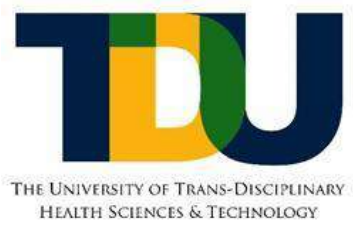
EXPLORING RIBOSOME MEDIATED TRANSLATION REGULATION IN MAMMALS

A THESIS TO BE SUBMITTED TO

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY



FOR THE AWARD OF THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY

NIVEDITA HARIHARAN

UNDER THE GUIDANCE OF

DR. DASARADHI PALAKODETI

INSTITUTE FOR STEM CELL SCIENCE AND REGENERATIVE MEDICINE

GKVK POST, BELLARY ROAD, BENGALURU - 560065

AUGUST 2023

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

DECLARATION BY THE CANDIDATE

I declare that this thesis entitled "***Exploring Ribosome Mediated Translation Regulation in Mammals***" submitted for the award of Doctor of Philosophy to THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY, Bengaluru, is my original work, conducted under the supervision of my guide ***Dr. Dasaradhi Palakodeti***. I also wish to inform that no part of the research has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged

I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Nivedita Hariharan.

Place: Bengaluru **Signature of the Candidate**

Date: 21/08/2023

Name of candidate: *NIVEDITA HARIHARAN*
Reg. No: *21117020488*
Month & Year of Admission: *November 2018*

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

CERTIFICATE

This is to certify that the work incorporated in this thesis ***"Exploring Ribosome Mediated Translation Regulation in Mammals"*** submitted by ***Nivedita Hariharan*** was carried out under my supervision. No part of this thesis has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged. I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Research Supervisor

Date: 21/08/2023

Name, designation & address details:

Dr. Dasaradhi Palakodeti

Associate Professor,

Institute for Stem Cell Science and Regenerative Medicine,

GKVK-Post, Bellary Road,

Bengaluru-560065.

NH_Thesis

ORIGINALITY REPORT

2%	2%	0%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.embopress.org Internet Source	2%

Exclude quotes	Off	Exclude matches	< 2%
Exclude bibliography	Off		

Acknowledgements

My journey through this Ph.D. has only been possible because of the unwavering support and guidance of many people - my family, my guru Dr. Dasaradhi Palakodeti, the wonderful ladies at Dolna (on-campus childcare) I dedicate this work to all of them. I am extremely grateful to my family - my husband (Venkat), my children (Samaya & Aaditya), my parents (Hariharan & Meenakshi), my brother (Sai Guru Prasath) for their omnipresent love and support for me. Special mention to Venkat, who has been my rock, in all shapes and forms, through life.

To any person embarking on a journey towards Ph.D., I would point them towards Das, as he is the best mentor anyone could ask for. I earnestly thank him for guiding me, for encouraging me, challenging me and giving me the space to develop as a problem-solver/thinker and not become just a technician. His empathy & generosity towards others - both personal and professional, his ability to interact positively both with preschoolers (including my kids) and high-level bureaucrats, and his passion for science never ceases to inspire me. I strive to be a better person, a better scientist and always make my guru proud of me.

I am very thankful to the founding members of the BLiSc community, who so thoughtfully built and shaped the on-campus child care facility - 'Dolna'. I believe women empowerment primarily needs good quality, accessible and affordable child-care. Kudos to all the ladies providing excellent care and support at Dolna, especially Ms. Teja and Ms. Thangam. They have been my backbone these past 5 years.

I would also like to thank Dr. Deepa Agashe, Dr. Sabarinathan Radhakrishnan for their guidance as my thesis committee members, Dr. Arati Ramesh, Dr. Tina Mukherjee for their support, all the members of my lab (especially Ms. Sumana Ghosh) who have been an integral part of this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 7

Synopsis

Exploring Ribosome mediated translation regulation in mammals

Introduction (*Chapter One of the thesis*)

Ribosomes, which are the hubs of protein synthesis (translation) have been typically viewed as passive components in translation regulation, with interest centred on intrinsic features of the mRNAs and *trans*-acting factors like RNA Binding Proteins (RBPs), non-coding RNAs. Basically composed of ribosomal RNAs (rRNAs) and proteins (RPs) forming two subunits (large subunit, LSU; small subunit, SSU), ribosomes possess domain-specific features. For instance, eukaryotic cytosolic ribosomes have nearly doubled in size than those seen in bacteria and archaea, which is mainly due to an increase in the number of RPs and the presence of multi-nucleotide insertions in eukaryotic rRNAs termed 'Expansion Segments' (ESs) (Veldman et al. 1981; Ware et al. 1983; Clark et al. 1984; Hassouna, Michot, and Bachellerie 1984; Brimacombe 1981). Therefore it is conceivable that changes in ribosomal protein composition and post-translational modifications, rRNA variations and modifications can give rise to a heterogenous pool of ribosomes, which can regulate and translate different subsets of mRNAs ('Ribosome Filter hypothesis') (Mauro and Edelman 2002). With the idea of 'ribosomal heterogeneity' gathering momentum of late, the traditional perception of ribosomes in a merely catalytic role in translation is changing. This notion is fairly nascent and still under debate, with only a handful of studies showing that ribosomes can indeed regulate translation (Díaz-López et al. 2019; Gómez Ramos et al. 2016; Fujii et al. 2018; Shankar et al. 2020).

Although first described at least four decades back, ESs are still a black box with very little known about their function. In the ribosome structure, ESs are present on the surface without perturbing the highly conserved core elements (Ben-Shem et al. 2011; Chandramouli et al. 2008a; Petrov et al. 2015; Anger et al. 2013a) and were initially considered to be non-functional insertions. These multi-nucleotide insertions of variable length are present at conserved positions along the eukaryotic rRNAs when compared to the bacterial rRNAs. While ESs of the SSU are of similar lengths across eukaryotes, those of the LSU can be variable in length. For instance, a LSU ES called ES7L is only around 100 nucleotides in *Saccharomyces cerevisiae* while it is more than 800 nucleotides in *Homo sapiens* (Parker et al. 2015). But not all LSU ESs necessarily show a size increase that correlates with organismal complexity. Further, even though 14 ESs from the SSU and 41 ESs from the LSU have been reported overall, some of them could be specific to subsets of clades (Yokoyama and Suzuki 2008). The species-specific diversification of ESs and their role in regulating translation has largely remained unaddressed. The nature and known potential functions of ESs are discussed in depth in

our review "The story of rRNA expansion segments: Finding functionality amidst diversity" published in WIREs RNA with DOI: https://doi.org/10.1002/wrna.1732 To understand the variability of the different LSU ESs across eukaryotes, we compared 28S rRNA sequences from species representing various clades across eukaryotes. Given that the size increase of the SSU rRNA is not as drastic among eukaryotes as that of the LSU rRNA, we decided to focus on the latter. From our comparison, we observed a LSU ES called ES30L, specific to endothermic vertebrates (mammals and birds). ES30L is highly GC-rich in *H. sapiens*, very flexible, and positioned near the ribosome's t-RNA and mRNA exit site. These aspects prompted us to focus on this rRNA segment and study its interactome, which could help us gain insights into its function.

Therefore, the primary objectives of our study have been -

1. To compare ES30L across eukaryotes.
2. To examine the potential of ES30L to bind to protein-coding transcripts in *H. sapiens*.
3. To investigate the protein interactome of ES30L in *H. sapiens*.

Materials and Methods (*Chapter Two of the thesis*)

The materials and methods used in this study are elaborated in the manuscript "ES30L: A large subunit rRNA expansion segment enriched in mammals and birds" published in bioRxiv with the DOI: https://doi.org/10.1101/2022.09.30.510333. This manuscript is currently under peer review.

Results

The results of this study will be elaborated over five chapters as mentioned below.

Comparison of ESs across eukaryotes (*Chapter Three of the thesis*)

We mined 28S rRNA sequences from public databases like Silva LSU r132 (26) and NCBI Nucleotide and constructed a multiple sequence alignment (MSA) using them. The major ESs in the human LSU rRNAs were marked in the MSA based on the coordinates taken from Parker et al., (Table 1) and were used to extract the concurrent regions from all the other species used in the MSA. We also mapped these expansions on the structure of the human ribosome. Our observations from this analysis are summarized below -

1. ESs are present on the ribosomal surface in both subunits (Figure 1)
2. ESs of the LSU such as ES5L, ES24L and ES27L show variability in length and/or base composition across bilaterians.
3. Other ESs such as ES9L, ES12L, ES26L are similar in length and base composition.
4. Some expansions contain insertions that are specific to a clade(s). For instance, ES30L, ES39L and ES15L have insertions present only in endothermic

vertebrates. Arthropods have some expansions that are unique to their clade such as those in ES31L. Some expansions are slightly longer in invertebrates than in vertebrates such as ES19L and ES20L.

The function of almost all of these expansions and the relevance of species-specific extensions are unknown. We focussed on ES30L, a highly GC-rich ES with mammal and bird-specific expansion that is an extension of L1-stalk.

Table 1 : Co-ordinates of the LSU ESs in *H. sapiens*.

ES	Co-ordinates			GC%
	U13369_7935-12969 (taken from Parker et al., 2015)	**Mapping in PDB:4UG0**	**Mapping in NR_003287.4**	
ES5L	114-156	114-159	114-159	78.57
ES7L	465-1265	474-1278	474-1278	83.75
ES9L	1384-1487	1397-1500	1397-1500	80.58
ES10L	1682-1711	1695-1724	1695-1724	75.86
ES12L	1793-1829	1806-1842	1806-1842	77.78
ES15L	2075-2256	2088-2269	2088-2269	83.98
ES19L	2439-2494	2452-2507	2452-2507	74.55
ES20L	2538-2575	2551-2588	2551-2588	70.27
ES24L	2685-2705	2698-2718	2698-2718	85
ES26L	2751-2764	2764-2777	2764-2777	61.54
ES27L	2875-3586	2888-3607	2888-3607	86.64
ES30L	3955-4013	3976-4035	3976-4035	91.8
ES31L	4057-4124	4079-4154	4079-4154	83.58
ES39L	4698-4905	4729-4940	4729-4940	82.61
ES41L	4983-4996	5018-5031	5018-5031	69.23

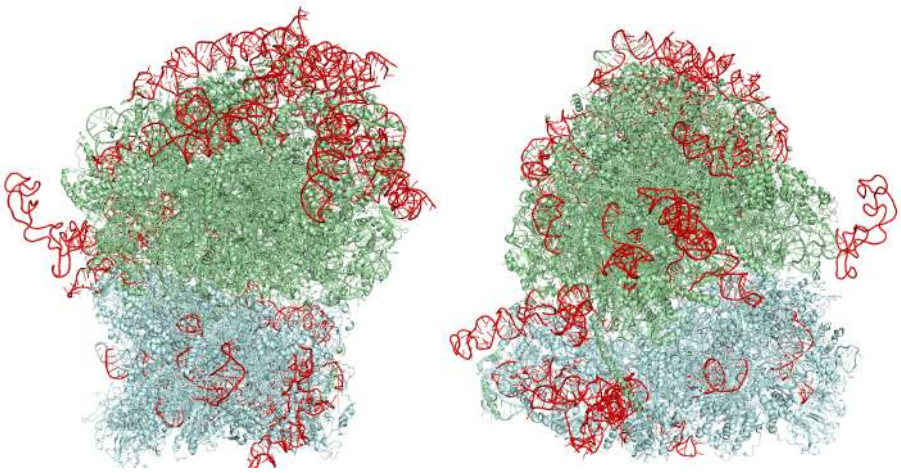


Figure 1: This figure shows the human ribosome structure (PDB 4UG0) with all the major ESs marked in red, LSU in green and SSU in cyan. ***(A)*** Front View ***(B)*** Back View

ES30L is a large subunit ES specific to endothermic vertebrates (*Chapter Four of the thesis*)

To compare the ES30L region across eukaryotes, we extracted the region corresponding to nucleotides 3975-4035 in the human 28S rRNA sequence from the MSA (Figure 2A). We also manually edited this section of the alignment to make it better suited to analyse the extent of conservation. This region showed the highest expansion in mammals (Figure 2C) and was highly GC-rich (~92% in *H. sapiens*). We observed that amphibians and arthropods also had short expansions in this region, but exhibited high variability in length within these two groups. Further, we also observed that while ES30L was GC-rich in chordates, it had among the lowest GC-richness in arthropods (Figure 2D). We used the edited alignment to calculate Gap-Adjusted Shannon's Entropy (GASE), which is plotted in Figure 2E and noted that even though ES30L was well conserved within both clades, the conservation was better among mammals than birds. Mapping this region on the structure of the human ribosome showed that ES30L extrudes as a disordered stretch from the L1-stalk and is usually not completely resolved in cryo-EM because of its tentacle-like nature (Anger et al. 2013a; Chandramouli et al. 2008a; Khatter et al. 2015; Natchiar et al. 2017). It is usually reconstructed with the help of secondary structure modelling (Figure 2B, 1C). Given its GC-richness and position on the ribosome, we hypothesized that it could bind to mRNAs and proteins and regulate translation.

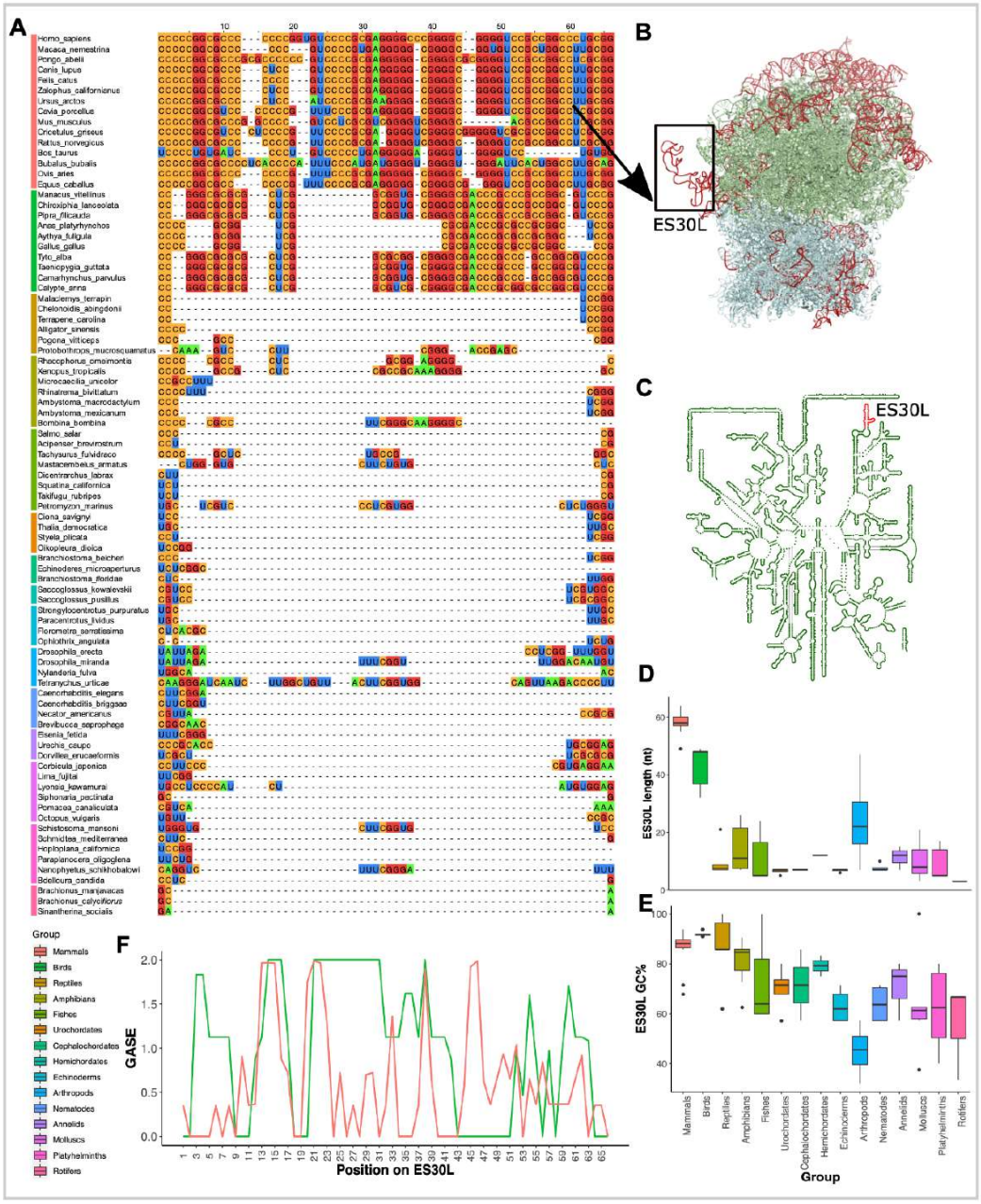


Figure 2: ES30L is a 28S rRNA expansion segment that is expanded largely in mammals and birds. This figure shows the comparison of the ES30L region from different clades (denoted by the coloured bars on the left of the MSA) across bilateria

(A) Multiple Sequence Alignment (MSA) of the ES30L region from 85 organisms across Bilateria.

(B) Front view of the human 80S ribosome (PDB: 4UG0 10.2210/pdb4UG0/pdb) (Khatter et al. 2015). The boxed region on the ribosome is ES30L and corresponds to the rRNA stretch shown in the MSA. It is important to note that the tertiary structure of ES30L is modelled based on both

secondary structure prediction and cryo-EM data and is not completely resolved because of its flexibility. The ribosome structure was rendered using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).
(C) Secondary structure of the *H. sapiens* 28S rRNA with the ES30L highlighted in red, created using the RiboVision web server (Bernier et al. 2014). The coordinates of ES30L and the other ESs were taken from Parker *et al*., (Parker et al. 2015).
(D) & (E) Box plots displaying the lengths and the GC% of the ES30L region among the sampled organisms across bilateria.
(F) Gap Adjusted Shannon's Entropy (GASE) plot for the ES30L region from mammalian (red) and bird (green) sequences shown in the MSA.

ES30L possesses complementarity to many protein-coding transcripts (*Chapter Five of the thesis*)

We developed a computational pipeline to probe the extent of complementarity prevalent between ES30L and protein-coding transcripts from *H. sapiens*. For this analysis, we used a 'sliding window' approach and extracted continuous overlapping 7 to 15 nucleotide stretches from ES30L. Transcript regions that exhibited complementarity (with only canonical Watson-Crick pairing) to these ES30L stretches either with zero or one non-terminal mismatch were considered for further analysis. Regions other than ESs in the 28S rRNA (termed as core segments or CSL) were used as a control to delineate the specificity of ES30L to transcripts. Our key observations from this set of analyses are :

1. Overall, ES30L had 7.25% higher pooled complementary hits with no mismatch to transcripts than CSL (paired Wilcoxon signed-rank test, $p < 0.05$; Figure 3A), with 84.54% of the transcripts possessing at least one stretch perfectly complementary to ES30L.
2. ES30L exhibited 36.12% more complementarity than CSL at lengths of 9 nucleotides or higher (Figure 3B). This was even higher when one mismatch was allowed, as more than 50% of the stretches with lengths of 7 or 8 got extended to 9 nucleotides or longer.
3. Complementary transcript stretches that are longer than 15 nucleotides were sparse with 1148 hits from 560 genes if one mismatch was allowed. This gene set included many that are involved in developmental pathways and neurogenesis among others.
4. Around two-thirds of the complementarity observed between ES30L and transcripts stemmed from the latter half of the ES (after position 25; Figure 3C), with the skew more pronounced for longer complementary stretches. With CSL, the complementarity observed was uniform across the segments (Figure 3D).
5. The density of complementary stretches in the transcripts was nearly three times higher in 5' UTR when compared to CDS and 3' UTR (Figure 3A), with an

enrichment around the start codon (Figure 3E-F). The differential density among the transcript sectors was not stark in the case of CSL.

6. Many transcripts possessed more than one stretch that is complementary to ES30L, with 425 of them containing more than 50 such stretches. Gene ontology analysis of this set showed an enrichment of genes involved in neuronal development and differentiation.

All the trends observed here hold true even when one mismatch was allowed in the complementarity analysis. Our results so far suggest that ES30L has a higher propensity to interact with transcripts, especially around the start codon.

A

Category	No. Of Complementary Hits	
	ES30L	CSL
All mRNA sectors	525190	489677
5' UTR (per 100nt)	1.0347	0.4759
CDS (per 100nt)	0.3437	0.2789
3' UTR (per 100nt)	0.2128	0.2826

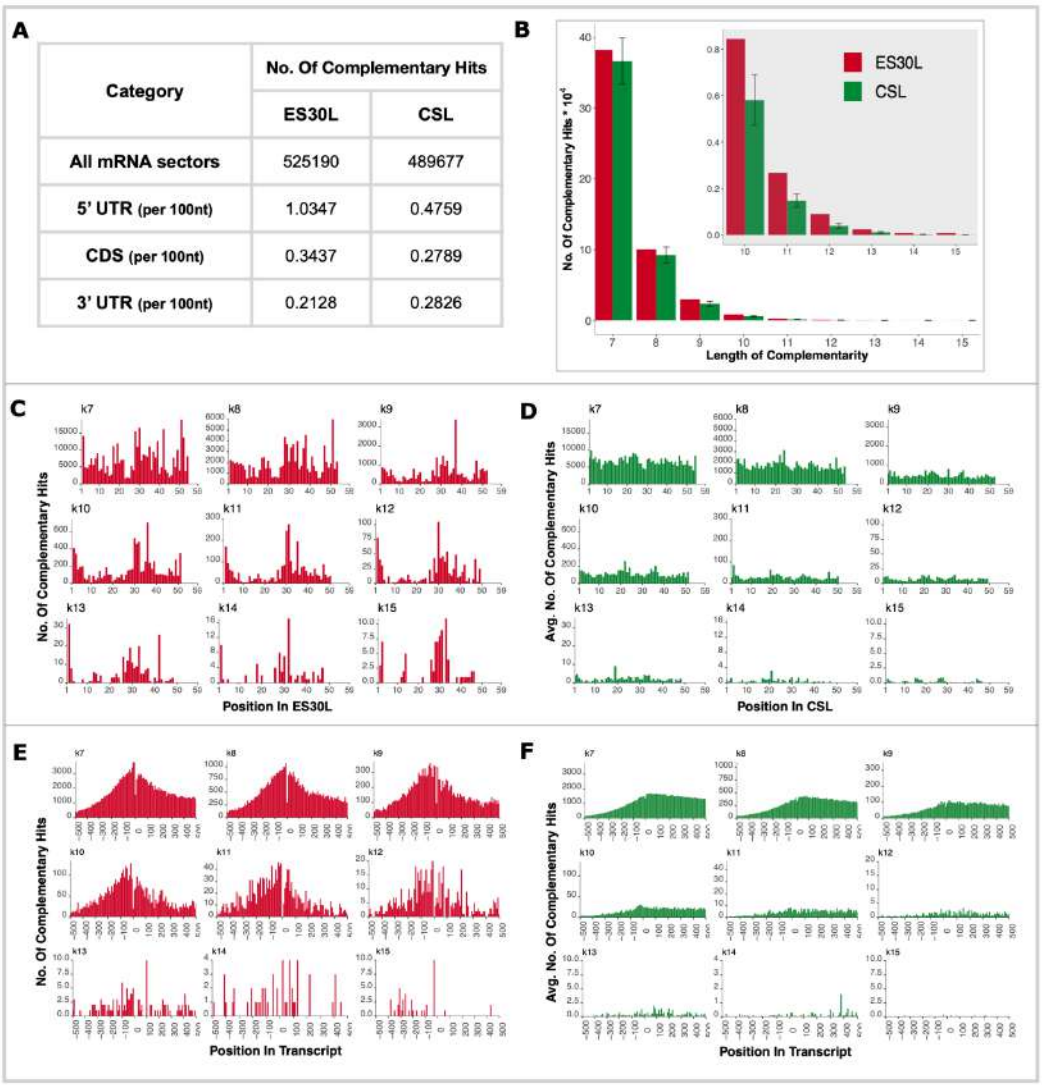


Figure 3: ES30L (from human 28S rRNA) possesses complementarity to many protein-coding human transcripts. This figure shows the summarized comparison between ES30L and CSL (core segment from the LSU; see *Methods*) regions with regard to their contiguous complementarity (zero mismatch) to protein-coding transcripts. In the case of CSL segments, all the data shown here is the average value from the ten segments used in the analysis.
(A) Table showing the total number of complementary hits in all mRNA sectors (5' UTR, CDS, 3' UTR) when matched to ES30L and CSL segments. The table also shows the density of complementary hits from each mRNA sector, which is the total complementary hits from that sector divided by the cumulative length of the sector from all the transcripts.
(B) Bar plot showing the number of complementary stretches ranging from a length of 7 to 15 or more nucleotides for ES30L (red) and CSL (green) fragments. The inset in this panel is the data for 10 nucleotides and higher, magnified for better clarity.
(C) & ***(D)*** Histogram showing the number of complementarity stretches to transcripts, arising from each position on ES30L (red) and CSL (green) fragments respectively.
(E) & ***(F)*** Histogram showing the number of complementarity stretches arising from each bin (bin size used here is 10) on the protein-coding transcripts, with ES30L (red) and CSL (green) fragments. The '0' on the x-axis denotes the start codon (TSS) and the negative numbers upstream of it, denoting positions in the 5' UTR. Only the 500 bases flanking the TSS have been shown in these plots. The complete histogram of this data is provided in Supplementary Figures S2-IC, S2-ID, S2-IIE, and S2-IIF.

ES30L could bind to protein-coding transcripts (*Chapter Six of the thesis*)

To test whether ES30L can interact with transcripts, we pulled down mRNAs from HEK293T cell lysates using a biotinylated ES30L mimic followed by high-throughput RNA sequencing. We used a random oligonucleotide of the same length as ES30L as a control in the *in vitro* pulldown assay. Concurrently, we re-analysed a published RNA-RNA interactome dataset (Aw et al. 2016) from HeLa cells to probe for ES30L binding to transcripts *in vivo*. Data from these analyses was overlaid with our bioinformatic predictions to check the extent of complementarity in these transcripts. Our results are summarized below :

1. Our sequencing data post the *in vitro* pulldown showed that 1550 transcripts were more than two-fold enriched ($q<0.05$) in the ES30L fraction over input (Figure 4A). Out of these, 100 transcripts showed a 50% higher fold enrichment in the ES30L fraction than CSL, when compared to input (Figure 4B). Gene set enrichment analysis revealed that many of these transcripts are involved in various aspects of RNA metabolism and translation (Figure 4C).
2. More than 90% of the transcripts selectively enriched in the ES30L fraction from our pulldown included at least one stretch (of any length) complementary to ES30L either with or without a mismatch.
3. From the *in vivo* RNA-RNA interactome dataset, we retrieved 193 inter and intra-molecular interactions involving the ES30L region, out of which 143 were with

protein-coding transcripts (Figure 4D). Nearly 65% of the 143 interacting transcript stretches encompass at least one complementary stretch to ES30L (Figure 4E).

4. The 143 *in vivo* interactions involve 118 genes, 25 of which are also present in our *in vitro* pulldown data. This overlap could vary because these data are from two different cell lines.
5. Since interactions involving RNAs can also be structure based, we further probed for putative IRES elements (Yang et al. 2021; Zhao et al. 2020) in the ES30L interacting mRNA stretches from the *in vivo* interactome data. Genes such as RPS11, FTH1 and NHP2 include ES30L-binding stretches which are part of predicted IRES elements. These regions have multiple short stretches with complementarity to ES30L.
6. Further, we did *in silico* analysis for putative structural motifs in 5' UTR stretches containing complementarity to ES30L from the transcripts in our pulldown. The results indicated that around 4-14% of the transcripts could form consensus motifs (Figure 4F), which contain binding motifs for proteins such as HNRNPs, SRSFs, DHX9, PTBP. Interestingly, binding motifs for these proteins are also present in ES30L.

Based on our observations so far, we hypothesize that ES30L could bind to complementary transcript regions, some of which may be part of IRES structures.

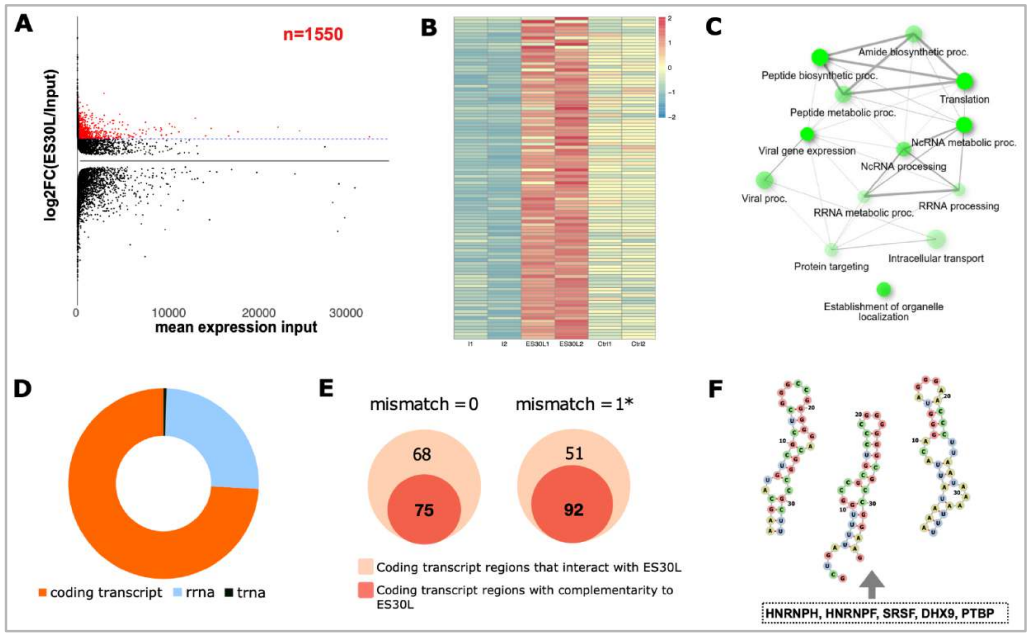


Figure 4: ES30L can potentially interact with protein-coding human transcripts. This figure summarises data that suggests that such an interaction is possible.

(A) MA plot showing the distribution of the fold change of all the transcripts between the ES30L fraction and the input, relative to the mean expression in the input. The points highlighted in red represent the 1550 transcripts that are more than two-fold enriched (q<0.05) in the ES30L fraction over the input.
(B) Heatmap showing the expression profiles (normalised read count) of the 100 transcripts that are more than two-fold enriched in the ES30L fraction over the input fraction and 50% higher fold change than the random fragment in the pulldown.
(C) Network plot showing the gene categories over-represented in the 100 transcripts that are more than two-fold enriched in the ES30L fraction over the input fraction and 50% higher fold change than the random fragment in the pulldown.
(D) Pie chart showing the distribution of the RNAs that interact with the ES30L region reported in Aw *et. al. (Aw et al. 2016)*.
(E) Venn diagram showing the overlap between the number of transcript regions interacting with the ES30L region reported in Aw *et. al.* *(Aw et al. 2016)* and the number of those transcript regions harbouring at least one stretch complementary to ES30L.
(F) RNA motifs obtained from *de novo* secondary structure prediction. The sequences used in this prediction were taken from the 100 transcripts from our pull-down and were 5' UTR regions which include stretches complementary to ES30L flanked by 20 nucleotides.

