Dissecting the function and molecular evolution of translated long non-coding RNAs

Isabel Jacqueline Birds

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Faculty of Biological Sciences

School of Molecular and Cellular Biology

August 28, 2023

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapters 2 and 3 of the thesis has appeared in publication as follows:

Cytoplasmic long non-coding RNAs are differentially regulated and translated during human neuronal differentiation. Katerina Douka, Isabel Birds, Dapeng Wang, Andreas Kosteletos, Sophie Clayton, Abigail Byford, Elton J. R. Vasconcelos, Mary J. O'Connell, Jim Deuchars, Adrian Whitehouse, and Julie L. Aspden. RNA, June 2021.

I was responsible for designing and performing experiments, acquiring, analysing and interpreting data, and drafting and revising the manuscript. Specifically, I quality checked, processed, and aligned Poly-Ribo-Seq data, and identified actively translated ORFs. I analysed the translational profile and efficiency of these ORFs, and calculated their amino acid compositions. I also performed sequence homology based conservational analysis.

Specifically in this thesis, Katerina Douka carried out Poly-Ribo-Seq on SH-SY5Y cells, and provided valuable expertise on the human central nervous system. Andreas Kosteletos analysed the transcription of translated lncRNAs in human tissues, and in neuronal diseases and disorders. Elton J. R. Vasconcelos analysed SH-SY5Y mass spectrometry datasets.

Acknowledgements

I have many people to thank for their support both leading up to and throughout this PhD, and I am deeply grateful for all of them.

To my parents and brother, thank you for your unwavering support, for encouraging me to pursue all of my interests, and for always believing I could do anything I put my mind to. I could not have achieved this without you.

To my partner Rob, thank you for being my safe place, and for always being on my team. Now we can finally start planning that wedding! Also thank you to our little dog Sadie, for being the best writing companion (and snack sharer) I could wish for.

To my supervisor Dr Julie Aspden, thank you for your encouragement, guidance, and invaluable mentorship. You have nurtured a wonderfully collaborative lab group, and I feel very lucky to have been part of it.

To my co-supervisor Prof. Mary O'Connell, thank you for taking me on in my undergraduate degree and introducing me to computational biology. Your expertise and guidance have had a lasting impact on my life.

To my co-supervisor Prof. David Westhead, thank you for your valuable input and guidance, and for thinking of me when exciting opportunities came your way.

To all of the members of the Aspden and O'Connell labs, and the wider Leeds community, thank you for being great colleagues, and more importantly friends. Special thanks to my fantastic collaborator Dr Katerina Douka, both for the Poly-Ribo-Seq data this thesis is based on, and all the time spent together puzzling over our work.

To my fellow co-operators at Open Data Services, thank you for being the most understanding and supportive colleagues I could ask for as I wrote up this thesis.

To my rugby families at The University of Leeds, Swinton Lionesses, and Didsbury Toc H, thank you for being my teammates both on and off the pitch. Rugby became an important part of my life at the start of this PhD, and you have supported me more than you know.

This work was undertaken on ARC3 and ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. Thank you to the HPC team, in particular Martin Callaghan, for their expertise and support.

Thank you to The Sir Richard Stapley Educational Trust and The University of Leeds Graduate School Scholarships Fund for supporting me financially throughout this PhD.

Abstract

Long non-coding RNAs (lncRNAs) are non-coding transcripts of at least 200 nucleotides, which are typically lowly expressed, often in a tissue and developmental-stage specific manner. They generally lack sequence conservation, and 40% of human lncRNAs are expressed in the brain. Despite forming $\sim$23% of the human transcriptome, only a small proportion of lncRNAs have been well characterised. In particular the function and potential translation of cytoplasmic lncRNAs remains poorly understood. This thesis aims to identify and characterise actively translated human neuronal lncRNAs using Poly-Ribo-Seq data, and to investigate conservation of their peptide products in non-human species.

242 actively translated lncRNA smORFs were detected in undifferentiated and differentiated human neuroblastoma cells (SH-SY5Y). These lncRNA smORFs exhibit triplet periodicity and translational efficiencies comparable to protein coding ORFs. The expression of these translated lncRNAs is significantly enriched in human brain tissue, throughout development time points from pre to postnatal, compared to untranslated cytoplasmic lncRNAs. The resulting peptides exhibit amino acid compositions similar to that of canonical proteins, but tend not to contain known domains.

27% of the translated lncRNA smORFs exhibited sequence conservation in non-human species, suggesting they are under selective constraint. On average, conserved, translated lncRNA smORF peptides were longer, and more likely to be validated by Ribo-Seq and mass spectrometry data from other human tissues and cell lines. One translated lncRNA smORF found to be conserved was in the antisense lncRNA LIPT2-AS1, which results in the synthesis of a 272 aa peptide. The LIPT2-AS1 lncRNA is highly expressed in neural cells and down regulated in large-cell medulloblastoma. LIPT2-AS1-smORF exhibited sequential and syntenous conservation across euarchontoglires, and the resulting peptide contains a DNA binding domain, suggesting a role in the regulation of gene expression.

Together, these results demonstrate that cytoplasmic lncRNAs are actively translated, biologically important in human neuronal differentiation, and are likely to have conserved functions.

Contents

List of Tables

List of Figures

Acronyms

A-site acceptor site.

aa amino acids.

ABCE1 ATP-binding cassette.

ASAP ATP synthase-associated peptide.

BED Browser Extensible Data.

CCDS consensus coding DNA sequence.

CDS coding DNA sequence.

ceRNA competing endogenous RNA.

circRNA circular RNA.

CLIP-seq cross-linking immunoprecipitation.

dORF downstream open reading frames.

DTEO differential translation efficiency ORFs.

DTO differentially transcribed ORFs.

E-site exit site.

ENE expression and nuclear retention element.

eRF1 eukaryotic peptide chain release factor subunit 1.

eRNA enhancer RNA.

FDR false discovery rate.

FPKM fragments per kilobase of exon per million reads mapped.

GOC Gene Order Conservation.

HGT horizontal gene transfer.

HMM Hidden Markov Model.

HNRNA heterogeneous nuclear RNA.

HoT Heads-or-Tails.

HtH helix-turn-helix.

IRES internal ribosome entry site.

KRAB Kruppel-associated box.

KSHV Kaposi's sarcoma-associated herpesvirus.

lincRNA long intergenic non-coding RNA.

lncRNA long non-coding RNA.

mascRNA MALAT1-associated small cytoplasmic RNA.

miRNA microRNA.

MLN myoregulin.

mRNA messenger RNA.

MSAs Multiple sequence alignments.

mTORC1 mammalian target of rapamycin complex 1.

MUSCLE MUltiple Sequence Comparison by Log-Expectation.

ncORF non-coding open reading frame.

ncRNA non-coding RNA.

NMD nonsense-mediated decay.

NPCs Neural progenitor cells.

nt nucleotides.

ORFs open reading frames.

P-site peptidyl site.

PAR promoter-associated RNA.

PCA principle component analysis.

piRNA Piwi-interacting RNA.

PLN phospholamban.

pncr003:2L putative noncoding RNA 003 in 2L.

Poly(A) polyadenylated.

Poly-Ribo-Seq Polysome profiling.

RA retinoic acid.

RBDs RNA-binding domains.

RBPs RNA-binding proteins.

rDNA ribosomal DNA.

Ribo-Seq ribosome profiling.

RPKM Reads per kilobase of exon per million reads mapped.

rRNA ribosomal RNA.

RRS ribosome release score.

RT-PCR reverse transcription polymerase chain reaction.

shRNA short hairpin RNA.

siRNA small interfering RNA.

SLN sarcolipin.

smORFs small open reading frames.

snoRNA small nucleolar RNA.

snRNA Small nuclear RNA.

SPAR/SPAAR small regulatory polypeptide of amino acid response.

SRA1 Steroid Receptor RNA Activator 1.

TE Translational efficiency.

TERC Telomerase RNA.

TF transcription factors.

TIS translation initiation site.

TMM trimmed mean of M-values.

TPM Transcripts per million.

tRNA transfer RNA.

uORFs upstream open reading frames.

UTR untranslated region.

vtRNA vault RNA.

Chapter 1

Introduction

1.1 Non-coding RNA

Approximately 80% of the $\sim$3.1 billion nucleotides that form the human genome are transcribed, but only $\sim$2.94% of these form messenger RNA (mRNA), which is subsequently translated to produce protein (ENCODE Project Consortium, 2012). This means that the vast majority of transcripts are classified as non-coding RNA (ncRNA). Historically our understanding of ncRNA functions has lagged behind that of mRNA, given the sheer diversity and abundance of these transcripts. However, in the last $\sim$20 years since the human genome was first sequenced (Lander et al., 2001), advances in high-throughput sequencing, multi-omics techniques, and computing power have massively advanced our knowledge, creating entirely new fields of study into ncRNA, once termed the "dark matter of the genome" (Blaxter, 2010).

