Rev. Mol. Cell Biol.* **12,** 505–516 (2011).

92. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19,** 327–341 (2018).

93. Müller-McNicoll, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA–protein complexes. *Nat. Rev. Genet.* **14,** 275–287 (2013).

94. Bousquet-Antonelli, C. & Deragon, J.-M. A comprehensive analysis of the La-motif protein superfamily. *RNA* **15,** 750–764 (2009).

95. Stavraka, C. & Blagden, S. The La-Related Proteins, a Family with Connections to Cancer. *Biomolecules* **5,** 2701–2722 (2015).

96. Harvey, R. F. *et al.* Trans-acting translational regulatory RNA binding proteins. *Wiley Interdiscip. Rev. RNA* **9,** e1465 (2018).

97. Lahr, R. M. *et al.* The La-related protein 1-specific domain repurposes HEAT-like repeats to directly bind a 5′TOP sequence. *Nucleic Acids Res.* **43,** 8077– 8088 (2015).

98. Lahr, R. M. *et al.* La-related protein 1 (LARP1) binds the mRNA cap, blocking eIF4F assembly on TOP mRNAs. *Elife* **6,** (2017).

99. Fonseca, B. D. *et al.* La-related Protein 1 (LARP1) Represses Terminal Oligopyrimidine (TOP) mRNA Translation Downstream of mTOR Complex 1 (mTORC1). *J. Biol. Chem.* **290,** 15996–16020 (2015).

100. Burrows, C. *et al.* The RNA binding protein Larp1 regulates cell division, apoptosis and cell migration. *Nucleic Acids Res.* **38,** 5542–5553 (2010).

101. Hopkins, T. G. *et al.* The RNA-binding protein LARP1 is a post-transcriptional regulator of survival and tumorigenesis in ovarian cancer. *Nucleic Acids Res.* **44,** 1227–1246 (2016).

102. Mura, M. *et al.* LARP1 post-transcriptionally regulates mTOR and contributes to cancer progression. *Oncogene* **34,** 5025–5036 (2015).

103. von der Haar, T., Ball, P. D. & McCarthy, J. E. G. Stabilization of Eukaryotic Initiation Factor 4E Binding to the mRNA 5′-Cap by Domains of eIF4G. *J. Biol. Chem.* **275,** 30551–30555 (2000).

104. Chan, S., Choi, E.-A. & Shi, Y. Pre-mRNA 3′-end processing complex assembly and function. *Wiley Interdiscip. Rev. RNA* **2,** 321–335 (2011).

105. Brook, M. & Gray, N. K. The role of mammalian poly(A)-binding proteins in co-ordinating mRNA turnover. *Biochem. Soc. Trans.* **40,** 856–864 (2012).

106. Smith, R. W. P., Blee, T. K. P. & Gray, N. K. Poly(A)-binding proteins are required for diverse biological processes in metazoans. *Biochem. Soc. Trans.* **42,** 1229–1237 (2014).

107. Wei, C.-C., Balasta, M. L., Ren, J. & Goss, D. J. Wheat Germ Poly(A) Binding Protein Enhances the Binding Affinity of Eukaryotic Initiation Factor 4F and (iso)4F for Cap Analogues †. *Biochemistry* **37,** 1910–1916 (1998).

108. Bi, X. & Goss, D. J. Wheat Germ Poly(A)-binding Protein Increases the ATPase and the RNA Helicase Activity of Translation Initiation Factors eIF4A, eIF4B, and eIF-iso4F. *J. Biol. Chem.* **275,** 17740–17746 (2000).

109. Svitkin, Y. V *et al.* The requirement for eukaryotic initiation factor 4A (elF4A) in translation is in direct proportion to the degree of mRNA 5' secondary structure. *RNA* **7,** 382–94 (2001).

110. Tarun, S. Z. & Sachs, A. B. Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. *EMBO J.* **15,** 7168–77 (1996).

111. Rajkowitsch, L., Vilela, C., Berthelot, K., Ramirez, C. V. & McCarthy, J. E. G. Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. *J. Mol. Biol.* **335,** 71–85 (2004).

112. Martineau, Y. *et al.* Poly(A)-Binding Protein-Interacting Protein 1 Binds to Eukaryotic Translation Initiation Factor 3 To Stimulate Translation. *Mol. Cell. Biol.* **28,** 6658–6667 (2008).

113. Karim, M. M. *et al.* A mechanism of translational repression by competition of Paip2 with eIF4G for poly(A) binding protein (PABP) binding. *Proc. Natl.Acad. Sci.* **103,** 9494–9499 (2006).

114. Braun, R. E. Post–transcriptional control of gene expression during spermatogenesis. *Semin. Cell Dev. Biol.* **9,** 483–489 (1998).

115. Collier, B., Gorgoni, B., Loveridge, C., Cooke, H. J. & Gray, N. K. The DAZL family proteins are PABP-binding proteins that regulate translation in germ cells. *EMBO J.* **24,** 2656–2666 (2005).

116. Reynolds, N. *et al.* Dazl binds in vivo to specific transcripts and can regulate the pre-meiotic translation of Mvh in germ cells. *Hum. Mol. Genet.* **14,** 3899– 3909 (2005).

117. Kawahara, H. *et al.* Neural RNA-binding protein Musashi1 inhibits translation initiation by competing with eIF4G for PABP. *J. Cell Biol.* **181,** 639–653 (2008).

118. Yanagiya, A., Delbes, G., Svitkin, Y. V., Robaire, B. & Sonenberg, N. The poly(A)-binding protein partner Paip2a controls translation during late spermiogenesis in mice. *J. Clin. Invest.* **120,** 3389–3400 (2010).

119. Delbes, G., Yanagiya, A., Sonenberg, N. & Robaire, B. PABP Interacting Protein 2A (PAIP2A) Regulates Specific Key Proteins During Spermiogenesis in the Mouse1. *Biol. Reprod.* **86,** 95 (2012).

120. Furuichi, Y. *et al.* Methylated, blocked 5 termini in HeLa cell mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 1904–8 (1975).

121. Zan-Kowalczewska, M. *et al.* Removal of 5'-terminal m7G from eukaryotic mRNAs by potato nucleotide pyrophosphatase and its effect on translation. *Nucleic Acids Res.* **4,** 3065–81 (1977).

122. Shatkin, A. J. & Manley, J. L. The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.* **7,** 838–842 (2000).

123. Moteki, S. & Price, D. Functional coupling of capping and transcription of mRNA. *Mol. Cell* **10,** 599–609 (2002).

124. Muthukrishnan, S., Morgan, M., Banerjee, A. K. & Shatkin, A. J. Influence of 5'-terminal m ^{7} G and 2'-O-methylated residues on messenger ribonucleic acid binding to ribosomes. *Biochemistry* **15,** 5761–5768 (1976).

125. Ramanathan, A., Robb, G. B. & Chan, S.-H. mRNA capping: biological functions and applications. *Nucleic Acids Res.* **44,** 7511–7526 (2016).

126. Werner, A. Predicting translational diffusion of evolutionary conserved RNA structures by the nucleotide number. *Nucleic Acids Res.* **39,** e17–e17 (2011).

127. Bélanger, F., Stepinski, J., Darzynkiewicz, E. & Pelletier, J. Characterization of hMTr1, a Human Cap1 2′- *O* -Ribose Methyltransferase. *J. Biol. Chem.* **285,** 33037–33044 (2010).

128. DONMEZ, G., Hartmuth, K. & Lührmann, R. Modified nucleotides at the 5' end of human U2 snRNA are required for spliceosomal E-complex formation.*RNA* **10,** 1925–1933 (2004).

129. Ramanathan, A., Robb, G. B. & Chan, S. H. mRNA capping: Biological functions and applications. *Nucleic Acids Research* **44,** 7511–7526 (2016).

130. Matsuo, H. *et al.* Structure of translation factor eIF4E bound to m7GDP and interaction with 4E-binding protein. *Nat. Struct. Biol.* **4,** 717–724 (1997).

131. Culjkovic, B., Topisirovic, I. & Borden, K. L. B. Controlling Gene Expression through RNA Regulons: The Role of the Eukaryotic Translation Initiation Factor eIF4E. *Cell Cycle* **6,** 65–69 (2007).

132. Carroll, M. & Borden, K. L. B. The Oncogene eIF4E: Using Biochemical Insights to Target Cancer. *J. Interf. Cytokine Res.* **33,** 227–238 (2013).

133. Culjkovic-Kraljacic, B. & Borden, K. L. B. Aiding and abetting cancer: mRNA export and the nuclear pore. *Trends Cell Biol.* **23,** 328–335 (2013).

134. Haghighat, A. & Sonenberg, N. eIF4g dramatically enhances the binding of eIF4E to the mRNA 5'-cap structure. *J. Biol. Chem.* **272,** 21677–21680 (1997).

135. Mader, S., Lee, H., Pause, A. & Sonenberg, N. The translation initiation factor eIF-4E binds to a common motif shared by the translation factor eIF-4 gamma and the translational repressors 4E-binding proteins. *Mol. Cell. Biol.* **15,** 4990–7 (1995).

136. Marcotrigiano, J., Gingras, A. C., Sonenberg, N. & Burley, S. K. Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G. *Mol. Cell* **3,** 707–16 (1999).