ES30L can interact with various RNA Binding Proteins (*Chapter Seven of the thesis*)

To identify the proteins that interact with ES30L, we did an *in vitro* pulldown of proteins from HEK293T cells using the same biotinylated RNA mimics as before, followed by mass spectrometry. We also re-analysed published CLIP-seq datasets (Van Nostrand et al. 2020; Porter et al. 2021) from K562 or HCT116 cells for proteins that were selectively enriched in the ES30L fraction in our pulldown to probe their interaction with ES30L *in vivo*. The results are summarized below -

1. In total fifty-eight proteins (Figure 5A) were more than 1.5 fold enriched (q<0.05) in the ES30L fraction when compared to input, which included various categories of RBPs such as a)Ribosome biogenesis factors (NKRF, NCL) b)Helicases (DDX21, DHX30, DHX9) c)G-quadruplex binding proteins (hnRNPF) d)Splicing factors (SRSF1, SRSF2, SRSF6, U2AF2) e)Ribosomal proteins f)Proteins involved in translation (PCBP1, HNRNPK, IGF2BP1, SRP14, RBM4).
2. Overall, gene ontology analysis of these proteins showed that many of them are involved in various aspects of RNA metabolism (Figure 5B).
3. Out of the 58 identified proteins, 42 were part of a published ribo-interactome (Simsek et al. 2017)(Figure 6A).
4. In our analysis of the CLIP-seq data DHX30, NPM1 showed enrichment in reads mapping to ES30L relative to the input, suggesting that they could potentially interact with ES30L *in vivo*. Other proteins like PCBP1, SRSF1, IGF2BP1, U2AF2 or HNRNPK did not show such enrichment in this dataset (Figure 6B).

5. Since *in vivo* molecular interactions could be influenced by their micro-environment, we compared CLIP-Seq data for PCBP1 from two cell types (K562 and HCT116). Our analysis indicated a preferential mapping of reads towards the latter half of ES30L in one cell type and not the other, hinting at the possibility that interactions of RBPs with ES30L could be context-dependent (Figure 6C).

Based on our *in silico* analysis, many proteins from our *in vitro* pulldown contain multiple RNA binding domains and have motifs on both ES30L and complementary transcript stretches. Therefore, interactions between ES30L and transcripts may be mediated by RNA-binding proteins.

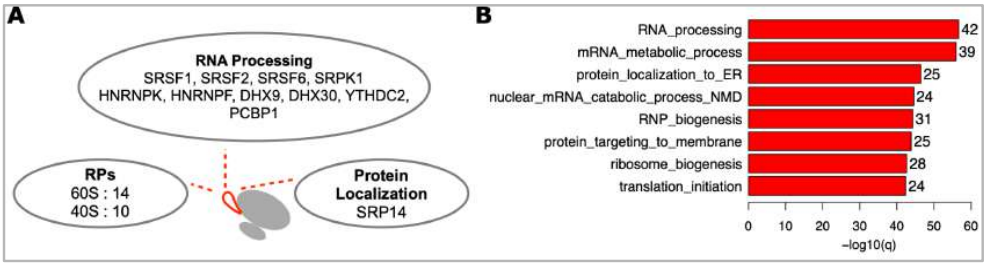


Figure 5: This figure gives an overview of the potential protein interactome of ES30L
(A) Illustration depicting the broad protein categories that were enriched more than two-fold (q<0.05) over the total protein fraction in our pulldown.
(B) Bar plot showing the GO (biological process) categories of the proteins that were enriched more than two-fold (q<0.05) over the total protein fraction in our pulldown.

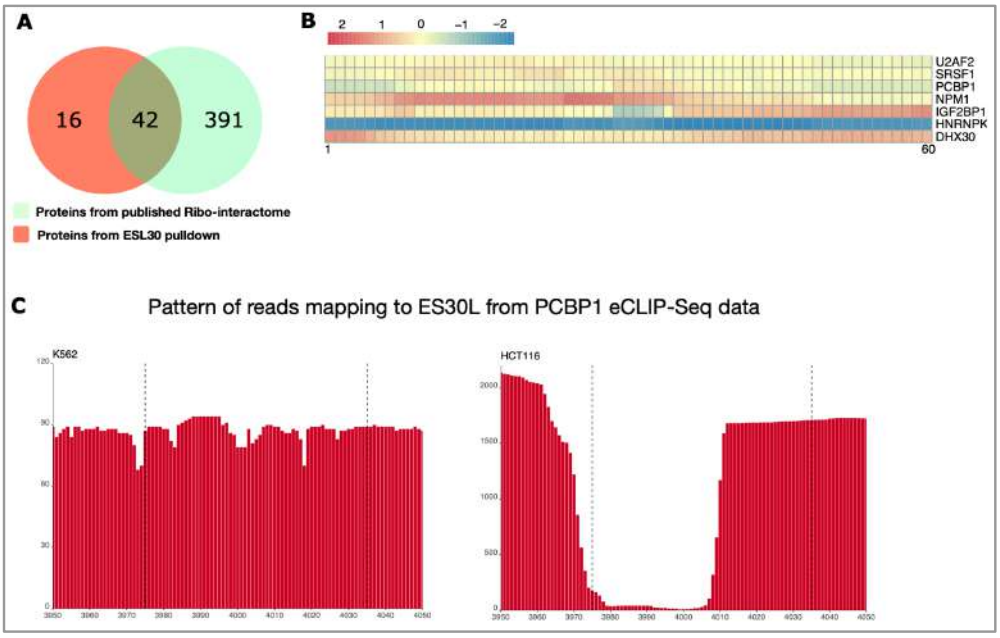


Figure 6: ES30L could interact with RBPs.

(A) Venn diagram showing the overlap between the proteins identified in our pulldown experiment and a published ribo-interactome ((Shi et al. 2017)).
(B) Heatmap showing the enrichment of reads (for the protein fraction over their size-matched input) in the ES30L region from published CLIP-seq data ((Van Nostrand et al. 2020)) for eight proteins that were identified in our pulldown.
(C) & (D) Coverage plot showing the read depth pattern across the ES30L stretch (U13369:3955-4013; Supplementary Table S1C) from two different studies - Van Nostrand et al. 2020 (*left*; K562 cell line) and Porter et al. 2021 (*right*; HCT116 cell line).

Conclusion and Future Directions (*Chapter Eight of the thesis*)

Some of the key findings from my work are:

1. ESs (especially of the LSU) may not be uniform (in terms of length and base composition) across eukaryotes and may contain species-specific variations.
2. ES30L has expanded the most in mammals, with the second-highest expansion in birds. It is well conserved in both these clades.
3. This segment has a high degree of complementarity to several protein-coding mRNAs, which are highly clustered around the start codon, with the highest density in the 5' UTR.
4. Our *in vitro* pulldown combined with our analysis of other data shows that ES30L can interact with mRNAs, which could involve base-pairing and structural 5' UTR motifs.
5. A similar *in vitro* pulldown of proteins indicates that ES30L could bind proteins that are involved in various aspects of RNA metabolism. Interestingly, some of these proteins have been shown to be important for cap-independent translation and based on our *in silico* analysis, also possess binding motifs on both ES30L and 5' UTR stretches in mRNAs from our RNA pulldown.

In summary, we think that ES30L can interact with secondary structural elements in the 5' UTR, RBPs and influence translation initiation, although such an influence could be either stabilising or inhibitory. This is one possible role of ES30L, though it could be involved in other cellular processes too. Our hypothesis is very difficult to test *in cellulo* using a knock-out strategy because rRNA is present as tandem repeats across multiple chromosomes in the human genome. Future strategies to establish the function of ES30L *in vivo* could be to design highly specific ES30L inhibitors that prevent all its interactions or knock-in the segment in to rRNA from other species. Considering that nothing is known about most ESs, we think it is important to develop techniques that are tailored to these enigmatic segments, as it could widen our understanding of ribosomes in general.

List of Publications

1. **Hariharan, N.**, Ghosh, S., & Palakodeti, D. (2023). **The story of rRNA expansion segments: Finding functionality amidst diversity.** *Wiley Interdisciplinary Reviews. RNA*, *14*(1), e1732.

2. **Hariharan, N.**, Ghosh, S., Nallan, A. N., Ramesh, A., Agashe, D., & Palakodeti, D. (2022). **Expansion Segment ES30L enriched in birds and mammals can potentially regulate protein synthesis.** *bioRxiv* 2022.09.30.510333; https://doi.org/10.1101/2022.09.30.510333

3. Sakr, R., Cattenoz, P. B., Pavlidaki, A., Ciapponi, L., Marzullo, M., **Hariharan, N.**, Mukherjee, T., & Giangrande, A. (2022). **Novel cell- and stage-specific transcriptional signatures defining Drosophila neurons, glia and hemocytes.** *bioRxiv* 2022.06.30.498263; https://doi.org/10.1101/2022.06.30.498263

4. Cattenoz, P. B., Sakr, R., Pavlidaki, A., Delaporte, C., Riba, A., Molina, N., **Hariharan, N.**, Mukherjee, T., & Giangrande, A. (2020). **Temporal specificity and heterogeneity of Drosophila immune cells.** *The EMBO Journal*, *39*(12), e104486.

Chapter 1

Introduction

1.1. Brief History of Ribosomes

Ribosomes are the hubs of protein synthesis and are ubiquitous to all domains of life. The serendipitous discovery of ribosomes occurred in the early 1940s when a lipid and ribonucleoprotein-rich fraction termed the 'microsome', capable of protein synthesis was isolated [1–4]. Microsomes were first visualized by electron microscopy and were reported to contain spherical ribonucleoprotein particles, which were later named 'Ribosomes' [5,6].

Figure 1.1 - Components of Ribosome

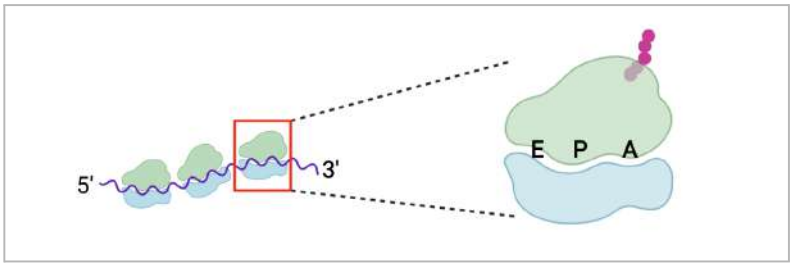


Figure 1.1 - The ribosomal LSU and SSU are colored green and blue respectively, with the emerging nascent polypeptide represented in pink. 'E' refers to tRNA exit site, 'P' to peptidyl transfer site and 'A' acceptor site.

Ribosomes contain a common core set of ribosomal proteins (RPs) and three or four RNAs (rRNAs) spread across two subunits (the large subunit or LSU and the small subunit or SSU; Table 1). For many years, work on the ribosomes was focused on RPs, since they were thought to be the catalytic component, whereas the rRNAs were relegated to the role of a scaffold for the RPs. The establishment of rRNA as the primary catalytic component of the ribosome came through the discovery of ribozymes, along with various other studies in the 1970s and 1980s [7–16]. The ribosomal subunits are assembled at the start codon on the mRNA to form the 80S ribosome, which then translocates towards the 3' end to synthesize a polypeptide (Figure 1.1). In the translating 80S ribosome, the LSU catalyzes peptide bond formation and forms the polypeptide exit tunnel, while the SSU associates with mRNA during initiation complex formation and mediates tRNA–mRNA interaction [17,18].

1.2. Ribosomes and translation regulation

Even with such important core functions in translation, no regulatory role was ascribed to the ribosome for a long time. Translation regulation, which determines the protein repertoire of a cell, has emerged as one of the key mechanisms for fine-tuning gene expression [19,20]. Typically, the *cis*-regulatory elements in mRNAs and the *trans*-acting translation factors have been considered to be the main drivers of translation regulation, with the ribosomes relegated to the role of a passive participant in this process (Figure 1.2.1). However, this perspective has been changing with studies indicating that ribosomes could regulate translation [21,22].

Figure 1.2.1 - Overview of translation regulation

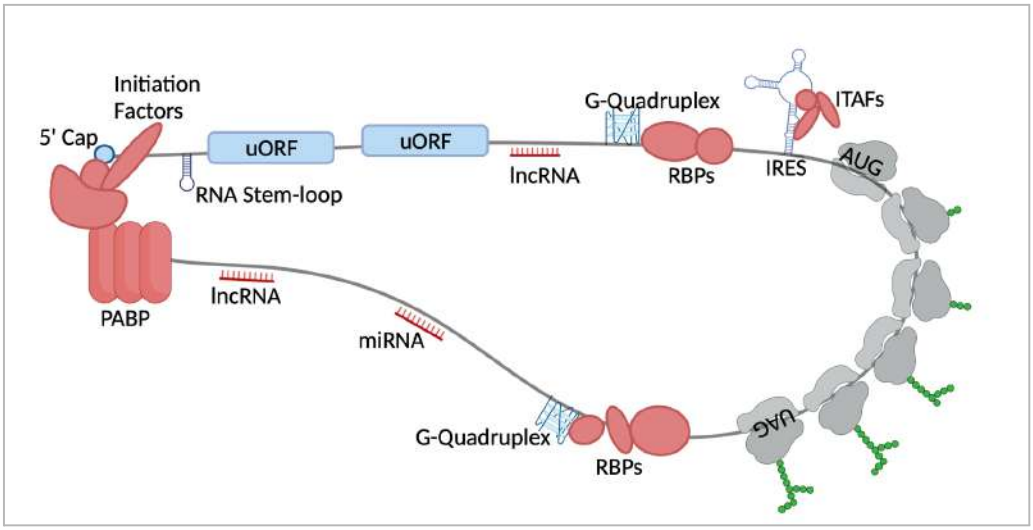


Figure 1.2.1 - Ribosomes are colored gray in this illustration, with the *cis* mRNA components in blue and the *trans* acting factors (proteins or small RNAs) in red.

Various studies have shown that the expression of the ribosomal components can be specific to a tissue or a developmental stage, bolstering the idea of 'specialized ribosomes' [23–28]. As an extension of this idea, Mauro and Edelman proposed that ribosomes may regulate translation of mRNA subsets via differential binding in their 'Ribosome Filter Hypothesis' [29]. The distinctive binding of the mRNAs to the ribosome could possibly occur because of variations either in RPs and/or rRNA (Figure 1.2.2). The first evidence of specialized ribosomes showed that those with RPS25 or RPL10a but not RPL22, preferentially translate different subsets of mRNAs [30]. One striking example of a tissue-specific phenotype caused by defects in various RPs is Diamond Blackfan Anaemia (DBA) [31]. DBA includes multiple congenital defects such as bone marrow failure, predisposition to tumors, craniofacial

abnormalities and heart problems that depend on underlying RP that is mutated. One explanation for this is that different tissues require 'specialized ribosomes' that include certain RPs to translate specific transcripts.

Figure 1.2.2 - Can Ribosomes regulate translation?

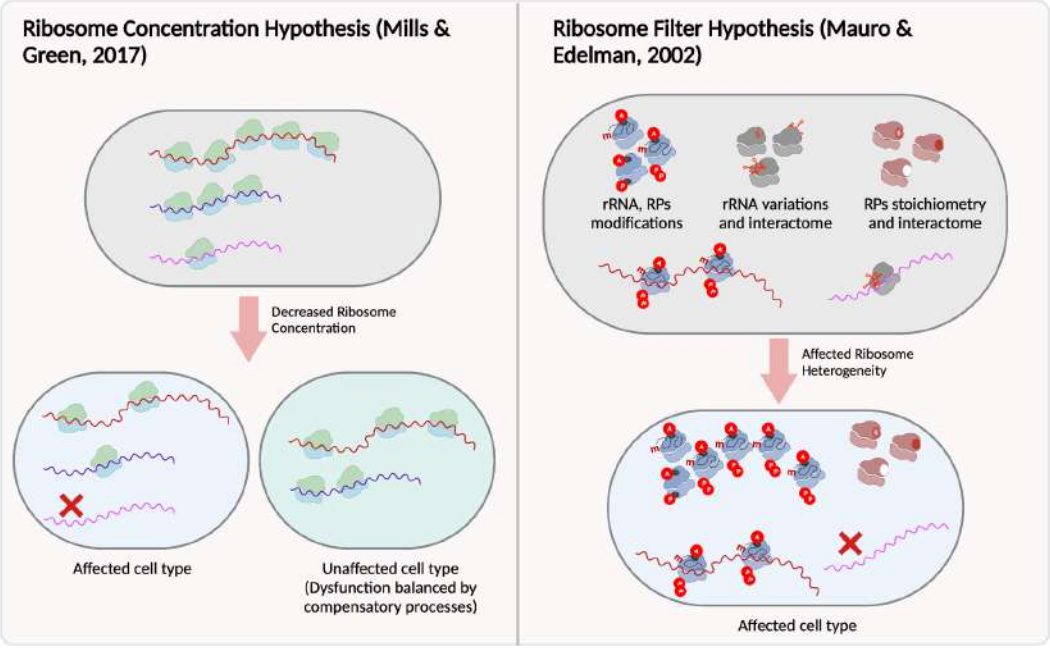


Figure 1.2.2 - This figure summarizes the current views on the way ribosomes can lead to tissue-specific phenotypes seen in Ribosomopathies. One perspective (*left*) is that of homogenous ribosomes, whose paucity can lead to poor translation of specific mRNAs that are critical only in certain tissues. The other view (*right*) is that of heterogeneity in ribosomes which can impact translation of subsets of mRNAs.

Alternatively, another hypothesis ('Ribosome Concentration Hypothesis') considers that a decreased ribosome concentration could lower global translation, with some transcripts more sensitive to the perturbation than others, causing the tissue-specific phenotype [32] (Figure 1.2.2). But this theory fails to explain the prevalence of distinct sets of congenital defects associated with different RP mutations. Although these contrasting views (Ribosome Filter Hypothesis vs. Ribosome Concentration Hypothesis) are yet to be resolved, the two scenarios could co-exist and would be an interesting line of future investigation.

1.3. Evolution of ribosomes

Mammalian ribosomes have nearly doubled in size compared to the bacterial ones (Figure 1.3) [33]. This size increase is attributable to an expansion in the number and

size of both RPs and rRNAs (Table 1). The advent of sequencing technology and the insight that rRNA is the catalytic center for protein synthesis spurred many groups to study the conservation and evolutionary differences of rRNA among different species [34–37].

Figure 1.3 - Evolution of ribosomes

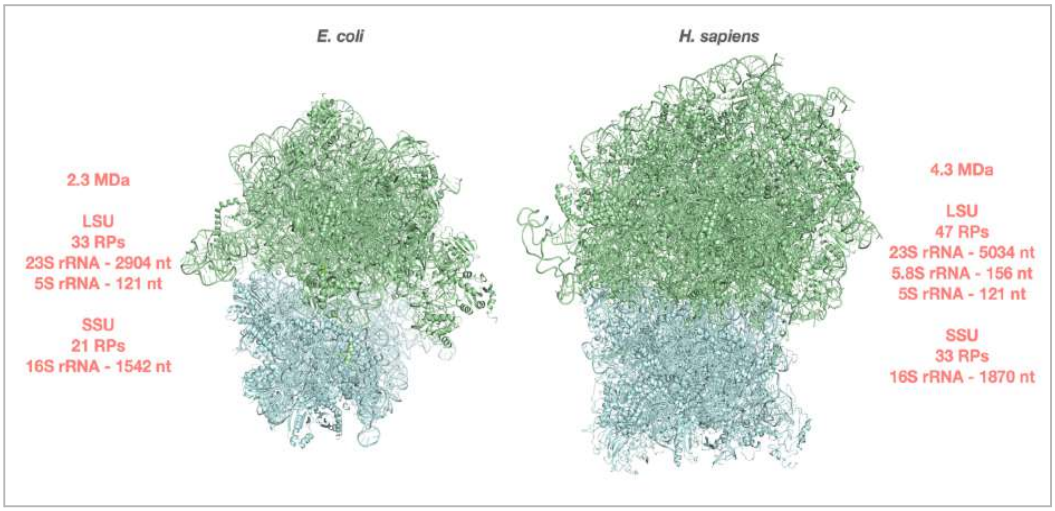


Figure 1.3 - The bacterial 70S ribosome (PDB: 7N1P) and the human 80S ribosome (PDB: 4UG0) structures used in this figure were rendered using PyMOL.

It was noted that the 25S/28S eukaryotic rRNA was longer than the 23S bacterial rRNA. For instance, the 23S rRNA in *Escherichia coli* was 2904 nt long, while the 25S rRNA in *Saccharomyces cerevisiae* was 3392 nt long. This increase in length was even higher in amphibians, birds, mammals, and could be attributed to GC-rich insertions at specific positions along the rRNA. Such insertions were also observed in the eukaryotic 18S rRNA relative to the 16S bacterial rRNA and were termed as expansion segments (ESs) [36]. ESs were also referred to as “insertions,” “divergent domains,” or “variable regions" in some of the early studies by other groups [35,38–41]. ESs possess many interesting features that make them good candidates to explore ribosome-centric translation regulation.

1.4. rRNA Expansion Segments

The earliest insights on ESs came from two parallel studies [42,43] that observed double and triple loops with the same morphology across vertebrates in rRNA electron micrographs. Later studies that compared rRNA sequences from *E. coli* (bacterium), *S. cerevisiae* (yeast), *Physarum polycephalum* (slime mold), *Mus musculus*

(mammal), *Rattus norvegicus* (mammal) and *Xenopus laevis* (amphibian)[34];[44];[35];[36];[37] showed that eukaryotic rRNAs contained multi nucleotide insertions at specific conserved positions relative to the bacterial rRNAs (Figure 1.4). Except for these insertions, the other rRNA regions were structurally conserved due to the presence of compensatory base-pair changes. Interestingly, deletions were also observed in mammalian and kinetoplastid (protozoan) mitochondrial rRNAs concurrent with the position of the eukaryotic expansions [45–47];[44,48] These observations led to the idea of a common structural core across bacteria, archaea and eukaryotes, over which insertions/deletions may occur.

Figure 1.4 - Secondary structure of rRNAs from E. coli and H. sapiens

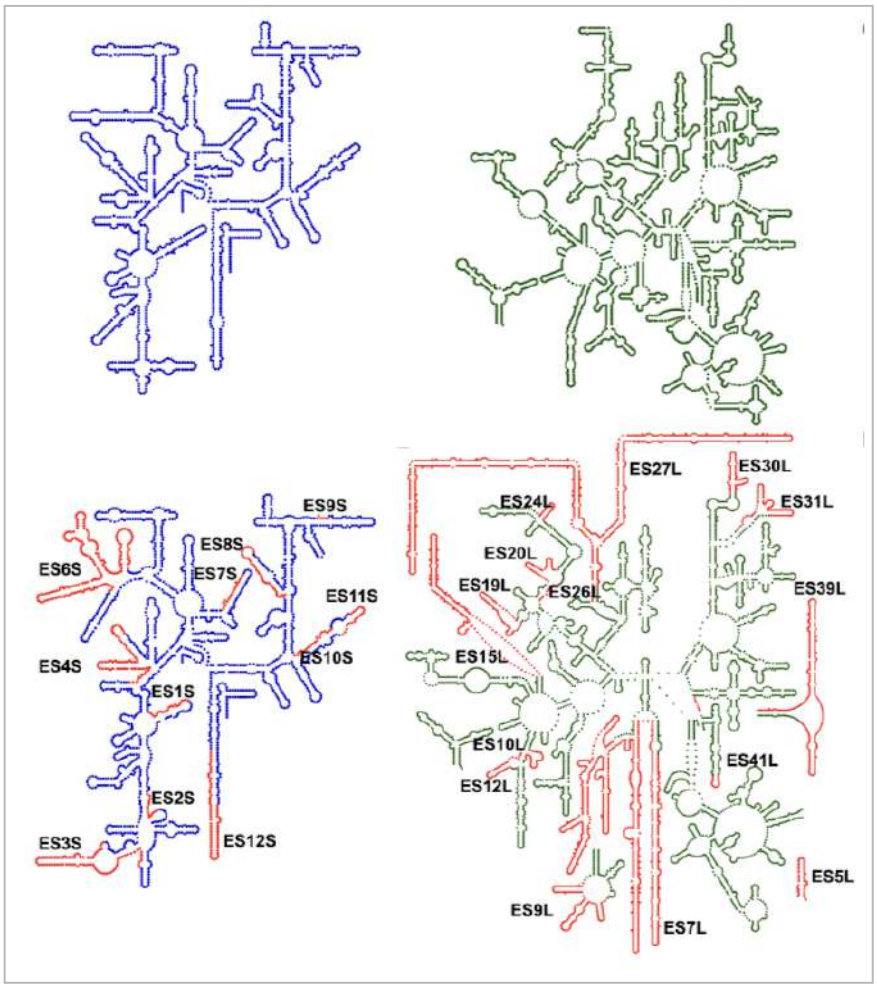


Figure 1.4 - Secondary structures of 16S/18S (blue) and 23S/28S (green) rRNAs from *E. coli* (top) and *H. sapiens* (bottom). The human ESs are colored in red.

Mainly considered to be a distinctive feature of eukaryotic ribosomes, rRNA insertions have also been identified in some bacteria and archaea [49–52]. The occurrence

of insertions in the 5S rRNA of bacterial and archaeal genomes does not correlate with any phylogenetic branching across these two superkingdoms [49]. This raises the possibility that such species-specific rRNA insertions could either confer some niche-specific adaptive advantage to those species or could be nonfunctional insertions undergoing neutral evolution. Some bacterial genomes harbor both the normal and expanded versions of the 5S rRNA [49]. Whether both of these versions get incorporated into functional ribosomes and whether they lead to any functional specialization is yet to be probed. It is also unclear whether these insertions are similar to the eukaryotic ESs. However, archaeal ribosomes harbor micro-expansions in the rRNA regions that coincide with eukaryotic ESs [53,54].

Initial studies estimated that the cytosolic ribosomes in eukaryotes have about 12 ESs in their SSU rRNA and 41 in the LSU rRNAs (ES1-3L and a part of ES4L in 5.8S rRNA and the rest in 25/28S rRNA) when compared with the *E. coli* ribosomes (Figure 1.4) [45,55]. However, it is important to note that only a subset of these segments may be present in a particular eukaryotic species [55,56]. Following the usual naming convention, ESs have been denoted as ESXN (where X represents the number of the ES, and N indicates either the LSU, L or the SSU, S).

1.4.1. On the evolution of ESs

Although the origin and evolution of ESs is still not completely clear, a few studies have attempted to unravel it. ESs could either be remnants of a longer ancestral rRNA sequence or they could be recent insertions over the more conserved rRNA core.

Figure 1.4.1.1 - Evolution of ribosomes (adapted from Anger et al., 2013)

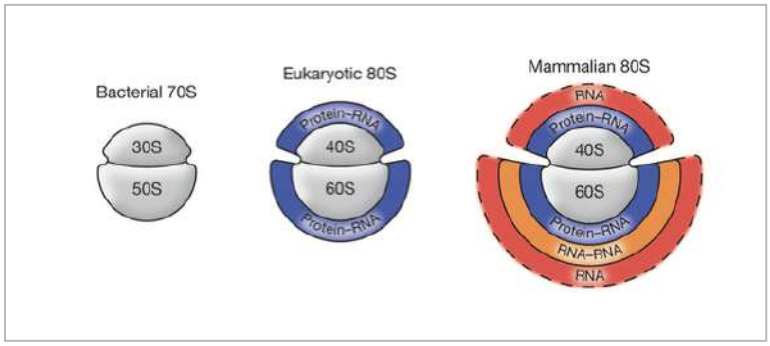


Figure 1.4.1.1 - The gray region represents the common structural core, with the layered accrual of RPs and ESs represented in colors.

According to the 'accretion model', the architecture of primitive ribosomes was laid out before the branching of kingdoms, on which iterative addition of rRNA fragments has occurred during evolution (Figure 1.4.1.1) [57–59]. Such an accrual could involve either formation of non-intrusive helical branches in ancestral helical trunks or elongation of an already existing helix (Figure 1.4.1.2) [54,60]. One study attempted a random insertion of a 31-nt long RNA segment into the rRNA of *E. coli* and mapped the sites of insertion [55]. They observed that in both 16S and 23S rRNA, most of the functional insertions exhibited a tendency to coincide with the regions of eukaryotic ESs, leading them to suggest that the ESs might have evolved from the prokaryotic core rRNA through a process of random genetic insertion and selection. Another important link in the evolution of these enigmatic segments is the micro expansions (5-20nt) reported in the archaeon *Pyrococcus furiosus*, some of which coincide with regions of eukaryotic ESs [53,54].

Figure 1.4.1.2 - Accretion of rRNA ESs (taken from Petrov et al., 2015)

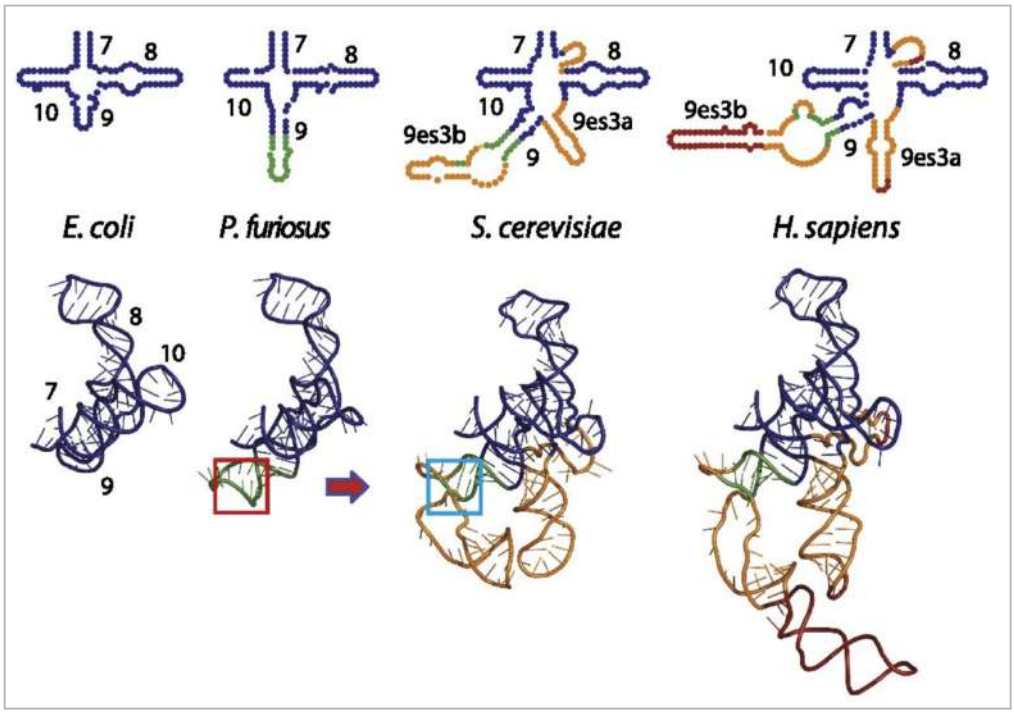


Figure 1.4.1.2 - This figure shows rRNA accretion (green, yellow and red) in helix 7-10 (ES3S) that does not perturb the common core (dark blue).