We now understand that ncRNAs play key roles in cellular processes at all stages of development and disease. Classification of ncRNAs is based on their function, localisation, biogenesis, size, and structure (Table 1.1). Small ncRNAs of ≤ 200 nucleotides are particularly well categorised, including transfer RNA (tRNA) which is a key component of the translational process. A large proportion of ncRNAs ($\sim$35.5%) are however annotated as long non-coding RNA (lncRNA); non-coding transcripts of at least 200 nt in length (Guttman and Rinn, 2012; Mattick and Rinn, 2015). There are currently approximately 20,000 annotated human lncRNA genes, corresponding to $\sim$58,000 transcripts (Frankish et al., 2021). The definition of lncRNAs is relatively arbitrary, making for a very heterogeneous group of transcripts, but does conveniently exclude the well defined smaller ncRNA. A small proportion of lncRNAs have been well characterised; the *Xist* transcript which mediates silencing of the X chromosome (Sahakyan, Yang, and Plath, 2018), and *HOTAIR* which epigenetically represses the *HOXD* locus (Hajjari and Salavaty, 2015) are notable examples. However, lncRNAs are generally poorly understood. This is partly due to the expression of lncRNAs, which is generally at lower levels than mRNA, and often in a tissue, cell type, and developmental stage specific manner (Jandura and Krause, 2017).

The basic features of many lncRNAs are similar to that of mRNAs, including a 5' cap and, for approximately 39% of currently annotated lncRNAs in humans, a 3' polyadenylated (Poly(A)) tail (Derrien et al., 2012). Indeed, lncRNAs were first grouped as a "collection of mRNA-like non-coding RNAs ... without defined ORFs" (Erdmann et al., 1999). A small number of alternate stabilisation mechanisms to polyadenylation have been identified in lncRNAs, including expression and nuclear retention element (ENE)-like structures as in the nuclear lncRNA *MALAT1* (Brown, Valenstein, et al., 2012). Originally discovered in the lncRNA *PAN* produced by Kaposi's sarcoma-associated herpesvirus (KSHV) (Conrad and

Table 1.1: **Summary of the classes of ncRNA.** A brief overview of definition and length of the ncRNA in each class.

Class	Definition	Length (nt)
Circular RNA (circRNA)	Single stranded circularised transcripts with a wide range of proposed regulatory functions (Hsu and Coca-Prados, 1979).	100 - 1000+
Enhancer RNA (eRNA)	ncRNA transcribed from DNA enhancer regions which regulate transcription (Santa et al., 2010).	<2000
Long non-coding RNA (lncRNA)	Non coding transcripts of >200 nt. Varied roles, only defined for a subset of lncRNAs (Kapranov et al., 2007).	>200
microRNA (miRNA)	Single stranded transcripts, form part of the RNA induced splicing complex. (Reinhart et al., 2000).	19 - 24
Piwi-interacting RNA (piRNA)	Characterised by a 5’ uridine and a 3’ 2’-O-methyl modification. Forms complexes with PIWI proteins, silencing transposable elements in germline cells (Girard et al., 2006).	24 - 31
Ribosomal RNA (rRNA)	RNA component of ribosomes, functions as a ribozyme (Hoagland et al., 1958).	121, 159, 1,869 & 5,070 in human
Small interfering RNA (siRNA)	Double stranded RNA, processed by Dicer. Induces the cleavage of complementary sequences via RNA interference.	88
Small nuclear RNA (snRNA)	Component of spliceosome (Hodnett and Busch, 1968).	∼150
Small nucleolar RNA (snoRNA)	Nucleolar RNAs processed from introns. Form complexes with proteins to guide modification of rRNAs, tRNAs, and snRNAs.	60 - 300
Telomerase RNA (TERC)	Provides template for telomere replication (Blackburn and Gall, 1978).	Varies
Transfer RNA (tRNA)	Pairs to mRNA codons to transfer amino acids during translation (Hoagland et al., 1958).	70 - 90
Vault RNA (vtRNA)	Forms part of the vault ribonucleoprotein complex (Kedersha and Rome, 1986).	88 - 98 in human

Steitz, 2005), the ENE structure contains a U-rich internal loop which forms a triple helix with the Poly(A) tail, protecting the RNA from degradation pathways. As *MALAT1* is cleaved by RNAse-P it does not have a Poly(A) tail, instead forming an ENE-like helix with an A-rich portion of the transcript. Another small subset of lncRNAs are the snoRNA-related lncRNAs which are processed by the snoRNA machinery, with a snoRNA sequence at either one or both ends of the transcript (Yin et al., 2012). The variation we observe in stability is reflected in the half-lives of mammalian lncRNAs, which can range from less than 30 minutes to over 48 hours (Clark et al., 2012), and along with control at the transcriptional level contributes to their dynamic expression profiles.

The majority of lncRNAs are transcribed by RNA polymerase II (Devaux et al., 2015) and spliced. LncRNA genes generally contain a smaller number of exons than protein coding genes, and a particularly high proportion have two exons (42%), compared to 6% of human protein coding genes (Derrien et al., 2012). Although initial studies predominantly identified nuclear lncRNAs (Derrien et al., 2012), some lncRNAs are now known to be enriched in the nucleus, some in the cytoplasm, and others shuttle between the two locations. Three main mechanisms exist which lead to lncRNAs being retained in the nucleus; i) specific nuclear retention motifs (Shukla et al., 2018), ii) retained introns due to insufficient splicing, and iii) tethering by proteins. For example the lncRNA *Xist* is anchored to the nuclear periphery by CIZ1 (Ridings-Figueroa et al., 2017). However, the majority of 5' capped, polyadenylated lncRNAs are exported to the cytoplasm.

Although termed non-coding, lncRNAs often contain open reading frames (ORFs). All that is required to form a potential ORF is an in-frame start and stop codon, meaning they can be found randomly throughout the genome (Ladoukakis et al., 2011). The ORFs found in lncRNAs are small open reading frames (smORFs) of $\sim$100 codons or fewer, meaning their coding potential is generally dismissed. Despite this, some lncRNAs associate with the translational machinery (van Heesch, van Iterson, et al., 2014), and an increasing number have been found to be translated (Aspden et al., 2014) and produce a functional peptide (Magny et al., 2013; Pauli et al., 2014; Nelson et al., 2016). A significant portion of lncRNAs are therefore mis-annotated, although it is not yet clear how many of the peptides produced by these lncRNAs are functional, or how many lncRNAs function both at the transcript and the peptide level.

1.2 Classification of lncRNAs

Given the heterogeneous nature of lncRNAs, defining informative subcategories can be problematic. As such, their genomic locations and context with respect to protein coding genes are often used for classification purposes (Figure 1.1, Table 1.2). Approximately 53% of lncRNA genes are long intergenic non-coding RNA (lincRNA), located in the genomic interval between protein coding genes, while intronic lncRNAs are found within their introns. Sense and antisense lncRNAs overlap one or more protein coding exons, and are transcribed from the sense or antisense strand relative to a protein coding gene. Sense lncRNAs are also commonly referred to as sense overlapping lncRNAs, clarifying that these lncRNAs overlap the exons of a protein coding gene, on the same strand. Bidirectional lncRNAs initiate transcription within 1kb of a promoter of a protein coding gene on the opposite strand (Ponting, Oliver, and Reik, 2009; Devaux et al., 2015). The ease with which sequencing reads can be attributed to an individual lncRNA varies between these classes. For example, lincRNAs can be identified in RNA-seq and Ribo-seq data sets with higher levels of confidence. This is because they don't overlap a protein coding gene, so their reads are unlikely to be attributed to a mRNA, or as is the case for intronic lncRNAs, to erroneous splicing of a mRNA.