137. Pause, A. *et al.* Insulin-dependent stimulation of protein synthesis by phosphorylation of a regulator of 5'-cap function. *Nature* **371,** 762–767 (1994).

138. Hay, N. & Sonenberg, N. Upstream and downstream of mTOR. *Genes Dev.* **18,** 1926–45 (2004).

139. Karim, M. M. *et al.* A quantitative molecular model for modulation of mammalian translation by the eIF4E-binding protein 1. *J. Biol. Chem.* **276,** 20750–7 (2001).

140. Wang, X., Li, W., Parra, J.-L., Beugnet, A. & Proud, C. G. The C terminus of initiation factor 4E-binding protein 1 contains multiple regulatory features that influence its function and phosphorylation. *Mol. Cell. Biol.* **23,** 1546–57(2003).

141. Gingras, A. C. *et al.* Hierarchical phosphorylation of the translation inhibitor 4E-BP1. *Genes Dev.* **15,** 2852–64 (2001).

142. Gosselin, P. *et al.* The translational repressor 4E-BP called to order by eIF4E: new structural insights by SAXS. *Nucleic Acids Res.* **39,** 3496–3503 (2011).

143. Kinkelin, K., Veith, K., Grunwald, M. & Bono, F. Crystal structure of a minimal eIF4E-Cup complex reveals a general mechanism of eIF4E regulation in translational repression. *RNA* **18,** 1624–1634 (2012).

144. Lukhele, S., Bah, A., Lin, H., Sonenberg, N. & Forman-Kay, J. D. Interaction of the Eukaryotic Initiation Factor 4E with 4E-BP2 at a Dynamic Bipartite Interface. *Structure* **21,** 2186–2196 (2013).

145. Igreja, C., Peter, D., Weiler, C. & Izaurralde, E. 4E-BPs require non-canonical 4E-binding motifs and a lateral surface of eIF4E to repress translation. *Nat. Commun.* **5,** 4790 (2014).

146. Paku, K. S. *et al.* A conserved motif within the flexible C-terminus of the translational regulator 4E-BP is required for tight binding to the mRNA cap- binding protein eIF4E. *Biochem. J.* **441,** 237–245 (2012).

147. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3,** 211–222 (2006).

148. Ayoubi, T. A. & Van De Ven, W. J. Regulation of gene expression by alternative promoters. *FASEB J.* **10,** 453–60 (1996).

149. Jeronimo, C. *et al.* Systematic Analysis of the Protein Interaction Network for the Human Transcription Machinery Reveals the Identity of the 7SK Capping Enzyme. *Mol. Cell* **27,** 262–274 (2007).

150. Shabalina, S. A., Spiridonov, A. N., Spiridonov, N. A. & Koonin, E. V. Connections between alternative transcription and alternative splicing in mammals. *Genome Biol. Evol.* **2,** 791–9 (2010).

151. Zhang, T., Haws, P. & Wu, Q. Multiple Variable First Exons: A Mechanism for Cell- and Tissue-Specific Gene Regulation. *Genome Res.* **14,** 79–89 (2003).

152. Kawaji, H. *et al.* Dynamic usage of transcription start sites within core promoters. *Genome Biol.* **7,** R118 (2006).

153. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38,** 626–635 (2006).

154. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100,** 15776–15781 (2003).

155. Hashimoto, S. *et al.* 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22,** 1146–1149 (2004).

156. Gowda, M. *et al.* Robust analysis of 5′-transcript ends (5′-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Res.* **34,** e126–e126 (2006).

157. Salimullah, M., Mizuho, S., Plessy, C. & Carninci, P. NanoCAGE: A High-Resolution Technique to Discover and Interrogate Cell Transcriptomes. *Cold Spring Harb. Protoc.* **2011,** pdb.prot5559-pdb.prot5559 (2011).

158. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).

159. Tsuchihara, K. *et al.* Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res.* **37,** 2249–2263 (2009).

160. Kapranov, P. *et al.* RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science.* **316,** 1484–1488 (2007).

161. Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. DBTBS: a database of

transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.* **36,** D93–D96 (2008).

162. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30,** 892–7 (2001).

163. Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods* **7,** 528–534 (2010).

164. Pozner, A. *et al.* Transcription-coupled translation control of AML1/RUNX1 is mediated by cap- and internal ribosome entry site-dependent mechanisms. *Mol. Cell. Biol.* **20,** 2297–307 (2000).

165. Blaschke, R. J. *et al.* Transcriptional and Translational Regulation of the Léri-Weill and Turner Syndrome Homeobox Gene *SHOX*. *J. Biol. Chem.* **278,** 47820–47826 (2003).

166. Calkhoven, C. F. *et al.* Translational control of SCL-isoform expression in hematopoietic lineage choice. *Genes Dev.* **17,** 959–64 (2003).

167. Wang, X., Hou, J., Quedenau, C. & Chen, W. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Mol. Syst. Biol.* **12,** 875 (2016).

168. Arrick, B. A., Lee, A. L., Grendell, R. L. & Derynck, R. Inhibition of translation of transforming growth factor-beta 3 mRNA by its 5' untranslated region. *Mol. Cell. Biol.* **11,** 4306–13 (1991).

169. Sobczak, K. & Krzyzosiak, W. J. Structural Determinants of BRCA1 Translational Regulation. *J. Biol. Chem.* **277,** 17349–17358 (2002).

170. Murray-Zmijewski, F., Lane, D. P. & Bourdon, J.-C. p53/p63/p73 isoforms: an orchestra of isoforms to harmonise cell differentiation and response to stress. *Cell Death Differ.* **13,** 962–972 (2006).

171. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4,** 170112 (2017).

172. Dieudonné, F.-X. *et al.* The effect of heterogeneous Transcription Start Sites (TSS) on the translatome: implications for the mammalian cellular phenotype. *BMC Genomics* **16,** 986 (2015).

173. Eliseev, B. *et al.* Structure of a human cap-dependent 48S translation pre-initiation complex. *Nucleic Acids Res.* **46,** 2678–2689 (2018).

174. Kozak, M. Nucleotide sequences of 5’-terminal ribosome-protected initiation regions from two reovirus messages. *Nature* **269,** 391–4 (1977).

175. Lazarowitz, S. G. & Robertson, H. D. Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes. *J. Biol. Chem.* **252,** 7842–9 (1977).

176. Curran, J. A. & Weiss, B. What Is the Impact of mRNA 5′ TL Heterogeneity on Translational Start Site Selection and the Mammalian Cellular Phenotype? *Front. Genet.* **7,** 156 (2016).

177. Merrick, W. C. eIF4F: A Retrospective. *J. Biol. Chem.* **290,** 24091–24099 (2015).

178. Zinoviev, A., Hellen, C. U. T. & Pestova, T. V. Multiple Mechanisms of Reinitiation on Bicistronic Calicivirus mRNAs. *Mol. Cell* **57,** 1059–1073 (2015).

179. Haimov, O., Sinvani, H. & Dikstein, R. Cap-dependent, scanning-free translation initiation mechanisms. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1849,** 1313–1318 (2015).

180. Moll, I., Hirokawa, G., Kiel, M. C., Kaji, A. & Bläsi, U. Translation initiation with 70S ribosomes: an alternative pathway for leaderless mRNAs. *Nucleic Acids Res.* **32,** 3354–3363 (2004).

181. Akulich, K. A. *et al.* Four translation initiation pathways employed by the leaderless mRNA in eukaryotes. *Sci. Rep.* **6,** 37905 (2016).

182. Andreev, D. E., Terenin, I. M., Dunaevsky, Y. E., Dmitriev, S. E. & Shatsky, I. N. A leaderless mRNA can bind to mammalian 80S ribosomes and direct polypeptide synthesis in the absence of translation initiation factors. *Mol. Cell. Biol.* **26,** 3164–9 (2006).

183. Krokowski, D. *et al.* Characterization of hibernating ribosomes in mammalian cells. *Cell Cycle* **10,** 2691–702 (2011).

184. Elfakess, R. & Dikstein, R. A Translation Initiation Element Specific to mRNAs with Very Short 5′UTR that Also Regulates Transcription. *PLoS One* **3,** e3094 (2008).

185. Elfakess, R. *et al.* Unique translation initiation of mRNAs-containing TISU element. *Nucleic Acids Res.* **39,** 7598–609 (2011).

186. Sonenberg, N., Morgan, M. A., Merrick, W. C. & Shatkin, A. J. A polypeptide in eukaryotic initiation factors that crosslinks specifically to the 5’-terminal cap in mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 4843–7 (1978).

187. Carberry, S. E., Rhoads, R. E. & Goss, D. J. A spectroscopic study of the binding of m7GTP and m7GpppG to human protein synthesis initiation factor 4E. *Biochemistry* **28,** 8078–83 (1989).

188. Niedzwiecka, A. *et al.* Biophysical Studies of eIF4E Cap-binding Protein: Recognition of mRNA 5′ Cap Structure and Synthetic Fragments of eIF4G and 4E-BP1 Proteins. *J. Mol. Biol.* **319,** 615–635 (2002).

189. Tomoo, K. *et al.* Crystal structures of 7-methylguanosine 5'-triphosphate (m(7)GTP)- and P(1)-7-methylguanosine-P(3)-adenosine-5',5'-triphosphate (m(7)GpppA)-bound human full-length eukaryotic initiation factor 4E: biological importance of the C-terminal flexible region. *Biochem. J.* **362,** 539– 44 (2002).