If a common core was sufficient for functionality, why were these insertions and accretions tolerated? This can potentially be explained by Constructive Neutral Evolution [61–63]. According to this theory, nucleotide insertions that are either mildly

advantageous or disadvantageous, can be tolerated by large cells that replicate slowly with small effective population sizes. This could be so because such insertions cause a very low energetic burden in cells that enable them to escape natural selection and spread due to random genetic drift. Therefore, such forces which shape the overall genome architecture might have permitted expansions in rRNA too, which may have then been stabilized because of possible gain of function [51]. Post stabilization, the sequences of the ESs could have been diversified by compensatory slippage occurring during genome replication [64]. This is reminiscent of the possible evolution of ribosomes, in which the fortuitous interaction of proteins that stabilized the sporadic defects in pure ribozymes, combined with natural selection could have resulted in the emergence of a more efficient ribosome [62,65].

There are some exceptions to the accretion model, in which homologous expansions can occur in parallel. For instance, ES39L in Asgard archaea are larger than in any other archaea while some even exhibit ES39L larger than that of most eukaryotes [51]. This could either mean that the Last Archaeal and Eukaryotic Common Ancestor had a larger ES39L, which could could have shrunk in most archaea and developed parallelly in Asgard archaea and eukaryotes or that eukaryotes originate from within archaea in which the Asgard archaea and eukaryotes share a close common ancestry. This finding opens up the question of whether ribosomal complexity is more deeply rooted than previously thought [51].

1.4.2. Characteristics of ESs

➢ The insertion points of eukaryotic ESs coincide across taxa, but their sequence and length may vary (Figure 1.4.2)[34];[44];[35];[36];[37]. ESs need not show a monotonic increase in size across taxa. For instance, ES39L is larger in euglenozoans (protozoa) and tetrapods (vertebrates) when compared with other eukaryotes [64,66].

➢ A few ESs such as ES7L, ES27L show a phylogenetically linked increase in size with increasing organismal complexity (bacteria < archaea < single celled and invertebrate eukaryotes < vertebrates) (Figure 8). For example, ES7L which is the largest ES, spans 20 nucleotides in bacteria, 80 nucleotides in archaea to about 210 nucleotides in invertebrates and over 870 nucleotides in mammals [57].

Figure 1.4.2 - Representative illustration of ESs across eukaryotes (modified from Hariharan et al., 2022)

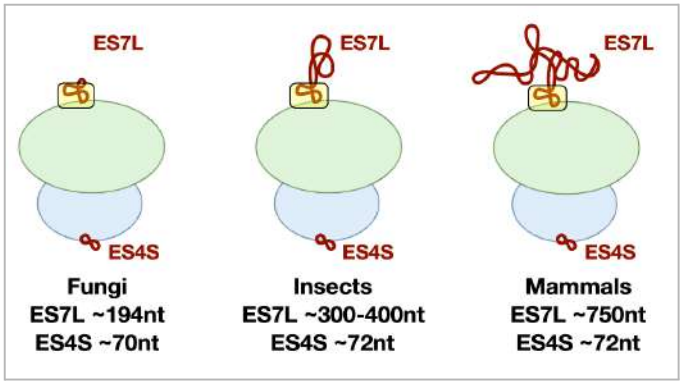


Figure 1.4.2 - This figure illustrates the higher rRNA expansion observed in the ribosomal LSU when compared to the SSU.

➢ Accretion is higher in the LSU rRNA than in the SSU rRNA (Figure 1.4.2) [60]. The *E. coli* 23S rRNA is around 2900 nucleotides long, whereas the 25S/28S rRNA has expanded by about 500 nucleotides in fungi and about 2400 nucleotides in mammals. On the other hand, the 16S rRNA has not expanded as drastically as it ranges from 1540 nucleotides in *E. coli* to about 1800 nucleotides in yeast and 1870 nucleotides in humans. The bulk of the expansion in the LSU rRNA occurs largely in the regions of ES7L and ES27L.

➢ Increase in the size of ES correlates with increase in repeats with a GC enrichment. ESs possess both inverted and direct repeats, which could have possibly arisen due to DNA strand slippage during replication. Compensatory slippage and selection may have resulted in stretches of same nucleotide repeats called homoiterons and other repetitive motifs, which also could have contributed to sequence diversity across taxa [64,66]. Such biased expansion may be one reason for the abundance of high GC-rich stretches in the ESs from vertebrate rRNAs. Contrastingly, ESs from single-celled and invertebrate eukaryotes have a lower prevalence of GC-rich stretches [64,66].

➢ The length and sequence variation in ESs can also occur within a species, including in different tissue types or developmental stages of an organism, though the extent of such variability and its relevance is yet to be understood [67,68].

1.4.3. Structure of ESs

➢ ESs are generally present on the ribosomal surface and are highly accessible. They do not perturb the conserved functional regions of the ribosome [59,69,70].

➢ ESs exhibit a different spatial distribution amongst the two ribosomal subunits. In LSU, ESs are generally present at the top, back and the sides without hindering the peptide exit tunnel and the conserved inter-subunit bridges. Whereas they are mostly concentrated in the foot region of the SSU [53,59,69,70].

➢ ESs are predicted to form remarkably conserved secondary structures within kingdoms, even when they diverge in their primary sequence [71]. For instance, ES7L exhibits a wide variability in size and sequence but contains a conserved "structural core" at its base (Figure 1.4.3) [57].

Figure 1.4.3 - ESs possess a conserved structural core (modified from Petrov et al, 2014)

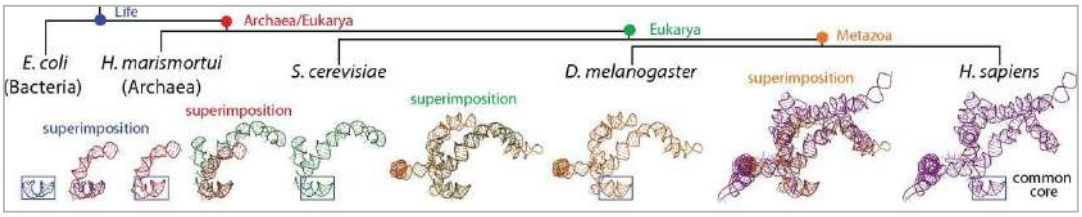


Figure 1.4.3 - The helical RNA enclosed in the boxes is the conserved core over which more recent expansions have been inserted in ES7L.

➢ ESs across eukaryotes have both helical and disordered single-stranded regions, which interact with RPs and other ESs. Such segments may also potentially associate with extra ribosomal proteins or mRNA [59,69,70,72]. As an example, ES3S and ES6S are predicted to form 7-9 bp helix across eukaryotes, which has been structurally shown in yeast. This helix is present in a region that is an important site for binding eukaryotic translation initiation factors [53,59,73,74].

➢ Species-specific changes in ESs can also lead to the formation of structures within the ribosome that are unique to the species. In the protozoan *Trypanosoma cruzi*, ES6S and ES7S are quite large and form a helical structure (turret), that is predicted to aid in translating the distinctly structured mRNAs in the organism [75]. The turret is unique to *T. cruzi* and is not observed in other eukaryotes.

➢ ESs (especially those from chordates) exhibit a propensity to form tertiary G-Quadruplexes. Expansions such as ES7L, ES3S, ES6S and ES12S of human ribosomes have been shown to form such structures *in vitro* [76–78]. The authors suggest that some of these structures can potentially switch between duplex and quadruplex conformations based on protein association and cellular conditions.

Concrete structures for some ESs, especially the distal ends of long ESs are either unavailable or are modeled based on secondary structure predictions and not resolved yet in cryo-EM reconstructions due to their flexible nature. Nevertheless, the insights that ESs can interact with extra-ribosomal components, form species-specific structures and even within a species, adopt different conformations widens the landscape of heterogeneous ribosomes, which could possibly aid in translation regulation.

1.4.4. Function of ESs

The functional relevance of ESs has been an ambiguous topic of discussion ever since their discovery. Early studies reported the occurrence of ESs in *S. cerevisiae* or *P. polycephalum* that were absent in *X. laevis*, which led to the perspective that they do not have a role in ribosome function. Moreover, even amongst the shared ESs, there was poor sequence or length conservation , which led to the view that ESs were tolerated because they do not perturb the core ribosome function [35,36,45]. However, despite the sequence divergence, homologous ESs possessed very similar secondary structures in the shared regions. Studies also revealed the presence of different evolutionary rates between the shared regions and the species-specific extensions in the ESs from different species [37,39]. These observations led to the speculation that ESs may have some role *in cellulo*.

Early functional studies on different ESs from *S. cerevisiae* (yeast) and *Tetrahymena thermophila* (protozoan) showed that their perturbation by insertion or deletion can lead to diverse outcomes. While ES9S, ES19L from yeast did not seem to affect ribosome biogenesis, an insertion into ES3S interfered with the formation of SSU [41,79]. In *T. thermophila*, ES7L and ES27L deletion resulted in ribosome biogenesis defects, but a concurrent replacement of the deleted ES27L stretch with the orthologous region

from *S. cerevisiae*, *Dictyostelium discoideum* (slime mold) or *Caenorhabditis elegans* (nematode) did not cause any such defect [40,80]. This suggested that the divergent stretches of the ES may still fold into similar tertiary structures across taxa, that allow them to function interchangeably.

Unfortunately systematic experimental studies on the nature and function of ESs have been sparse. Overall, the known and potential functions of a few ESs that have been studied are summarized below (Figure 1.4.4) [81] -

➢ Ribosome Biogenesis : ESs can act as binding sites for several factors involved in folding, assembly, nuclear export and subunit maturation during ribosome biogenesis in yeast [56,82–86]. While the involvement of some expansions like ES7L and ES27L have been ascertained only in yeast through structural and mutational studies, extensive probing needs to be extended to the other ESs as well across taxa.

➢ Translation & Protein Folding : Some ESs are present as tentacle-like extensions on the ribosome and may potentially act as docking sites for various factors, which in turn may enable fine-tuning translation. For instance, ES6S can act as a binding site for some translation initiation factors and also help in the scanning of mRNA for proper positioning of the ribosome [87–89].

Expansions such as ES7L and ES27L can associate with ribosome-interacting proteins for several processes like initiator methionine cleavage from the nascent polypeptide [90], acetyl or selenocysteine incorporation in nascent polypeptides [91,92].

ES7L may also be involved in co-translational folding of the nascent polypeptides as they may bind chaperones [93,94]. ESs can also possibly bind to mRNAs and small RNAs based on computational prediction and some experimental data [72,95–97]. For instance, the human ES9S has been shown to bind to the 5’ UTR of Hoxa9 mRNA, while the yeast ES6S is predicted to have two binding sites for the snoRNA, snR30. One study has reported the prevalence of rRNA derived fragments (rRFs) in human and mouse, the potential of such fragments to regulate gene expression is as yet unexplored [98].

Figure 1.4.4 - Known and potential functions of ESs (taken from Hariharan et al., 2022)

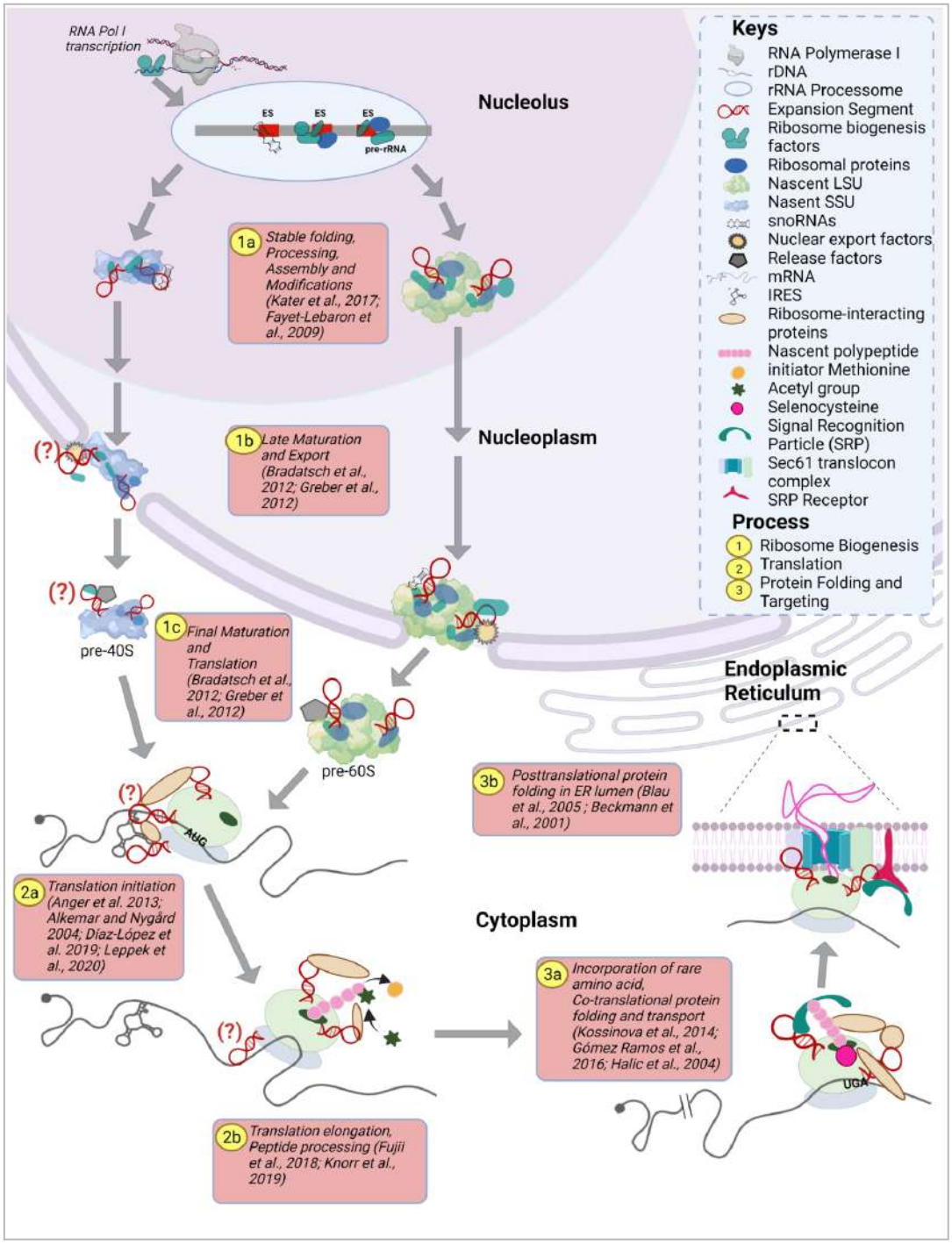


Figure 1.4.4 - Figure summarizing the known and potential functions played by ESs in a cell.
(?) indicates potential interaction that is yet to be supported by experimental data.
1. Ribosome Biogenesis: ESs may play a role in the different stages of ribosome biogenesis.
2. Translation: ESs play myriad roles in protein synthesis.
3. Protein folding and Targeting: ESs may help in the co-translational folding of emerging polypeptides. This figure has been created with BioRender.com

➢ ER binding & Protein Translocation : ES27L may play a role in the attachment of ribosomes to the endoplasmic reticulum (ER) membrane. It has been shown to contact a ribosome-associated ER membrane protein, which may help in stabilizing the binding [99]. In yeast, ES27L and ES24L have been reported to bind to the Sec61 translocon complex, which transports nascent proteins to the ER [100]. In wheatgerm-canine ribosome nascent chain complexes, ES24L, ES39L along with RPL16 have been shown to interact with Signal Recognition Particle [101].

1.5. The gap in the knowledge

Although ESs were first reported at least four decades back, we still do not possess even a rudimentary understanding of what most ESs do. While the influence of certain ESs in ribosome biogenesis and ER association is better studied, their potential in translation is under-explored. Even in such studies, only a few expansions such as ES21L, ES7L and ES6S from primarily yeast have emerged prominent. Considering that very little (if any) is known about most ESs with in yeast and all the other eukaryotes, it is therefore important to address two main categories of questions - 1) What does an ES do within a species? 2)How conserved is this ES across eukaryotes and does its variability lead to any functional diversification?

Since the complexity of translation has increased in eukaryotes, we wanted to understand whether and how ESs contribute to this increased complexity. Since the bulk of the rRNA expansion occurs in the LSU, I was interested in comparing the LSU ESs from diverse eukaryotes. From this comparison, we observed a rRNA segment called ES30L that was specific to endothermic vertebrates (birds and mammals).

Given its location on the ribosome near the tRNA exit site (extended from the L1-stalk), its GC-richness and tentacle-like nature reported in ribosome structural studies, we hypothesized that it would be a good candidate that could associate with both mRNA and ribosome-interacting proteins. Therefore, I was also keen on exploring the role of this ES in translation regulation.

1.6. Aim and Objectives

With the afore-mentioned ideas, the primary objectives of our study have been -

1. To compare 28S rRNA sequences centered on the major ESs across eukaryotes.
2. To understand the variability of ES30L among the endothermic vertebrates.
3. To examine the potential of ES30L to bind to protein-coding transcripts in *H. sapiens* using both computational and experimental approaches.
4. To investigate the protein interactome of ES30L in *H. sapiens*.

1.7. References

The references cited are included at the end of the thesis.

Chapter 2

Materials and Methods

Comparison of ES30L among bilaterians

28S rRNA sequences of all the organisms used in this study (85 organisms across 15 phylogenetic groups; refer to Supplementary Table S1A for sequence details) were mined from public databases like Silva LSU r132[102] and NCBI Nucleotide. Multiple Sequence Alignment (MSA) of the rRNAs was built using Clustal Omega[103] and visualized using Jalview 2.11.2.2[104]. The ESs of the human 28S rRNA were marked in the MSA based on the co-ordinates taken from Parker et al.,[105]. The region of the MSA corresponding to ES30L was selected and manually edited in Jalview to improve the alignment. This corrected alignment was used to calculate and plot GC percentage, length and Gap-Adjusted Shannon's Entropy (GASE)[106] for mammals and birds using custom scripts written in python, bash, awk and R. The ES30L region was also marked in the published human ribosome structure[107] (10.2210/pdb4UG0/pdb) downloaded from PDB. The ribosome structure was rendered using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC). The secondary structure of LSU rRNA either as whole or in parts (of the ES30L region) from H. sapiens and the other organisms used in this study was created using the RiboVision web server[108].

Complementary analysis between ES30L and human protein-coding transcripts

Human protein-coding transcript sequences (83129) were downloaded from GENCODE (v0.29)[109]. The human ES30L sequence was extracted from the NCBI reference sequence NR_003287.4 (coordinates: 3975-4035). Ten randomly picked 60 nt stretches from CSL (regions in the 28S rRNA other than the ESs; ES boundaries were taken from Parker et al.[110]) were extracted and also used as a control for the complementarity analysis. The ES30L and the CSL sequences were reverse complemented (only canonical Watson-Crick A:U & G:C antisense pairing considered in this analysis) and used to obtain shorter stretches that range from 7 to 15 nucleotides using the sliding overlapping window approach using custom scripts. For instance, the first 7 nucleotide sequence would span from position 1 to 7 in the reference, the next sequence would span from position 2 to 8 and so on. These shorter

stretches from ES30L and CSL were then mapped to the GENCODE transcripts using SeqMap v1.0.13[111], allowing up to one mismatch in the mappings.

The data presented in this study with mismatch '0' refers to contiguous mapping with zero mismatches. For mismatch '1', data with zero mismatches at lengths of 7 or 8 nucleotides and up to 1 non-terminal mismatch at lengths 9 and above was considered. Only the longest contiguous match (either with 0 or 1 mismatch, as appropriate) from a transcript-ES30L region were considered for further analysis, with the redundant overlapping hits removed using custom awk scripts. Complementary stretches of length 16 nucleotides or more were obtained by extending the overlapping 15 nucleotide matches using custom awk script. All downstream analyses and plotting was done using custom written scripts in bash, awk, python and R[112]. Gene ontology analysis was done using the shinyGO v0.76 web server[113]. Statistical tests used in this study to compare the number of complementary stretches have been referred to from Parker et al.[114]. Briefly, Wilcoxon signed-rank test at $p<0.05$ was used to test the significance of difference in complementarity to transcripts between ES30L and CSL fragments. To characterize the difference in the extent of complementarity from lengths of 7 to 15 or more nucleotides, linear regression was used.

RNA and Protein pulldown from HEK293T cell lysates with biotinylated RNA oligos

HEK293T cells were cultured in DMEM (Dulbecco's Modified Eagle Medium) media supplemented with 10% FBS and 1% gentamicin. Cells were grown at 37°C with 5% CO2 for 48 hours and were harvested when they reached 80% confluence. Both the RNA oligos (Supplementary Table S3-I) were synthesized and obtained from Integrated DNA Technologies and included a 5' biotin tag. The 5'-biotinylated RNA oligos were used in pulldown experiments, following the protocol adapted from Krishna et al.[115]. Briefly, 3 µg of the folded biotinylated RNA oligonucleotides in 100 µl of RNA structure buffer (10 mM Tris pH 7.0, 0.1 M KCl, 10 mM MgCl2), was incubated with 1 mg of pre-cleared HEK293T cell extracts in cell lysis buffer (20 mM Tris–HCl pH 7.5, 150 mM NaCl, 1.8 mM MgCl2, 0.5% NP40, protease inhibitor cocktail for mammalian cells (no EDTA), 1 mM DTT, 80 U/ml RNase inhibitor) at

room temperature for two hours (100mg of the cell lysate was set aside as input). Post incubation, Dynabeads M-280 Streptavidin™ (11206D, Invitrogen) beads were added to the mixture and incubated for one hour at room temperature. Protein was eluted from the streptavidin beads using SDS buffer (250mM Tris pH 6.8, 10% SDS, 5% beta-mercaptoethanol) followed by acetone precipitation. Incase of RNA, TRIzol (Invitrogen)-based extraction from the beads was done. The protein pellets were resuspended in 25mM ammonium bicarbonate and the RNA pellets in nuclease free water. The total protein/RNA was also extracted from the 'input' using either acetone or TRIzol-based methods as appropriate. The isolated protein and RNA samples were sent for mass spectrometry and high-throughput sequencing respectively.

RNA-sequencing and data analysis

The isolated RNA was analyzed on microfluidic RNA gel electrophoresis with the Agilent 2100 Bioanalyzer before RNA sequencing. Sequencing libraries were prepared using NEBNext® Ultra™ II Directional RNA Library Prep after poly(A) selection from the isolated RNA and sequenced on an Illumina Hiseq 2500 machine (n=2; 2 biological replicates each from from ES30L fraction, random fragment fraction and input). Post sequencing, 45 to 64 million single end (1*50 bp) reads were obtained. Adapters were trimmed from the reads using cutadapt v1.8.3[116] (-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC) and were mapped to the GENCODE human transcriptome v0.29 using hisat2 v2.1.0[117] (default parameters used). Around 80-85% of the reads mapped to the reference transcriptome. The raw read counts mapping to the transcripts were obtained using featurecounts v1.5.0-p1[118]. Normalization of read counts and differential expression analysis was done using the DESeq2 v1.10.1[119] package in R[120]. Transcripts which were two-fold or more upregulated ($q<0.05$) in the ES30L fraction over the input were considered for further analysis. Plots were generated using the R packages pheatmap[121] and ggplot2[122]. Gene ontology analysis was done using the shinyGO v0.76 webserver[123]. Custom bash and awk scripts were used for all downstream analysis.

LC-MS/MS

The isolated protein was digested using sequencing grade trypsin and the Tryptic peptides were dissolved in 0.1% formic acid (FA) and 2% acetonitrile (ACN). They

were then analyzed on a Thermo EASY-nLC™ 1200 nano System coupled to a Thermo Scientific Orbitrap Fusion Tribrid Mass Spectrometer. Each peptide sample was injected into Pepmap™ 100; (75µmx 2cm ; Nanoviper C18, 3µm ; 100Å) via the auto-sampler of the nano System. Peptides eluted from the peptide trap were separated in a EASY SPRAY PEPMAP RSLC C18 3µm; 50cm x 75µm; 100Å at 45°C. Mobile phase A (0.1% FA in H2O) and mobile phase B (0.1% FA in 80% ACN) were used to establish a 60-min gradient at a flow rate of 250 nl/min. Peptides were then analyzed on mass spec with an EASY nanospray source (Thermo Fisher, MA) at an electrospray potential of 1.9 kV. A full MS scan (375–1,700 m/z range) was acquired at a resolution of 120000, a maximum injection time of 50ms. Dynamic exclusion was set as 20s. MS2 fragmentation was done using HCD technique. Resolution for HCD spectra was set to 30,000 at m/z 100. The AGC setting of full MS scan and MS2 were set as 4E5 and 5E4, respectively. The 10 most intense ions above a 5,000 counts threshold were selected for HCD fragmentation with a maximum injection time of 54 ms. Isolation width of 1.2 was used for MS2. Single and unassigned charged ions were excluded from LC-MS/MS. For HCD, stepped up collision energy (5%) was set to 30%.

LC-MS/MS data analysis

The raw data (n=4; 2 technical and 2 biological replicates each from ES30L fraction, random fragment fraction and input) was converted to the mascot generic file format using Proteome Discoverer[124] version 1.4 (Thermo Electron, Bremen, Germany) for deisotoping the LC-MS/MS spectra. The spectral search was performed using a Mascot server[125] (version 2.4.1; Matrix Science, Boston, MA) against the human UniProt database[126] (downloaded on August 10, 2018), with peptide mass tolerance of 10 ppm and fragment mass tolerance of 0.6 Da. Two missed trypsin cleavage sites per peptide were tolerated. Carbamidomethylation (C) was set as a fixed modification, while oxidation (M) and acetylation (N) were variable modifications.

Label-free quantification of proteins was based on emPAI values of each identified protein reported by Mascot. All further analysis was done using customized bash, awk and R scripts. The emPAI values were then normalized[127,128] to compare the relative protein abundance across samples. Only those proteins that were detected in all the

four replicates of the ES30L fraction were considered for further analysis and the undetected values from the other samples were replaced with zero. For those proteins that were undetected in all the four replicates in random fragment and input fractions, the value for the first replicate was replaced by 0.000001. Fold change between the pulldown fractions relative to input was computed and Students one-tailed t-test was used to test the significance of fold enrichment, followed by correction for multiple testing using false discovery rate estimation. The proteins that were more than 1.5 fold enriched in the ES30L fraction over input ($q<0.05$) were probed for gene set enrichment using the shinyGO v0.76 web server[129]. The reported ribo-interactome protein list used in this study was taken from Simsek et al.[130].

Mining and re-analysis of published Splash-Seq and putative IRES datasets

The list of interactions from the *in vivo* RNA-RNA interactome data[131] was downloaded from GitHub (https://csb5.github.io/splash/). The interactions which included 28S rRNA segments spanning the bin positions 3900 and 4000 (corresponding to ES30L) as one of the interactants were selected for downstream analysis. The reference transcriptome used by the authors was downloaded from their GitHub page and the sequences for interacting transcript regions from the selected interactions were extracted using bedtools v2.25.0[132]. The extent of complementarity to ES30L was analyzed in these extracted transcript stretches, using the same pipeline used for the GENCODE (v0.29)[133] transcriptome.

To check whether any of the transcript regions interacting with ES30L (based on SPLASH-seq data) were part of the putative IRES elements used in this study, putative IRESs from human mRNAs were downloaded from two databases - IRESbase[134] and Human IRES Atlas[135]. The sequences for the putative IRES regions were extracted from the transcripts using bedtools v2.25.0[136] and transcript mapping was done using a nucleotide level blast v2.2.31[137]. The extracted IRES stretches were also checked for complementarity to ES30L using the same pipeline elaborated before. Customized bash and awk scripts were used for data parsing and analysis.

In silico prediction of secondary structural motifs

To probe for the presence of potential structural motifs in the 5‘ UTR regions of transcripts that were selectively enriched in the ES30L fraction over input from our

RNA pulldown, we considered the 100 transcripts that showed a 50% higher enrichment in ES30L than in the random fragment. Out of the 100, 53 transcripts possessed 229 stretches which were perfectly complementary to ES30L in their 5' UTR. Sequences that included 20nt flanking these 229 transcript stretches were extracted using bedtools v2.25.0[136]. Any structural motifs present among these sequences were predicted using the covariance model based tool CMfinder v0.2[138]. ATtRACT web server[139] was used to predict binding sites for RBPs on the RNA motifs obtained from the *in silico* analysis. The secondary structures for the predicted RNA motifs were rendered using the online tool, forna[140]. Data parsing was done using custom written bash and awk scripts.

Data mining and re-analysis of published CLIP-Seq datasets

Published eCLIP-Seq data (fastq.gz files) from K562 cells for RBPs (DHX30, IGF2BP1, NPM1, PCBP1, SRSF1, HNRNPK, U2AF2) were downloaded from ENCODE database (Accession IDs were retrieved from Van Nostrand et al. [141,142]. For each protein, we considered only the first mate data for two eClip replicates along with one size-matched input for our analysis. The downloaded reads were processed using cutadapt v2.10[143] to trim adapters and mapped to a 28S rDNA reference (U13369.1:7935-12969) using hisat2 v2.1.0[144]. The read depth per position of the reference was computed using samtools v1.7[145] and was normalized to per million reads. Fold change (log2 scale) was calculated between the protein pulldown samples and their respective input using custom scripts. In order to compare the binding pattern of PCBP1 between cell types, we downloaded published easyCLIP-Seq data (n=2; size matched input unavailable) from the GEO database (GSE131210) from another study[146,147]) done in HCT116 cells. Read depth across 28S rRNA for this dataset was obtained by following the same pipeline used for the other eCLIP-seq dataset. Plots were generated using the R packages pheatmap [148] and ggplot2 [149].

Data availability

The RNA-Seq transcriptomic data have been submitted to the NCBI SRA database under the BioProject accession *PRJNA883809*. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [150] partner repository with the dataset identifier *PXD036986*.

Chapter 3

Comparison of ESs across eukaryotes

3.1. Introduction

The limited extent of studies on rRNA ESs have primarily focussed on those from either humans or yeast. The variability of these expansions in most of the eukaryotes still lack any systematic documentation. To address this lacuna, we have compared these regions across eukaryotes at the nucleotide level. We also wanted to understand whether any changes in sequence translates to the structure. Because ribosome structures are available for only a handful of eukaryotes, here we have considered human and drosophila ribosomes for the structural comparison. Since the number and lengths of the LSU ESs from the few studied species are higher than the SSU ones, we decided to focus our analysis on the LSU expansions.

3.2. Comparison of 28S rRNA ESs across bilateria

We downloaded and curated 28S rRNA sequences from databases like Silva LSU r132 and NCBI Nucleotide and constructed a multiple sequence alignment (MSA) using them.

Figure 3.2.1 - ESs in the human ribosome

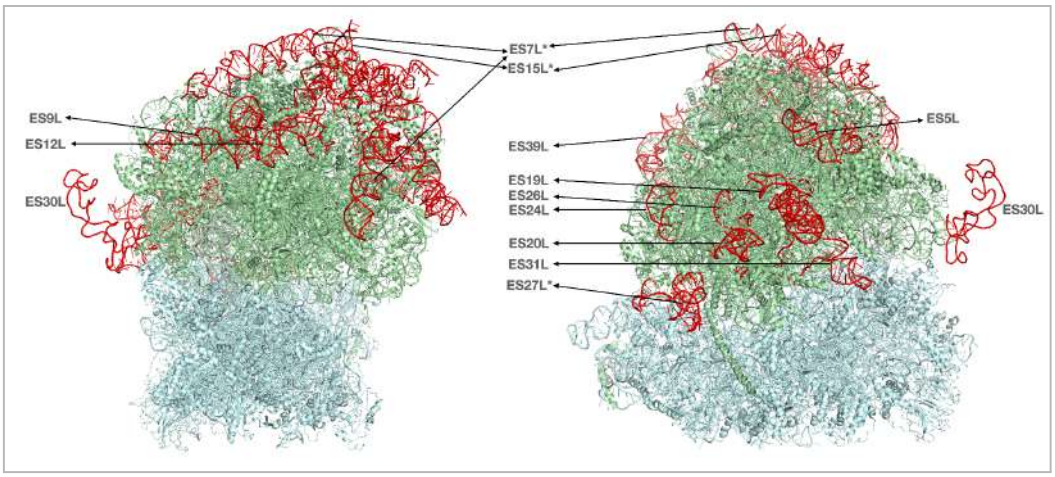


Figure 3.2.1 - This figure shows the human ribosome structure (PDB 4UG0) with all the major LSU ESs marked in red, LSU in green and SSU in cyan. ***(left)*** Front View ***(right)*** Back View. * indicates that the expansion is not completely present.