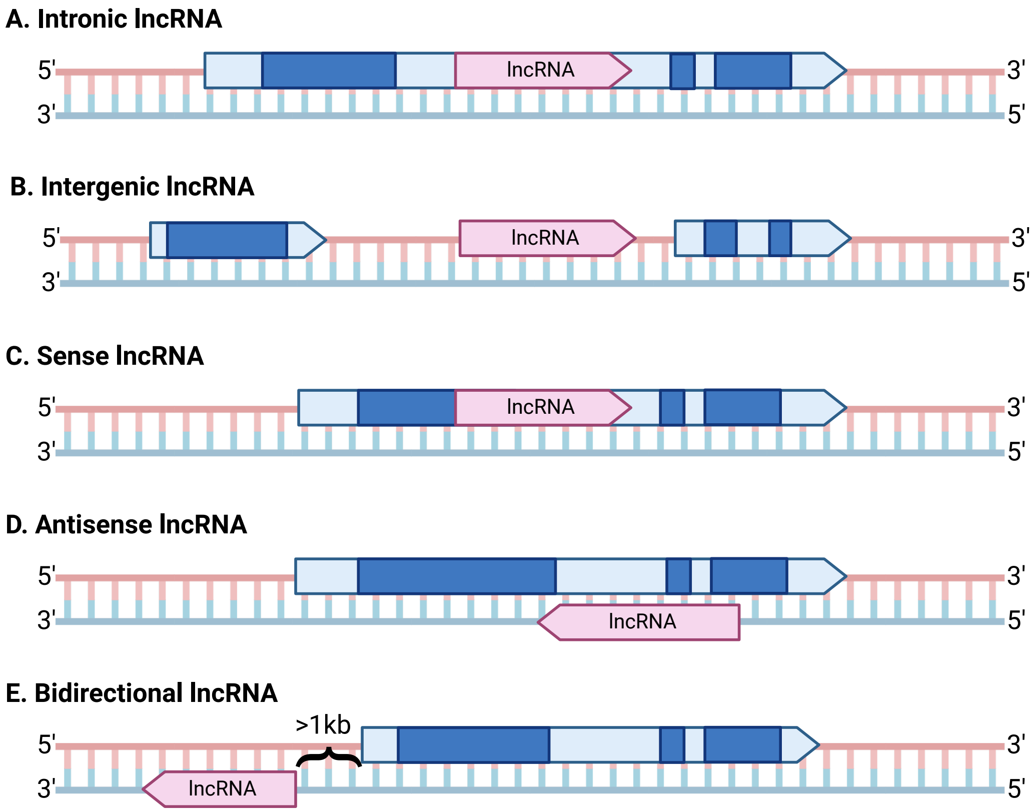


Figure 1.1: **Classes of lncRNAs based on their location and context relative to neighbouring protein coding genes. A.** Intronic lncRNAs; within the introns of protein coding genes. **B.** Intergenic lncRNA; in the genomic interval between protein coding genes. **C.** Sense overlapping lncRNA; overlap one or more exons of a protein coding gene on the sense strand relative to that gene. **D.** Antisense lncRNA; overlap one or more exons of a protein coding gene on the antisense strand relative to that gene. **E.** Bidirectional lncRNA; initiate transcription within 1kb of a promoter of a protein coding gene on the opposite strand. Pink arrows represent lncRNAs genes. Blue arrows represent protein coding genes; light blue sections are introns and dark blue sections are exons. Created with BioRender.com.

Table 1.2: **Number of annotated human lncRNAs by class.** From Gencode release 30, as later releases use a generic long non-coding RNA biotype (Frankish et al., 2021).

Class	**Genes**	**Transcripts**
Intergenic	7,696	14,933
Antisense	5,611	11,636
Intronic	890	950
Sense	179	368
Bidirectional	91	329

1.3 Conservation of lncRNAs

Protein coding genes are highly conserved across species, due to strong selective pressure to maintain the correct reading frame and functional amino acid sequence. LncRNAs generally lack sequence conservation or even orthologs in other species, and are under weaker selective pressure than coding sequences, although the selective pressure is stronger than that on neutrally evolving sequences such as introns (Young et al., 2012; Haerty and Ponting, 2013). Over 70% of lncRNAs have no sequentially conserved orthologs in species which diverged at least 50 million years ago (Hezroni, Koppstein, et al., 2015), and approximately 30% of human lncRNAs are thought to be primate specific (Derrien et al., 2012). Taken together, a reasonable conclusion may be that the majority of lncRNAs are non functional, "junk DNA" (Palazzo and Lee, 2015).

However, the sequence conservation found in protein coding genes is not the only level at which conservation can be observed. Indeed, tRNAs are under strong pressure to maintain their secondary structure, allowing for significant sequence variation in some portions of the transcript, while other features are highly conserved (Lin, Chan, and Lowe, 2019). Where lncRNAs are found to exhibit sequence conservation across species, it is often in short modules, such as the conserved repetitive regions observed in *Xist* (Johnsson et al., 2014; Brockdorff, 2018). If a lncRNA is functioning as a decoy, for example, a short sequence may be all that is required in order to bind and sequester its target, the conservation of which is likely to be missed if the rest of the transcript has diverged. Further, some lncRNA sequence conservation may be being missed due to the methods used in comparative analysis to score sequence similarity. LncRNA genes are smaller than protein coding genes, with a median length of 5,011 nt in humans compared to 23,288 nt for protein coding genes (Zerbino et al., 2018). This means that lncRNAs are often unable to obtain high sequence similarity scores (Gandhi, Caudron-Herger, and Diederichs, 2018), despite the potential for strong constraint of small portions of their sequence. Methods are coming on stream to address this issue, such as SEEKR (Kirk et al., 2018) which evaluates the sequence similarity of functionally related lncRNAs based on short sequence motifs (k-mers), instead of focusing on linear sequence homology.

LncRNA conservation should therefore be approached as multidimensional, considering sequence, expression, synteny, secondary and tertiary RNA structures, and function (Diederichs, 2014). LncRNAs exhibit tissue/cell type and developmental stage specific expression (Ponting, Oliver, and Reik, 2009), and this specificity is conserved to a level comparable to mRNA, as are their promoters (Necsulea et al., 2014; Hezroni, Koppstein, et al., 2015). Syntenic conservation relates to the conservation of genomic context. If the act of transcribing a given lncRNA affects neighbouring genes, the location of this lncRNA is key to its function and may be highly conserved, while the sequence of the lncRNA itself is of little relevance (Diederichs, 2014). LncRNAs may also form functional secondary or tertiary structures, which only require the maintenance of short modules of sequence conservation. However, these structures can be difficult to elucidate and study. Given the length of lncRNAs, bioinformatic methods to predict structure can be ineffective, and often overlook non Watson-Crick interactions (Blythe, Fox, and Bond, 2016). These are interactions other than the canonical Watson Crick G·C and A·U pairings, which include but are not limited to G·U, G·A, A·U, U·U, and G·A pairings (Olson et al., 2009). Approaches to deal with this issue currently include the use of "crowdsourcing"; using RNA folding games and human users to solve complex RNA folding problems (Lee, Kladwang, et al., 2014; Koodli et al., 2019).

Interestingly, conservation of lncRNAs may also vary between species, e.g. strong purifying selection was found to act in the exonic sequences of a set of intergenic lncRNAs in *D. melanogaster*, with little or no conservation seen across the entirety of the sequence in a set of human lncRNAs (Haerty and Ponting, 2013).

1.4 Functions of lncRNAs

For the majority of annotated lncRNAs, their functions are not known, nor are their mechanisms of action. Indeed, a large proportion of lncRNAs are thought to be non-functional (Goudarzi et al., 2019). However, a small proportion of the ∼58,000 human lncRNA transcripts (Frankish et al., 2021) still represents a large population, which function via interactions with RNA, DNA, proteins, or combinations of the three.

Detailed investigation into a single lncRNA with a known phenotype or disease association can reveal a function, mechanism of action, and perhaps conserved function in other species. This is difficult to achieve at scale, especially given the lack of an obvious sequence-function relationship in many cases. One method commonly termed “guilt by association” has been employed, in which lncRNAs are clustered with mRNAs based on expression patterns (Lefever et al., 2017). The clusters are then assigned gene ontology and disease association terms based on the mRNAs, which are assumed to also apply to the lncRNAs in each cluster. Improvements upon this approach integrate more data to gain a more realistic picture of possible functions, for example by including where the transcripts localise in the cell (Uszczynska-Ratajczak et al., 2018). Although this approach is broad and lacks precision, it allows for initial investigations of large groups of lncRNAs, and can provide guidance to search for certain domains or features.

1.4.1 Nuclear functions of lncRNA

Initial research into lncRNA function focused on the nucleus, and some of the best studied lncRNAs exhibit nuclear localisation and function. Here lncRNAs can be subdivided into three broad categories; transcription-only, *cis*-acting and *trans*-acting lncRNAs. Transcription-only lncRNAs are those for which the act of transcription is functional, affecting regulatory elements overlapped by the lncRNA and therefore altering the expression of neighbouring genes. *Cis*-acting lncRNAs regulate genes at their site of transcription via a wide range of mechanisms, while *trans*-acting lncRNAs are transported to and act elsewhere in the genome (Kopp and Mendell, 2018).

1.4.1.1 Regulation via lncRNA transcription

Transcription-only lncRNAs have no function beyond the regulation of neighbouring genes via their transcription, which can have an activating or repressive effect. The transcription of the paternally expressed mammalian lncRNA *Airn* represses the *Igf2r* gene (Figure 1.2A), as it overlaps the *Igf2r* promoter in an antisense orientation and reduces RNA polymerase II recruitment (Latos et al., 2012). This also represses two other genes which form a cluster with *Igf2r*; *Slc22a2* and *Slc22a3*, despite not overlapping them (Sleutels, Zwart, and Barlow, 2002). The resulting *Airn* transcript is highly unstable and poorly conserved, and truncation of the transcript by insertion of polyadenylation cassettes had no impact on function (Latos et al., 2012). This demonstrates that the transcript product is not important to function, only the act of transcription (Seidl, Stricker, and Barlow, 2006).