190. Tomoo, K. *et al.* Structural basis for mRNA Cap-Binding regulation of eukaryotic initiation factor 4E by 4E-binding protein, studied by spectroscopic, X-ray crystal structural, and molecular dynamics simulation methods. *Biochim. Biophys. Acta - Proteins Proteomics* **1753,** 191–208 (2005).

191. Marcotrigiano, J., Gingras, A. C., Sonenberg, N. & Burley, S. K. Cocrystal structure of the messenger RNA 5' cap-binding protein (eIF4E) bound to 7- methyl-GDP. *Cell* **89,** 951–61 (1997).

192. O'Leary, S. E., Petrov, A., Chen, J. & Puglisi, J. D. Dynamic recognition of the mRNA cap by Saccharomyces cerevisiae eIF4E. *Structure* **21,** 2197–207 (2013).

193. Stripecke, R., Oliveira, C. C., McCarthy, J. E. & Hentze, M. W. Proteins binding to 5' untranslated region sites: a general mechanism for translational regulation of mRNAs in human and yeast cells. *Mol. Cell. Biol.* **14,** 5898–909 (1994).

194. Lindqvist, L., Imataka, H. & Pelletier, J. Cap-dependent eukaryotic initiation factor-mRNA interactions probed by cross-linking. *RNA* **14,** 960–9 (2008).

195. Lambert, N. *et al.* RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol. Cell* **54,** 887– 900 (2014).

196. Zinshteyn, B., Rojas Duran, M. F. & Gilbert, W. V. Translation initiation factor eIF4G1 preferentially binds yeast transcript leaders containing conserved oligo-uridine motifs. *RNA* rna.062059.117 (2017). doi:10.1261/rna.062059.117

197. Tamarkin-Ben-Harush, A., Vasseur, J. J., Debart, F., Ulitsky, I. & Dikstein, R. Cap-proximal nucleotides via differential eIF4E binding and alternative promoter usage mediate translational response to energy stress. *Elife* **6,** (2017).

198. Kozak, M. An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell Biol.* **115,** 887–903 (1991).

199. Ivanov, I. P., Loughran, G., Sachs, M. S. & Atkins, J. F. Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc. Natl. Acad. Sci. U. S. A.* **107,** 18056–60 (2010).

200. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10,** 748 (2014).

201. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36,** W70-4 (2008).

202. Stark, M. R. & Rader, S. D. Efficient splinted ligation of synthetic RNA using RNA ligase. *Methods Mol. Biol.* **1126,** 137–149 (2014).

203. Gandin, V. *et al.* Polysome Fractionation and Analysis of Mammalian Translatomes on a Genome-wide Scale. *J. Vis. Exp.* 1–10 (2014). doi:10.3791/51455

204. Krieg, P. A. *A laboratory guide to RNA : isolation, analysis, and synthesis*. (Wiley-Liss, 1996).

205. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335,** 552–7 (2012).

206. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12,** R18 (2011).

207. *The FASTQ format*.

208. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

209. Michel, A. M. *et al.* RiboGalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol.* **13,** 316–319 (2016).

210. Sidrauski, C., McGeachy, A. M., Ingolia, N. T. & Walter, P. The small molecule ISRIB reverses the effects of eIF2α phosphorylation on translation and stress granule assembly. *Elife* **4,** e05033 (2015).

211. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–7 (2013).

212. Beckert, B. & Masquida, B. *Synthesis of RNA by In Vitro Transcription*. 29–41 (Humana Press, 2011). doi:10.1007/978-1-59745-248-9_3

213. Tabor, S. Expression Using the T7 RNA Polymerase/Promoter System. in *Current Protocols in Molecular Biology* **Chapter 16,** Unit16.2 (John Wiley & Sons, Inc., 2001).

214. Kochetkov, S. ., Rusakova, E. . & Tunitskaya, V. . Recent studies of T7 RNA polymerase mechanism. *FEBS Lett.* **440,** 264–267 (1998).

215. TriLink | Long RNA Synthesis, Longmer RNA. Available at: https://www.trilinkbiotech.com/mRNA/longrna.asp. (Accessed: 18th January 2018)

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

217. Wang, X.-Q. & Rothnagel, J. A. 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Res.* **32,** 1382–91 (2004).

218. Weill, L., Belloc, E., Bava, F. A. & Méndez, R. Translational control by changes in poly(A) tail length: Recycling mRNAs. *Nature Structural and Molecular Biology* **19,** 577–585 (2012).

219. Michel, A. M. & Baranov, P. V. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip. Rev. RNA* **4,** 473– 490 (2013).

220. Gebauer, F., Preiss, T. & Hentze, M. W. From cis-regulatory elements to complex RNPs and back. *Cold Spring Harb. Perspect. Biol.* **4,** a012245 (2012).

221. Usman, N. & Cedergren, R. Exploiting the chemical synthesis of RNA. *Trends Biochem. Sci.* **17,** 334–339 (1992).

222. Cranston, J. W., Silber, R., Malathi, V. G. & Hurwitz, J. Studies on ribonucleic acid ligase. Characterization of an adenosine triphosphate-inorganic pyrophosphate exchange reaction and demonstration of an enzyme-adenylate complex with T4 bacteriophage-induced enzyme. *J. Biol. Chem.* **249,** 7447–56 (1974).

223. Sugino, A., Snopek, T. J. & Cozzarelli, N. R. Bacteriophage T4 RNA ligase. Reaction intermediates and interaction of substrates. *J. Biol. Chem.* **252,** 1732– 1738 (1977).

224. Uhlenbeck, O. C. & Gumport, R. I. T4 RNA Ligase. *Enzymes* **15,** 31–58 (1982).

225. Deana, A., Celesnik, H. & Belasco, J. G. The bacterial enzyme RppH triggers messenger RNA degradation by 5??? pyrophosphate removal. *Nature* **451,** 355–358 (2008).

226. Kershaw, C. J. & O'Keefe, R. T. Splint ligation of RNA with T4 DNA ligase. *Methods Mol. Biol.* **941,** 257–269 (2012).

227. Stark, M. R., Pleiss, J. A., Deras, M., Scaringe, S. A. & Rader, S. D. An RNA ligase-mediated method for the efficient creation of large, synthetic RNAs. *RNA* **12,** 2014–9 (2006).

228. Cerritelli, S. M. & Crouch, R. J. Ribonuclease H: the enzymes in eukaryotes. *FEBS J.* **276,** 1494–1505 (2009).

229. Uchiyama, Y. *et al.* DNA-Linked RNase H for Site-Selective Cleavage of RNA1. *Bioconjugate Chem* **5,** 327–332 (1994).

230. Lapham, J. & Crothers, D. M. RNase H cleavage for processing of in vitro transcribed RNA for NMR studies and RNA ligation. *RNA* **2,** 289–296 (1996).

231. Lima, W. F. & Crooke, S. T. Cleavage of single strand RNA adjacent to RNA-DNA duplex regions by Escherichia coli RNase H1. *J. Biol. Chem.* **272,** 27513– 6 (1997).

232. Kurreck, J. Design of antisense oligonucleotides stabilized by locked nucleic acids. *Nucleic Acids Res.* **30,** 1911–1918 (2002).

233. Lee, T.-S., Wong, K.-Y., Giambasu, G. M. & York, D. M. Bridging the Gap Between Theory and Experiment to Derive a Detailed Understanding of Hammerhead Ribozyme Catalysis. in 25–91 (2013). doi:10.1016/B978-0-12-381286-5.00002-0

234. Hoehlig, K., Bethge, L. & Klussmann, S. Stereospecificity of Oligonucleotide Interactions Revisited: No Evidence for Heterochiral Hybridization and Ribozyme/DNAzyme Activity. *PLoS One* **10,** e0115328 (2015).

235. Firth, A. E. & Patrick, W. M. GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in randomized libraries. *Nucleic Acids Res.* **36,** W281–W285 (2008).

236. Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M. F. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Appl. Environ. Microbiol.* **71,** 8966–8969 (2005).

237. Sipos, R. *et al.* Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* **60,** 341–350 (2007).

238. Kapranov, P. From transcription start site to cell biology. *Genome Biol.* **10,** 217 (2009).

239. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22,** 1760–1774 (2012).

240. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI reference sequences: Current status, policy and new initiatives. *Nucleic Acids Res.* **37,** (2009).

241. Castrignanò, T. *et al.* ASPicDB: A database resource for alternative splicing analysis. *Bioinformatics* **24,** 1300–1304 (2008).

242. (DGT), T. F. C. and the R. P. and C. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).

243. Balagopal, V. & Parker, R. Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. *Curr. Opin. Cell Biol.* **21,** 403–408 (2009).

244. Chassé, H., Boulben, S., Costache, V., Cormier, P. & Morales, J. Analysis of translation using polysome profiling. *Nucleic Acids Res.* **45,** gkw907 (2016).

245. Malone, R. W., Felgner, P. L. & Verma, I. M. Cationic liposome-mediated RNA transfection. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 6077–81 (1989).

246. Dalby, B. *et al.* Advanced transfection with Lipofectamine 2000 reagent: primary neurons, siRNA, and high-throughput applications. *Methods* **33,** 95– 103 (2004).

247. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7,** 1534–50 (2012).

248. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–94 (1998).

249. Chen, D. & Patton, J. T. Reverse Transcriptase Adds Nontemplated Nucleotides to cDNAs During 5′-RACE and Primer Extension. *Biotechniques* **30,** 574–582 (2001).

250. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9,** 72–74 (2012).

251. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–90 (2004).

252. Chung, B. Y. *et al.* The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* **21,** 1731– 45 (2015).

253. Zou, K. H., Tuncali, K. & Silverman, S. G. Correlation and Simple Linear Regression. *Radiology* **227,** 617–628 (2003).

254. Gallie, D. R., Ling, J., Niepel, M., Morley, S. J. & Pain, V. M. The role of 5'-ABP concentration on cap and poly(A) tail function during translation in Xenopus oocytes. *Nucleic Acids Res.* **28,**2943–2953 (2000).

255. Pelletier, J. & Sonenberg, N. Insertion mutagenesis to increase secondary structure within the 5' noncoding region of a eukaryotic mRNA reduces translational efficiency. *Cell* **40,** 515–26 (1985).

256. Roy, S. *et al.* Redefining the transcriptional regulatory dynamics of classically and alternatively activated macrophages by deepCAGE transcriptomics. *Nucleic Acids Res.* **43,** 6969–6982 (2015).

257. Milligan, J. F. & Uhlenbeck, O. C. [5] Synthesis of small RNAs using T7 RNA polymerase. in 51–62 (1989). doi:10.1016/0076-6879(89)80091-6

258. Nelissen, F. H. T. *et al.* Fast production of homogeneous recombinant RNA—towards large-scale production of RNA. *Nucleic Acids Res.* **40,** e102–e102 (2012).

259. Baronti, L., Karlsson, H., Marušič, M. & Petzold, K. A guide to large-scale RNA sample preparation. *Anal. Bioanal. Chem.* **410,** 3239–3252 (2018).

260. Brandariz-Fontes, C. *et al.* Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Sci. Rep.* **5,** 8056 (2015).

261. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10,** n/a-n/a (2014).

262. Liang, S. *et al.* Polysome-profiling in small tissue samples. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx940

263. Rojas-Duran, M. F. & Gilbert, W. V. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18,** 2299–305 (2012).

264. Keys, H. R. Juxtacap Nucleotide Sequence Modulates eIF4E Binding and Translation . *bioRxiv Biochem.* (2017). doi:10.1101/165142

265. Kuge, H. & Richter, J. D. Cytoplasmic 3' poly(A) addition induces 5' cap ribose methylation: implications for translational control of maternal mRNA. *EMBO J.* **14,** 6301–10 (1995).

266. Kuge, H., Brownlee, G. G., Gershon, P. D. & Richter, J. D. Cap ribose methylation of c-mos mRNA stimulates translation and oocyte maturation in Xenopus laevis. *Nucleic Acids Res.* **26,** 3208–14 (1998).

267. Lodish, H. F. Model for the regulation of mRNA translation applied to haemoglobin synthesis. *Nature* **251,** 385–388 (1974).

268. Rosen, H., Segnis, G. Di & Kaempfer, R. Translational Control by Messenger RNA Competition for Eukaryotic Initiation Factor 2 *. *J. Biol. Chem.* **257,** 946– 952 (1982).

269. Thillier, Y. *et al.* Synthesis of 5' cap-0 and cap-1 RNAs using solid-phase chemistry coupled with enzymatic methylation by human (guanine-N7)-methyl transferase. *RNA* **18,** 856–868 (2012).

270. Spencer, E., Loring, D., Hurwitz, J. & Monroy, G. Enzymatic conversion of 5'-phosphate-terminated RNA to 5'-di- and triphosphate-terminated RNA. *Proc.Natl. Acad. Sci. U. S. A.* **75,** 4793–7 (1978).

271. Vassilenko, K. S., Alekhina, O. M., Dmitriev, S. E., Shatsky, I. N. & Spirin, A. S. Unidirectional constant rate motion of the ribosomal scanning particle during eukaryotic translation initiation. *Nucleic Acids Res.* **39,** 5555–5567 (2011).

272. Ma, X. M. & Blenis, J. Molecular mechanisms of mTOR-mediated translational control. *Nat. Rev. Mol. Cell Biol.* **10,** 307–318 (2009).

273. Tomoo, K. *et al.* Structural basis for mRNA Cap-Binding regulation of eukaryotic initiation factor 4E by 4E-binding protein, studied by spectroscopic, X-ray crystal structural, and molecular dynamics simulation methods. *Biochim. Biophys. Acta - Proteins Proteomics* **1753,** 191–208 (2005).

274. Zuberek, J. *et al.* Phosphorylation of eIF4E attenuates its interaction with mRNA 5' cap analogs by electrostatic repulsion: intein-mediated protein ligation strategy to obtain phosphorylated protein. *RNA* **9,** 52–61 (2003).

275. Hiremath, L. S., Webb, N. R. & Rhoads, R. E. Immunological detection of the messenger RNA cap-binding protein. *J. Biol. Chem.* **260,** 7843–9 (1985).

276. Zinshteyn, B., Rojas-Duran, M. F. & Gilbert, W. V. Translation initiation factor eIF4G1 preferentially binds yeast transcript leaders containing conserved oligo-uridine motifs. *RNA* **23,** 1365–1375 (2017).

277. Gross, J. D. *et al.* Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell* **115,** 739–750 (2003).

Appendix

Table 1: Characteristics of oligos ordered from Trilink Technologies.

Short Oligo number	**O.D(A_{260})**	**Extinction coefficient**	**Molecular weight (g)**
sO1	69.2	515.8	16229.1
sO2	27.2	515.8	16229.1
sO3	63.5	457.1	14334.9

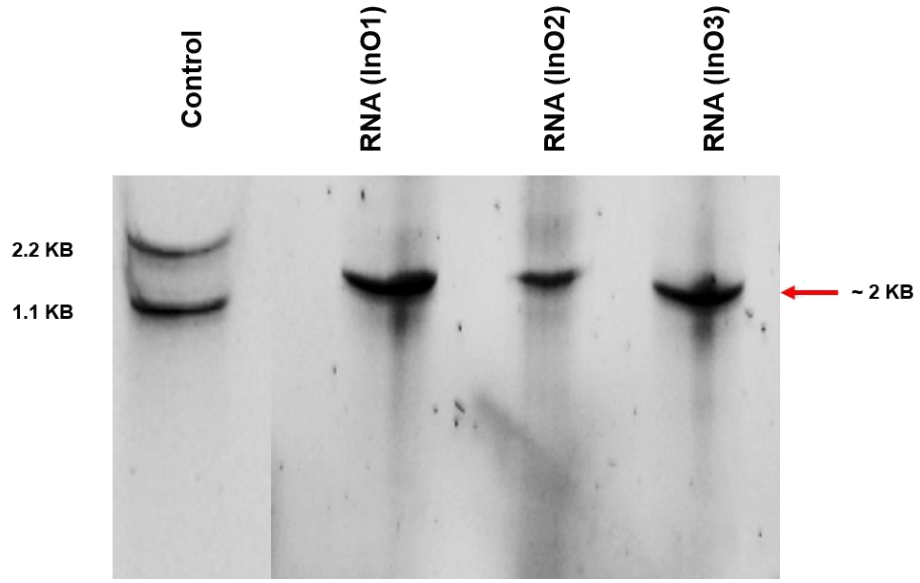


Figure 1: Confirming the quality of RNA using an 8% PAGE UREA gel. RNA generated from in vitro transcription using AmpliScribe™ T7-Flash™ Transcription Kit in an 8% PAGE urea gel shows a clear band without any denaturation bands in lower sizes. RNA generated from lnOs 1,2 and 3 respectively are shown against a control luciferase template containing two RNA molecules of sizes 1.1 and 2.2 KB respectively.

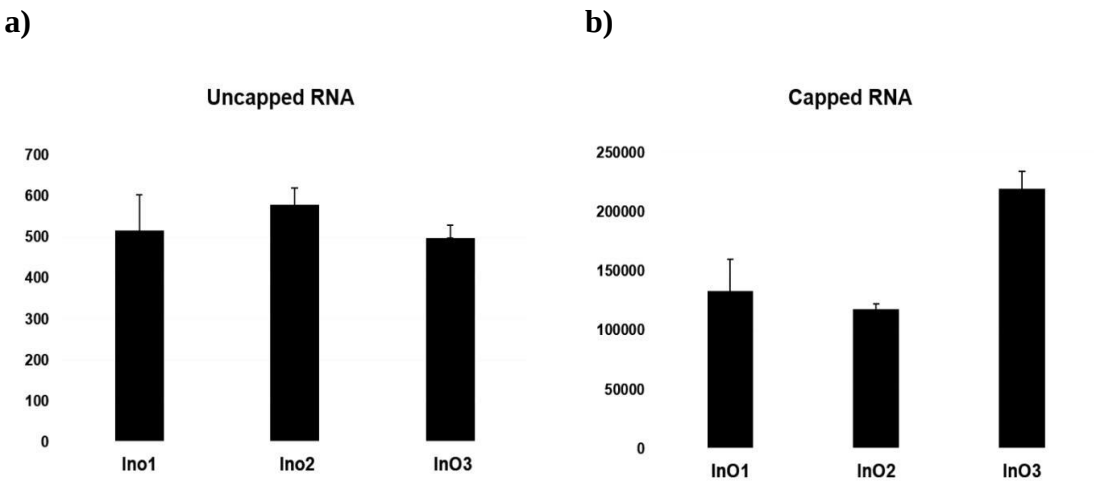


Figure 2: Luciferase light units in the reporter assay to measure transfection efficiencies of a) capped and b) uncapped lnO RNA. HEK293T cells were transfected with lnO RNA at 8-10 µg/6 well plate and firefly luciferase light units were measured to evaluate transfection efficiencies. Transfection efficiencies are tabulated for each lnO RNA. Capped RNA is transfected over 300-fold better than uncapped RNA.