The major ESs in the human LSU rRNAs were marked in the MSA based on the coordinates taken from Parker et al., (Table 1) and were used to extract the concurrent regions from all the other species used in the MSA. We also mapped these expansions on the structure of the human and drosophila ribosomes (Figure 3.2.1, 3.2.2). Our main observations are highlighted below -

1. ESs were present on the ribosomal surface in both subunits. The distal ends of some expansions such as ES7L, ES27L, ES15L, ES6S are absent in the structures (especially in the human ribosome). Some segments like ES30L are included in the reconstructed structure through secondary structure predictions and modelling data. This is because of their tentacle-like nature, which makes structural data reconstructions extremely difficult.

Figure 3.2.2 - ESs in the drosophila ribosome

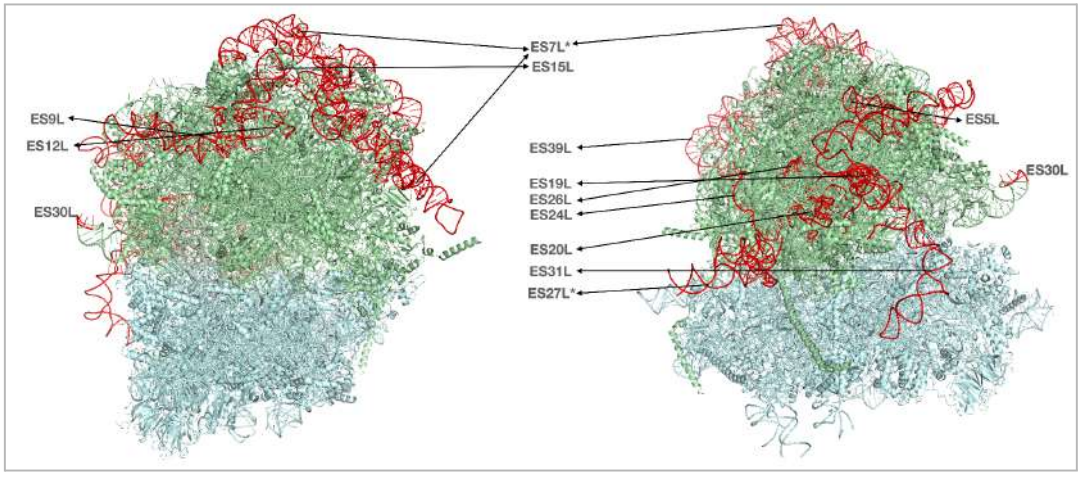
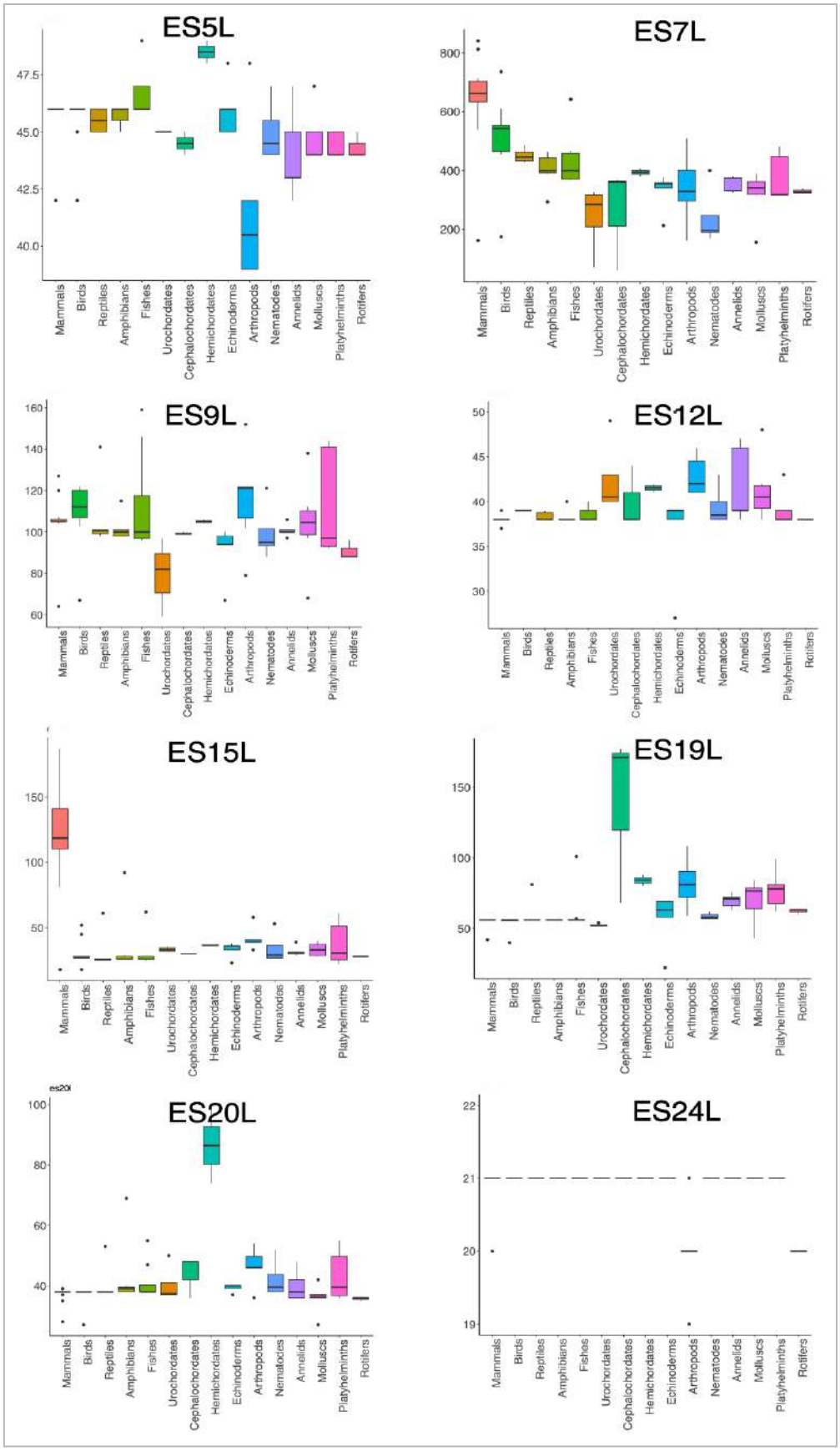


Figure 3.2.2 - This figure shows the drosophila ribosome structure (PDB 4V6W) with all the major LSU ESs marked in red, LSU in green and SSU in cyan. ***(left)*** Front View ***(right)*** Back View. * indicates that the expansion is not completely present.

2. LSU expansions showed a spectrum of length variability across the eukaryotic clades considered in our analysis (Figure 3.2.3). A few LSU ESs such as ES7L, ES27L show a phylogenetically linked increase in their sizes. Some expansions are slightly longer in invertebrates than in vertebrates such as ES19L. Segments like ES9L, ES12L, ES20L, ES26L are quite conserved in their lengths. Length variability in some expansions such as ES5L are observable only in certain clades.

Figure 3.2.3 - Length Distribution of ESs across eukaryotes



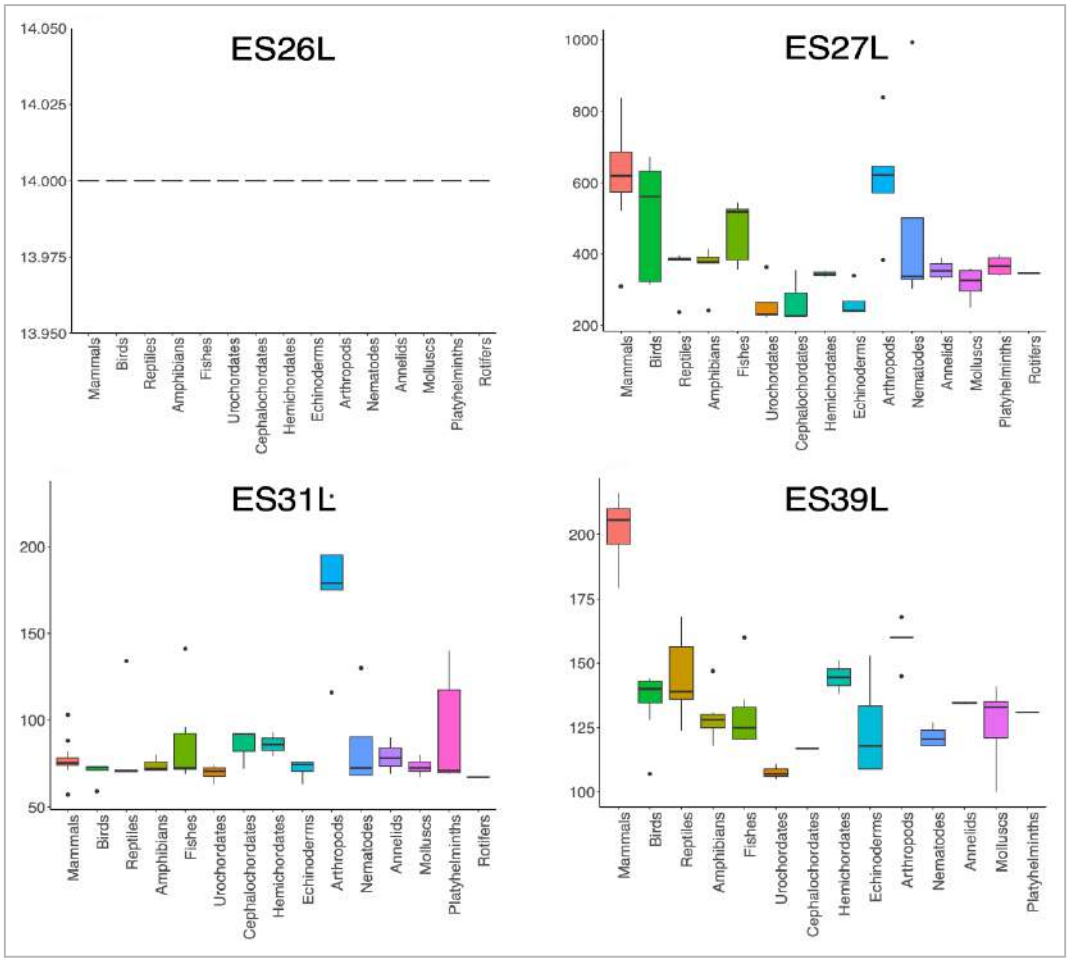


Figure 3.2.3 - These box plots illustrate the length distribution of the main LSU ESs across various eukaryotic clades.

3. We observed that the human ESs have a higher GC-content on average than the core rRNA segments (Figure 3.2.4; Table 2.2). Across eukaryotes, chordates possessed more GC-rich ESs than the invertebrate eukaryotes (Figure 3.2.5). Arthropods had amongst the lowest GC-richness in their expansions, which could probably be because their genomes are inherently AT-rich.

Figure 3.2.4 - GC-content of ESs and the core rRNA regions in the human 28S rRNA

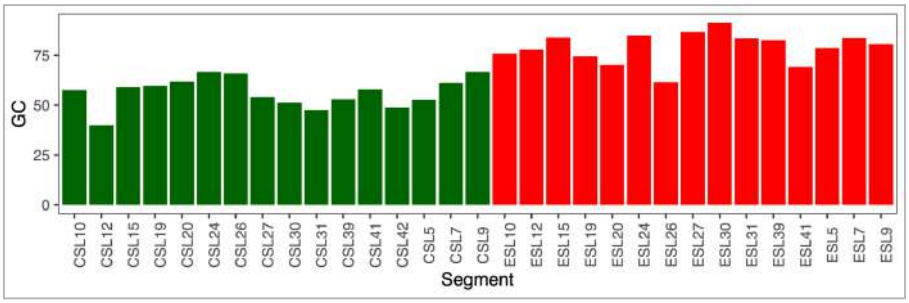
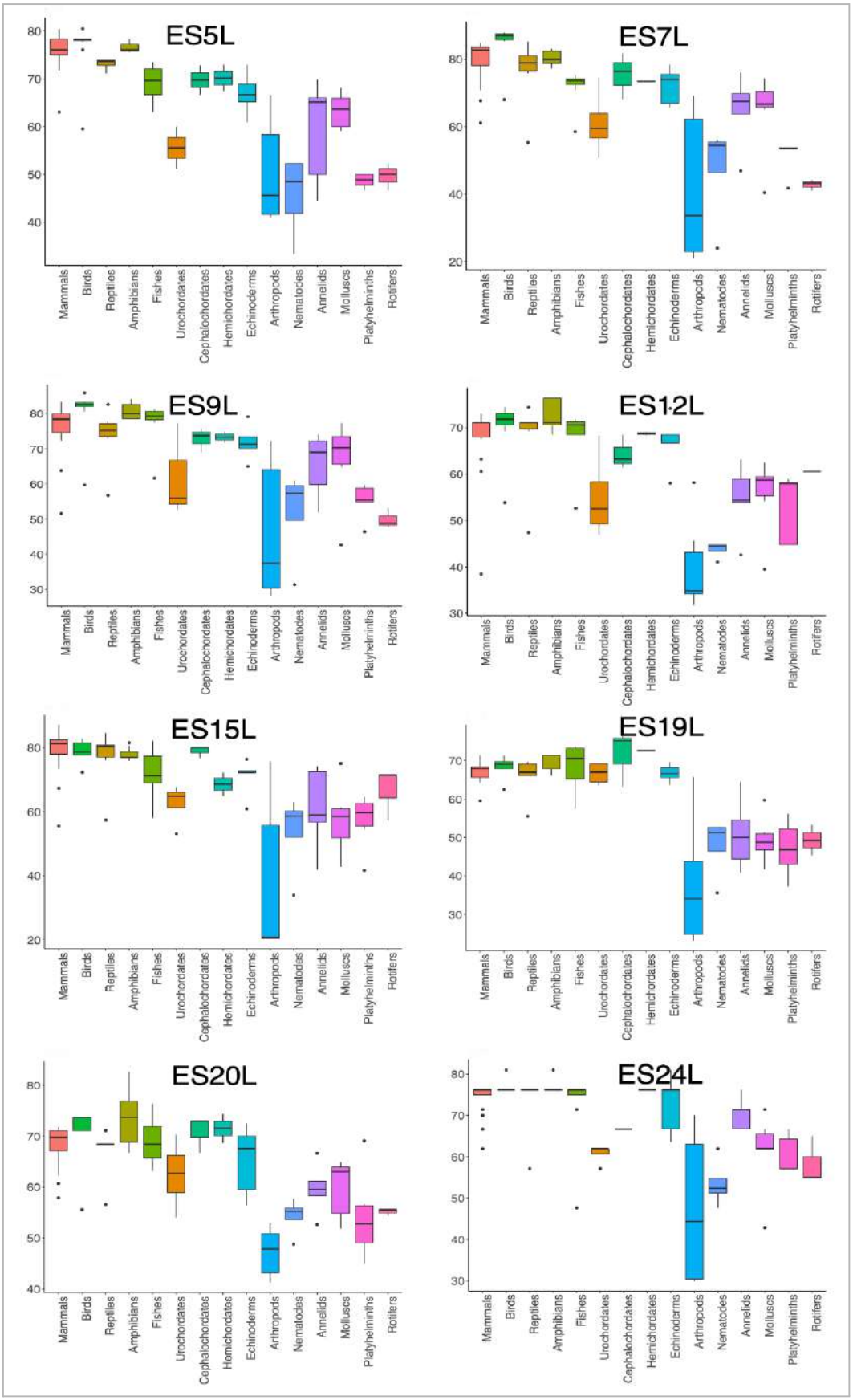


Figure 3.2.4 - This plot shows the GC content of the various 28S rRNA segments.

Figure 3.2.5 - GC-content Distribution of ESs across eukaryotes



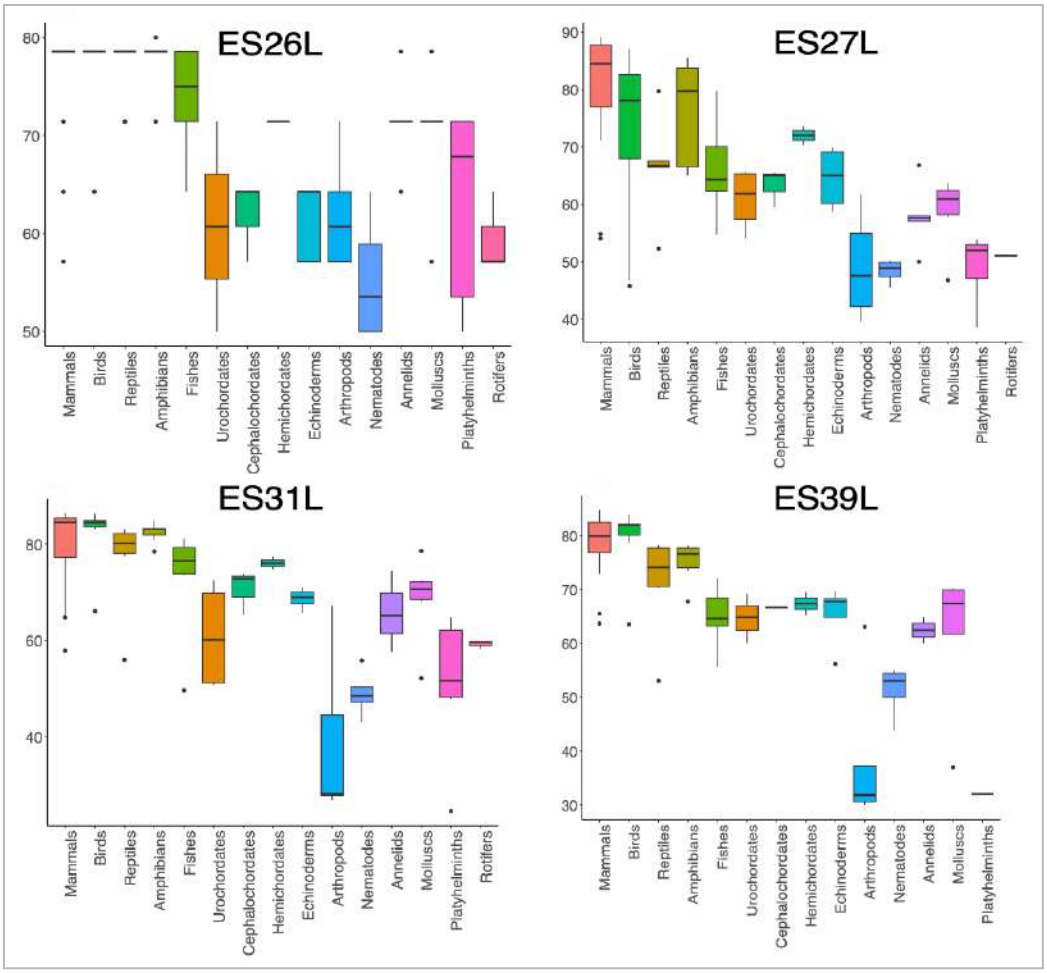


Figure 3.2.5 - These box plots illustrate the GC-content distribution of the main LSU ESs across various eukaryotic clades.

4. Some expansions contained insertions that were specific to a clade(s). For instance, ES30L, ES39L and ES15L have insertions present mostly in mammals (Figure 3.2.6, Figure 4.2.1, Figure 3.2.8). Arthropods have some expansions that are unique to their clade such as those in ES31L (Figure 3.2.7, Figure 3.2.8).

5. Even though expansions exhibited variability in their nucleotide sequence among different clades, their tertiary structure was quite conserved. Most of the ESs in the human and drosophila ribosomes contained a similar core element. Any clade-related insertions seemed to be extended from this core (Figure 3.2.8).

Figure 3.2.6 - ES15L possesses insertion unique to mammals

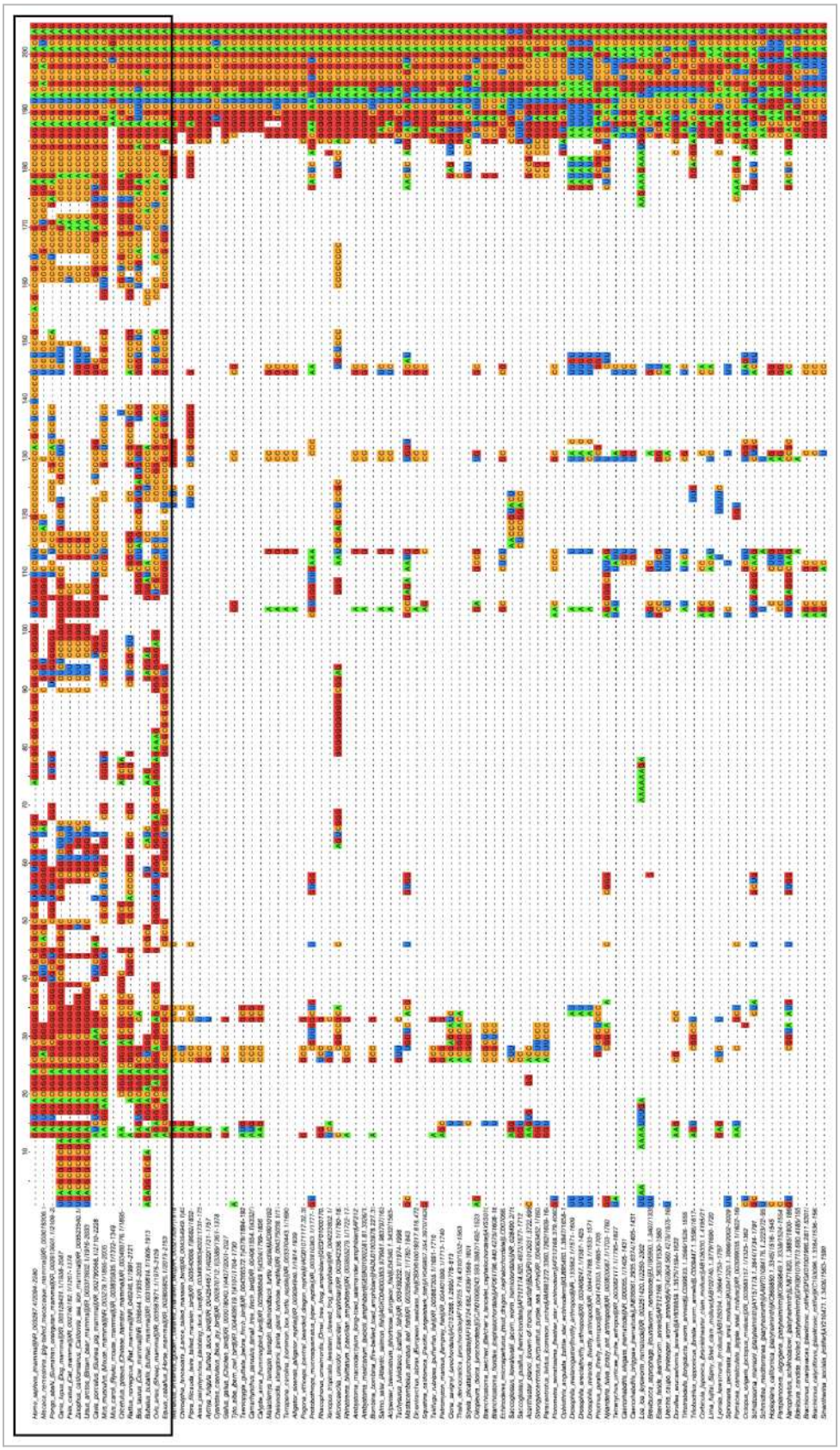


Figure 3.2.6 - This is a part of the MSA showing stretch from ES15L that is unique to mammals (enclosed in the box).

Figure 3.2.7 - ES31L has an arthropod-specific insertion

Figure 3.2.7 - This is a part of the MSA showing stretch from ES31L that is unique to arthropods (enclosed in the box).

Figure 3.2.8 - Structures of the human and drosophila LSU ESs

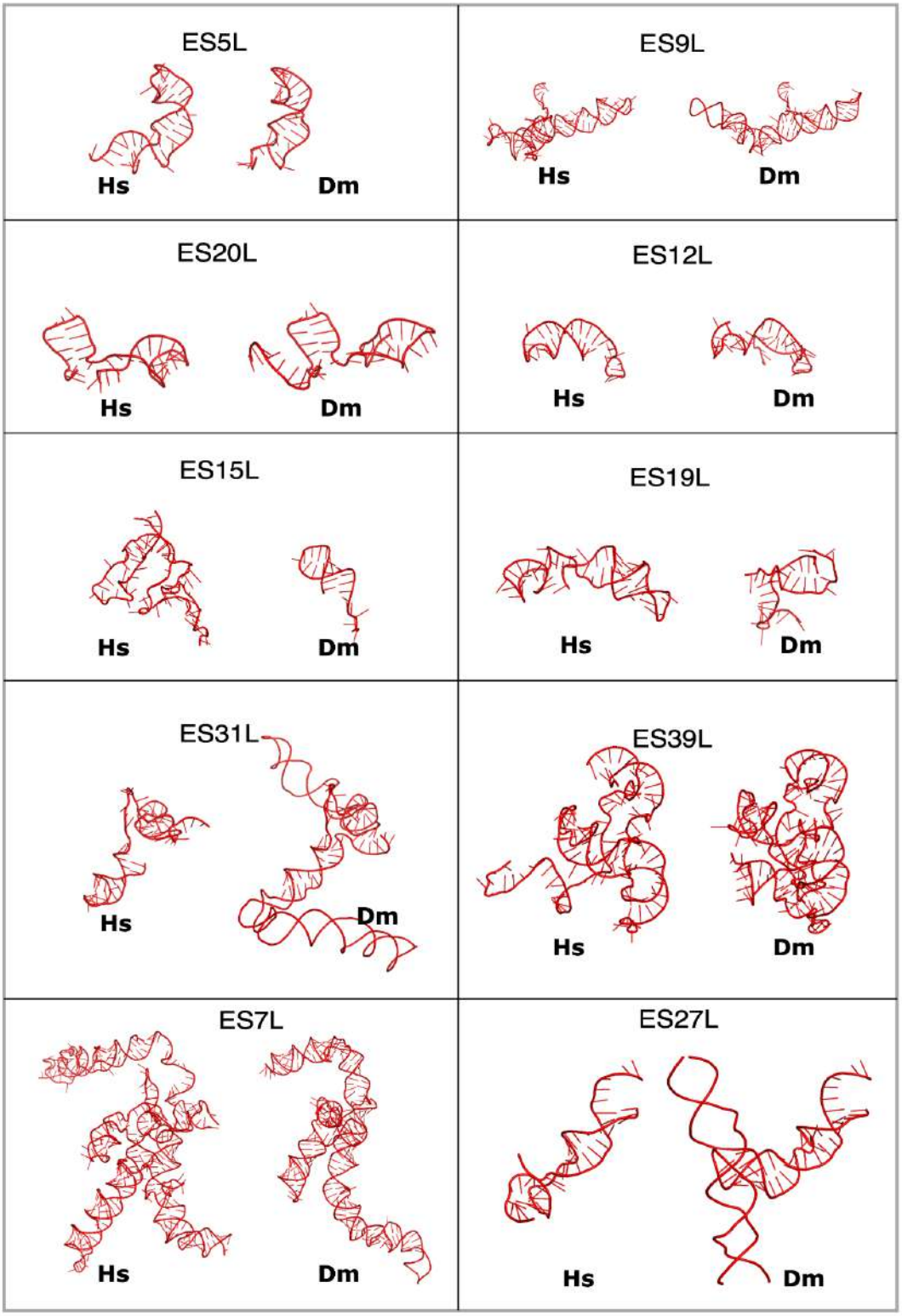


Figure 3.2.5 - This figure displays the structures of the major LSU ESs from *Homo sapiens* (denoted Hs in the figure; PDB 4UG0) and *Drosophila melanogaster* (denoted Dm in the figure; PDB 4V6W). The distal ends of some of the segments such as ES27L, ES15L, ES39L are unavailable because of their high conformational flexibility.

3.3. Summary and Discussion

So far, we have shown that ESs can possess species-specific variations and accretions across eukaryotes. Our interest was in understanding whether ESs can contribute to the complex process of translation regulation. We focussed on ES30L, a highly GC-rich LSU expansion that was present only in endothermic vertebrates (birds and mammals).

Chapter 4

ES30L is a large subunit ES specific to endothermic vertebrates

4.1. Introduction

ES30L is a GC-rich LSU ES that is present only in endothermic vertebrates among eukaryotes (Figure 4.2.1). In the human ribosome, ES30L protrudes from helix 38 of the 28S rRNA, which is part of the L1-stalk (Figure 4.1). This segment is highly flexible because of which it is usually modeled using secondary structure predictions in the cryo-EM reconstruction of ribosome structures. But studies have reported this expansion to appear tentacle-like in ribosomes [59,69].

Figure 4.1 - Structure of the human ribosome

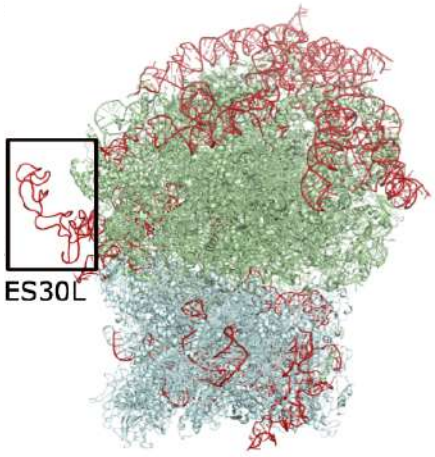


Figure 4.1 - Cryo-EM structure of the human ribosome (PDB: 4UG0) showing the LSU (green), SSU (cyan) and ESs (red). ES30L is enclosed in a box.

The L1-stalk has been shown to interact with the Internal Ribosome Entry Site (IRES) of some viral mRNAs [151,152]. Since ES30L is proximal to the L1-stalk and is highly flexible, we wanted to study its conservation among the endothermic vertebrates and probe its potential to interact with mRNAs and proteins in humans.

4.2. ES30L is present only in endothermic vertebrates and is best conserved in mammals

Sequence comparison of the ES30L stretch from various organisms from different clades across eukaryotes is shown in Figure 4.2.1.

Figure 4.2.1 - MSA of ES30L across eukaryotes

Figure 4.2.1 - Alignment of ES30L from various clades across eukaryotes. This is a stretch from our multiple sequence alignment of 28S rRNA sequences.

From the multiple sequence alignment, we observed that the ES30L stretch exhibited the highest expansion in mammals and birds. This region was about 3 nucleotides long in rotifers and increased to about 40-60 nucleotides in birds and mammals. Other

clades like amphibians, fishes and arthropods also had shorter insertions in this stretch, but exhibited a high intra-group variability in length (Figure 4.2.2; Table 2.1).