1.4.1.2 Regulation of chromatin structure

LncRNAs can regulate chromatin structure *in cis* and *in trans*; in this section an example of each mode is provided. Perhaps the best studied example is *Xist*, a key component of X chromosome inactivation in placental mammals (Brown, Ballabio, et al., 1991) (Figure 1.2B). To produce an equal dosage of X-linked genes in males (XY) and females (XX), one of a females X chromosomes is inactivated, referred to as Xi. *Xist* is expressed exclusively from the Xi, and acts *in cis* to coat the chromosome and recruit a number of protein partners which establish and maintain gene silencing. Ultimately the Xi is condensed to form a Barr body (Lyon, 1962; Borensztein et al., 2017), and anchored to the nuclear periphery. Mutations in

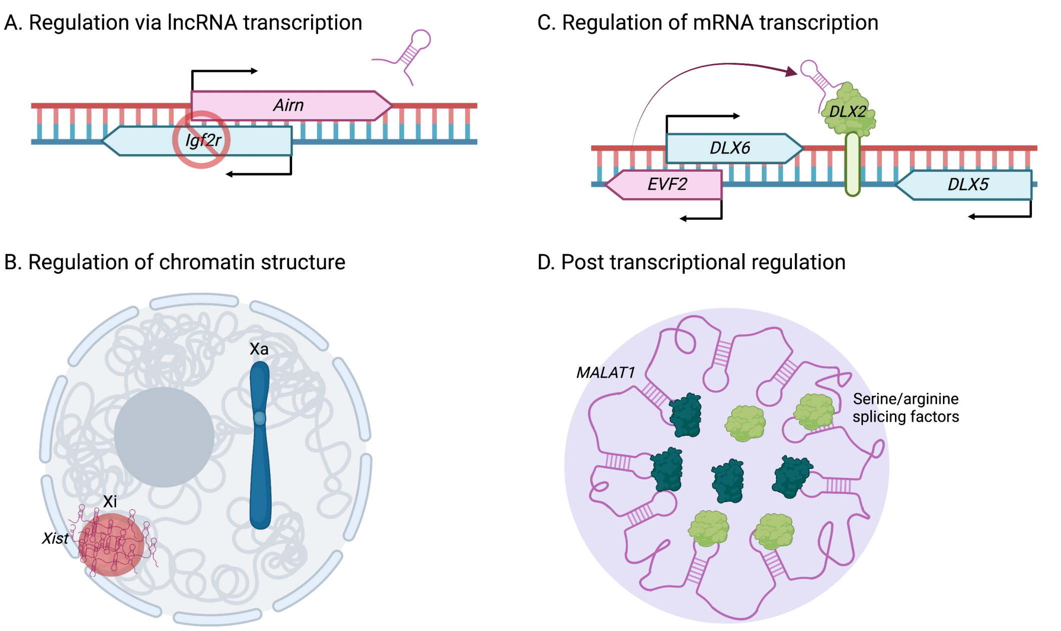

A. Regulation via lncRNA transcription
Airn
Igf2r
C. Regulation of mRNA transcription
DLX2
DLX6
EVF2
DLX5
B. Regulation of chromatin structure
Xa
Xi
Xist
D. Post transcriptional regulation
MALAT1
Serine/arginine
splicing factors

Figure 1.2: **Known nuclear functions of lncRNAs. A.** Regulation via lncRNA transcription; transcription of the lncRNA *Airn* represses *Igf2r* transcription. **B.** Regulation of chromatin structure; *Xist* coats the Xi chromosome, establishing and maintaining gene silencing. **C.** Regulation of mRNA transcription; *Evf-2* recruits *DLX2* to the *DLX5/6* enhancer, increasing their transcription. **D.** Post transcriptional regulation; *MALAT1* sequesters splicing factors to nuclear speckles, regulating alternative splicing of endogenous pre-mRNAs. Created with BioRender.com.

the *Xist* promoter lead to extremely skewed inactivation patterns in human (Plenge et al., 1997).

The sequence of the antisense lincRNA *HOTAIR* is poorly conserved within mammals (He, Liu, and Zhu, 2011), but its function in human has been well studied. Acting *in trans*, *HOTAIR* forms a molecular scaffold between *PRC2* via a 5' domain and *LSD1* via a 3' domain, recruiting them to genes in the *HOXD* cluster (Schorderet and Duboule, 2011). *HOXD* target genes are then downregulated via H3K27 trimethylation by *PRC2*, and H3K4 demethylation by *LSD1* (Rinn et al., 2007). This important role in epigenetic regulation has led to the implication of *HOTAIR* dysregulation in a range of human cancers (Bhan and Mandal, 2015).

1.4.1.3 Regulation of mRNA transcription

The transcription of mRNA can be regulated by lncRNAs via interaction with transcription factors (TF), having activating and repressive effects *in cis* and *in trans*. *In cis*, a lncRNA may hybridise to a TF binding site, blocking TF binding and repressing transcription. Alternatively, the lncRNA may recruit TFs to a nearby binding site, enhancing transcription. One such example is *Evf-2*, which is antisense to the TF *DLX6* (Figure 1.2C). *Evf-2* forms a complex with the *DLX2* TF, increasing the transcriptional activity of the *DlX5/6* enhancer (Feng et al., 2006). This interaction has been demonstrated by immunoprecipitation of *DLX2 in vivo* in rat embryonic nuclear extracts, followed by reverse transcription polymerase chain reaction (RT-PCR) of the *Evf-2* transcript (Feng et al., 2006). *In trans*, lncRNAs can affect the localisation of TFs and other proteins, either enhancing or repressing their access to binding sites.

1.4.1.4 Post transcriptional regulation

LncRNAs can also affect the post transcriptional regulation of mRNAs in the nucleus, regulating their splicing, transport and degradation. Originally identified in lung cancer, in particular in tumours which subsequently metastasised (Ji, Diederichs, et al., 2003), *MALAT1* is a highly abundant lncRNA with a well conserved sequence in vertebrates, which has been linked to a range of functions and diseases (Arun, Aggarwal, and Spector, 2020). One function of *MALAT1* is *in trans* as a decoy, sequestering splicing factors to nuclear speckles (Figure 1.2D). Specifically, *MALAT1* binds to serine/arginine splicing factors, regulating alternative splicing in a set of endogenous pre-mRNAs (Tripathi et al., 2010).

1.4.2 Cytoplasmic functions of lncRNA

Although nuclear lncRNAs are more extensively documented, many lncRNAs are enriched in the cytoplasm (Ulitsky and Bartel, 2013), where they act *in trans* to modulate cellular function. Recent work in the human K652 cell line (chronic myelogenous leukaemia) detected 54% of expressed lncRNAs in the cytoplasm (Carlevaro-Fita et al., 2016). Cytoplasmic lncRNAs are generally 5'capped, spliced and polyadenylated (Fuke and Ohno, 2008).

1.4.2.1 Regulation of mRNA translation

A number of lncRNAs have been implicated in mRNA translation, including *BACE1-AS*, an antisense lncRNA which associates with the mRNA *BACE1*, increasing its stability (Figure 1.3A). The base-pairing between the transcripts also masks the binding site of miR-485-5p, a miRNA which represses *BACE1* translation (Faghihi, Zhang, et al., 2010). This upregulates the translation of *BACE1* which has been implicated in heart disease and the progression of Alzheimer's disease (Greco et al., 2017). This mechanism was first elucidated when expression of *BACE1-AS* was found to be elevated in the brains of Alzheimer's patients (Faghihi, Modarresi, et al., 2008). Further, the knockdown of *BACE1-AS* was found to result in the reduction of *BACE1* mRNA, and overexpression results in increased *BACE1* mRNA levels. *BACE1-AS* can also function as a competing endogenous RNA (ceRNA) as it shares miRNA

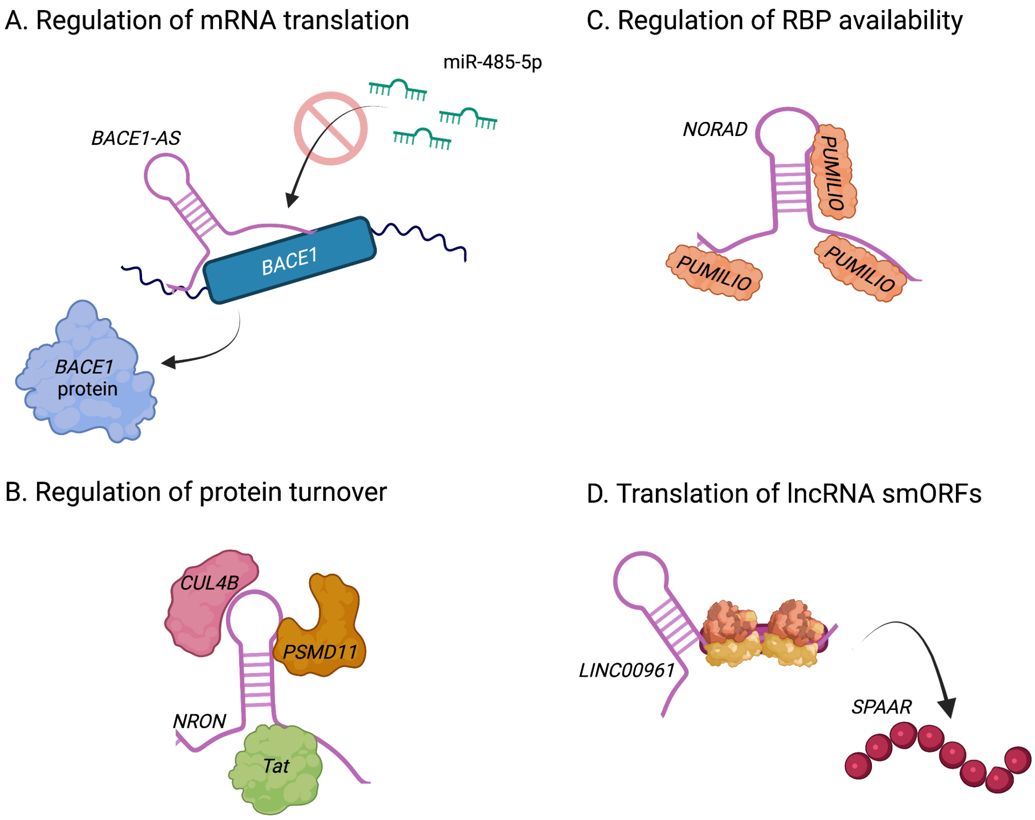

A. Regulation of mRNA translation
miR-485-5p
BACE1-AS
BACE1
BACE1 protein
C. Regulation of RBP availability
NORAD
PUMILIO
PUMILIO
PUMILIO
B. Regulation of protein turnover
CUL4B
PSMD11
NRON
Tat
D. Translation of lncRNA smORFs
LINC00961
SPAAR