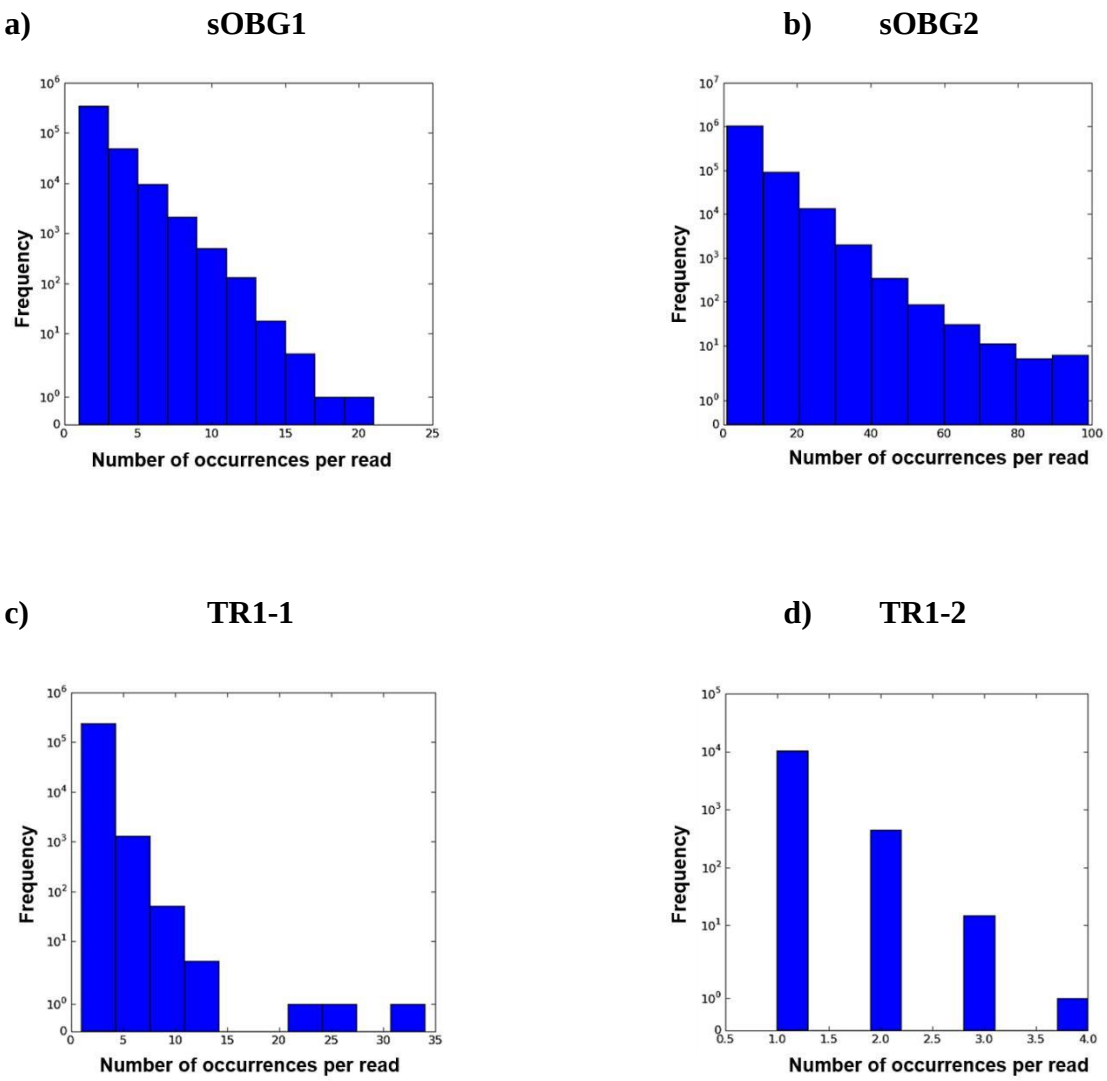


Figure 3: Frequency of the number of occurrences of reads in samples (a-d) as shown in the figure.

Illumina MiSeq has an output of 25 million reads per lane. We used 4 samples in the lane giving rise to unequal distribution. The expected number of reads upon equal distribution would result in 6.25 million reads per sample. Each library has a possibility of 410 unique variants. If the total library size is 6.25 million reads, it leads to a maximum of 6 occurrences of each unique variant. However, the number of reads per sample differed significantly from each other as seen in table 3.1 leading to a different number of occurrences for each possible unique variant in the library. Multiple occurrences of the same unique variant can potentially indicate biases introduced due to technicalities in the protocol.

Table 2: Raw data information on NGS data using Hi-seq3000 on two sequencing lanes shown as a and b respectively.

a)

Sample Name	Clean reads	Clean bases	Read length(bp)	Q20 (%)	GC (%)
lnO2	39608144	1940799056	49	95.71	44.97
sOBG3	24550751	1202986799	49	77.12	47.05
lnO3	30896905	1513948345	49	95.97	45.53
TR2-1	29490071	1445013479	49	95.98	44.50
TR3-2	38460630	1884570870	49	95.56	44.39
TR1-2	11992123	587614027	49	75.93	43.96
lnO1	34413338	1686253562	49	95.67	44.49
TR2-2	52453046	2570199254	49	96.29	43.66
TR3-1	36363104	1781792096	49	96.30	44.41

b)

Sample Name	Clean reads	Clean bases	Read length(bp)	Q20 (%)	GC (%)
PR1-1	9912887	485731463	49	69.20	46.84
PR1-2	28519077	1397434773	49	91.23	44.06
PR2-1	24205021	1186046029	49	90.24	44.77
PR2-2	2371855	116220895	49	63.68	46.32
PR3-1	29634484	1452089716	49	90.90	44.63
PR3-2	39403259	1930759691	49	90.76	44.54
sOBG1	29481565	1444596685	49	90.48	45.15
sOBG2	31542397	1545577453	49	90.70	44.47
TR1-1	69652844	3412989356	49	90.04	43.83