Figure 4.2.2 - Length and GC-content distribution of ES30L

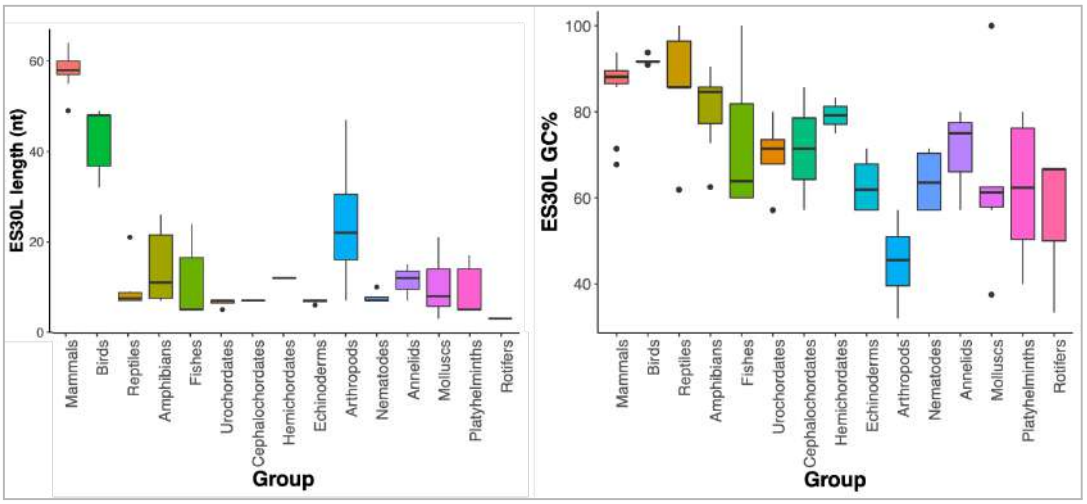


Figure 4.2.2 - These plots display the average length (*left*) and GC-content (*right*) distribution of ES30L across eukaryotes.

Also, while ES30L was GC-rich in chordates (~92% in *H. sapiens*), it had among the lowest GC-richness among arthropods (Figure 4.2.2; Table 2.1). We edited the ES30L region in the MSA manually and used it to calculate Gap-Adjusted Shannon's Entropy (GASE), in order to compare the conservation pattern across endothermic vertebrates. We saw that even though ES30L was well conserved within both clades, the conservation was better among mammals than birds (Figure 4.2.3).

Figure 4.2.3 - ES30L conservation across mammals and birds

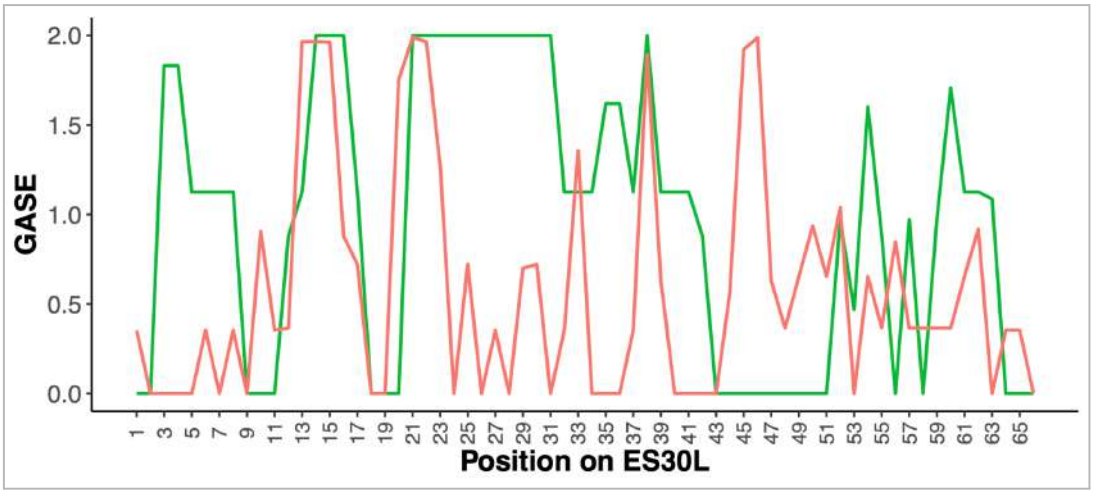


Figure 4.2.3 - Gap Adjusted Shannon's Entropy (GASE) across ES30L for birds (*green*) and mammals (*red*). An entropy value of 0 represents perfect conservation and that of 2 represents none.

4.3. Summary and Discussion

In summary, we have shown that ES30L is a 91% GC-rich 61nt ES that is present only in endothermic vertebrates among eukaryotes. In fact, ES30L is the most GC-rich expansion segment in the human 28S rRNA (Table 2.2). Prior work shows that the equivalent region in *E. coli* is amenable to nucleotide insertions [55]. So its stabilization only in endothermic vertebrates raises a possibility that ES30L could have gained some functional relevance in this group.

Figure 4.3 - Secondary structure of ES30L across domains

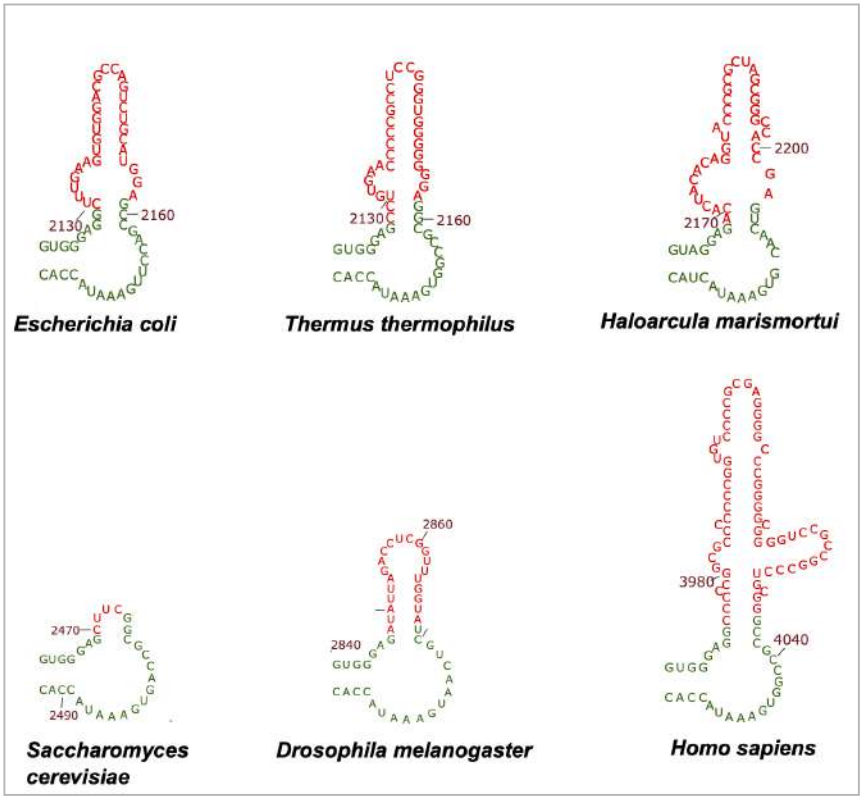


Figure 4.3 - Secondary structure of the ES30L region (red) from organisms across domains.

Interestingly while mostly absent in unicellular and invertebrate eukaryotes, this region spans about 30 nt long in the bacteria *E. coli* and *Thermus thermophilus*, 35 nt long in the archaeon *Haloarcula marismortui* (Figure 4.2.1 & Figure 4.3). Whether this rRNA stretch was part of the last common archaeal and eukaryotic ancestor and was lost in unicellular and invertebrate eukaryotes or if it independently evolved in bacteria and endothermic vertebrates needs further phylogenetic analyses.

Chapter 5

ES30L possesses complementarity to many protein-coding transcripts

5.1. Introduction

According to 'Ribosome Filter Hypothesis', ribosomes can bind differentially to various transcripts that subsequently influences downstream translation efficiency (Figure 1.2.2). Although some studies have explored mRNA:rRNA interactions either computationally or experimentally, hardly a handful have explored this angle in the context of ESs. Therefore, studying rRNA ES:mRNA interactions is important. Many rRNA ESs could potentially be good candidates for interactions with transcripts because they are solvent-exposed and contain structurally dynamic regions [96].

One computational study using human and yeast 18S rRNA sequences led to the perspective that rRNA have surface-exposed 'sticky regions' that can interact with transcripts and regulate their translation [96]. One of the ways this interaction could be facilitated is via base-pairing. Corroborative to this idea, another bioinformatics report revealed the presence of extensive complementarity between the long LSU ESs and human mRNAs [95]. Given its proximity to the L1-stalk in the ribosome and its tentacle-like nature, we were interested in investigating the extent and pattern of complementarity between human ES30L and protein-coding transcripts.

5.2. ES30L possess a higher degree of complementarity to transcripts than the core rRNA segments

To investigate the complementarity between ES30L and protein-coding transcripts, we probed for contiguous matches starting from a length of 7 nucleotides allowing for 0 or 1 mismatch. We used ten randomly selected 60 nt stretches from the core 28S rRNA segments (denoted 'CSL') as a control to check for specificity between ES30L and the transcripts. Only canonical Watson-Crick base pairs (A:U & G:C) were considered in our analysis. We observed that the transcripts possessed higher complementarity to ES30L than the CSL segments both with 0 and 1 mismatch (7.25% and 20.85% more in ES30L than CSL; $p<0.05$ in a paired Wilcoxon test) (Figure 5.2.1A, C; 5.4.1A, B). The slope of the linear regression on the length vs the number of complementary hits was steeper for the CSL segments than for ES30L both

without a mismatch (Figure 5.2.1B, D; R^2>0.99). This indicates that there is more complementarity at longer lengths between transcripts and ES30L than with the CSL segments suggesting a higher potential for ES30L to bind to protein-coding transcripts.

Figure 5.2.1 - Overall complementarity between ES30L and protein-coding transcripts

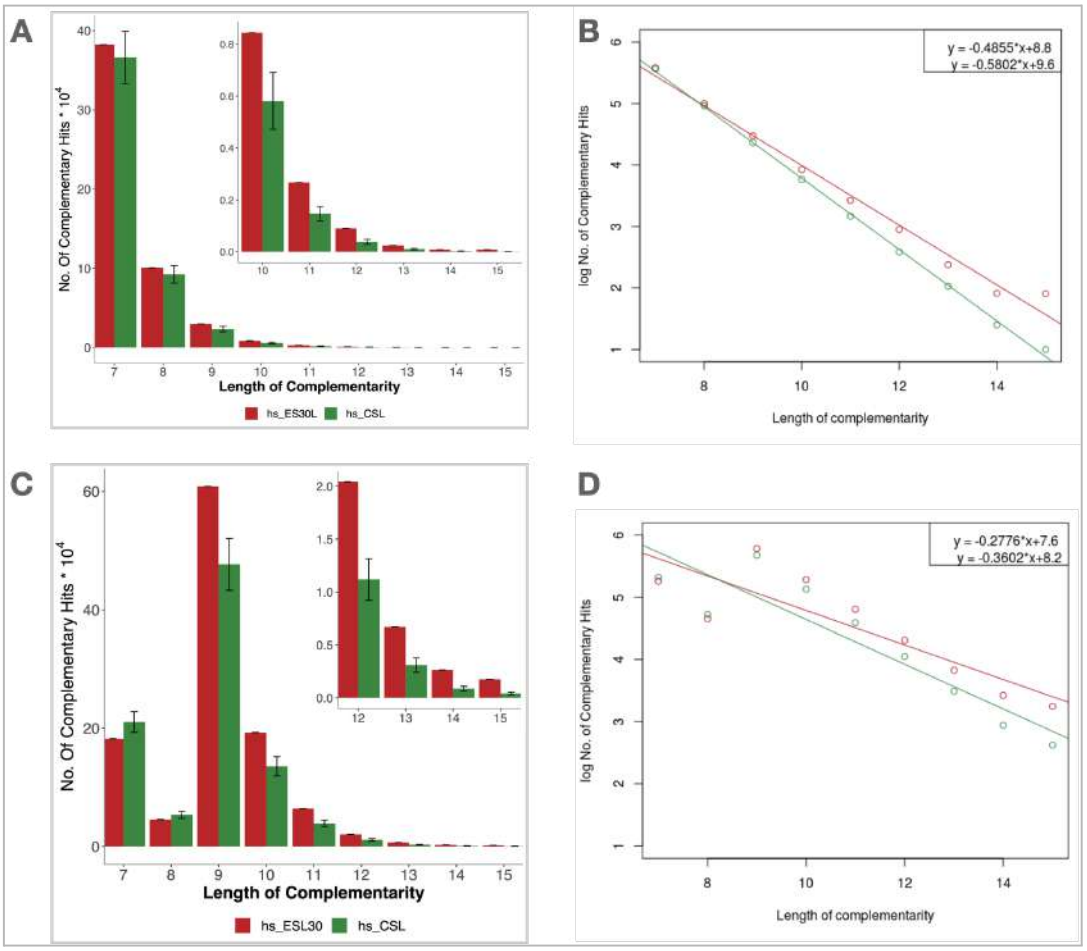


Figure 5.2.1 - (A) Bar plot showing the number of complementary hits at varying lengths with 0 mismatch.
(B) Linear Regression plot between length and lof of number of complementary hits with 0 mismatch.
(C) Bar plot showing the number of complementary hits at varying lengths with 1 mismatch.
(D) Linear Regression plot between length and lof of number of complementary hits with 1 mismatch.

Overall, ES30L showed higher complementarity to transcripts than the CSL segments at longer nucleotides stretches (Figure 5.2.1A, C). To elaborate, with 7 and 8 nt stretches, ES30L and CSL segments showed a similar level of perfect complementarity to the transcripts, with the proportion being 5.3% higher in ES30L.

This difference increases to 36.12% in ES30L over CSL segments at lengths of 9 nt or more. Transcript stretches spanning 15 nt or more that are perfectly complementary to ES30L were rare, with only 53 such hits detected in our analysis. This number was markedly lower at only 2 hits in case of CSL segments.

Figure 5.2.2 - Complementarity between ES30L and transcripts with 0 or 1 mismatch

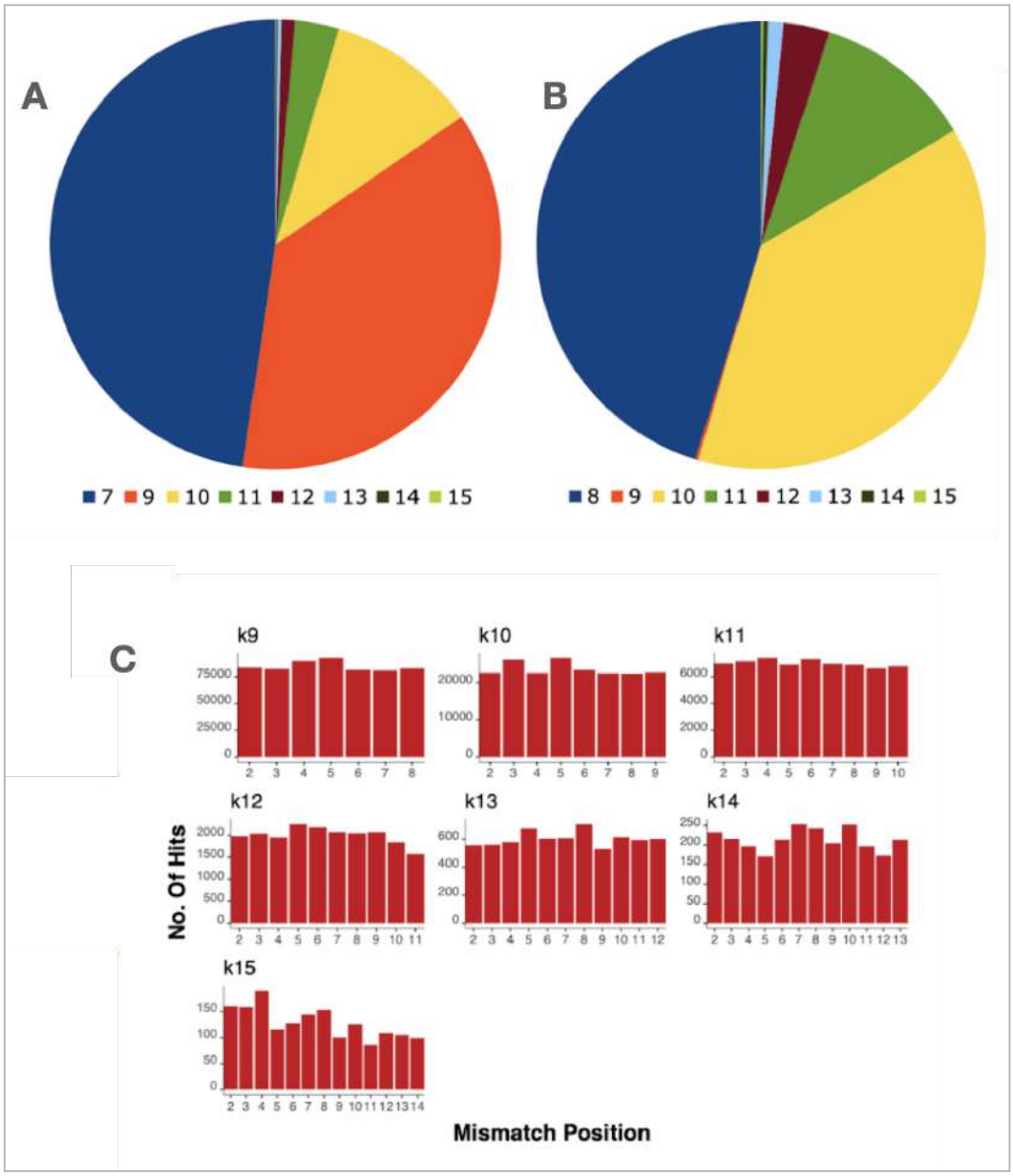


Figure 5.2.2 - (A) Pie chart displaying the distribution of 7 nt complementary stretches with 0 mismatch to longer stretches after the inclusion of 1 mismatch.
(B) Pie chart displaying the distribution of 8 nt complementary stretches with 0 mismatch to longer stretches after the inclusion of 1 mismatch.
(C) Histogram showing the number of hits with a mismatch at each position for complementary stretches of varying lengths. In this analysis, only non-terminal mismatches were considered.

With the inclusion of a mismatch, the pooled hits at lengths of 7 and 8 nucleotides was 13.6% lower in ES30L than the CSL segments (Figure 5.2.1A, C). This is because more than 50% of the complementary hits between ES30L and transcripts get extended to longer stretches (Figure 5.2.2A, B). We did not observe any positional preference for the mismatch with the complementary stretch (Figure 5.2.2C). Similar to the trend observed with 0 mismatch, ES30L showed 34.47% more complementarity at lengths of 9 nt or more than CSL segments (Figure 5.2.1A, C). We observed a drastic difference between ES30L and CSL for complementary stretches that were equal to or longer than 15 nt, with 1148 hits in the case of ES30L versus 80 hits on an average for the CSL segments.

Figure 5.2.3 - Gene Ontology of the genes with long complementary stretches to ES30L

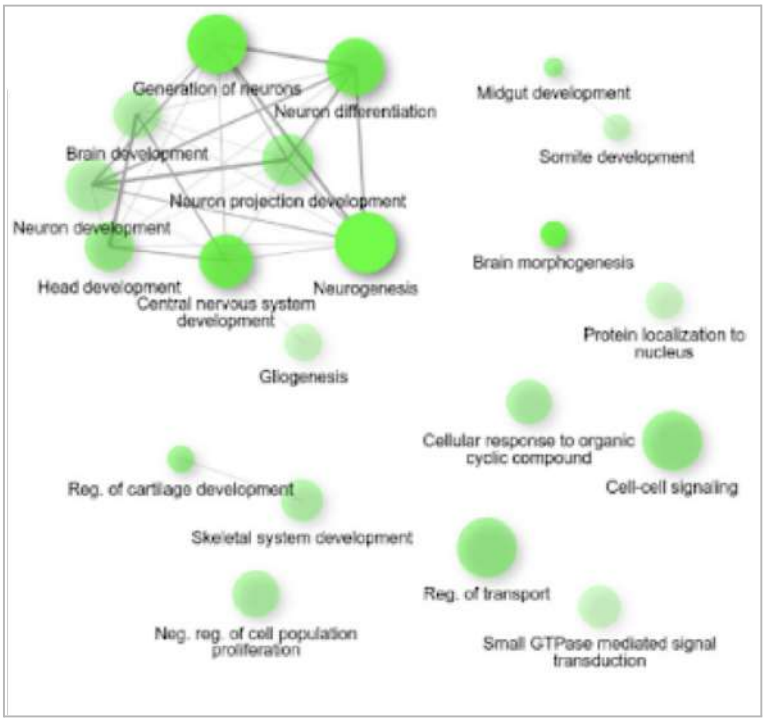


Figure 5.2.3 - This network plot displays the biological processes that the genes with 15 nt or longer stretches of complementarity to ES30L are known to be involved in.

The 1148 complementary transcript stretches were part of 560 genes. Further, gene set enrichment analysis of these 560 genes showed that they are involved in developmental pathways and neurogenesis among others (Figure 5.2.3). Exploring the occurrence and biological significance of interactions between ES30L and developmental genes *in vivo*, would be an exciting avenue of further investigation.

5.3. The complementarity to transcripts stems more from the second half of ES30L

We next wanted to check whether the complementarity to transcripts were preferential to any region in ES30L. Our analysis revealed that, in total, around 66% of the complementarity stemmed from the latter half of this segment (after position 25 in the 60 nt stretch). This percentage ranged from 62% for 7 nt stretches to nearly 76% for 15 nt complementary stretches (Figure 5.3A).

Figure 5.3 - Distribution of complementarity across ES30L

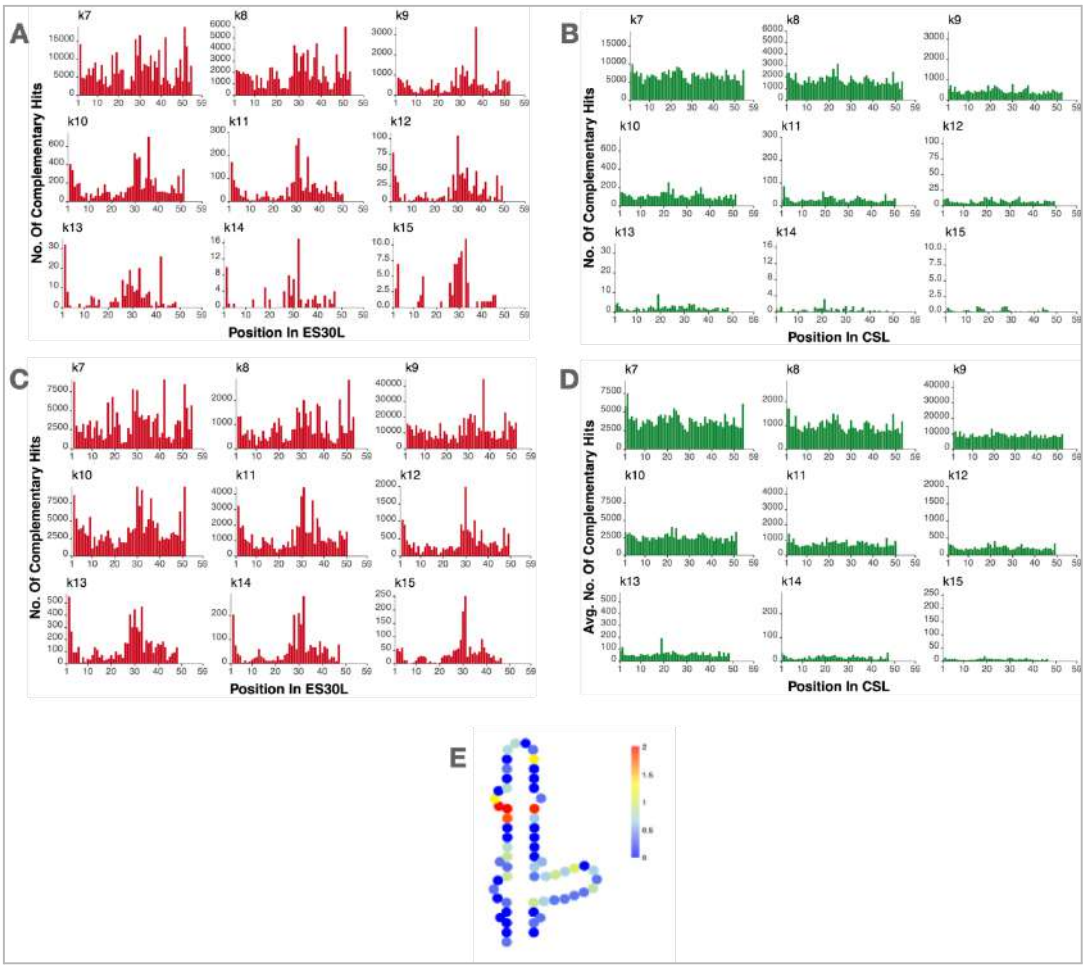


Figure 5.3 - (A) & (B) Histogram showing the spread of complementarity to transcripts across ES30L (*red*) and CSL segments (*green*) with 0 mismatch.
(C) & (D) Histogram showing the spread of complementarity to transcripts across ES30L (*red*) and CSL segments (*green*) with 1 mismatch.
(E) Secondary structure of human ES30L color-coded by entropy. An entropy value of 0 (*blue*) represents perfect conservation and that of 2 (*orange*) represents no conservation.

A similar trend was observed with stretches that included one mismatch (61% for 7 nt stretches to about 72% for 15 nt stretches post the 25th position on ES30L) (Figure

5.3C). In case of CSL segments, such a pattern was not observed as the complementarity was uniform across the segments, at all lengths either with or without a mismatch (Figure 5.3B, D). Of particular note was our prior observation that the second half of ES30L appeared well conserved in mammals, with many conserved stretches of homoiteronic guanines (Supplementary Figure 5.3E). The correlation between this sequence conservation in human ES30L and the enrichment of complementarity from this region is an interesting aspect that is yet to be explored.

5.4. The density of transcript stretches complementarity to ES30L is the highest in the 5' UTR with an enrichment around the start codon

We also wanted to check if there was a pattern in the distribution of complementarity to ES30L across protein-coding transcripts. Our analysis revealed that the density of perfectly complementary hits to ES30L was at least three to six times higher in 5' UTR than in CDS and 3' UTR of transcripts (Figure 5.4.1A). The results were similar even after the inclusion of a mismatch (Figure 5.4.1B). This could possibly be due to the GC-richness of ES30L and the 5' UTR regions in eukaryotes [153]. In the case of CSL segments, this difference between the other transcript sectors was not as stark as that for ES30L, either with or without a mismatch (Figure 5.4.1A, B).

Figure 5.4.1 - Density of complementary hits across the transcript sectors

Mismatch = 0

Category	No. Of Complementary Hits	
	ES30L	CSL
All mRNA sectors	525190	489677
5' UTR (per 100nt)	1.0347	0.4759
CDS (per 100nt)	0.3437	0.2789
3' UTR (per 100nt)	0.2128	0.2826

Mismatch = 1

Category	No. Of Complementary Hits	
	ES30L	CSL
All mRNA sectors	1124509	930524
5' UTR (per 100nt)	1.9605	0.7108
CDS (per 100nt)	0.6605	0.4948
3' UTR (per 100nt)	0.3958	0.5155

Figure 5.4.1 - Table showing the total number of complementary hits that are 7nt to 15nt long across transcript sectors with 0 (A) and 1 mismatch (B). The tables also include the number of hits for every 100 nt across each transcript sector.

Overall, the complementarity to ES30L (both with and without mismatch) showed the highest enrichment around the start codon (Figure 5.4.2A, C). Even in the region flanking the start codon, the 50 nt upstream stretch in the 5' UTR exhibited a higher

propensity to possess complementarity than the downstream stretch in the CDS (4.17% in 5' UTR vs 3.98% in CDS of total hits with 0 mismatch).

With the CSL segments, the enrichment was higher at 2.33% of total hits in the first 50 nt in CDS than the 50 nt upstream stretch in 5' UTR, which had 1.99% (Figure 5.4.2B). This trend was persistent even when one mismatch was allowed in the complementary stretches (Figure 5.4.2D). These observations suggest that ES30L had a higher probability of binding to the 5' UTR or early CDS stretches in transcripts, although interactions further downstream is also possible. We speculate that any interactions in the CDS or 3' UTR may contribute to translational pausing, mRNA targeting and decay.

Figure 5.4.2 - Distribution of complementarity across the protein-coding transcripts

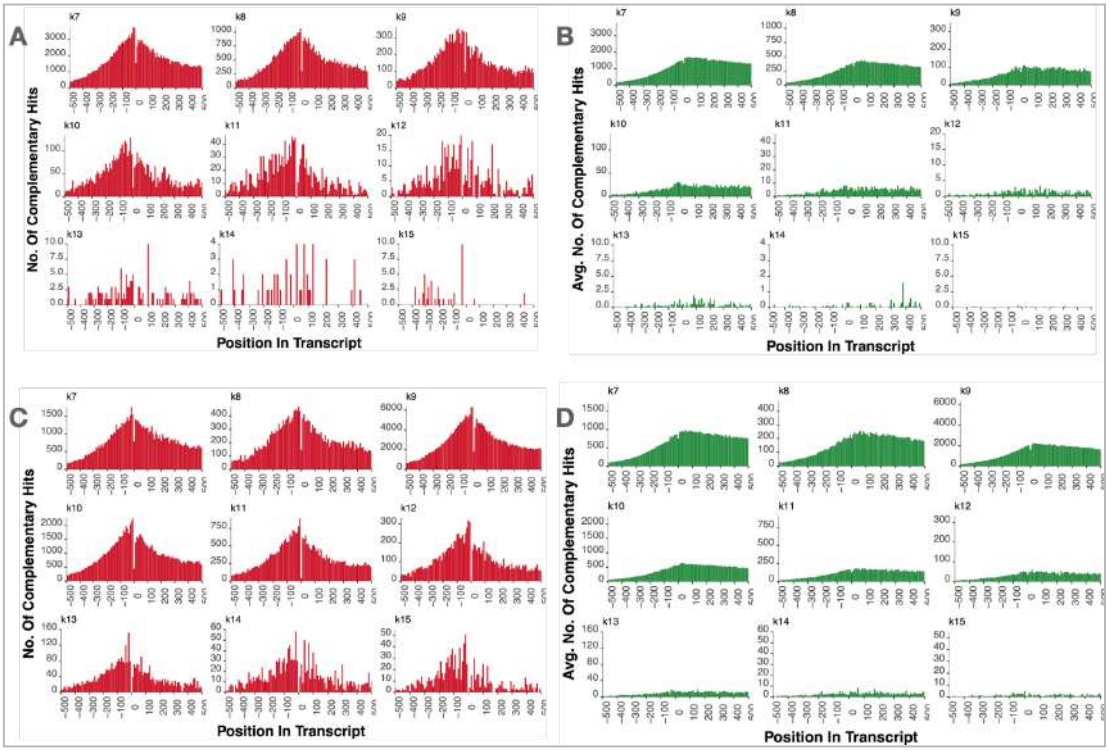


Figure 5.4.2 - (A) & (B) Histogram showing the spread of complementarity to ES30L (*red*) and CSL segments (*green*) with 0 mismatch, across transcripts.
(C) & (D) Histogram showing the spread of complementarity to ES30L (*red*) and CSL segments (*green*) with 1 mismatch, across transcripts.
Note: In this figure, '0' on the x-axis denotes the start codon, with negative coordinates indicating 5' UTR. The 500 nucleotides flanking the start codon have been shown in the plot to highlight the difference in both pattern and magnitude of complementarity to transcripts observed between ES30L and the CSL segments.