Figure 1.3: **Potential cytoplasmic functions of lncRNAs. A.** Regulation of mRNA transcription; base-pairing between *BASE1-AS* and *BACE1* increases translation of*BACE1*. **B.** Regulation of protein turnover; *NRON* forms a complex with *CUL4B*, *PSMD11*, and *Tat*, promoting the degradation of *Tat*. **C.** Regulation of RBP availability; *NORAD* sequesters *PUMILIO* proteins, limiting their repressive effect on target mRNA. **D.** Translation of lncRNA smORFs; *LINC00961* is translated to produce *SPAAR*. Created with BioRender.com.

target sites with *BACE1*, sequestering miRNAs which target the mRNA and preventing its degradation (Zeng et al., 2019).

1.4.2.2 Regulation of protein turnover

As well as regulating mRNA stability, lncRNAs can regulate protein stability and turnover. The expression of lncRNA *NRON* alters significantly upon HIV-1 infection (Imam et al., 2015), and has been found to form a complex with the ubiquitin/proteasome components *CUL4B* and *PSMD11*, and the HIV-1 regulatory protein *Tat* (Figure 1.3B) (Li, Chen, et al., 2016). This promotes the degradation of *Tat*, potentially suppressing viral transcription.

1.4.2.3 Regulation of RBP and miRNA availability

Other cytoplasmic lncRNAs function as decoys, competing with mRNA for RNA-binding proteins (RBPs), circRNAs, and microRNAs (Schmitz, Grote, and Herrmann, 2016; Noh et al., 2018). *HULC*, for example, acts as a decoy for miR-372, reducing the repressive effect of miR-372 on the mRNA *PRKACB* (Wang, Liu, et al., 2010). The conserved mammalian lncRNA *NORAD* is key to genomic stability, with knockdown triggering aneuploidy in cell lines (Lee, Kopp, et al., 2016). *NORAD* contains a high number of *PUMILIO* binding sites, sequestering a large portion of *PUMILIO* proteins and limiting their repressive effect on mRNAs, which include those involved in DNA repair and mitosis (Figure 1.3C).

1.4.2.4 Translation of lncRNA smORFs

A proportion of lncRNAs, despite being "non-coding", contain actively translated smORFs, some of which produce functional peptides. The lncRNA *LINC00961* is conserved between human and mouse, and contains a smORF which encodes the peptide 90 aa *SPAAR* (Figure 1.3D) (Matsumoto, Pasut, et al., 2017). *SPAAR* contains a transmembrane domain, and sequesters subunits of the v-ATPase complex at the lysosomal membrane, inhibiting mTORC1 activation. mTORC1 controls protein synthesis and cell growth, and following muscle injury *SPAAR* is thought to be downregulated in order to increase mTORC1 activation (Matsumoto, Clohessy, and Pandolfi, 2017).

1.5 Eukaryotic translation

Eukaryotic translation consists of four main stages; initiation, elongation, termination and recycling (Kapp and Lorsch, 2004). Each of these steps are tightly regulated, and aberrations can lead to disease, such as the various ribosomopathies caused by mutations in components of the ribosome (Kang et al., 2021). However, this model should be taken as a basic understanding of translation, which can be deviated from to allow for more complex translational regulation and non-canonical translation.

Translation initiation describes the assembly of the 80S ribosome at the start codon of the mRNA, and occurs canonically via the 5’ cap dependent pathway (Figure 1.4A). Initially, the ternary complex (eIF2·GTP·Met-tRNA$_i$) binds the 40S ribosomal subunit, forming the 43S pre-initiation complex. The 43S is loaded at the 5’ cap of an mRNA, and begins scanning in the 3’ direction through the 5’ UTR. Upon encountering an AUG start codon in a suitable sequence context, the anticodon of the Met-tRNA$_i$ binds to the start codon, halting scanning and triggering the joining of the 60S ribosomal subunit to form the 80S ribosome. In vertebrates the optimum sequence context is GCC(A/G)CC**AUG**G; the Kozak consensus sequence (Kozak, 1986). Weaker Kozak sequences; i.e. divergence from the optimal sequence, increase the likelihood of ‘leaky scanning’, where the ribosome moves past a start codon without initiating translation. Translation can also initiate from alternate start codons, which in eukaryotes can occur at all single-base substitutions from AUG, bar AGG and AAG (Peabody, 1989).

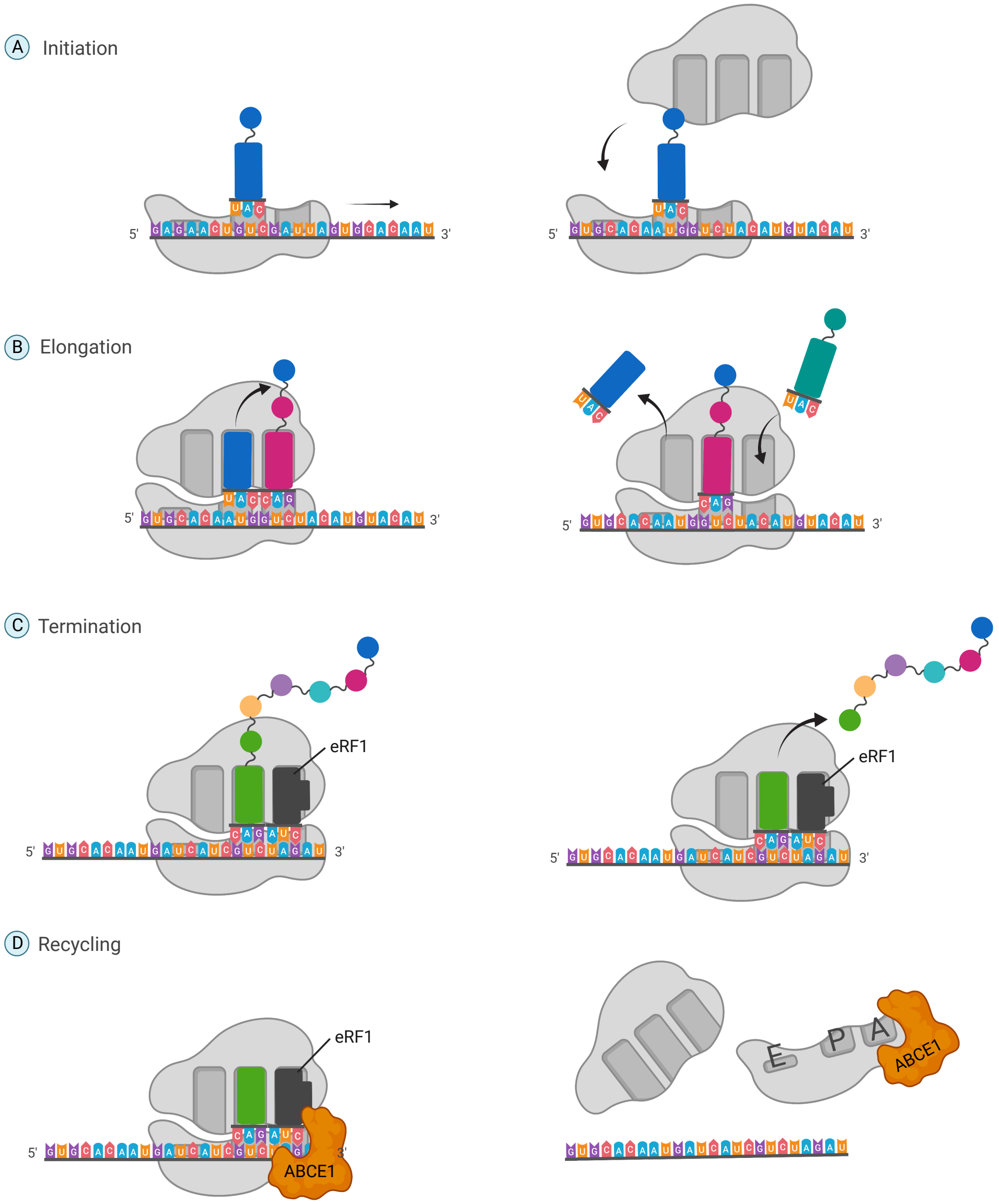


Figure 1.4: **Key steps in eukaryotic translation. A.** Initiation; the 43S pre-initiation complex is loaded at the 5’ cap, scans in the 3’ direction, and upon encountering the start codon joins with the 60S subunit to form the 80S ribosome. **B.** Elongation; tRNA enters the ribosome, peptide bonds are catalysed between their amino acids to grow the peptide chain, and the ribosome moves along the transcript. **C.** Termination; the ribosome encounters a stop codon, and the completed polypeptide chain is released by eRF1. **D.** Recycling; ABCE1 catalyses recycling of the 80S complex into 40S and 60S subunits. The E-site, P-site, and A-sites are labeled on the 40S. Adapted from Protein Translation, by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

Following initiation the ribosome begins elongation (Figure 1.4B). tRNAs enter the ribosome at the acceptor site (A-site), where they bind to their matching codon, and a peptide bond is catalysed between the two amino acids at the A-site. The ribosome moves to the next codon in the 3' direction and the empty tRNA is released via the exit site (E-site), allowing the cycle to repeat again. A key aspect of this process is maintenance of the reading frame, and prevention of frameshift events. Although the frequency of these errors is extremely low in the general population, there are a subset of mRNAs for which frameshift events occur at a significantly higher frequency (Atkins et al., 2000). In the *E.coli prfB* gene, translation of the full-length termination release factor 2 protein requires a +1 frameshift, avoiding an in frame stop codon (Márquez et al., 2004). This frameshift occurs $\sim 30\%$ of the time, compared to approximately 1/30,000 incorporated amino acids in the wider mRNA population (Weiss et al., 1988; Jørgensen and Kurland, 1990).