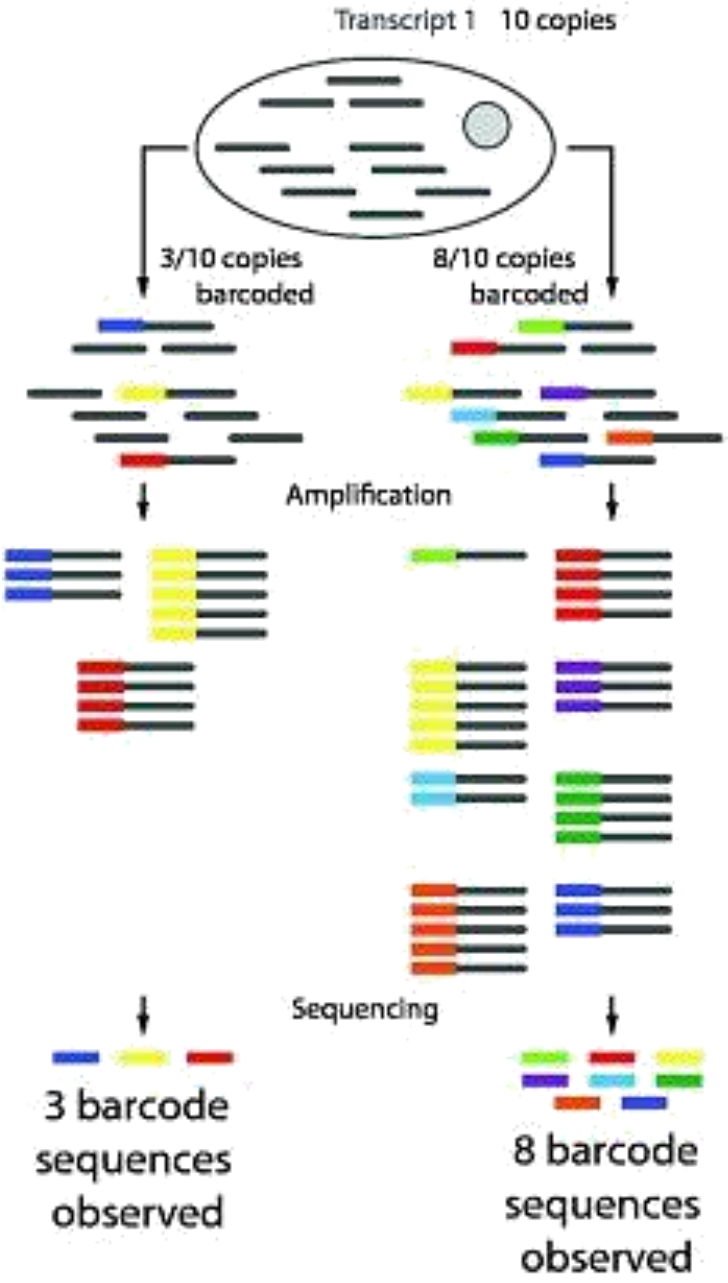
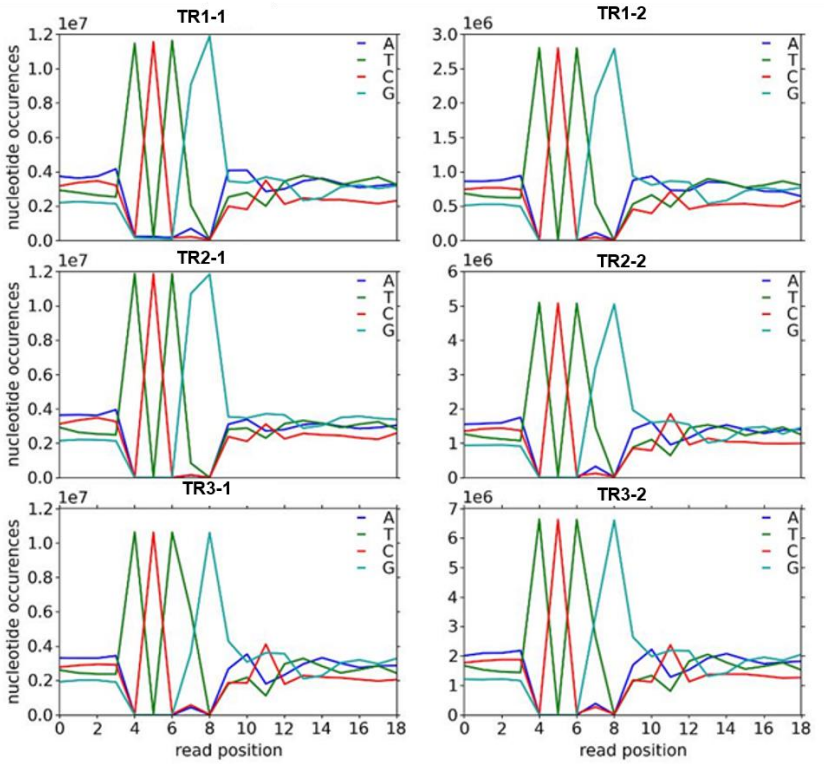


Figure 4: The principle of using UMI to remove PCR duplicates. In this figure, if three out of the ten transcript molecules in total are labelled with the unique identifier (i.e. barcoded), only three barcodes will be observed in the sequencing data. Converting eight out of the ten molecules leads to the identification of eight barcodes in the sequencing reads. A UMI can become saturated if the number of transcripts copies exceeds the number of possible UMI combinations.

a)


TR1-1
TR1-2
TR2-1
TR2-2
TR3-1
TR3-2
nucleotide occurences
read position
A
T
C
G

b)

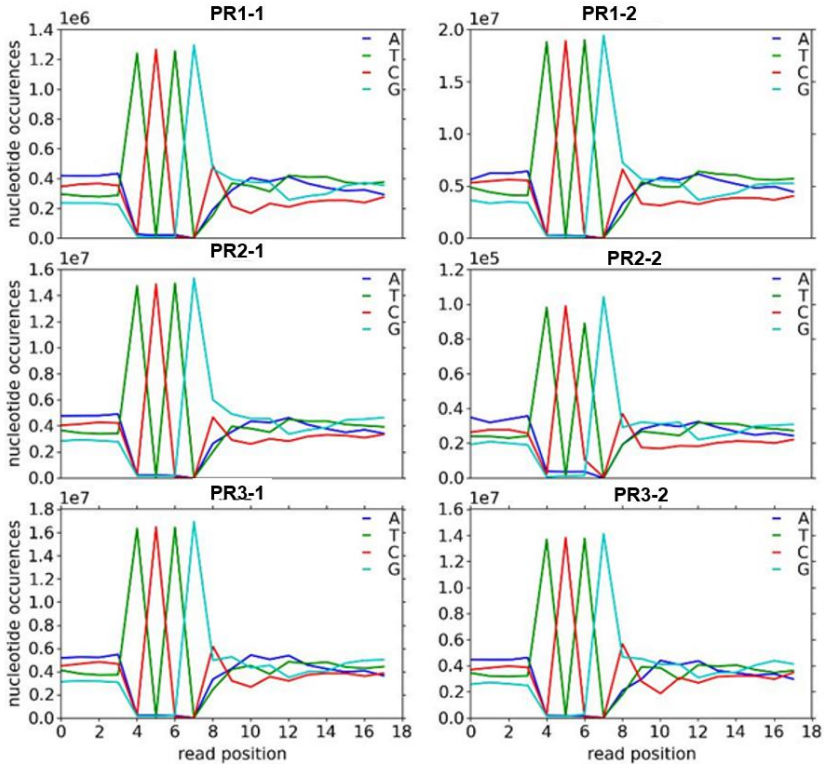

PR1-1
PR1-2
PR2-1
PR2-2
PR3-1
PR3-2
nucleotide occurences
read position
A
T
C
G

Figure 5: Frequency of occurrence of nucleotides at positions containing random nucleotides in position 8-18. An extra nucleotide was found to be added 5' proximal to the G (following the UMI sequence). The major fraction of this extra nucleotide is G and a smaller fraction of these are T. As the extra nucleotide is accounted for the reads that are 19 nucleotides in length are included in this analysis. In both cases, the sequence considered includes the last 11 nucleotides, (G including the 10nt randomer).

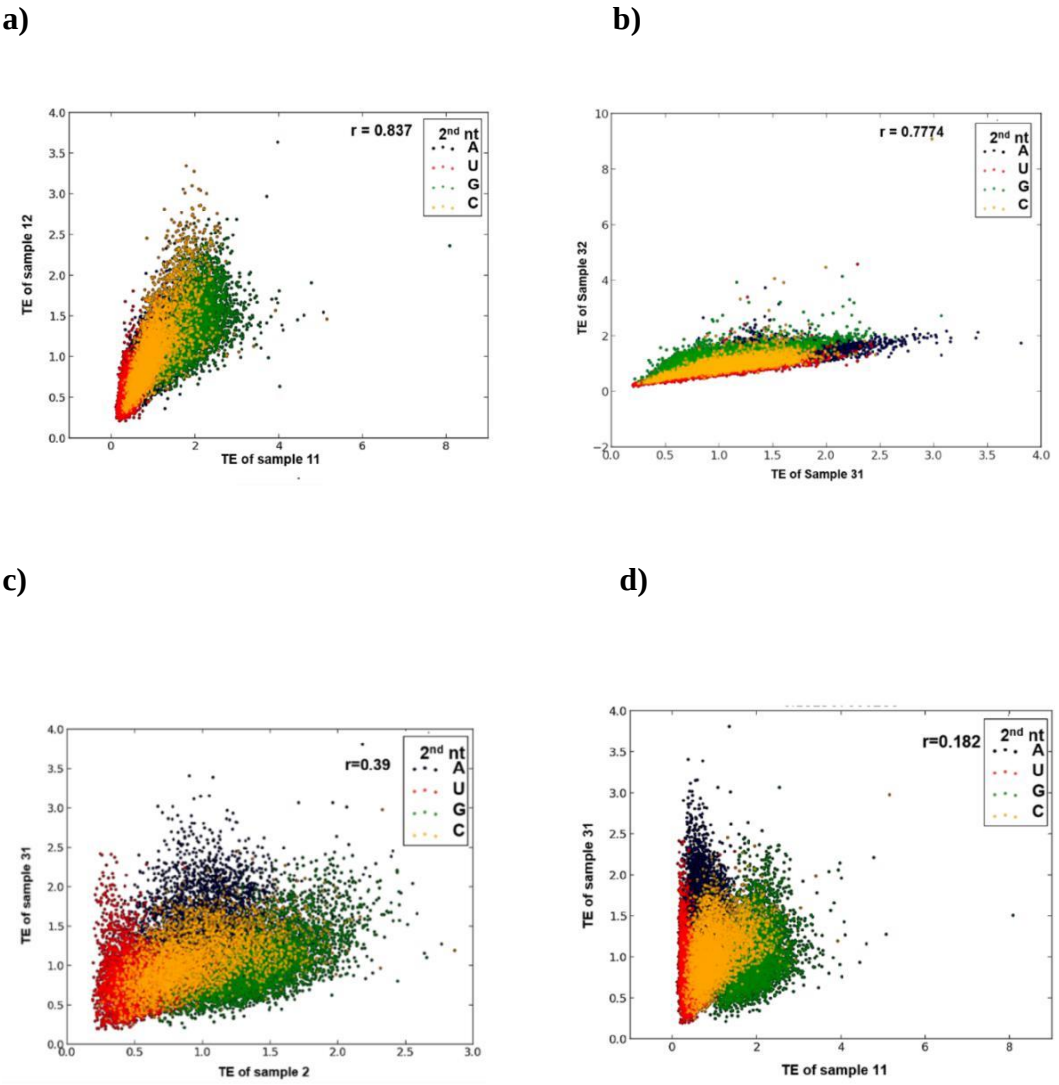


Figure 6: Correlation plot between different libraries showing the importance of position 2 of E5S. Correlations of measurements of TIRES produced in position 2 were plotted. TE was calculated based on each position from 2-11 by taking the ratios of PR/TR of each sample. Pearson's correlation was calculated between the TE of the following samples: a) 1-1vs 1-2 b) 3-1 vs 3-2 c) 3-1 vs 2 and d) 1-1 vs 3-1.