5.5. The number of stretches complementary to ES30L varies among transcripts

Out of the 83129 transcripts used in this analysis, 84.54% has at least one stretch complementary to ES30L, with the number increasing to 91.77% post the inclusion of a mismatch. With the CSL segments, nearly 90% of transcripts on an average have at least one complementary stretch, with the proportion going up to 96.57% with one mismatch. Many of these transcripts have multiple complementary stretches to ES30L (Figure 5.5B). Our data showed that 36 transcripts had over 100 stretches and 389 transcripts that possessed between 50 to 100 such stretches.

With one mismatch, this number was higher (514 transcripts with more than 100 stretches and 3301 transcripts with 50 to 100 stretches). Gene ontology analysis revealed that the genes that had a hit frequency of 25-50 and 50 or higher, both contained genes that are involved in neuronal development and differentiation (Figure 5.5A, C). Based on our prior data, genes that had complementary stretches that were 15 nt or longer also included genes that were implicated in neuronal development. Therefore, we think delineating any potential influence of ES30L on these genes in neurons merits further in depth experimental probing.

Figure 5.5 - Transcripts with multiple complementary stretches to ES30L

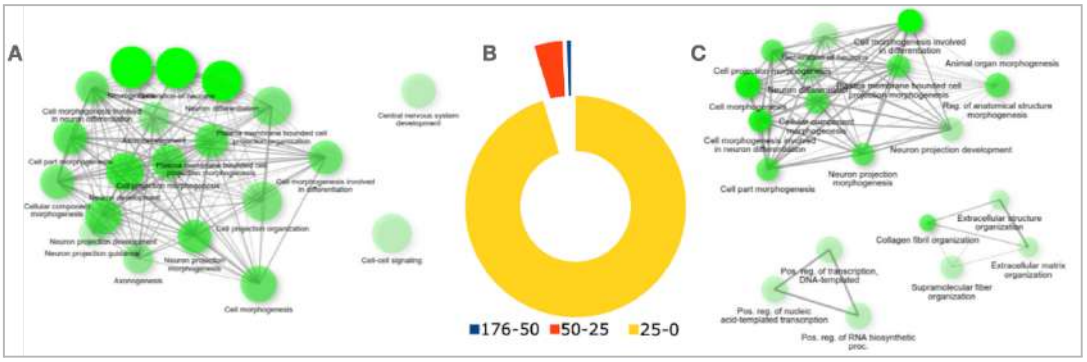


Figure 5.5 - (A) Network plot of biological processes for genes that possess between 25 and 50 stretches perfectly complementary to ES30L.
(B) Pie chart depicting the spread of transcripts containing multiple complementary stretches to ES30L.
(C) Network plot of biological processes for genes that possess between 50 and 176 stretches perfectly complementary to ES30L.

5.6. Summary and Discussion

Our analysis shows that the human ES30L exhibits extensive complementarity to many protein-coding transcripts, which is higher than that seen in the control segments. Given the pervasiveness of the complementarity observed, it is unclear

what could impart selectivity to any potential interactions between the transcripts and ES30L. Our speculation is that two factors may contribute: 1)Number of complementary stretches present in a transcript, which could increase the possibility of an interaction. 2)Length of the complementary stretches. From our data, we observe that nearly 30% of the transcripts possess a minimum of one complementary stretch that is 9 nt or longer, which increases further with the inclusion of a mismatch. This raises the possibility that some transcripts could contain even longer stretches with multiple noncontiguous contact points between ES30L and transcript segments, which could possibly aid in binding. This aspect could be particularly relevant if we factor in the single nucleotide polymorphisms and other variations that could be present in the interacting RNA stretches. Thus, our data so far indicates that ES30L possesses a good potential to bind to transcripts, which we wanted to explore further.

Chapter 6

ES30L could bind to protein-coding transcripts

6.1. Introduction

Experimental evidence for mRNA:rRNA ES interactions is sparse and is an area that requires extensive investigation. So far, only a few ESs such as ES6S, ES7S, ES9S have been reported to interact directly with mRNAs and these interactions have been shown to involve both base-pairing and secondary structural elements. To elaborate, human ES6S has been implicated to interact with an IRES element in the 5' UTR of Rbm3 transcripts both in various mammalian cell lines and in cell free translation systems. The study speculated that a region within the IRES element that was complementary to the expansion could be involved in the interaction [97].

The first concrete evidence of an ES:mRNA interaction was a cryo-EM structure of human ES9S binding to a stem-loop element in the 5' UTR of Hoxa9 mRNA, which regulates the IRES-dependent translation of the downstream open reading frame [72]. Very interestingly, the yeast ES9S that varied from its orthologous expansion in humans, did not bind to the mRNA, thereby opening up the possibility of species-specific differences in the function of ESs. Further, the ES9S has also been shown to interact with multiple mRNAs across 5' UTR, CDS and 3' UTR in mouse embryonic stem cells, which may potentially involve base complementarity [154].

Another example of a structure-based interaction is that of human ES7S binding to HCV (Hepatitis C Virus) IRES, shown through cryo-EM [155]. There is no experimental data on such interactions for any of the other ESs. Since our bioinformatic analyses revealed the presence of complementarity to ES30L in many protein-coding transcripts and given the flexible tentacle-like nature of ES30L, we wanted to check whether ES30L could bind to protein-coding transcripts.

6.2. *In vitro* pulldown of transcripts from HEK293T lysates using ES30L mimic

To test whether ES30L can interact with mRNAs, we performed *in vitro* RNA pulldown with a biotinylated ES30L mimic in HEK293T cell lysates. As a control, we used a biotinylated random fragment of the same length. The sequences used in the

pulldown are included in Table 2.4. Both these samples, along with the total RNA (input) were prepared in duplicates. Poly(A) selection was done on the isolated RNA, followed by high-throughput sequencing on Illumina HiSeq 2500 and RNA-Seq data analysis.

Figure 6.2.1 - Transcripts enriched in the ES30L fraction in the in vitro pulldown

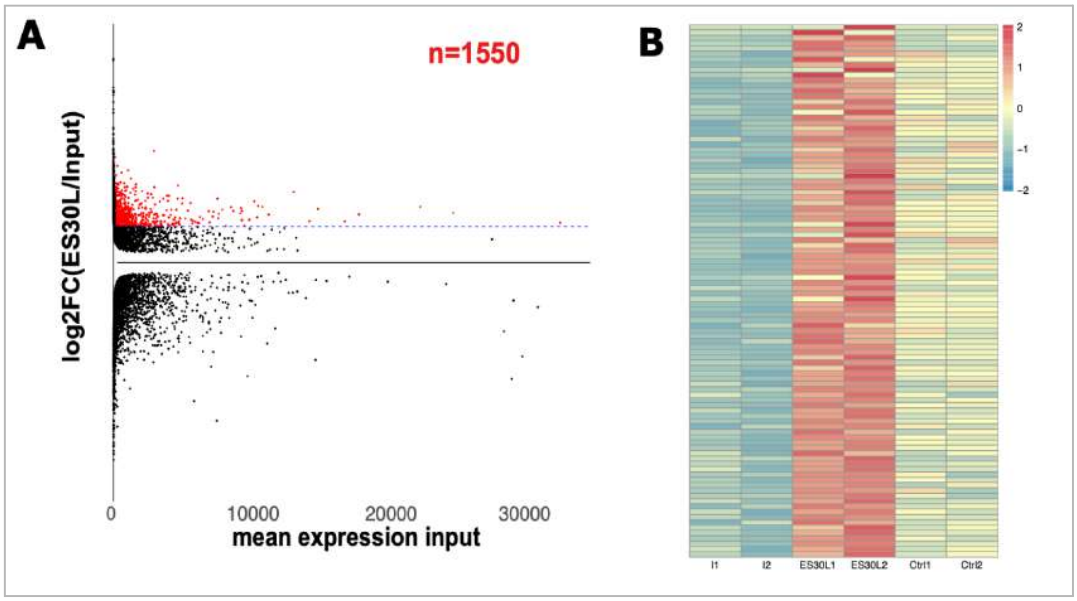


Figure 6.2.1 - (A) MA plot highlighting the 1550 transcripts that show a more than 2-fold enrichment in the ES30L fraction over input (*red*).
(B) Heatmap showing the expression profile of the 100 transcripts out of the 1550, that show a 50% higher enrichment over input in ES30L when compared to the control.

Differential expression analysis of protein-coding transcripts between pulldown fractions and input showed that 1550 transcripts were more than two-fold enriched (FDR<0.05) in the ES30L fraction over input (Figure 6.2.1A). Out of the 1550, 100 transcripts had a 50% higher enrichment over input than the control fragment (Figure 6.2.1B; Table 6.2).

These 100 transcripts contained genes that are known to be involved in several RNA related processes such as RNA metabolism and translation (Figure 6.2.2; Table 6.2). We next wanted to overlap this pulldown data with our complementarity data from our bioinformatic analysis.

Figure 6.2.2 - Gene ontology of transcripts that are differentially enriched in ES30L

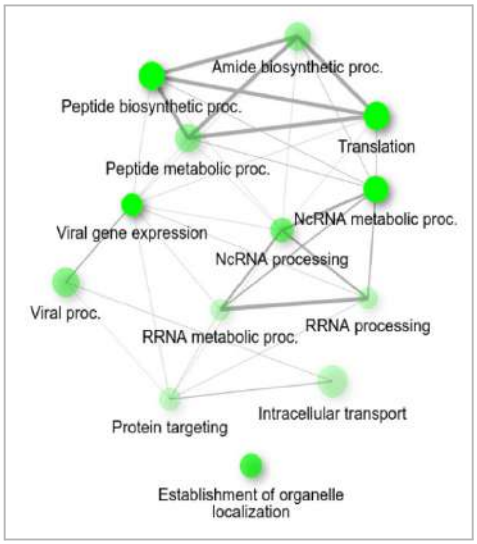


Figure 6.2.2 - Network plot of biological processes for 100 transcripts that show a higher enrichment in ES30l vs control.

6.3. Interaction of transcripts with ES30L could involve base pairing

Probing for complementary stretches in the 100 differentially enriched transcripts from the ES30L fraction revealed that 91% of them included at least one perfectly complementary site to ES30L (93% when 1 mismatch is allowed). The different lengths of such stretches is illustrated in Figure 6.3.

Figure 6.3 - Complementarity to ES30L in the differentially enriched transcripts

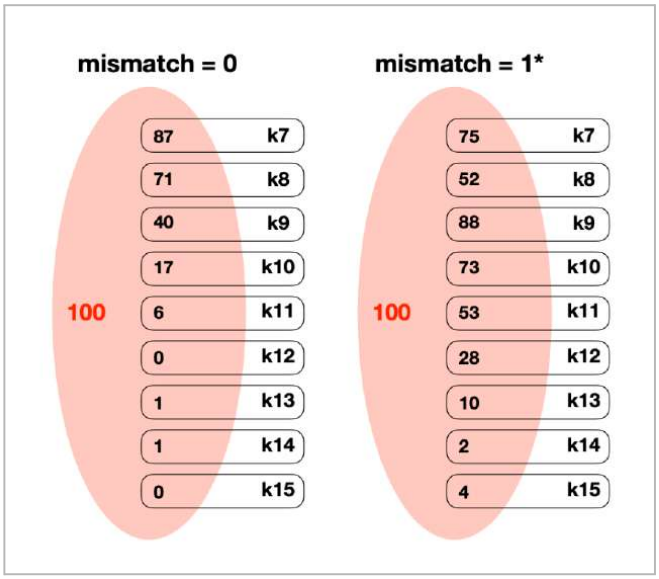


Figure 6.3 - Venn diagram enumerating the number of complementary stretches to ES30L of varying lengths in the 100 differential enriched transcripts in the *in vitro* pulldown.

Most of these transcripts possessed more than one complementary stretch, with 34% of them having more than 10 such stretches. Our data from this experiment shows that

some transcripts could bind to ES30L *in vitro*. The involvement of base-pairing in such interactions is correlative and needs to be probed further. Therefore, we next wanted to investigate whether ES30L can bind to transcripts *in vivo* and whether such interactions could also involve base-pairing.

6.4. Checking the ES30L binding to transcripts in HeLa cells by reanalysing a published RNA-RNA interactome dataset

We re-analyzed a published *in vivo* RNA-RNA interactome data to check for the presence of interactions between ES30L and other RNAs [156]. The study has reported pairwise RNA-RNA interaction dataset in HeLa cells, from which we were able to mine 193 interactions involving the ES30L region. These 193 interactions included the association of ES30L with mostly protein-coding transcripts (143 interactions involving 118 genes), other regions of the rRNA and one with tRNA (Figure 6.4). These interactions were present across 5' UTR, CDS as well as 3' UTR.

Figure 6.4 - In vivo interactions between ES30L and other RNAs

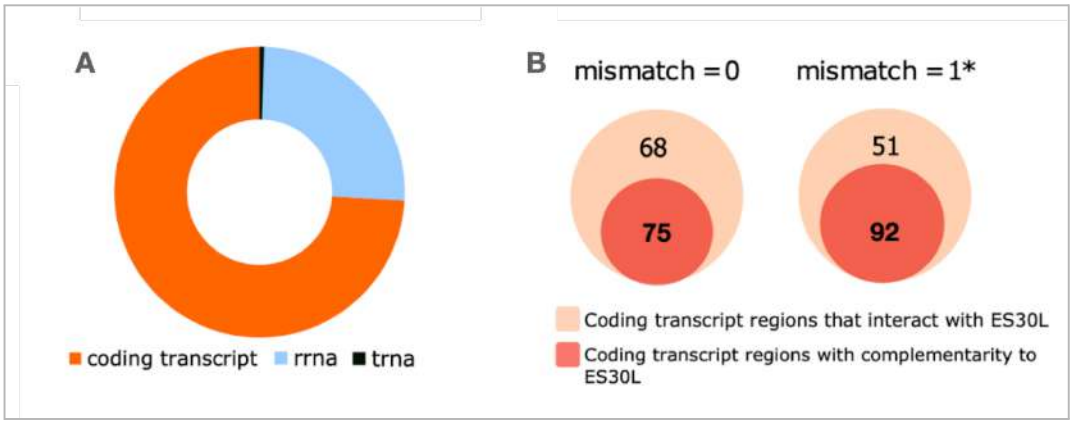


Figure 6.4 - (A) Pie chart displaying the RNA categories that ES30L binds to in the *in vivo* interactome data.
(B) Venn diagram illustrating the presence of complementarity to ES30L in the 143 transcript stretches associated with ES30L that were mined from the *in vivo* interactome.

We extracted the 143 transcript regions reported to bind to ES30L and checked for presence of complementarity to ES30L in these sequences. We noted that 75 of these regions included at least one stretch that was perfectly complementary to ES30L, which increased to 92 post the inclusion of a mismatch (Figure 6.4B). Further, out of the 118 coding genes, 25 were also enriched in the ES30L fraction in our *in vitro* pull down. The overlap could possibly vary, because the interactome and our pulldown employ different techniques and contain data from different cell lines. Even in these

25 genes, around 68-80% had at least one complementary region to ES30L with or without a mismatch, which could be higher when more mismatches are allowed. This observation bolsters our hypothesis that interaction of ES30L to transcripts could involve base-pairing.

6.5. A few transcript regions that bind to ES30L could be part of putative Internal Ribosomal Entry Sites

Interactions between RNAs could also involve secondary structural elements. To check for this, we downloaded putative IRES elements reported in the 118 protein-coding genes from the interactome data [134,135]. Our analysis showed that 14 of these genes contain putative IRES elements in their 5' UTR. We also observed that a part of IRES stretch in 3 out of the 14 genes interacted with ES30L based on the interactome data and also included complementarity to ES30L (Figure 6.5).

Figure 6.5 - ES30L may potentially interact with IRES of some transcripts

Gene	Predicted IRES	Putative ES30L Interacting region from SPLASH-seq	Complementarity in the interacting region
RPS11	1-174	100-174	Yes
NHP2	1-174	101-173	
FHT2	24-235	78-177	

Figure 6.5 - Figure tabulating the details of the transcript stretches interacting with ES30L, that are part of putative IRES elements. The interacting stretches also include complementarity to ES30L.

Given the context-dependence of *in vivo* RNA-RNA interactions, we wanted to probe the potential of ES30L to bind to the other putative IRESs. Towards this, we checked for complementarity to ES30L in all the downloaded IRESs, which revealed the presence of multiple such stretches in more than 50% of the sequences. Together, our analysis hints that ES30L can potentially bind to complementary transcript regions, some of which may also be a part of putative IRES structures.

6.6. Interaction with ES30L may involve secondary structural motifs in the 5' UTR of the transcripts

We next wanted to investigate any potential RNA motifs in the 5' UTR of the 100 transcripts differentially enriched in our pulldown. We extracted the stretches perfectly

complementary to ES30L along with 20 flanking nucleotides in the 5' UTR of these transcripts and performed *in silico* structure prediction using CMfinder. Out of the 100, 53 transcripts had 229 complementary stretches in their 5' UTR.

Figure 6.6 - Predicted RNA motifs in the 5' UTR of enriched transcripts from in vitro pulldown

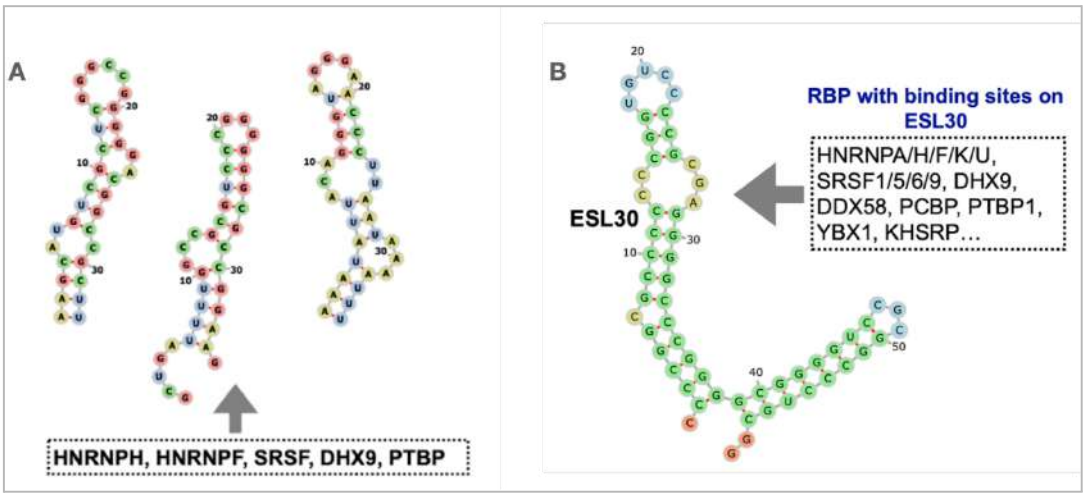


Figure 6.6 - (A) This figure includes consensus RNA motifs obtained from our *in silico* prediction using 5' UTR complementary stretches to ES30L from enriched transcripts. These motifs also contain binding sites for various RNA Binding Proteins (RBPs), listed in the figure.
(B) Predicted secondary structure of ES30L, which contains binding motifs for various RBPs.

Our *in silico* predictions indicated that 3 stem-loop RNA motifs could potentially form in these regions (Figure 6.6A). The first RNA motif was present in about 12% of the transcripts with the second and third ones present in 4% and 14% of the transcripts respectively. These RNA motifs also included binding sites for various RNA Binding Proteins (RBPs) such as HNRNPs, SRSFs, DHX9, PTBP. The protein binding motifs were obtained using the ATtRACT web server. Interestingly, ES30L also had binding motifs for the same set of RBPs (Figure 6.6B).

6.7. Summary and Discussion

In summary, our results show that ES30L can bind to protein-coding transcripts *in vitro*, which may potentially be mediated by base-pairing. This notion is further supported by an unbiased, already published RNA-RNA interactome data [156], which shows interactions involving ES30L *in vivo* as well. Some of these interactions may also involve secondary structural elements, as has been the case so far, for the reported interactions involving ESs. In fact, to reiterate, the L1-stalk, out of which ES30L protrudes out, is known to interact with IRES elements in a few mRNAs [151,152].

However, the *in cellulo* impact of these interactions needs further probing. One other observation of note, was the presence of binding motifs for the same RBPs on both ES30L and the transcripts enriched in our pulldown. This prompted us to investigate the protein interactome of ES30L.

Chapter 7

ES30L can interact with various RNA Binding Proteins

7.1. Introduction

The presence of complementarity between two RNA molecules alone may not be sufficient for an intermolecular interaction, because base-pairing would require the steric availability of the complementary stretches. Moreover, *trans*-acting factors like RNA Binding Proteins (RBPs) may also play a role in regulation of such interactions. In fact, several RBPs (~430) could interact with the ribosomes, form the 'ribo-interactome' and could influence translation regulation [157]. Some of these Ribosome Associated Proteins (RAPs) have been shown to interact with the tentacle-like arms of the large ESs like ES27L, ES7L, ES6S and are implicated to participate in processes like translation initiation, fidelity, peptide processing and targeting (Section 1.4.4). The association of RAPs to most other ESs is unknown. Many RBPs are known to bind to RNAs at GC-rich motifs and since ES30L is 92% GC-rich, we think it could bind to RBPs.

7.2. *In vitro* pulldown of proteins from HEK293T lysates using ES30L mimic

We pulled down proteins from HEK293T cell lysates using the biotinylated RNA molecules (ES30L mimic and random fragment) that we had used earlier (Table 2.4). The isolated proteins were quantified by mass spectrometry. Both these samples along with the total protein (input) were processed in 4 replicates (2 biological and 2 technical). Our analysis showed that fifty-eight proteins (Figure 7.2A; Table 7.2) that were more than 1.5 fold enriched (FDR<0.05) in the ES30L fraction when compared to input, but were not enriched in the control.

Figure 7.2 - Protein Interactome of ES30L

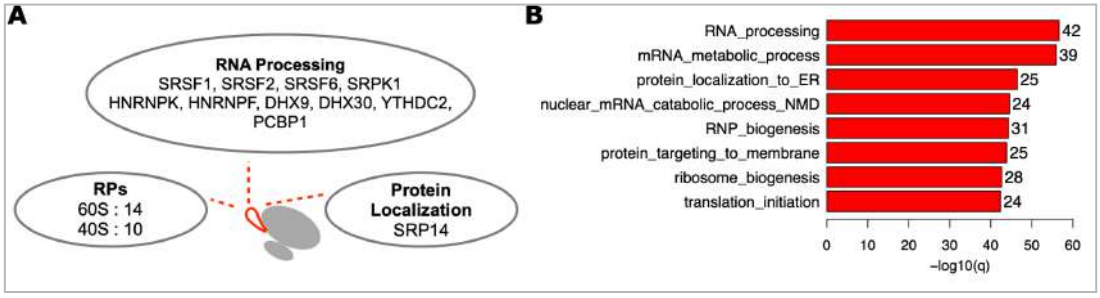


Figure 7.2 - (A) Illustration showing the broad categories of proteins that were differentially enriched in the ES30L fraction when compared to the control.

(B) Bar plot showing the gene ontology (biological processes) of the 58 proteins enriched in our *in vitro* pulldown.

These proteins belonged to various categories such as: 1)Ribosome biogenesis factors (NKRF, NCL) 2)Helicases (DDX21, DHX30, DHX9) 3)G-quadruplex binding proteins (hnRNPF) 4)Splicing factors (SRSF1, SRSF2, SRSF6, U2AF2) 5)Ribosomal proteins 6)Proteins involved in translation (PCBP1, HNRNPK, IGF2BP1, SRP14, RBM4). Overall, a gene ontology analysis of these proteins showed that many of them are involved in various aspects of RNA metabolism (Figure 7.2B).

7.3. Several proteins that could interact with ES30L are part of a published ribo-interactome

To check if any of the 58 proteins were reported to be RAPs, we overlapped our list with the 430 proteins from a published ribo-interactome data [158]. Our comparison showed that 42 out of the 58 proteins can bind to ribosomes (Figure 7.3). Curiously, we also noticed that RAPs enriched in our pulldown data such as the HNRNPs, SRSFs were predicted to have binding motifs both on ES30L and transcript 5' UTR stretches in our prior informatic analysis (Figure 6.6), reinforcing our speculation that these proteins could bind to both RNAs.

Figure 7.3 - Many of the potential ES30L interacting proteins are RAPs

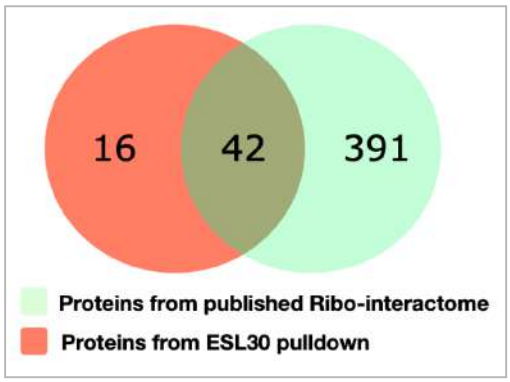


Figure 7.3 - This venn diagram highlights the overlap between the 58 enriched proteins from our pulldown and the published ribo-interactome.

7.4. ES30L can bind to RBPs in vivo based on reanalysis of eCLIP-seq data from K562 cells

We re-analyzed published CLIP-seq data from K562 cells for 7 RBPs that were differentially enriched in ES30L over input [159]. CLIP-seq data captures *in vivo* interactions between RNAs and the protein of interest. So we were interested in

probing whether these RBPs bind to ES30L stretch in the *in vivo* dataset. Our data showed that proteins such as DHX30, NPM1 showed enrichment in reads mapping to ES30L relative to input, suggesting that they could potentially interact with ES30L *in vivo* (Figure 7.4). The other proteins such as PCBP1, SRSF1, IGF2BP1, U2AF2 or HNRNPK were not enriched in ES30L over input in this dataset.

Figure 7.4 - Pattern of potential binding of RBPs on ES30L

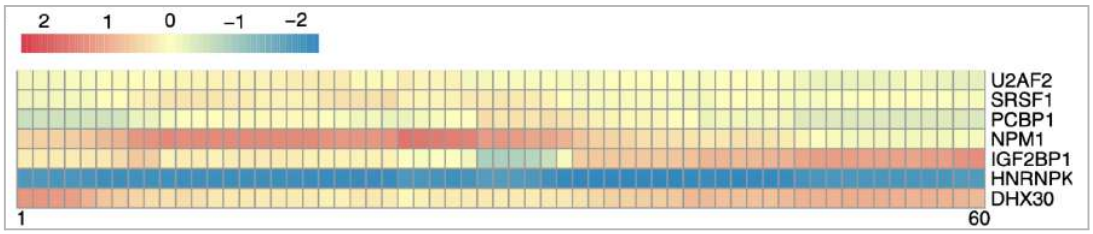


Figure 7.4 - This heatmap shows the fold change between the RNA binding protein and the control samples across ES30L.

7.5. The interaction of ES30L to proteins may be context-dependent

Molecular interactions are sensitive to their *in vivo* micro-environment and therefore we wanted to check whether RBPs can exhibit different binding to ES30L in different cell types.

Figure 7.5 - Potential binding pattern of PCBP1 on ES30L in K562 vs HCT116 cell lines

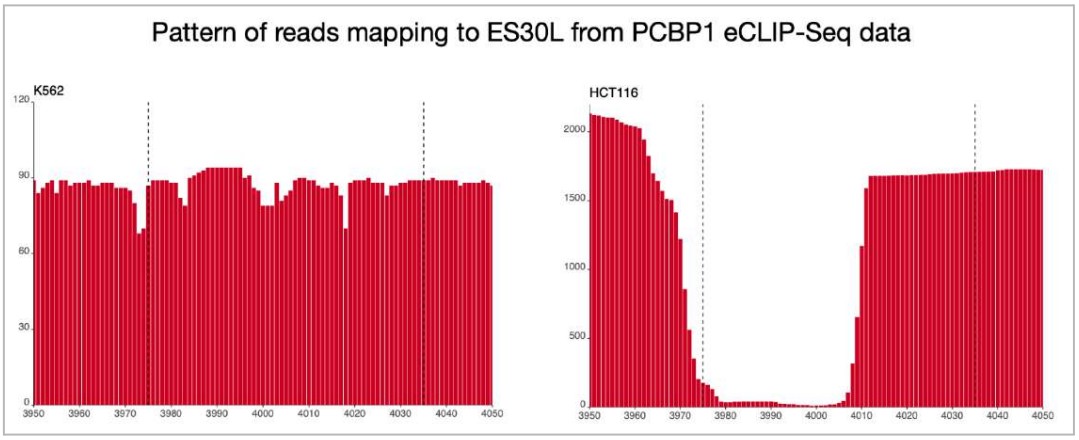


Figure 7.5 - These bar plots show the pattern of reads (indicating PCBP1 binding) mapping across the ES30L stretch in K562 (*left*) and HCT116 (*right*) cell lines.

Towards this, we considered CLIP-Seq data for PCBP1 from two K562 and HCT116 cells. Since data from input was unavailable from the study in HCT116 cell line, we used the read mapping pattern as a proxy to detect any potential difference in protein binding. Our analysis of the data from HCT116 cells [160], indicated a preferential

mapping of reads towards the latter half of ES30L (coordinates corresponding to nucleotides 4005-4035), with much fewer reads mapping to the first half (Figure 7.5). In contrast, read mapping data from K562 cells [161] showed no such mapping pattern along ES30L (Figure 7.5), hinting at the possibility that ES30L:PCBP1 interactions could be context dependent. This needs to be experimentally probed for various RBPs across cell types.

7.6. Summary and Discussion

To summarize, we have identified that many RBPs that are known to be involved in translation regulation can potentially bind to ES30L *in vitro*. Most of these proteins are known to associate with ribosomes and a few of them possibly interact with ES30L *in vivo*. Although, the interaction of such RBPs with ES30L inside the cell could be context-dependent. This could be especially important, as secondary structure predictions and structural modeling indicate that ES30L folds in to a helical structure [162–166]. Since we currently have no understanding of how stable the secondary structure of ES30L would be *in vivo*, we think the RBPs may regulate its accessibility for any potential interactions. Our notion is supported by the presence of helicases like YTHDC2, DDX21, DHX30, DHX9 which are selectively enriched in the ES30L fraction in our pulldown. DHX30 can impact global translation [167,168], while DHX9, DHX36 and DDX21 enable the translation of mRNAs by interacting with RNA G-Quadruplexes (RGs) in their 5' UTR in various human cell lines [169,170]. Another study [171–177] has proposed that DHX36 and hnRNPF (also part of our pulldown data) act in a concerted manner to regulate the translation of mRNAs via RG remodeling in glioblastoma. It is known that RGs can also form in rRNA in vivo in HeLa cells [178] and computational prediction shows that ES30L can also form RGs [179]. Detailed experimental analyses would be required to explore the impact of RGs in ES30L:mRNA binding and the role of RBPs in this context.