Elongation can be slowed by particular amino acid sequence combinations, either due to some peptide bonds forming at a slower rate (Wohlgemuth et al., 2008), or due to certain codons being less optimal and therefore having a smaller pool of tRNAs available for elongation (Ikemura and Ozeki, 1983). The ribosome may also be stalled by secondary structures in the mRNA, such as stem loops or pseudoknots (Dinman, 2012).

When the ribosome reaches the end of an ORF, signified by a stop codon (UAA, UAG, or UGA) in the A-site, translation is terminated (Figure 1.4C). The completed polypeptide chain is released following hydrolysis of the bond linking the chain to the P-site tRNA by eukaryotic peptide chain release factor subunit 1 (eRF1). As in initiation, the efficiency of termination is affected by the sequence context, and the combination of the stop codon and the first nucleotide of the 3' UTR increase or decrease the likelihood of stop codon readthrough (Schuller and Green, 2018).

Finally, the 80S complex must be recycled back into 60S and 40S subunits, ready to carry out another round of translation (Figure 1.4D). The ATP-binding cassette (ABCE1) binds to the ribosome, dissociating the 80s into free subunits, and remaining bound to the 40s (Schuller and Green, 2018).

Further translational regulation occurs at the level of the ribosome. Far from being a homogenous pool of translational machinery, recent evidence has revealed heterogeneity in the composition of ribosomes, termed specialised ribosome (Guo, 2018). Although this is a relatively new area of research, several mechanisms of action for translation regulation have been hypothesised including altering 40S recruitment to the transcript, start codon selection, elongation rate, and stop codon recognition (Norris, Hopes, and Aspden, 2021).

1.6 Non-canonical translation

The canonical view of eukaryotic translation describes an ORF of 100 codons or more, with an AUG start codon, in a protein coding transcript. However, advances in bioinformatic, sequencing, and peptidomic techniques have expanded our understanding of translation to reveal a diverse range of non-canonical ORFs. These include upstream open reading frames (uORFs) found in the 5'UTR, upstream of the start codon of the main ORF, downstream open reading frames (dORF), found in the 3'UTR, and smORFs (sometimes referred to as sORFs) of <100 codons found in lncRNAs, circRNAs, and miRNAs. Work to explore the small ORFeome found that 10% of mouse proteins are shorter than 100 aa (Frith et al., 2006), and hundreds of potential functional smORFs have been identified in *Drosophila* (Ladoukakis et al., 2011).

Translation initiation can also occur independently of the 5' cap, via internal ribosome entry site (IRES) mediated initiation. These highly structured elements initiate cap-independent protein synthesis by interacting with RBPs and ribosomes, allowing downstream initiation of translation (Karginov et al., 2017). Research on IRES elements has mainly focused on viral gene expression, due to the necessity of reducing reliance on translation initiation factors in viral protein translation. However, in conditions that may compromise cap dependent translation, such as the stress conditions found in inflammatory breast cancer, cellular IRESs have been found to be both present and functional (Silvera et al., 2009; Komar and Hatzoglou, 2011). Approximately 100 cellular IRES elements have now been experimentally characterised in eukaryotes (Kolekar et al., 2016).

1.6.1 Upstream ORFs

Over 50% of human mRNAs have at least one AUG upstream of the main ORF in the 5' untranslated region (UTR), creating a potential uORF. Fewer uORFs are found in in eukaryotes than would be expected by chance (Iacono, Mignone, and Pesole, 2005; Lawless et al., 2009), and existing uORFs are under selective pressure (Churbanov et al., 2005). This is because the presence of uORFs can have a repressive effect on the main ORF (Andreev et al., 2022) reducing translational efficiency by an average of 30-48% in human, mouse and zebrafish (Chew et al., 2013). This repression occurs via several mechanisms, including stalling of the ribosome on the uORF, which can block additional scanning ribosomes or induce mRNA decay (Meijer and Thomas, 2002). The ribosome may also continue scanning and reinitiate at the main ORF, although this is considered to be inefficient, particularly if there is a small distance between the uORF and main ORF (Kozak, 1987; Barbosa, Peixeiro, and Romão, 2013). The majority of uORFs have an adequate or weak translation initiation site (TIS) context (Iacono, Mignone, and Pesole, 2005), meaning they may also be skipped by the translational machinery, termed leaky scanning (Kozak, 1980). Under stress conditions levels of leaky scanning across uORFs can increase, upregulating translation of the main ORF (Renz, Valdivia-Francia, and Sendoel, 2020).

As uORFs often function via their presence and translational context, their sequence can be poorly conserved, as the resulting peptide is not functional (Iacono, Mignone, and Pesole, 2005). However, mass spectrometry analysis has found evidence of stable peptides produced from uORFs (Oyama et al., 2004; Slavoff et al., 2013) which can act as another layer of regulation of the main ORF, or have other functions (summarised in Renz, Valdivia-Francia, and Sendoel, 2020).

1.6.2 Evidence for non-canonical ORFs

The technique of ribosome profiling (Ribo-Seq) provides evidence to support the translation of non-canonical ORFs. Initially developed in 2009 (Ingolia, Ghaemmaghami, et al., 2009), Ribo-Seq is a high throughput method which identifies a snapshot of the actively translated RNA in a sample by establishing the sequences of all RNA bound by ribosomes. To perform Ribo-Seq, a sample of cells or tissue is treated with a translational inhibitor and lysed. Following isolation of the cytoplasm, RNaseI digestion is performed, which digests all RNA in the sample not protected by a bound ribosome or RNA-binding protein. The lysate is subjected to ultracentrifugation on a sucrose cushion to select the ribosome bound RNA fragments, called ribosome footprints. Ribosome footprints are generally 28-32 nucleotides in length, although this varies between organism, cell/tissue type, conditions, and experimental protocol (Aspden et al., 2014; Ingolia, Brar, et al., 2013). The RNA fragments are then purified to remove the ribosomes and other bound complexes, and processed to undergo RNA sequencing.

Choice of experimental protocol and conditions can have a major impact on the quality and expected outputs from Ribo-Seq data. The translational inhibitor used to stall the ribosomes affects the position of ribosome footprint across the ORF; the commonly used inhibitor cycloheximide creates a build up of Ribo-Seq reads around the start codon and first 15 nt of the ORF in human (Douka, Agapiou, et al., 2022). Ribo-Seq has been adapted for and used in many species, cell types, and tissues (Ingolia, Ghaemmaghami, et al., 2009; Ingolia, Lareau, and Weissman, 2011; Chew et al., 2013; Aspden et al., 2014; Chothani, Adami, Widjaja, et al., 2022). There is however a high likelihood of false positives arising from Ribo-Seq data (Gelhausen et al., 2022; Prensner, Abelin, et al., 2023). RNA bound by ribosomes is not necessarily being translated, as scanning 40S ribosomal subunits can also create footprints, or a single 80S ribosome may be spuriously bound. A key step in the analysis of Ribo-Seq data is therefore the selection of reads exhibiting triplet periodicity. This is the signature movement of the translating ribosomes across the ORF, three nucleotides at a time, in frame with the start codon.