Table 3: Positive interactions (Synergy value >$2^{0.12}$) of di-nucleotides in the E5S influencing TIRES.

Position of 1st nucleotide	1st nucleotide	Position of 2nd nucleotide	2nd Nucleotide	Synergy value
2	G	3	A	1.25
3	C	4	C	1.15
3	T	5	C	1.13
4	G	5	A	1.13
4	C	5	C	1.12
3	G	5	T	1.11
4	A	5	G	1.11
3	C	4	T	1.10
4	C	5	T	1.10
5	G	6	A	1.10
3	A	4	G	1.10
3	T	4	G	1.09
3	G	7	G	1.08

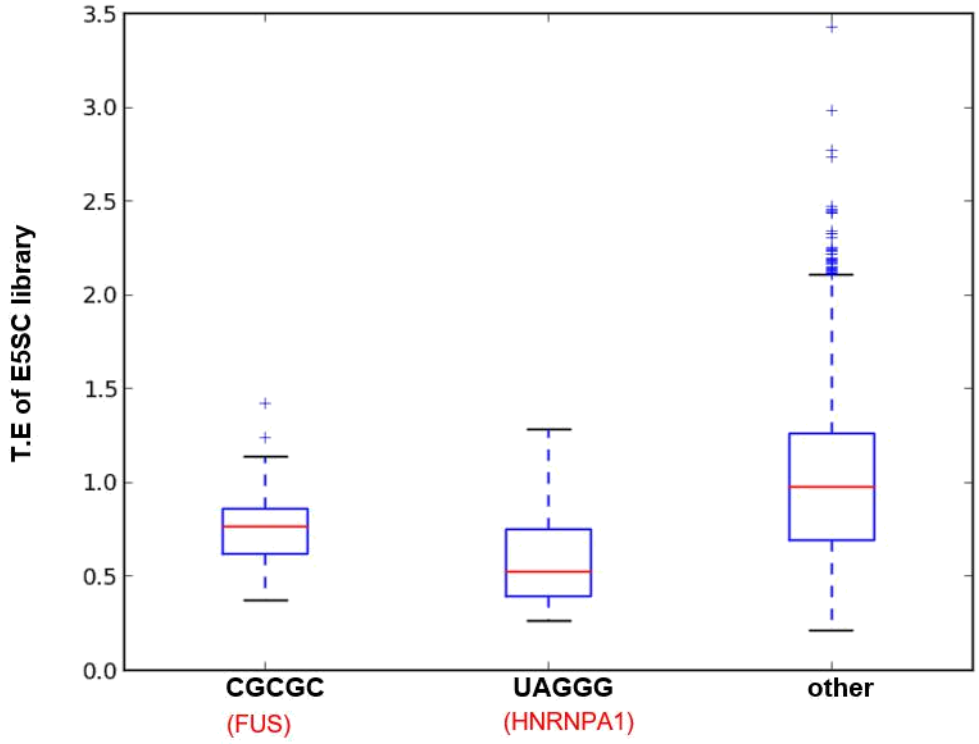


Figure 7: The motifs of RBPs FUS and HNRNPA1 highly expressed in HEK293T cells were downregulated in the E5S library. The motifs highly represented by RBPs FUS and HNPRNPA1 were taken from the RBP compendium in humans 211. These motifs were scanned across the E5S context in comparison to all other 5nt motifs present in the library and represented as a box plot.

a)

PANTHER GO-Slim Molecular Function	Homo sapiens (REF) #	Client Text Box Input (▽ Hierarchy NEW! ⓘ) #	expected	Fold Enrichment	+/-	raw P value	FDR
damaged DNA binding	27	15	4.09	3.67	+	1.76E-04	2.11E-03
↳nucleic acid binding	1625	359	246.20	1.46	+	7.88E-11	2.52E-09
translation elongation factor activity	19	9	2.88	3.13	+	7.67E-03	4.60E-02
↳RNA binding	385	126	58.33	2.16	+	1.15E-12	5.51E-11
endodeoxyribonuclease activity	22	10	3.33	3.00	+	6.30E-03	4.17E-02
↳nuclease activity	162	43	24.54	1.75	+	1.72E-03	1.38E-02
↳catalytic activity	4217	805	638.90	1.26	+	3.59E-11	1.38E-09
DNA-directed RNA polymerase activity	50	22	7.58	2.90	+	1.31E-04	1.68E-03
↳transferase activity	1325	268	200.75	1.34	+	1.42E-05	2.49E-04
nucleotidyltransferase activity	78	33	11.82	2.79	+	3.80E-06	8.10E-05
single-stranded DNA binding	48	19	7.27	2.61	+	8.47E-04	7.75E-03
aminoacyl-tRNA ligase activity	43	17	6.51	2.61	+	1.65E-03	1.44E-02
↳ligase activity	227	67	34.39	1.95	+	5.13E-06	9.84E-05
ubiquitin-protein ligase activity	80	25	12.12	2.06	+	3.27E-03	2.42E-02
structural constituent of ribosome	126	38	19.09	1.99	+	4.31E-04	4.60E-03
↳structural molecule activity	506	104	76.66	1.36	+	5.19E-03	3.69E-02
mRNA binding	142	36	21.51	1.67	+	7.38E-03	4.57E-02
Unclassified	11852	1692	1795.66	.94	-	6.11E-04	6.18E-03
ligand-gated ion channel activity	135	6	20.45	.29	-	6.51E-04	6.25E-03
↳ion channel activity	336	15	50.91	.29	-	2.74E-08	6.58E-07
↳transmembrane transporter activity	832	94	126.05	.75	-	5.44E-03	3.73E-02
signal transducer activity	960	41	145.45	.28	-	7.76E-23	7.45E-21
G-protein coupled receptor activity	309	11	46.82	.23	-	3.25E-09	8.90E-08
↳receptor activity	1128	46	170.90	.27	-	8.12E-28	1.56E-25
cation channel activity	117	4	17.73	.23	-	3.78E-04	4.27E-03
voltage-gated ion channel activity	88	3	13.33	.23	-	2.72E-03	2.09E-02
phospholipase activity	86	1	13.03	.08	-	9.92E-05	1.36E-03
↳lipase activity	116	5	17.57	.28	-	1.67E-03	1.39E-02
antigen binding	104	1	15.76	.06	-	1.44E-05	2.30E-04
glutamate receptor activity	39	0	5.91	< 0.01	-	7.38E-03	4.72E-02
cytokine receptor binding	40	0	6.06	< 0.01	-	7.70E-03	4.48E-02
↳cytokine activity	109	2	16.51	.12	-	5.24E-05	7.74E-04
↳receptor binding	865	49	131.05	.37	-	3.91E-15	2.50E-13

b)

PANTHER GO-Slim Biological Process	Homo sapiens (REF) #	Client Text Box Input (▽ Hierarchy NEW! ⓘ) #	expected	Fold Enrichment	+/-	raw P value	FDR
gluconeogenesis	23	4	.32	12.53	+	4.71E-04	1.64E-02
↳monosaccharide metabolic process	76	6	1.05	5.69	+	9.15E-04	2.23E-02
↳primary metabolic process	4753	90	65.96	1.36	+	1.19E-03	2.07E-02
↳metabolic process	5878	114	81.57	1.40	+	4.54E-05	2.77E-03
regulation of carbohydrate metabolic process	23	4	.32	12.53	+	4.71E-04	1.44E-02
RNA localization	85	8	1.18	6.78	+	4.16E-05	5.08E-03
nuclear transport	123	10	1.71	5.86	+	1.52E-05	3.70E-03
nucleobase-containing compound transport	116	8	1.61	4.97	+	3.13E-04	1.27E-02
protein targeting	171	11	2.37	4.64	+	4.42E-05	3.59E-03
mRNA splicing, via spliceosome	178	8	2.47	3.24	+	4.20E-03	4.88E-02
↳mRNA processing	244	10	3.39	2.95	+	2.74E-03	3.51E-02
↳RNA metabolic process	1570	37	21.79	1.70	+	1.68E-03	2.40E-02
↳nucleobase-containing compound metabolic process	2797	60	38.81	1.55	+	6.82E-04	1.85E-02
cytoskeleton organization	404	14	5.61	2.50	+	1.98E-03	2.68E-02
↳organelle organization	1212	31	16.82	1.84	+	1.42E-03	2.17E-02
↳cellular component organization	1964	43	27.25	1.58	+	3.21E-03	3.92E-02
↳cellular component organization or biogenesis	2099	47	29.13	1.61	+	1.15E-03	2.15E-02
biosynthetic process	1745	41	24.22	1.69	+	1.26E-03	2.06E-02
nitrogen compound metabolic process	2524	58	35.03	1.66	+	1.31E-04	6.39E-03
single-multicellular organism process	1665	9	23.11	.39	-	9.50E-04	2.11E-02
↳multicellular organismal process	1684	9	23.37	.39	-	9.59E-04	1.95E-02

Figure 8: GO enrichment using of genes obtained from polysomes of MCF7 cell lysates whose TSS information was obtained from NanoCAGE 86 a) Genes containing GG and GC in the nucleotide positions 1 and 2 and b) Genes starting with U in position 1. It was observed that GG and GC occurred at the highest frequency in the NanoCAGE data obtained from polysomes and genes that began with U (T) in position 1 had the lowest frequency. The selected genes (a and b) were used as an input gene list for GO analysis and PANTHER-go slim tool was used to elucidate the molecular functions.