Several of the differentially enriched proteins in our pulldown, are known ITAFs (IRES Trans-Acting Factors), which can exert either a stimulatory or an inhibitory influence on IRES-mediated translation. For instance, PCBP1/2 along with hnRNPK have been shown to stimulate IRES-mediated translation of c-myc mRNA in vitro in rabbit reticulocyte lysates [180,181]. Other proteins from our analysis like SRSF1 [182],

SRSF6 [183], YTHDC2 [184], NCL [185], IGF2BP1 [186], SRP14 [187], RBM4 [188,189] have also been shown to activate translation of their cognate mRNAs. Conversely, other proteins from our data like DDX21 [190] and hnRNPK [191] have been shown to inhibit viral IRESs and negatively regulate translation in vivo in mammalian cell lines. Moreover, our bioinformatic analysis also reveals the presence of binding sites for many of these proteins in both ES30L and 5' UTR stretches from mRNAs enriched in our RNA pulldown. Such RBPs are known to possess multiple RNA binding domains [192], reinforcing our speculation that they could bind to both ES30 and mRNA to enable an interaction.

Chapter 8

Discussion and Conclusion

8.1. Key Observations from the study

Some of the key findings from my work are:

1. The large subunit ESs can vary both in terms of length and base composition across eukaryotes and may also contain species-specific variations.

2. ES30L has expanded the most in mammals, with the second-highest expansion in birds, but is absent in other eukaryotes. It is well conserved in both these clades.

3. This segment has a high degree of complementarity to several protein-coding transcripts, which are highly clustered around the start codon, with the highest density in the 5' UTR.

4. Our *in vitro* pulldown combined with our re-analysis of other published datasets shows that ES30L can interact with mRNAs, which could potentially involve base-pairing and 5' UTR stem-loop motifs.

5. A similar *in vitro* pulldown of proteins indicates that ES30L could bind proteins that are involved in various aspects of RNA metabolism. Interestingly, some of these proteins are ITAFs and are known to be involved in cap-independent translation.

8.2. Summary

ES30L is a highly flexible LSU ES that is enriched only in the endothermic vertebrates. Even though the distal ends of a few long ESs like ES7L or ES27L are also tentacle-like, they are positioned towards the back of the ribosome and are better accessible to the peptide exit point on the ribosome. Whereas, the tentacle-like ES30L is perfectly positioned as an extension of the L1-stalk, which is the side where E-site tRNA and the mRNA exit the ribosome. This aspect combined with its GC-rich composition, makes it a good candidate to interact with both transcripts and proteins, which is explored and highlighted by our work.

By collating our data with existing literature, we hypothesize that ES30L can bind to secondary structural elements of some mRNAs via base-pairing with the involvement of ITAFs, in a context-dependent manner. Such an interaction may influence translation initiation, although the effect could be either stabilizing or inhibitory.

8.3. Future Directions

Our hypothesis is quite challenging to test at present because rRNA is distributed as tandem repeats across five chromosomes in the human genome. Therefore knocking the expansion out or modifying it, are currently not possible. A knock-in strategy in to the rRNA of another species (say yeast) may also not truly reflect the *in vivo* relevance in mammals. This is because we currently lack any understanding of how ES30L and its interacting partners may have co-evolved in mammals.

One strategy could be to design a highly specific inhibitor to block ES30L, though a paucity of information on sequence variability and modifications present on the segment could present a bottleneck. But, if an inhibitor is successfully generated, it could be used to study the *in vivo* changes in protein synthesis under various conditions. Another strategy would be to probe the structure of ES30L in various molecular complexes and organisms, as it may provide insights into the dynamic conformational changes at different stages of translation. Given the potential of ESs to participate in diverse cellular processes, it is imperative that we expand the repertoire of techniques available to study these enigmatic segments.

Tables

Table 1.1: Ribosome size and overall composition across kingdoms

Type of ribosome		**Size of ribosome**	**Size and Composition of Large subunit (LSU)**	**Size and Composition of Small subunit (SSU)**
Cytosolic	Bacterial (e.g. *Escherichia coli*)	70S	50S (23S rRNA,5S rRNA + 33 proteins)	30S (16S rRNA +21 proteins)
	Archaeal	70S	50S (23S rRNA, 5S rRNA + 35–40 proteins depending on the organism)	30S (16S rRNA + 25–28 proteins depending on the organism)
	Eukaryote (*Saccharomyces cerevisiae*)	80S	60S (25S rRNA, 5S rRNA, 5.8S rRNA + 45 proteins)	40S (18S rRNA+33 proteins)
	Eukaryote (e.g. *Homo sapiens*)	80S	60S (28S rRNA, 5S rRNA, 5.8S rRNA + 46 proteins)	40S (18S rRNA+33 proteins)
Mitochondrial	Kinetoplastids (e.g. *Trypanosoma brucei)*	50S	40S (12S rRNA +77 proteins)	28-30S (9S rRNA +56 proteins)
	Mammal (e.g. *Homo sapiens*)	55S	39S (16SrRNA, CP tRNA, 52 proteins)	28S (12S rRNA+30 proteins)
	Fungi (e.g. *Saccharomyces Cerevisiae*)	74S	54S (21S rRNA+46 proteins)	37S (15S rRNA + ~36 proteins)
	Higher plants (e.g. *Brassica oleracea* var. *botrytis*, or cauliflower)	78S	50S (26SrRNA, 5S rRNA+ 45 proteins)	33S (18S rRNA+37 proteins)
Chloroplast	Plants	70S	50S (23S rRNA, 5S rRNA,4.5S rRNA +31 ribosomal protein)	30S (16S rRNA+ 21 ribosomal protein)

Table 2.1: 28S rRNA sequences mined for the multiple sequence alignment

Organism	Common name_Clade	Accession
Homo_sapiens	human_mammal	NR_003287.4
Macaca_nemestrina	pigtailed_macacaque_mammal	XR_003018306.1
Pongo_abelii	Sumatran_orangutan_mammal	XR_002913601.1
Canis_lupus	Dog_mammal	XR_003129431.1
Felis_catus	Cat_mammal	XR_002740182.1
Zalophus_californianus	California_sea_lion_mammal	XR_003523540.1
Ursus_arctos	Brown_bear_mammal	XR_003313602.1
Cavia_porcellus	Guinea_pig_mammal	XR_002790568.1
Mus_musculus	Mouse_mammal	NR_003279.1
Cricetulus_griseus	Chinese_hamster_mammal	XR_003488776.1
Rattus_norvegicus	Rat_mammal	NR_046246.1
Bos_taurus	Cow_mammal	NR_036644.1
Bubalus_bubalis	buffalo_mammal	XR_003109814.1
Ovis_aries	Sheep_mammal	XR_003587871.1
Equus_caballus	Horse_mammal	XR_002805825.1
Manacus_vitellinus	golden_collared_manakin_bird	XR_003213097l7918
Chiroxiphia_lanceolata	Lance_tailed_manakin_bird	XR_004354949.1l4282
Pipra_filicauda	wire_tailed_manakin_bird	XR_003540666.1l8682
Anas_platyrhynchos	dabbling_duck_bird	XR_003493880l4022
Aythya_fuligula	tufted_duck_bird	XR_004254457.1l4020
Gallus_gallus	Chicken_bird	XR_003078040.1l4437
Tyto_alba	barn_owl_bird	XR_004409519.1l4107
Taeniopygia_guttata	zebra_finch_bird	XR_004366572.1l4378
Camarhynchus_parvulus	small_tree_finch_bird	XR_004061554.1l4332
Calypte_anna	hummingbird_bird	XR_003988849.1l4324
Malaclemys_terrapin_terrapin	diamondback_turtle_reptile	MDXI01019247.4129.7973
Chelonoidis_abingdonii	pinta_giant_tortoise_reptile	XR_004375038.1
Terrapene_carolina	common_box_turtle_reptile	XR_003370443.1

Alligator_sinensis	alligator_reptile	XR_003094940.1
Pogona_vitticeps	central_bearded_dragon_reptile	HACI01071117.32.3921
Protobothrops_mucrosquamatus	viper_reptile	XR_003810363.1
Rhacophorus_omeimontis	Omei_tree_frog_amphibian	GEGF01000170.623.4124
Xenopus_tropicalis	western_clawed_frog_amphibian	XR_004223802.1
Microcaecilia_unicolor	caecilian_amphibian	XR_003942336.1
Rhinatrema_bivittatum	caecilian_amphibian	XR_003855279.1
Ambystoma_macrodactylum	longtoed_salamander_amphibian	AF212178.10.3557
Ambystoma_mexicanum	axolotl_amphibian	GEBK01002281.81.3326
Bombina_bombina	firebellied_toad_amphibian	HADU01003978.227.3157
Salmo_salar	Atlantic_salmon_fish	AGKD04000353.979981.983797
Acipenser_brevirostrum	shortnose_sturgeon_fish	U34340.1.3432
Tachysurus_fulvidraco	catfish_fish	XR_003439232.1
Mastacembelus_armatus	eel_fish	XR_003296402.1
Dicentrarchus_labrax	European_seabass_fish	CBXY010019217.816.4723
Squatina_californica	Pacific_angelshark_fish	AY049857.254.3970
Takifugu_rubripes	pufferfish_fish	XR_003887305.1
Petromyzon_marinus	lamprey_fish	XR_004401899.1
Ciona_savignyi	urochordate	AC137268.1.2895
Thalia_democratica	urochordate)AF158725.718.4310	2711-2717
Styela_plicata(urochordate	AF158724.682.4336	2755-2761
Oikopleura_dioica	urochordate	GCJN01008243.593.3987
Branchiostoma_belcheri	Belcher's_lancelet_cephalochordate	AYSS01013736.1.2778
Branchiostoma_floridae	cephalochordate	AF061796.440.4275
Echinoderes_microaperturus	mud_dragon_cephalochordate	LC007066.1.3248
Saccoglossus_kowalevskii	acorn_worm_hemichordate	NR_028490.2
Saccoglossus_pusillus	hemichordate	AF212174.620.4673

Strongylocentrotus_purpuratus	purple_sea_urchin_echinoderm	XR_004063452.1
Paracentrotus_lividus	common_urchin_echinoderm	AM981272.3360.7262
Florometra_serratissima	feather_star_echinoderm	AF212168.379.4060
Ophiothrix_angulata	brittle_star_echinoderm	KU895460.1.3847
Drosophila_erecta(fruitfly_arthropod	XR_003408247.1	3078-3096
Drosophila_miranda	fruitfly_arthropod	XR_004474788.1
Nylanderia_fulva	crazy_ant_arthropod	XR_003820087.1
Tetranychus_urticae	mite_arthropod	XR_003082117.1
Caenorhabditis_elegans	nematode	NR_000055.1
Caenorhabditis_briggsae	nematode	JN636061.2935.6374
Necator_americanus	hookworm_nematode	XR_001230651.1
Brevibucca_saprophaga	roundworm_nematode	EU195990.1.3407
Eisenia_fetida	earthworm_annelid	AF212166.1.3633
Urechis_caupo	innkeeper_worm_annelid	AF342804.560.4275
Dorvillea_erucaeformis	annelid	AY838859.1.3575
Corbicula_japonica	brackishwater_clam_mollusc	AB126330.1.4195
Lima_fujitai	Spiny_Shell_clam_mollusc	AB102740.1.3779
Lyonsia_kawamurai	mollusc	AB126334.1.3964
Siphonaria_pectinata	sea_snail_mollusc	KU341320.1.3866
Pomacea_canaliculata	apple_snail_mollusc	XR_003099035.1
Octopus_vulgaris	octopus_mollusc	XR_003883157.1
Schistosoma_mansoni	platyhelminth	AY157173.1.3844
Schmidtea_mediterranea	platyhelminth	AAWT01084176.1.2223
Hoploplana_californica	platyhelminth	KC869850.1.3345
Paraplanocera_oligoglena	platyhelminth	KC869849.1.3338
Nanophyetus_schikhobalowi	platyhelminth	LN871820.1.3885
Bdelloura_candida	triclad_platyhelminth	GFKB01076713.680.4185
Brachionus_manjavacas	planktonic_rotifer	GFGK01001986.2811.6301
Brachionus_calyciflorus	planktonic_rotifer	KC193096.1.3384
Sinantherina_socialis	rotifer	AY210471.1.3406

Table 2.2: Coordinates of the LSU ESs in *H. sapiens*

ES	Co-ordinates			GC%
	U13369_7935-12969 (taken from Parker et al., 2015)	**Mapping in PDB:4UG0**	**Mapping in NR_003287.4**	
ES5L	114-156	114-159	114-159	78.57
ES7L	465-1265	474-1278	474-1278	83.75
ES9L	1384-1487	1397-1500	1397-1500	80.58
ES10L	1682-1711	1695-1724	1695-1724	75.86
ES12L	1793-1829	1806-1842	1806-1842	77.78
ES15L	2075-2256	2088-2269	2088-2269	83.98
ES19L	2439-2494	2452-2507	2452-2507	74.55
ES20L	2538-2575	2551-2588	2551-2588	70.27
ES24L	2685-2705	2698-2718	2698-2718	85
ES26L	2751-2764	2764-2777	2764-2777	61.54
ES27L	2875-3586	2888-3607	2888-3607	86.64
ES30L	3955-4013	3976-4035	3976-4035	91.8
ES31L	4057-4124	4079-4154	4079-4154	83.58
ES39L	4698-4905	4729-4940	4729-4940	82.61
ES41L	4983-4996	5018-5031	5018-5031	69.23

Table 2.3: 28S rRNA core segment (CSL) sequences used in the complementarity analysis

S. No.	Sequence	Reference	Coordinates		Part of
			From	To	
1	GATCAGACGTGGCGACCCGCTGAATTTAAGCATATTAGTCAGCGGAGGAAAAGAAACTAA	U13369.1:7935-12969	11	71	CSL5
2	TAAACTCCATCTAAGGCTAAATACCGGCACGAGACCGATAGTCAACAAGTACCGTAAGGG	U13369.1:7935-12969	307	367	CSL7
3	TTGAAACACGGACCAAGGAGTCTAACACGTGCGCGAGTCGGGGGCTCGCACGAAAGCCGC	U13369.1:7935-12969	1306	1366	CSL9
4	ATAGGGGCGAAAGACTAATCGAACCATCTAGTAGCTGGTTCCCTCCGAAGTTTCCCTCAG	U13369.1:7935-12969	1608	1668	CSL10
5	CGAATCAACTAGCCCTGAAAATGGATGGCGCTGGAGCGTCGGGCCCATACCCGGCCGTCG	U13369.1:7935-12969	2010	2070	CSL15
6	AGCAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGAAGGGTTCCATGTGAACAGC	U13369.1:7935-12969	2350	2410	CSL19
7	CGGGAGCCCCGGGGAGAGTTCTCTTTTCTTTGTGAAGGGCAGGGCGCCCTGGAATGGGTT	U13369.1:7935-12969	2594	2654	CSL24
8	ACAGCCTCTGGCATGTTGGAACAATGTAGGTAAGGGAAGTCGGCAAGCCGGATCCGTAAC	U13369.1:7935-12969	2800	2860	CSL27
9	GGCGGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTAGTCTGGCACGGTGAAG	U13369.1:7935-12969	3864	3924	CSL30
10	GAAAAGTTACCACAGGGATAACTGGCTTGTGGCGGCCAAGCGTTCATAGCGACGTCGCTT	U13369.1:7935-12969	4347	4407	CSL39

Table 2.4: Synthesized biotinylated RNA oligonucleotides used in the *in vitro* pulldowns

Sequence Name	RNA Sequence	Length (nt)	GC content (%)	Modification
ES30L mimic	GGAGGCCCCCGGCGCCCCCCCGGUGUCCCCGCGAGGGGCCCGGGGCGGGGUCCGCGGCCCUGCGGGCCGCC	71	91	5’ Biotin tag
Random Fragment	CGUACAGCACAUAUAUCUGCUCUACAAGUGAUGCAGUUGUGUCGUAGCGUGCAGCCCGAUCGUAGUAUCAU	71	50	5’ Biotin tag

Table 6.2: Transcripts from the RNA pulldown that are more than two-fold (linear scale) enriched in ES30L over input (q< 0.05) and also show a more than 50% enrichment in ES30L when compared to the random fragment over input.

		ES30L/Input		RandomFragment/Input	
Transcript ID	Gene Name	log2FC	q value	log2FC	q value
ENST00000525675.1	C11orf98	5.625644387	0.000897679	2.374248887	0.260419144
ENST00000637264.1	POLG	4.833613533	0.035072547	1.613148121	0.627894569
ENST00000536055.5	ATAD3A	4.725786249	0.030781998	3.04270951	0.228382906
ENST00000563819.5	DHX38	4.715010436	0.033191499	2.018030772	0.497061906
ENST00000535788.1	UBB	4.628041272	1.84E-12	-0.98160684	0.470226185
ENST00000499869.6	WDR1	3.552796203	0.048457348	2.789373512	0.151694941
ENST00000579584.5	MYO1D	3.339214701	0.008789307	2.70673074	0.044959346
ENST00000421699.6	GOLGA2	3.292730578	0.008789751	2.412124929	0.080267447
ENST00000520915.5	NPIPB13	3.099019792	0.001540721	1.583247808	0.180406265
ENST00000447319.5	HMGN5	3.020900442	0.009874456	1.702906344	0.223246136
ENST00000618991.4	FGF8	2.988985384	0.016151732	1.892116119	0.186880705
ENST00000581012.6	UTP6	2.839524798	0.004746597	2.217056605	0.038952201
ENST00000619601.1	GAPDH	2.833010814	0.006849433	1.740758828	0.151236822
ENST00000531205.5	SIGIRR	2.744336434	0.003282824	1.646663028	0.128383333
ENST00000588025.5	FARSA	2.741350038	0.001385226	2.103347105	0.021612891
ENST00000444620.2	ARID4B	2.689806383	0.018257878	1.391462365	0.321928797
ENST00000354694.11	PES 1	2.660282659	0.011044434	1.917363609	0.095558136
ENST00000473616.5	NEBL	2.617506364	0.008943298	1.529586144	0.193879515
ENST00000429697.1	KHK	2.605211021	0.001060058	1.858393354	0.031810887
ENST00000389142.9	OS9	2.565675525	0.001470816	1.759337536	0.048120069
ENST00000342572.12	CYBC1	2.548249685	0.037961324	1.774255073	0.197755139
ENST00000478547.1	EIF3D	2.464898471	0.00745697	1.173599734	0.308825672
ENST00000356309.7	NCAPG2	2.363551134	0.015180187	1.774189935	0.093222792
ENST00000483966.3	HNRNPU	2.355892357	0.042919759	1.441175472	0.289863501
ENST00000645982.1	CTNNB1	2.34502979	0.007495426	1.448059736	0.151574607
ENST00000443831.1	LRP2	2.324688014	0.000157526	1.650903332	0.012740072
ENST00000264331.8	TOP2B	2.283029081	0.000500746	1.653902006	0.019428408

ENST00000360190.8	CDK5RAP2	2.260730645	0.013997193	1.267062831	0.247580761
ENST00000565504.5	TRAPPC2L	2.239843397	0.000828779	1.095336967	0.173240868
ENST00000403080.5	NFASC	2.220298874	0.013136493	1.141271273	0.298520737
ENST00000409711.1	RPL31	2.184931737	0.002099341	1.201530122	0.139558885
ENST00000368302.3	LAMTOR2	2.101912705	0.048129389	0.563233723	0.716677279
ENST00000305560.10	SPATA5L1	2.099769411	0.00415852	1.149239754	0.186394651
ENST00000230431.10	DNPH1	2.05256158	0.00032948	1.421489294	0.022207003
ENST00000521512.5	ANXA6	2.015860793	0.000277739	1.338352873	0.028217347
ENST00000404914.7	ATG4B	2.015279918	0.002878827	1.363672779	0.068937331
ENST00000547665.5	MARS	2.00208098	1.43E-23	1.277000733	7.21E-10
ENST00000547773.5	NAP1L1	2.001505966	0.000410972	1.134512758	0.080542295
ENST00000525235.1	ARFIP2	1.979020568	0.009552112	1.313740497	0.125893072
ENST00000249923.7	COPB1	1.944027046	0.011577205	1.358034487	0.109497934
ENST00000553878.5	CKB	1.936805001	0.010997061	1.347758002	0.109412646
ENST00000276416.10	BIN3	1.934834466	0.01457945	1.241438132	0.168075087
ENST00000253047.6	TMEM160	1.870529145	5.56E-29	0.77728043	1.27E-05
ENST00000557529.1	SHMT2	1.832192318	0.020382338	1.230201569	0.165539101
ENST00000419277.5	HLA_DPA1	1.805342054	0.001344889	1.197454225	0.05470351
ENST00000566283.1	TCF25	1.804080932	4.79E-05	1.155443669	0.017838906
ENST00000397442.5	CPNE1	1.796624105	0.041801583	1.020099153	0.33227517
ENST00000625194.3	ADSL	1.792390805	0.004353524	0.774641828	0.329673787
ENST00000437329.5	RPN2	1.776354051	0.028748684	0.957456001	0.328668345
ENST00000409860.1	C4orf48	1.664345723	0.000136672	0.229692205	0.747018055
ENST00000524552.5	HSPA8	1.661697236	1.27E-06	1.023164399	0.006285622
ENST00000360840.9	MRPL37	1.649535987	0.026339793	0.876997363	0.326518516
ENST00000595358.5	CCDC61	1.640144862	6.43E-05	0.966450733	0.035899944
ENST00000334678.7	RPS19BP1	1.633536248	1.90E-33	1.046228824	3.78E-14
ENST00000357886.8	CDK5RAP1	1.632363924	0.00213362	1.031560564	0.084656599
ENST00000399775.2	ENHO	1.625577504	9.21E-29	0.793387356	2.52E-07
ENST00000444548.5	NDUFA10	1.624129258	0.034978863	1.033798249	0.24032021
ENST00000382584.8	TTC14	1.613682733	2.24E-09	0.969759603	0.000814019

ENST00000444703.1	PPAN	1.611166072	0.001324542	0.89520969	0.125500897
ENST00000580632.5	NAT9	1.514724236	0.032480928	0.590318188	0.526405887
ENST00000372638.3	CITED4	1.485597739	3.93E-13	0.80114319	0.000280963
ENST00000373896.7	PHF19	1.460932605	0.007825534	0.701976479	0.299568723
ENST00000331454.3	TAB1	1.460680974	0.017147517	0.81735728	0.258877628
ENST00000371486.3	CPT2	1.441835134	0.019848332	0.324514458	0.724247432
ENST00000472354.1	RPL13	1.421928545	0.00873448	0.813242526	0.198424071
ENST00000221922.10	CCDC9	1.401845971	0.000683032	0.790685787	0.095531092
ENST00000568706.1	NUBP2	1.400703124	9.46E-05	0.60953514	0.165014193
ENST00000592741.5	PRKCSH	1.395794994	0.020038751	0.68693432	0.349183192
ENST00000392332.7	HIBCH	1.39335785	0.034716985	0.320473104	0.740631431
ENST00000409413.3	PRR35	1.385627122	0.024823076	0.189626196	0.846466149
ENST00000310473.7	UCP2	1.373359628	6.46E-09	0.593100275	0.030165633
ENST00000467712.1	PUSL1	1.369165679	0.014238015	0.726720619	0.277986794
ENST00000600467.6	RPS19	1.366838022	9.57E-05	0.52921182	0.230964973
ENST00000585892.5	DNM2	1.348194517	0.045537564	0.485597035	0.589273963
ENST00000215570.7	TIMM13	1.320877604	3.35E-21	0.678799502	3.80E-06
ENST00000448192.1	DDX56	1.285098531	0.004601561	0.629843605	0.25150052
ENST00000405717.7	SEC14L2	1.282352765	0.011222497	0.489748297	0.460037002
ENST00000491994.1	SKIV2L	1.280767585	0.003461949	0.644734476	0.220706007
ENST00000552360.3	INAFM1	1.269624183	3.29E-09	0.684208581	0.003613484
ENST00000376957.6	SRM	1.242995772	4.14E-18	0.650770869	1.60E-05
ENST00000541682.6	HES7	1.239418117	0.048696621	0.427425448	0.612923816
ENST00000525158.1	TIMM10	1.192551244	0.040085289	-0.117156961	0.90436916
ENST00000263100.7	A1BG	1.187321418	1.54E-10	0.372682739	0.100493245
ENST00000371542.3	LRRC26	1.169892147	0.011718692	0.235841778	0.73447814
ENST00000296238.3	CAMK2N2	1.164778545	0.000718336	0.357270874	0.444676209
ENST00000454650.1	NUDT1	1.147750967	0.000452377	0.490051707	0.225188615
ENST00000304952.10	HES4	1.138572536	0.003155773	0.44756135	0.36589347
ENST00000398393.5	MPP3	1.1190474	0.035070104	0.46815642	0.490259702
ENST00000409138.6	PSRC1	1.117478635	0.042897363	0.195273568	0.815764789

ENST00000371443.5	MRPL41	1.104456009	4.71E-12	0.240430107	0.251655988
ENST00000447866.5	FDPS	1.103837774	0.04882131	0.429987963	0.556095769
ENST00000314067.10	IFT43	1.102901357	0.012488711	0.466707089	0.4041064
ENST00000559651.1	ANKRD9	1.099747929	0.005016794	0.382847789	0.46336638
ENST00000355951.2	TCAF1	1.098301807	0.047239427	0.388525973	0.59780238
ENST00000471387.6	YIF1A	1.096673783	4.35E-05	0.417780943	0.211906238
ENST00000262947.7	MYDGF	1.095155363	1.89E-14	0.390531944	0.016604453
ENST00000587541.5	MADCAM1	1.089556539	0.001180568	0.43146312	0.312766044
ENST00000590253.2	G6PC3	1.088941656	0.000725864	0.334634594	0.44474295
ENST00000317479.7	PXMP2	1.077178306	0.002217426	0.391183519	0.394512715
ENST00000218230.5	PCSK1N	1.062833356	1.07E-10	0.439475533	0.018408435
ENST00000357429.10	C7orf50	5.587305863	0.010445894	5.387856632	0.014317647

Table 7.2 : Details of the proteins from the pulldown experiment that are more than 1.5 fold (linear scale) enriched in ES30L over input with q< 0.05.

Gene Name	Uniprot Accession	ES30L/Input		RandomFragment/Input		Reported role in
		log2FC	q value	log2FC	q value	
DDX21	Q9NR30	20.29843069	0.033599478	-4.89541737	NA	Translation regulation (McRae et al. 2020); ITAF (Abdullah et al. 2021)
GAPDH	E7EUT5	19.32262356	0.099961369	-3.32248553	0.30912911	
GPATCH4	A0A0A0MRK1	18.38261516	0.021514319	-5.538559015	NA	
KRT14	P02533	19.48159843	0.021514319	-5.538559015	0.30912911	
KRT2	P35908	21.58319986	0.028290899	-5.143518168	0.30912911	
MYL6	G3V1V0	20.38718118	0.043827631	-4.51201549	0.30912911	
NKRF	O15226-2	19.02090265	0.028290899	-5.143518168	0.30912911	
RBM14	Q96PK6	19.17121248	0.048376047	-4.369563304	0.30912911	
RBM25	P49756	18.69925997	0.044507575	-4.489805292	NA	
SART1	O43290	19.17513749	0.021514319	-5.538559015	NA	
TOP1	P11387	18.98896283	0.045442881	-4.459801888	NA	
DHX30	Q7L2E3-2	9.813793996	0.048376047	-4.369563304	0.30912911	Global translation (Lessel et al. 2017; Bosco et al. 2021)
YTHDC2	Q9H6S0	8.20877456	0.048376047	-4.369563304	0.30912911	ITAF (Kim and Siddiqui 2021)
SRPK1	H3BLV9	6.380402658	0.033599478	-4.89541737	0.30912911	
SRP14	H0YLA2	5.171388934	0.021514319	-5.538559015	0.30912911	ITAF (Khoury et al. 2021)
HIST1H1C	P16403	4.846392248	0.028290899	-5.143518168	0.30912911	
H1FX	Q92522	4.838675866	0.042240424	-4.565231874	0.30912911	
SRRM1	Q8IYB3-2	4.433329087	0.033599478	-4.89541737	0.30912911	
BCAS2	O75934	4.402857972	0.021514319	-5.538559015	0.30912911	

SRSF6	Q13247-3	4.177299409	0.033599478	-4.89541737	0.31618654	ITAF (Sanford et al. 2004)
RPS15	A0A0B4J2B4	3.501835359	0.048376047	-4.369563304	0.31618654	
SRSF2	J3KP15	3.496875374	0.02155607	-5.535762013	0.31618654	
SNRNP70	P08621-2	3.375448653	0.021514319	-5.538559015	0.31618654	
RPL37A	C9J4Z3	3.331669204	0.048376047	-4.369563304	0.31618654	
SRSF1	J3KTL2	3.229523267	0.021514319	-5.538559015	0.42796134	ITAF (Sanford et al. 2004)
NPM1	P06748-2	3.117680566	0.003522241	-8.14929066	0.30912911	
U2AF1L5	P0DN76	3.112994452	0.040097501	-4.640343864	0.31618654	
RPL28	H0YKD8	3.004956822	0.033599478	-4.89541737	0.05990003	
RPL13A	M0QYS1	2.699294768	0.02155607	-5.535762013	0.31618654	
DHX9	Q08211	2.695654641	0.048376047	-4.369563304	0.30912911	Translation Regulation (Murat et al. 2018)
RPL27A	E9PJD9	2.693739958	0.038126079	-4.713078022	0.30912911	
RPLP2	P05387	2.682995838	0.028290899	-5.143518168	0.38099831	
RPS21	Q8WVC2	2.626570086	0.045073073	-4.471590376	0.30912911	
U2AF2	P26368-2	2.581029545	0.021514319	-5.538559015	0.02751552	
RPL34	P49207	2.524322333	0.033599478	-4.89541737	0.01817957	
RPS14	A0A2R8Y811	2.492496933	0.021514319	-5.538559015	0.06148136	
RPL6	Q02878	2.450596499	0.048376047	-4.369563304	0.4455471	
RPS24	A0A087WUS0	2.443609254	0.043075449	-4.536990355	0.02751552	
RPL14	E7EPB3	2.386764761	0.033599478	-4.89541737	0.31618654	
RPL35A	P18077	2.367871634	0.040097501	-4.640343864	0.02864022	
RPS18	P62269	2.366676347	0.028290899	-5.143518168	0.33070065	
RPL12	P30050	2.324811212	0.021514319	-5.538559015	0.0445189	
RPS28	P62857	2.30483653	0.028290899	-5.143518168	0.34340273	
RPS4X	P62701	2.293318183	0.028290899	-5.143518168	0.30912911	

NCL	P19338	2.286787507	0.021514319	-5.538559015	0.30912911	ITAF (Hung et al. 2014; Han et al. 2021)
RPS12	P25398	2.238280791	0.045073073	-4.471590376	0.31618654	
IGF2BP1	Q9NZI8	2.1366178	0.021514319	-5.538559015	0.30912911	ITAF (Weinlich et al. 2009)
RPL32	D3YTB1	2.091964678	0.021514319	-5.538559015	0.02751552	
RPL30	E5RI99	2.086190899	0.021514319	-5.538559015	0.30912911	
RPS5	M0R0F0	1.991262774	0.045073073	-4.471590376	0.30912911	
RPL19	J3KTE4	1.964188304	0.048607974	-4.362663187	0.31618654	
PRPF19	Q9UMS4	1.92431007	0.038126079	-4.713078022	0.30912911	
RPL24	C9JNW5	1.696521759	0.028290899	-5.143518168	0.30912911	
BANF1	O75531	1.586396911	0.045980279	-4.44284097	0.13656574	
RPSA	A0A0C4DG17	1.542108733	0.046687043	-4.420833974	0.38606014	
HNRNPF	P52597	1.400082744	0.031816729	-4.974070666	0.44076763	Translation regulation (Herviou et al. 2020)
PCBP1	Q15365	1.070677288	0.021514319	-5.538559015	0.02751552	ITAF (Pickering et al. 2003; Evans et al. 2003)
HNRNPK	P61978-2	0.8099289967	0.033599478	-4.89541737	0.30912911	ITAF (Evans et al. 2003; Liu et al. 2020)

References

1. Claude, A. CONCENTRATION AND PURIFICATION OF CHICKEN TUMOR I AGENT. *Science* **87**, 467–468 (1938).
2. Claude, A. PARTICULATE COMPONENTS OF NORMAL AND TUMOR CELLS. *Science* **91**, 77–78 (1940).
3. Claude, A. THE CONSTITUTION OF MITOCHONDRIA AND MICROSOMES, AND THE DISTRIBUTION OF NUCLEIC ACID IN THE CYTOPLASM OF A LEUKEMIC CELL. *J. Exp. Med.* **80**, 19–29 (1944).
4. Keller, E. B., Zamecnik, P. C. & Loftfield, R. B. THE ROLE OF MICROSOMES IN THE INCORPORATION OF AMINO ACIDS INTO PROTEINS. *Journal of Histochemistry & Cytochemistry* vol. 2 378–386 Preprint at https://doi.org/10.1177/2.5.378 (1954).
5. Palade, G. E. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.* **1**, 59–68 (1955).
6. Dintzis, H. M. The wandering pathway to determining N to C synthesis of proteins: Some recollections concerning protein structure and biosynthesis. *Biochem. Mol. Biol. Educ.* **34**, 241–246 (2006).
7. Crick, F. H. The genetic code. *Sci. Am.* **207**, 66–74 (1962).
8. Crick, F. H. C. The origin of the genetic code. *Journal of Molecular Biology* vol. 38 367–379 Preprint at https://doi.org/10.1016/0022-2836(68)90392-6 (1968).
9. Noller, H. F. & Chaires, J. B. Functional modification of 16S ribosomal RNA by kethoxal. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 3115–3118 (1972).
10. Shine, J. & Dalgarno, L. The 3′-Terminal Sequence of *Escherichia coli* 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Proceedings of the National Academy of Sciences* vol. 71 1342–1346 Preprint at https://doi.org/10.1073/pnas.71.4.1342 (1974).
11. Steitz, J. A. & Jakes, K. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 4734–4738 (1975).
12. Herr, W., Chapman, N. M. & Noller, H. F. Mechanism of ribosomal subunit association: discrimination of specific sites in 16 S RNA essential for association activity. *J. Mol. Biol.* **130**, 433–449 (1979).
13. Ofengand, J., Liou, R., Kohut, J., Schwartz, I. & Zimmermann, R. A. Covalent

crosslinking of transfer ribonucleic acid to the ribosomal P site. Mechanism and site of reaction in transfer ribonucleic acid. *Biochemistry* vol. 18 4322–4332 Preprint at https://doi.org/10.1021/bi00587a010 (1979).