A wide range of computational methods have been developed to analyse Ribo-Seq data (reviewed in (Kiniry, Michel, and Baranov, 2019)), in particular to detect and measure active translation. The first measure developed was Translational efficiency (TE), from the original Ribo-Seq study (Ingolia, Ghaemmaghami, et al., 2009). This measure simply divides the density of ribosome footprints by RNA-seq reads to calculate TE, and early works defined actively translated transcripts as those with high TE (Ruiz-Orera, Messeguer, et al., 2014). Alternately, the ribosome release score (RRS) focused on the release of ribosomes at the stop codon, causing a decrease in Ribo-Seq coverage over the 3'UTR (Guttman, Russell, et al., 2013). RRS is defined as the ratio of the number of Ribo-Seq reads in the ORF and the 3'UTR, normalised by length and the ratio of RNA-Seq reads. However, the RRS does not account for cases where multiple ORFs are actively translated. Another strategy involves the profiling of distinct Ribo-Seq read patterns around start and stop codons, which forms part of the ORF-RATER (Fields et al., 2015) and riboHMM (Raj et al., 2016) approaches. Depending on the translational inhibitor used in the Ribo-Seq protocol, these profiles can vary widely. Later methods including ORFscore (Bazzini et al., 2014), RibORF (Ji, Song, et al., 2015), and Ribotaper (Calviello, Mukherjee, et al., 2016) have focused on the sub-codon resolution on the Ribo-Seq data, scoring ORFs based on the number of in frame reads and coverage across the ORF. A standardised annotation of non-canonical human ORFs identified using Ribo-Seq is currently being developed, the majority of which were identified using methods which rely on the sub-codon resolution of Ribo-Seq reads (Mudge et al., 2022). Recent work has also looked to expand these annotations to include more data from primary biological material, as the majority of human Ribo-Seq studies have been performed on cell lines, due to the need for large amounts of primary material for this protocol (Chothani, Adami, Widjaja, et al., 2022).

The Ribo-Seq protocol has also been developed further to improve the detection of active translation. Polysome profiling (Poly-Ribo-Seq) is a variation of Ribo-Seq which includes a polysome fractionation step, allowing RNA bound by multiple ribosomes (polysomes) to be specifically selected and sequenced (Figure 1.5) (Aspden et al., 2014). A downside of this technique is the loss of particularly small smORFs of 80 nt or less, which are too small to be bound by multiple ribosomes. Poly-Ribo-Seq has revealed populations of cytoplasmic lncRNAs associating with translated transcripts, or undergoing translation themselves (Aspden et al., 2014).

Evidence of non-canonical translation is also gathered by focusing on the peptidome. This is key, as active translation of an ORF does not prove that the resulting peptide is stable, and it may be degraded shortly after production. Used in combination, bioinformatic predictions and evidence of translation from Ribo-Seq can aid peptide discovery by mass spectrometry, allowing for the design of custom protein databases (Menschaert et al., 2013; Bazzini et al., 2014; Aspden et al., 2014). However, as the traditional definition of protein coding ORFs is 100 codons or more, many projects have an artificial cut-off of 100 aa as an effort to reduce false positives. Further, technological limitations mean that validating peptides under 50 aa in length is particularly challenging, and specific mass spectrometry techniques are required to study them (Slavoff et al., 2013; Khitun and Slavoff, 2019). In particular, short peptides are generally lowly abundant, so specific protocols to enrich for short peptides are required (Fabre, Combier, and Plaza, 2021). The choice of enzyme when digesting peptides is also

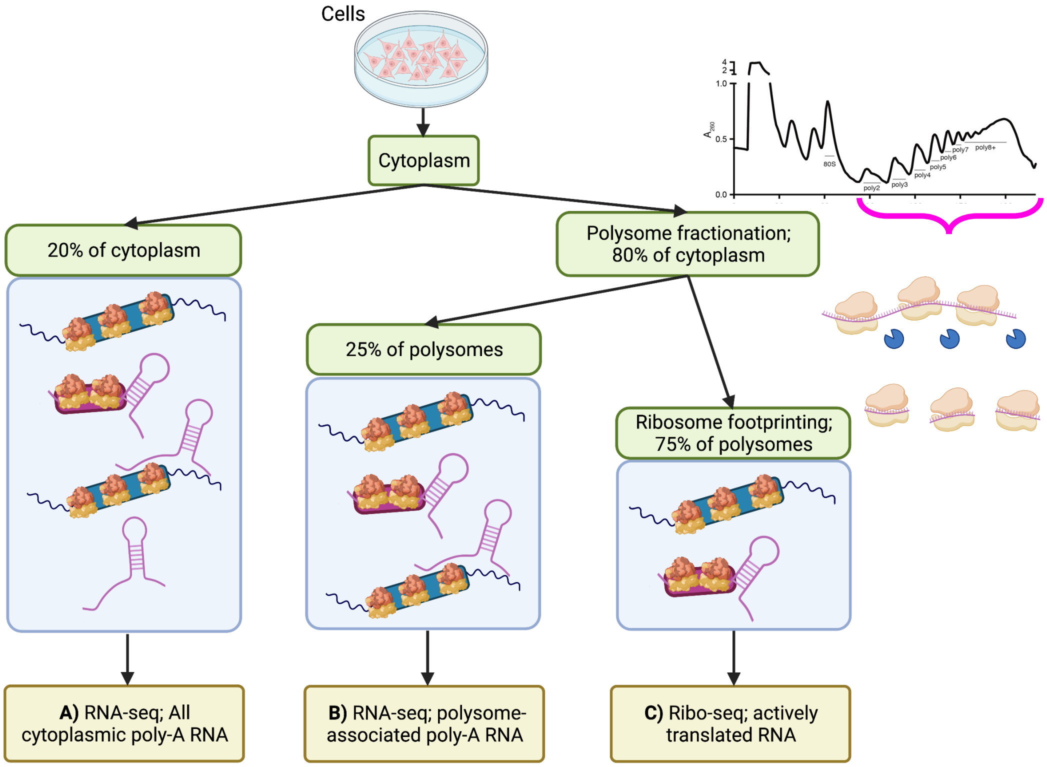

Cells
Cytoplasm
20% of cytoplasm
Polysome fractionation; 80% of cytoplasm
25% of polysomes
Ribosome footprinting; 75% of polysomes
A) RNA-seq; All cytoplasmic poly-A RNA
B) RNA-seq; polysome-associated poly-A RNA
C) Ribo-seq; actively translated RNA
4
2
1.0
0.5
0.0
A_{260}
80S
poly2
poly3
poly4
poly5
poly6
poly7
poly8+

Figure 1.5: **Schematic of Poly-Ribo-Seq analysis.** Cells are lysed, and the cytoplasm isolated. 20% of the cytoplasm is Poly-A selected and sequenced using RNA-seq, to give **A)** All cytoplasmic poly-A RNA. This consists of polysome-associated translated mRNA and lncRNA, polysome-associated (not translated) lncRNA, and cytosolic lncRNAs which do not interact with the translational machinery. The remaining 80% of cytoplasm undergoes sucrose density gradient ultracentrifugation to select polysome fractions; those labelled highlighted in pink on the schematic. 25% of the merged polysome fractions are Poly-A selected and sequenced using RNA-seq to give **B)** Polysome-associated poly-A RNA. This consists of polysome-associated translated mRNA and lncRNA, and polysome-associated (not translated) lncRNA. The remaining 75% of the merged polysome fraction is sequenced using Ribo-seq, in which the polysome fractions are treated with RNaseI to isolate polysome protected fragments, identifying **C)** Actively translated mRNA and lncRNA. Created with BioRender.com.

key, as commonly used enzymes such as trypsin may create fragments which are too small to be captured by the mass spec analysis (Kim, Zhong, and Pandey, 2016).

1.6.3 Translation of lncRNA ORFs

Many transcripts were originally annotated as lncRNAs because they lack ORFs >100 aa. However, the majority of lncRNAs contain in-frame start and stop codon pairs, creating potential smORFs with an average size of 24 codons (Couso and Patraquim, 2017). Multiple smORFs are often found on a single transcript in overlapping or polycistronic arrangements, with a median of 6 smORFs per lncRNA. LncRNA smORFs are the third most abundant class of smORFs after intergenic smORFs and uORFs (Couso and Patraquim, 2017). Previously thought to be untranslated and too small to be functional, ribosome profiling studies have detected translated lncRNA smORFs with patterns of ribosome footprinting indicative of translation (Ingolia, Lareau, and Weissman, 2011; Bazzini et al., 2014; Aspden et al., 2014). However, ribosome occupancy is not sufficient to demonstrate the production of functional peptides, and extensive work has been carried out to characterise these translational events.

Barriers to the characterisation of lncRNA smORFs include their size, which leads to low conservation scores from the majority of sequence based bioinformatic methods, and their assumed lack of coding potential (Andrews and Rothnagel, 2014). However, we do see examples of conserved smORFs, such as *Sarcolamban* in *Drosophila*. Originally annotated as *pncr003:2L*, this gene encodes two smORFs with lengths of 28 and 29 aa, which produce transmembrane peptides found to regulate calcium transport in the heart by interacting with the sarcoplasmic reticulum calcium pump, lowering its affinity for calcium (Magny et al., 2013). These peptides are thought to be functional homologues to the Sarcolipin and Phospholamban peptides found in vertebrates, revealing a conserved ancestral smORF peptide family. A further barrier is the use of non-canonical start codons by many smORFs, meaning they are often overlooked (Pueyo, Magny, and Couso, 2016). The peptides produced by lncRNA smORFs are also smaller and generally of lower abundance and stability than those from canonical ORFs; in combination with their small size this can make their detection by methods such as mass spectrometry problematic.