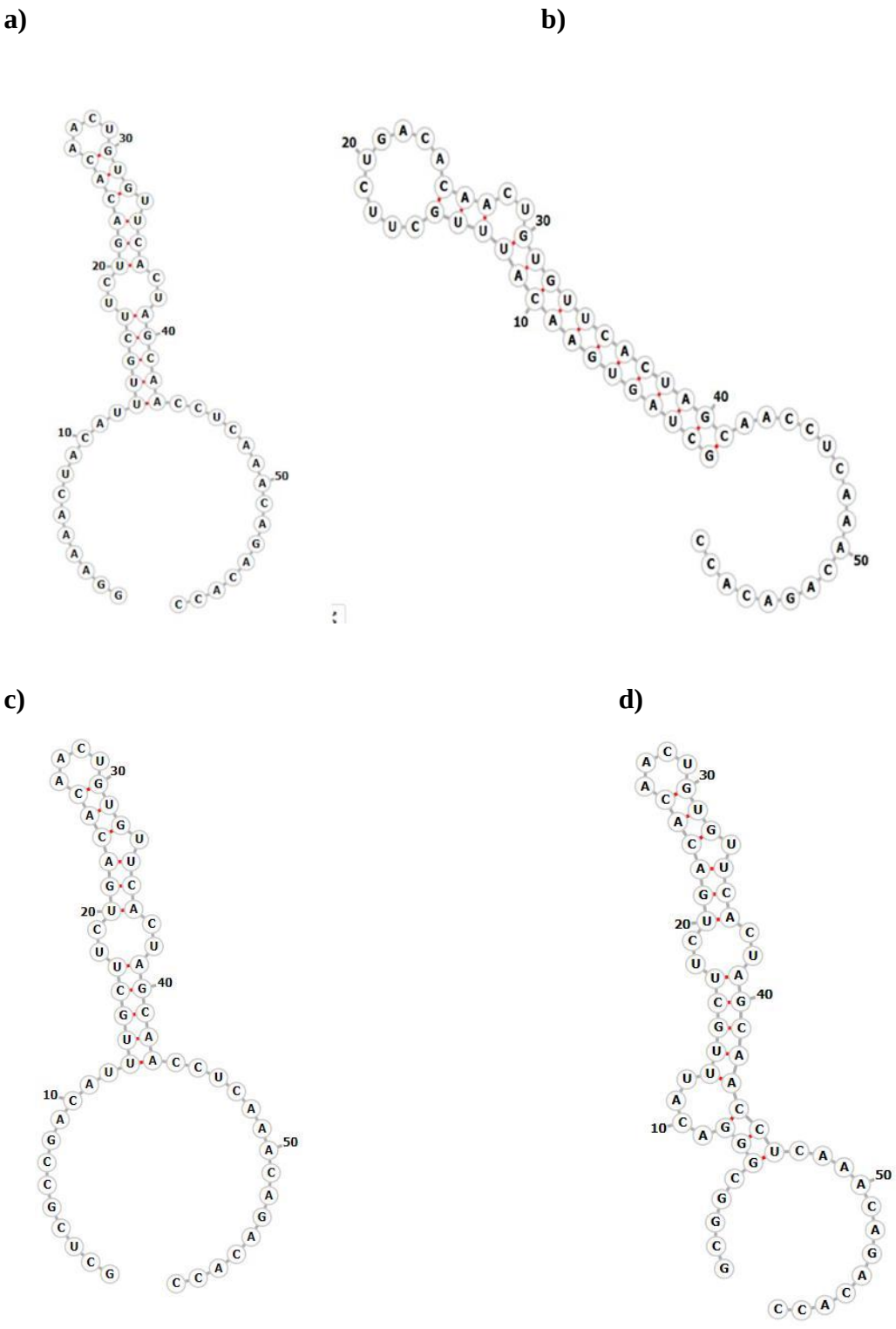


Figure 9: RNA structure prediction of model RNA containing GC percentages of E5S with the 5'TL of lnO RNA of a) 0-40% (GC%: 37.5%, GGAAAACU) b) 40-70% (GC%: 50%, GCUAGUGA) c) 70-90% (GC%: 87.5%, GCUCGCCG) and d) 90-100% (GC%:100%, GCGGCGGG)

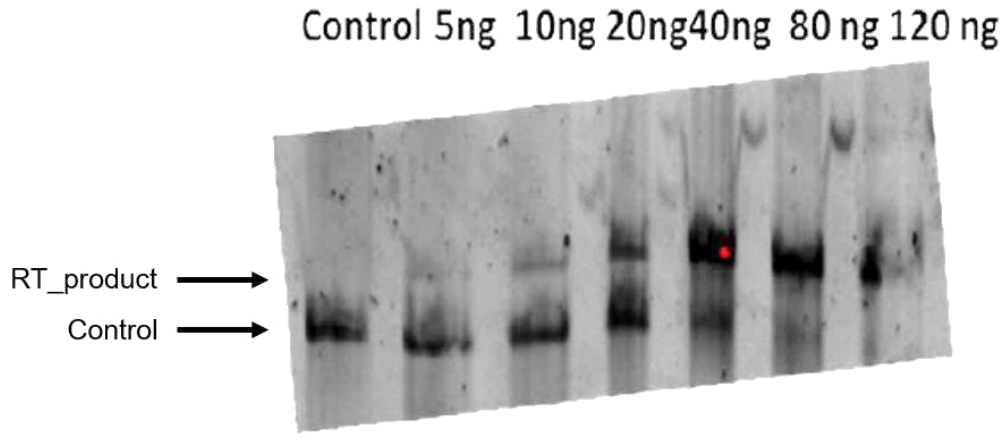
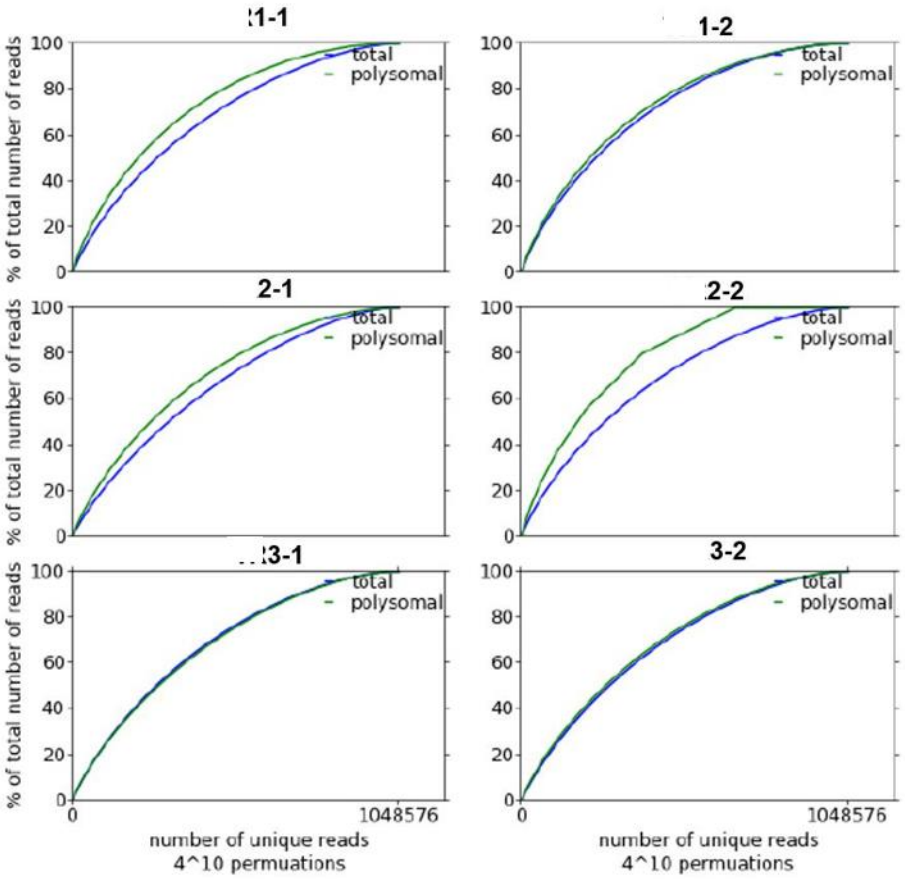


Figure 10: Quantification of the minimum amount of IVT RNA required for library preparation


1-1
1-2
2-1
2-2
3-1
3-2
total
polysomal
% of total number of reads
0
20
40
60
80
100
0
1048576
number of unique reads
4^10 permuations