14. Kearsey, S. E. & Craig, I. W. Altered ribosomal RNA genes in mitochondria from mammalian cells with chloramphenicol resistance. *Nature* **290**, 607–608 (1981).
15. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).
16. Zaug, A. J., Grabowski, P. J. & Cech, T. R. Autocatalytic cyclization of an excised intervening sequence RNA is a cleavage-ligation reaction. *Nature* **301**, 578–583 (1983).
17. Aitken, C. E. & Lorsch, J. R. A mechanistic overview of translation initiation in eukaryotes. *Nat. Struct. Mol. Biol.* **19**, 568–576 (2012).
18. Schluenzen, F. *et al.* Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* **102**, 615–623 (2000).
19. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* vol. 147 789–802 Preprint at https://doi.org/10.1016/j.cell.2011.10.002 (2011).
20. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
21. Xue, S. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat. Rev. Mol. Cell Biol.* **13**, 355–369 (2012).
22. Genuth, N. R. & Barna, M. The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. *Mol. Cell* **71**, 364–374 (2018).
23. Ramagopal, S. Induction of cell-specific ribosomal proteins in aggregation-competent nonmorphogenetic Dictyostelium discoideum. *Biochem. Cell Biol.* **68**, 1281–1287 (1990).
24. Williams, M. E. & Sussex, I. M. Developmental regulation of ribosomal protein L16 genes in Arabidopsis thaliana. *The Plant Journal* vol. 8 65–76 Preprint at https://doi.org/10.1046/j.1365-313x.1995.08010065.x (1995).
25. Whittle, C. A. & Krochko, J. E. Transcript profiling provides evidence of functional divergence and expression networks among ribosomal protein gene paralogs in Brassica napus. *Plant Cell* **21**, 2203–2219 (2009).
26. Gunderson, J. H. *et al.* Structurally Distinct, Stage-Specific Ribosomes Occur in *Plasmodium*. *Science* vol. 238 933–937 Preprint at https://doi.org/10.1126/science.3672135 (1987).
27. Kongsuwan, K. *et al.* A Drosophila Minute gene encodes a ribosomal protein. *Nature*

vol. 317 555–558 Preprint at https://doi.org/10.1038/317555a0 (1985).

28. Willig, T. N. *et al.* Mutations in ribosomal protein S19 gene and diamond blackfan anemia: wide variations in phenotypic expression. *Blood* **94**, 4294–4306 (1999).
29. Mauro, V. P. & Edelman, G. M. The ribosome filter hypothesis. *Proceedings of the National Academy of Sciences* vol. 99 12031–12036 Preprint at https://doi.org/10.1073/pnas.192442499 (2002).
30. Shi, Z. *et al.* Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. *Molecular Cell* vol. 67 71–83.e7 Preprint at https://doi.org/10.1016/j.molcel.2017.05.021 (2017).
31. Danilova, N. & Gazda, H. T. Ribosomopathies: how a common root can cause a tree of pathologies. *Dis. Model. Mech.* **8**, 1013–1026 (2015).
32. Mills, E. W. & Green, R. Ribosomopathies: There's strength in numbers. *Science* **358**, (2017).
33. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nature Structural & Molecular Biology* vol. 19 560–567 Preprint at https://doi.org/10.1038/nsmb.2313 (2012).
34. Veldman, G. M. *et al.* The primary and secondary structure of yeast 26S rRNA. *Nucleic Acids Res.* **9**, 6935–6952 (1981).
35. Ware, V. C. *et al.* Sequence analysis of 28S ribosomal DNA from the amphibian Xenopus laevis. *Nucleic Acids Res.* **11**, 7795–7817 (1983).
36. Clark, C. G., Tague, B. W., Ware, V. C. & Gerbi, S. A. Xenopus laevis 28S ribosomal RNA: a secondary structure model and its evolutionary and functional implications. *Nucleic Acids Res.* **12**, 6197–6220 (1984).
37. Hassouna, N., Michot, B. & Bachellerie, J. P. The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Res.* **12**, 3563–3583 (1984).
38. Gorski, J. L., Gonzalez, I. L. & Schmickel, R. D. The secondary structure of human 28S rRNA: the structure and evolution of a mosaic rRNA gene. *J. Mol. Evol.* **24**, 236–251 (1987).
39. Michot, B. & Bachellerie, J.-P. Comparisons of large subunit rRNAs reveal some eukaryote-specific elements of secondary structure. *Biochimie* vol. 69 11–23 Preprint at https://doi.org/10.1016/0300-9084(87)90267-7 (1987).
40. Sweeney, R., Chen, L. & Yao, M. C. An rRNA variable region has an evolutionarily conserved essential role despite sequence divergence. *Mol. Cell. Biol.* **14**, 4203–4215 (1994).

41. Musters, W. *et al.* The conserved GTPase center and variable region V9 from Saccharomyces cerevisiae 26S rRNA can be replaced by their equivalents from other prokaryotes or eukaryotes without detectable loss of ribosomal function. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 1469–1473 (1991).
42. Wellauer, P. K., Dawid, I. B., Kelley, D. E. & Perry, R. P. Secondary structure maps of ribosomal RNA. *Journal of Molecular Biology* vol. 89 397–407 Preprint at https://doi.org/10.1016/0022-2836(74)90527-0 (1974).
43. Schibler, U., Wyler, T. & Hagenbüchle, O. Changes in size and secondary structure of the ribosomal transcription unit during vertebrate evolution. *J. Mol. Biol.* **94**, 503–517 (1975).
44. Brimacombe, R. Secondary structure and evolution of ribosomal RNA. *Nature* vol. 294 209–210 Preprint at https://doi.org/10.1038/294209a0 (1981).
45. Gerbi, S. A. The evolution of eukaryotic ribosomal DNA. *Biosystems.* **19**, 247–258 (1986).
46. Tomal, A., Kwasniak-Owczarek, M. & Janska, H. An Update on Mitochondrial Ribosome Biology: The Plant Mitoribosome in the Spotlight. *Cells* **8**, (2019).
47. Mankin, A. S. & Kopylov, A. M. A SECONDARY STRUCTURE MODEL FOR MITOCHONDRIAL 12S RIBOSOMAL-RNA-AN EXAMPLE OF ECONOMY IN RIBOSOMAL-RNA STRUCTURE. *Biochem. Int.* **3**, 587–593 (1981).
48. Branlant, C. *et al.* Primary and secondary structures of Escherichia coli MRE 600 23S ribosomal RNA. Comparison with models of secondary structure for maize chloroplast 23S rRNA and for large portions of mouse and human 16S mitochondrial rRNAs. *Nucleic Acids Res.* **9**, 4303–4324 (1981).
49. Stepanov, V. G. & Fox, G. E. Expansion segments in bacterial and archaeal 5S ribosomal RNAs. *RNA* **27**, 133–150 (2021).
50. Kushwaha, A. K. & Bhushan, S. Unique structural features of the Mycobacterium ribosome. *Prog. Biophys. Mol. Biol.* **152**, 15–24 (2020).
51. Penev, P. I. *et al.* Supersized Ribosomal RNA Expansion Segments in Asgard Archaea. *Genome Biol. Evol.* **12**, 1694–1710 (2020).
52. Tirumalai, M. R., Kaelber, J. T., Park, D. R., Tran, Q. & Fox, G. E. Cryo-electron microscopy visualization of a large insertion in the 5S ribosomal RNA of the extremely halophilic archaeon Halococcus morrhuae. *FEBS Open Bio* **10**, 1938–1946 (2020).
53. Armache, J.-P. *et al.* Cryo-EM structure and rRNA model of a translating eukaryotic 80S ribosome at 5.5-A resolution. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 19748–19753 (2010).
54. Bowman, J. C., Petrov, A. S., Frenkel-Pinter, M., Penev, P. I. & Williams, L. D. Root of

the Tree: The Significance, Evolution, and Origins of the Ribosome. *Chem. Rev.* **120**, 4848–4878 (2020).

55. Yokoyama, T. & Suzuki, T. Ribosomal RNAs are tolerant toward genetic insertions: evolutionary origin of the expansion segments. *Nucleic Acids Res.* **36**, 3539–3551 (2008).
56. Ramesh, M. & Woolford, J. L., Jr. Eukaryote-specific rRNA expansion segments function in ribosome biogenesis. *RNA* **22**, 1153–1162 (2016).
57. Petrov, A. S. *et al.* Evolution of the ribosome at atomic resolution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10251–10256 (2014).
58. Caetano-Anollés, G. Ancestral Insertions and Expansions of rRNA do not Support an Origin of the Ribosome in Its Peptidyl Transferase Center. *Journal of molecular evolution* vol. 80 162–165 (2015).
59. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80–85 (2013).
60. Petrov, A. S. *et al.* History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15396–15401 (2015).
61. Stoltzfus, A. On the Possibility of Constructive Neutral Evolution. *Journal of Molecular Evolution* vol. 49 169–181 Preprint at https://doi.org/10.1007/pl00006540 (1999).
62. Muñoz-Gómez, S. A., Bilolikar, G., Wideman, J. G. & Geiler-Samerotte, K. Constructive Neutral Evolution 20 Years Later. *J. Mol. Evol.* **89**, 172–182 (2021).
63. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15690–15695 (2015).
64. Hancock, J. M. & Dover, G. A. 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Res.* **18**, 5949–5954 (1990).
65. Lukeš, J., Archibald, J. M., Keeling, P. J., Doolittle, W. F. & Gray, M. W. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63**, 528–537 (2011).
66. Parker, M. S., Sallee, F. R., Park, E. A. & Parker, S. L. Homoiterons and expansion in ribosomal RNAs. *FEBS Open Bio* **5**, 864–876 (2015).
67. Parks, M. M. *et al.* Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci Adv* **4**, eaao0665 (2018).
68. Locati, M. D. *et al.* Expression of distinct maternal and somatic 5.8S, 18S, and 28S rRNA types during zebrafish development. *RNA* vol. 23 1188–1199 Preprint at https://doi.org/10.1261/rna.061515.117 (2017).
69. Chandramouli, P. *et al.* Structure of the mammalian 80S ribosome at 8.7 A resolution. *Structure* **16**, 535–548 (2008).

70. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–1529 (2011).
71. Hancock, J. M., Tautz, D. & Dover, G. A. Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of Drosophila melanogaster. *Mol. Biol. Evol.* **5**, 393–414 (1988).
72. Leppek, K. *et al.* Gene- and Species-Specific Hox mRNA Translation by Ribosome Expansion Segments. *Mol. Cell* **80**, 980–995.e13 (2020).
73. Alkemar, G. & Nygård, O. A possible tertiary rRNA interaction between expansion segments ES3 and ES6 in eukaryotic 40S ribosomal subunits. *RNA* **9**, 20–24 (2003).
74. Alkemar, G. & Nygård, O. Secondary structure of two regions in expansion segments ES3 and ES6 with the potential of forming a tertiary interaction in eukaryotic 40S ribosomal subunits. *RNA* **10**, 403–411 (2004).
75. Gao, H., Ayub, M. J., Levin, M. J. & Frank, J. The structure of the 80S ribosome from Trypanosoma cruzi reveals unique rRNA components. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10206–10211 (2005).
76. Mestre-Fos, S. *et al.* Profusion of G-quadruplexes on both subunits of metazoan ribosomes. *PLoS One* **14**, e0226177 (2019).
77. Mestre-Fos, S. *et al.* G-Quadruplexes in Human Ribosomal RNA. *J. Mol. Biol.* **431**, 1940–1955 (2019).
78. Mestre-Fos, S., Ito, C., Moore, C. M., Reddi, A. R. & Williams, L. D. Human Ribosomal G-Quadruplexes Regulate Heme Bioavailability. Preprint at https://doi.org/10.1101/2020.04.15.042721.
79. Musters, W., Boon, K., van der Sande, C. A., van Heerikhuizen, H. & Planta, R. J. Functional analysis of transcribed spacers of yeast ribosomal DNA. *EMBO J.* **9**, 3989–3996 (1990).
80. Jeeninga, R. E. *et al.* Variable regions V13 and V3 of Saccharomyces cerevisiae contain structural features essential for normal biogenesis and stability of 5.8S and 25S rRNA. *RNA* **3**, 476–488 (1997).
81. Hariharan, N., Ghosh, S. & Palakodeti, D. The story of rRNA expansion segments: Finding functionality amidst diversity. *Wiley Interdiscip. Rev. RNA* e1732 (2022).
82. Bradatsch, B. *et al.* Structure of the pre-60S ribosomal subunit with nuclear export factor Arx1 bound at the exit tunnel. *Nat. Struct. Mol. Biol.* **19**, 1234–1241 (2012).
83. Greber, B. J., Boehringer, D., Montellese, C. & Ban, N. Cryo-EM structures of Arx1 and maturation factors Rei1 and Jjj1 bound to the 60S ribosomal subunit. *Nat. Struct. Mol. Biol.* **19**, 1228–1233 (2012).

84. Wang, X. *et al.* Coevolution of ribosomal RNA expansion segment 7L and assembly factor Noc2p specializes the ribosome biogenesis pathway between Saccharomyces cerevisiae and Candida albicans. *Nucleic Acids Res.* **49**, 4655–4667 (2021).
85. Ohmayer, U. *et al.* Studies on the Coordination of Ribosomal Protein Assembly Events Involved in Processing and Stabilization of Yeast Early Large Ribosomal Subunit Precursors. *PLoS One* **10**, e0143768 (2015).
86. Fayet-Lebaron, E., Atzorn, V., Henry, Y. & Kiss, T. 18S rRNA processing requires base pairings of snR30 H/ACA snoRNA to eukaryote-specific 18S sequences. *EMBO J.* **28**, 1260–1270 (2009).
87. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80–85 (2013).
88. Alkemar, G. & Nygård, O. Secondary structure of two regions in expansion segments ES3 and ES6 with the potential of forming a tertiary interaction in eukaryotic 40S ribosomal subunits. *RNA* **10**, 403–411 (2004).
89. Díaz-López, I., Toribio, R., Berlanga, J. J. & Ventoso, I. An mRNA-binding channel in the ES6S region of the translation 48S-PIC promotes RNA unwinding and scanning. *Elife* **8**, (2019).
90. Fujii, K., Susanto, T. T., Saurabh, S. & Barna, M. Decoding the Function of Expansion Segments in Ribosomes. *Mol. Cell* **72**, 1013–1020.e6 (2018).
91. Knorr, A. G. *et al.* Ribosome–NatA architecture reveals that rRNA expansion segments coordinate N-terminal acetylation. *Nature Structural & Molecular Biology* vol. 26 35–39 Preprint at https://doi.org/10.1038/s41594-018-0165-y (2019).
92. Kossinova, O., Malygin, A., Krol, A. & Karpova, G. The SBP2 protein central to selenoprotein synthesis contacts the human ribosome at expansion segment 7L of the 28S rRNA. *RNA* **20**, 1046–1056 (2014).
93. Gómez Ramos, L. M. *et al.* Yeast rRNA Expansion Segments: Folding and Function. *J. Mol. Biol.* **428**, 4048–4059 (2016).
94. Ghosh, A., Williams, L. D., Pestov, D. G. & Shcherbik, N. Proteotoxic stress promotes entrapment of ribosomes and misfolded proteins in a shared cytosolic compartment. *Nucleic Acids Res.* **48**, 3888–3905 (2020).
95. Parker, M. S., Balasubramaniam, A., Sallee, F. R. & Parker, S. L. The Expansion Segments of 28S Ribosomal RNA Extensively Match Human Messenger RNAs. *Front. Genet.* **9**, 66 (2018).
96. Pánek, J., Kolár, M., Vohradský, J. & Shivaya Valásek, L. An evolutionary conserved pattern of 18S rRNA sequence complementarity to mRNA 5' UTRs and its implications

for eukaryotic gene translation regulation. *Nucleic Acids Res.* **41**, 7625–7634 (2013).

97. Chappell, S. A. & Mauro, V. P. The Internal Ribosome Entry Site (IRES) Contained within the RNA-binding Motif Protein 3 (Rbm3) mRNA Is Composed of Functionally Distinct Elements. *Journal of Biological Chemistry* vol. 278 33793–33800 Preprint at https://doi.org/10.1074/jbc.m303495200 (2003).
98. Guan, L. & Grigoriev, A. Computational meta-analysis of ribosomal RNA fragments: potential targets and interaction mechanisms. *Nucleic Acids Res.* **49**, 4085–4103 (2021).
99. Blau, M. *et al.* ERj1p uses a universal ribosomal adaptor site to coordinate the 80S ribosome at the membrane. *Nat. Struct. Mol. Biol.* **12**, 1015–1016 (2005).
100. Beckmann, R. *et al.* Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372 (2001).
101. Halic, M. *et al.* Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* **427**, 808–814 (2004).
102. Quast, C. *et al.* SILVA Databases. *Encyclopedia of Metagenomics* 626–635 Preprint at https://doi.org/10.1007/978-1-4899-7478-5_250 (2015).
103. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
104. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
105. Parker, M. S., Sallee, F. R., Park, E. A. & Parker, S. L. Homoiterons and expansion in ribosomal RNAs. *FEBS Open Bio* **5**, 864–876 (2015).
106. Bernier, C. R., Petrov, A. S., Kovacs, N. A., Penev, P. I. & Williams, L. D. Translation: The Universal Structural Core of Life. *Mol. Biol. Evol.* **35**, 2065–2076 (2018).
107. Khatter, H., Myasnikov, A. G., Natchiar, S. K. & Klaholz, B. P. Structure of the human 80S ribosome. *Nature* **520**, 640–645 (2015).
108. Bernier, C. R. *et al.* RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss.* **169**, 195–207 (2014).
109. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* vol. 47 D766–D773 Preprint at https://doi.org/10.1093/nar/gky955 (2019).
110. Parker, M. S., Sallee, F. R., Park, E. A. & Parker, S. L. Homoiterons and expansion in ribosomal RNAs. *FEBS Open Bio* **5**, 864–876 (2015).
111. Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396 (2008).

112. R Development Core Team. *The R Reference Manual: Base Package*. (Network Theory., 2003).
113. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
114. Parker, M. S., Balasubramaniam, A., Sallee, F. R. & Parker, S. L. The Expansion Segments of 28S Ribosomal RNA Extensively Match Human Messenger RNAs. *Front. Genet.* **9**, 66 (2018).
115. Krishna, S. *et al.* Dynamic expression of tRNA-derived small RNAs define cellular states. *EMBO Rep.* **20**, e47789 (2019).
116. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
117. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
118. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2013).
119. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
120. R Development Core Team. *The R Reference Manual: Base Package*. (Network Theory., 2003).
121. Kolde, R. pheatmap: Pretty Heatmaps. R package version 1.0. 12. *CRAN. R-project. org/package= pheatmap* Preprint at (2019).
122. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).
123. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
124. Orsburn, B. C. Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **9**, (2021).
125. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
126. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).
127. Arike, L. & Peil, L. Spectral counting label-free proteomics. *Methods Mol. Biol.* **1156**, 213–222 (2014).

128. Ternan, N. G., Jain, S., Graham, R. L. J. & McMullan, G. Semiquantitative analysis of clinical heat stress in Clostridium difficile strain 630 using a GeLC/MS workflow with emPAI quantitation. *PLoS One* **9**, e88960 (2014).
129. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
130. Simsek, D. *et al.* The Mammalian Ribo-interactome Reveals Ribosome Functional Diversity and Heterogeneity. *Cell* **169**, 1051–1065.e18 (2017).
131. Aw, J. G. A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cell* **62**, 603–617 (2016).
132. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
133. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* vol. 47 D766–D773 Preprint at https://doi.org/10.1093/nar/gky955 (2019).
134. Zhao, J. *et al.* IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. *Genomics Proteomics Bioinformatics* **18**, 129–139 (2020).
135. Yang, T.-H., Wang, C.-Y., Tsai, H.-C. & Liu, C.-T. Human IRES Atlas: an integrative platform for studying IRES-driven translational regulation in humans. *Database* **2021**, (2021).
136. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
137. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
138. Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445–452 (2006).
139. Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database* **2016**, (2016).
140. Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377–3379 (2015).
141. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
142. Porter, D. F. *et al.* easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. *Nat. Commun.* **12**, 1569 (2021).

143. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
144. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
145. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
146. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
147. Porter, D. F. *et al.* easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. *Nat. Commun.* **12**, 1569 (2021).
148. Kolde, R. pheatmap: Pretty Heatmaps. R package version 1.0. 12. *CRAN. R-project. org/package= pheatmap* Preprint at (2019).
149. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).
150. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).
151. Boehringer, D., Thermann, R., Ostareck-Lederer, A., Lewis, J. D. & Stark, H. Structure of the hepatitis C virus IRES bound to the human 80S ribosome: remodeling of the HCV IRES. *Structure* **13**, 1695–1706 (2005).
152. Pisareva, V. P., Pisarev, A. V. & Fernández, I. S. Dual tRNA mimicry in the Cricket Paralysis Virus IRES uncovers an unexpected similarity with the Hepatitis C Virus IRES. *Elife* **7**, (2018).
153. Shabalina, S. A., Ogurtsov, A. Y., Rogozin, I. B., Koonin, E. V. & Lipman, D. J. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.* **32**, 1774–1782 (2004).
154. Leppek, K., Byeon, G. W., Fujii, K. & Barna, M. VELCRO-IP RNA-seq reveals ribosome expansion segment function in translation genome-wide. *Cell Rep.* **34**, 108629 (2021).
155. Yamamoto, H. *et al.* Molecular architecture of the ribosome-bound Hepatitis C Virus internal ribosomal entry site RNA. *EMBO J.* **34**, 3042–3058 (2015).
156. Aw, J. G. A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cell* **62**, 603–617 (2016).
157. Simsek, D. *et al.* The Mammalian Ribo-interactome Reveals Ribosome Functional Diversity and Heterogeneity. *Cell* **169**, 1051–1065.e18 (2017).

158. Simsek, D. *et al.* The Mammalian Ribo-interactome Reveals Ribosome Functional Diversity and Heterogeneity. *Cell* **169**, 1051–1065.e18 (2017).
159. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
160. Porter, D. F. *et al.* easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. *Nat. Commun.* **12**, 1569 (2021).
161. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
162. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80–85 (2013).
163. Chandramouli, P. *et al.* Structure of the mammalian 80S ribosome at 8.7 A resolution. *Structure* **16**, 535–548 (2008).
164. Khatter, H., Myasnikov, A. G., Natchiar, S. K. & Klaholz, B. P. Structure of the human 80S ribosome. *Nature* **520**, 640–645 (2015).
165. Natchiar, S. K., Myasnikov, A. G., Kratzat, H., Hazemann, I. & Klaholz, B. P. Visualization of chemical modifications in the human 80S ribosome structure. *Nature* **551**, 472–477 (2017).
166. Gorski, J. L., Gonzalez, I. L. & Schmickel, R. D. The secondary structure of human 28S rRNA: the structure and evolution of a mosaic rRNA gene. *J. Mol. Evol.* **24**, 236–251 (1987).
167. Lessel, D. *et al.* De Novo Missense Mutations in DHX30 Impair Global Translation and Cause a Neurodevelopmental Disorder. *Am. J. Hum. Genet.* **101**, 716–724 (2017).
168. Bosco, B. *et al.* DHX30 Coordinates Cytoplasmic Translation and Mitochondrial Function Contributing to Cancer Cell Survival. *Cancers* **13**, (2021).
169. Murat, P. *et al.* RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. *Genome Biol.* **19**, 229 (2018).
170. McRae, E. K. S. *et al.* An RNA guanine quadruplex regulated pathway to TRAIL-sensitization by DDX21. *RNA* **26**, 44–57 (2020).
171. Herviou, P. *et al.* hnRNP H/F drive RNA G-quadruplex-mediated translation linked to genomic instability and therapy resistance in glioblastoma. *Nat. Commun.* **11**, 2661 (2020).
172. Evans, J. R. *et al.* Members of the poly (rC) binding protein family stimulate the activity of the c-myc internal ribosome entry segment in vitro and in vivo. *Oncogene* vol. 22 8012–8020 Preprint at https://doi.org/10.1038/sj.onc.1206645 (2003).
173. Notari, M. *et al.* A MAPK/HNRPK pathway controls BCR/ABL oncogenic potential by

regulating MYC mRNA translation. *Blood* **107**, 2507–2516 (2006).

174. Pickering, B. M., Mitchell, S. A., Evans, J. R. & Willis, A. E. Polypyrimidine tract binding protein and poly r(C) binding protein 1 interact with the BAG-1 IRES and stimulate its activity in vitro and in vivo. *Nucleic Acids Res.* **31**, 639–646 (2003).
175. Sanford, J. R., Gray, N. K., Beckmann, K. & Cáceres, J. F. A novel role for shuttling SR proteins in mRNA translation. *Genes Dev.* **18**, 755–768 (2004).
176. López-Ulloa, B., Fuentes, Y., Pizarro-Ortega, M. S. & López-Lastra, M. RNA-Binding Proteins as Regulators of Internal Initiation of Viral mRNA Translation. *Viruses* **14**, (2022).
177. Murat, P. *et al.* RNA G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. *Genome Biol.* **19**, 229 (2018).
178. Mestre-Fos, S., Ito, C., Moore, C. M., Reddi, A. R. & Williams, L. D. Human ribosomal G-quadruplexes regulate heme bioavailability. *J. Biol. Chem.* **295**, 14855–14865 (2020).
179. Mestre-Fos, S. *et al.* Profusion of G-quadruplexes on both subunits of metazoan ribosomes. *PLoS One* **14**, e0226177 (2019).
180. Pickering, B. M., Mitchell, S. A., Evans, J. R. & Willis, A. E. Polypyrimidine tract binding protein and poly r(C) binding protein 1 interact with the BAG-1 IRES and stimulate its activity in vitro and in vivo. *Nucleic Acids Res.* **31**, 639–646 (2003).
181. Evans, J. R. *et al.* Members of the poly (rC) binding protein family stimulate the activity of the c-myc internal ribosome entry segment in vitro and in vivo. *Oncogene* vol. 22 8012–8020 Preprint at https://doi.org/10.1038/sj.onc.1206645 (2003).
182. Sanford, J. R., Gray, N. K., Beckmann, K. & Cáceres, J. F. A novel role for shuttling SR proteins in mRNA translation. *Genes Dev.* **18**, 755–768 (2004).
183. Swanson, C. M., Sherer, N. M. & Malim, M. H. SRp40 and SRp55 promote the translation of unspliced human immunodeficiency virus type 1 RNA. *J. Virol.* **84**, 6748–6759 (2010).
184. Kim, G.-W. & Siddiqui, A. N6-methyladenosine modification of HCV RNA genome regulates cap-independent IRES-mediated translation via YTHDC2 recognition. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
185. Han, S. *et al.* Nucleolin Promotes IRES-Driven Translation of Foot-and-Mouth Disease Virus by Supporting the Assembly of Translation Initiation Complexes. *J. Virol.* **95**, e0023821 (2021).
186. Weinlich, S. *et al.* IGF2BP1 enhances HCV IRES-mediated translation initiation via the 3′UTR. *RNA* vol. 15 1528–1542 Preprint at https://doi.org/10.1261/rna.1578409 (2009).
187. Khoury, G. *et al.* The RNA-Binding Proteins SRP14 and HMGB3 Control HIV-1 Tat

mRNA Processing and Translation During HIV-1 Latency. *Front. Genet.* **12**, 680725 (2021).

188. Lin, J.-C., Hsu, M. & Tarn, W.-Y. Cell stress modulates the function of splicing regulatory protein RBM4 in translation control. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2235–2240 (2007).
189. Niu, K., Zhang, X., Song, Q. & Feng, Q. G-Quadruplex Regulation of VEGFA mRNA Translation by RBM4. *International Journal of Molecular Sciences* vol. 23 743 Preprint at https://doi.org/10.3390/ijms23020743 (2022).
190. Abdullah, S. W. *et al.* DDX21, a Host Restriction Factor of FMDV IRES-Dependent Translation and Replication. *Viruses* **13**, (2021).
191. Liu, W. *et al.* hnRNP K Is a Novel Internal Ribosomal Entry Site-Transacting Factor That Negatively Regulates Foot-and-Mouth Disease Virus Translation and Replication and Is Antagonized by Viral 3C Protease. *J. Virol.* **94**, (2020).
192. Kramer, K. *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat. Methods* **11**, 1064–1070 (2014).