The pool of identified lncRNA smORF peptides has expanded rapidly (Aspden et al., 2014; Ruiz-Orera, Messeguer, et al., 2014; van Heesch, Witte, et al., 2019; Patraquim et al., 2022; Barczak et al., 2023), but only a select few peptides have been verified and studied in detail, as the the heterogeneity and technical challenges presented by lncRNAs make this difficult to achieve at scale. The mouse lncRNA *MyolncR4*, for example, encodes the 56 aa trans-membrane peptide *LEMP* which is highly conserved in vertebrates (Wang, Fan, et al., 2020). *LEMP* localises to the mitochondria, and has a role in skeletal muscle formation and regeneration. *LEMP* shares 71% aa identity between mammals and zebrafish, where its elimination via knockdown and knockout both result in impaired muscle development, which can be restored by the mouse ortholog (Wang, Fan, et al., 2020). In human, the *LEMP* peptide is referred to as *mitoregulin* from the gene *LINC00116*, where it localises to the inner mitochondrial membrane and binds cardiolipin (Stein et al., 2018).

1.7 The emergence of new genes

The emergence of new genes and their products is essential in the evolution of novel phenotypes, allowing existing species to adapt to their environment, and in some cases leading to speciation (Kaessmann, 2010; Chen, Krinsky, and Long, 2013). New protein coding genes can arise via an array of molecular mechanisms, including gene duplication, exon shuffling, fission, fusion, and horizontal gene transfer (Chen, Krinsky, and Long, 2013). The majority emerge from existing coding sequences via gene duplication, either by DNA-based duplication or RNA-based retrotransposition (Long et al., 2013). Following DNA based-duplication, the most probable outcome is that the duplicate becomes a non-functional pseudogene, and is subsequently lost from the genome (Ohno, 1972). Alternatively, both copies of the gene may be retained, preserving the ancestral function and increasing production of its protein product (Cotton and Page, 2005), although this process is complex as increased expression is not necessarily advantageous (Rogozin, 2014). Yet another outcome can be subfunctionalisation, wherein multiple functions of a single ancestral gene are split between the two copies. Finally, the duplicate may acquire a novel function via new regulatory elements which alter its expression or mutations which produce a novel protein, termed neofunctionalisation. Following RNA-based retrotransposition, in which genes undergo reverse transcription and insertion into the genome, the duplicate is more likely to evolve a novel function (Kaessmann, 2010). These mechanisms are not mutually exclusive, and may act in combination to produce novel genes. The emergence of new genes is a key aspect of evolution, contributing to the vast array of adaptations and phenotypes observed in organisms today.

1.7.1 Origins of lncRNAs

The evolutionary origins of lncRNA, remain unclear. Given their heterogeneity it is unlikely that they all emerged the same way. Thus far, five hypotheses have been proposed which may explain their origins (Figure 1.6) (Ponting, Oliver, and Reik, 2009):

i) Mutation of a protein coding gene

Mutations can occur in protein coding genes which disrupt the frame of the ORF without affecting expression of the transcript, thus creating a non coding gene (Figure 1.6i). A key example of this process is the *Xist* lncRNA, which evolved via a combination of mechanisms, including the pseudogenisation of the protein coding gene *Lnx3* (Duret et al., 2006). Up to 5% of conserved mammalian lncRNAs are thought to have originated from protein coding genes (Hezroni, Ben-Tov Perry, et al., 2017), and recent work identified 33 novel human lncRNAs which arose from protein coding gene loss events (Wen et al., 2023).

ii) Chromosomal rearrangement

Following chromosomal rearrangements, sequences that were previously well separated can move into close proximity, creating a potential non coding gene (Figure 1.6ii). As the eukaryotic transcriptional machinery is relatively promiscuous, if this region of the chromosome is in a permissive chromatin environment transcription may then initiate (Palazzo and Koonin, 2020). This could lead to a transcript which is only occasionally expressed at low levels, or further changes could produce a functional lncRNA transcript.

iii) Duplication of a non-coding gene

As with protein coding genes, non-coding genes can also be duplicated by DNA duplication or RNA-based retrotransposition. Retrotransposition can produce a functional retrogene (Figure 1.6iii), or a non-functional retropseudogene (Ponting, Oliver, and Reik, 2009).

iv) Tandem duplication

New lncRNAs can also emerge from local tandem duplication events which increase the length of existing genes, producing lncRNAs containing neighbouring repeats (Figure 1.6iv). The nuclear lncRNA *Kcnq1ot1*, for example, contains a large repeat rich region which spans more than half of the transcript (Pandey et al., 2008).

v) Transposable element insertion

Transposable elements are portions of DNA that can travel within the genome. They have contributed massively to the human genome; $\sim 45\%$ is recognisably derived from transposable elements, with the true percentage likely higher as more ancient transposable elements

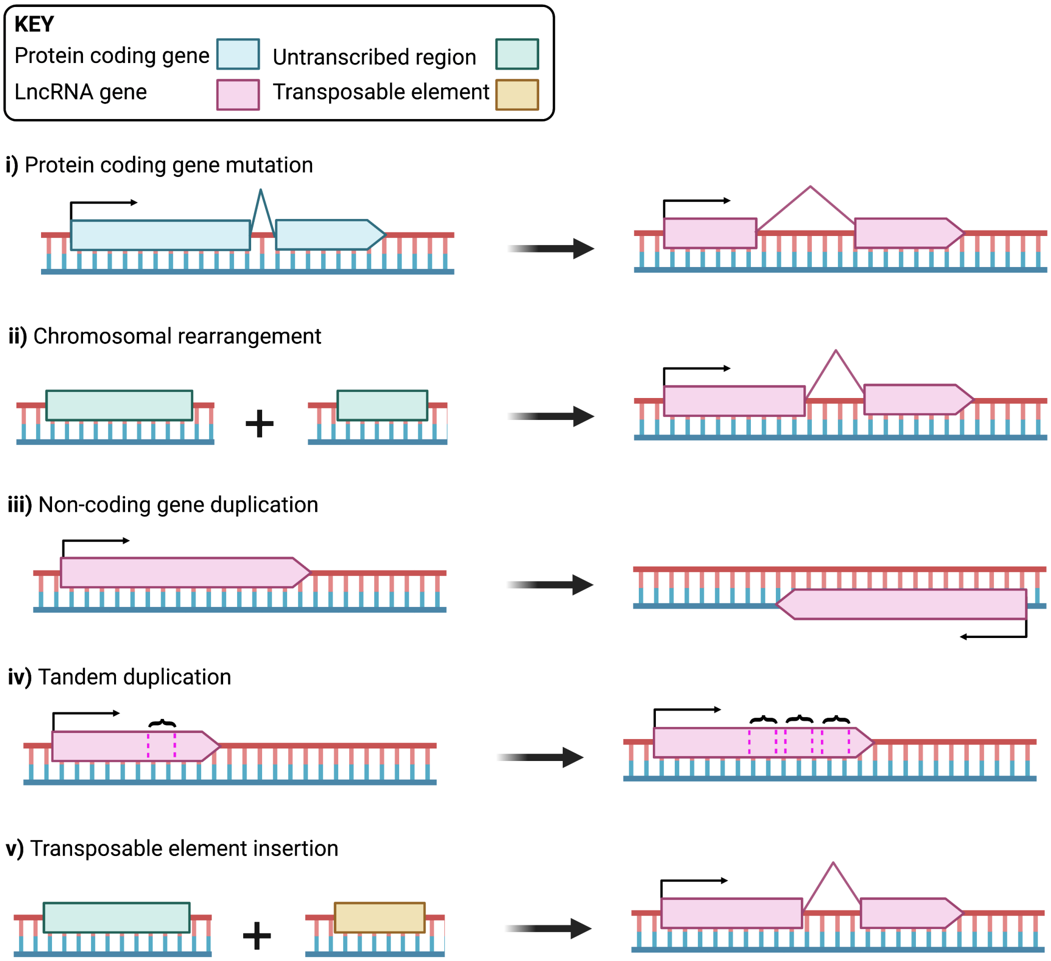

KEY
Protein coding gene
Untranscribed region
LncRNA gene
Transposable element
i) Protein coding gene mutation
ii) Chromosomal rearrangement
+
iii) Non-coding gene duplication
iv) Tandem duplication
v) Transposable element insertion
+

Figure 1.6: **Possible origins of lncRNAs i)** Mutations in protein coding genes can cause frame disruptions, creating a non-coding gene. **ii)** Chromosomal rearrangements can move to previously separate, untranscribed regions into close proximity, creating a multi-exon non-coding gene. **iii)** Duplication of a non-coding gene via retrotransposition can generate a new non-coding gene. **iv)** Tandem duplication events can create neighbouring repeats within an existing gene. **v)** Transposable elements can be co-opted to create a new gene. Blue arrows represent protein coding genes, and pink arrows lncRNA genes. Green boxes represent untranscribed regions, and yellow transposable elements. Created with BioRender.com.

have diverged beyond recognition (Cordaux and Batzer, 2009). They are also particularly prevalent in lncRNAs, unlike in protein coding genes, occurring in over 66% of lncRNA transcripts and accounting for $\sim 30\%$ of lncRNA sequence in human (Kapusta et al., 2013). Transposable elements are often incorporated into existing lncRNAs, such as *Xist* which sequentially gained and assimilated lineage-specific transposable elements throughout evolution (Elisaphenko et al., 2008). Novel ncRNAs can also emerge from transposable element insertions (Figure 1.6v), as they often contain functional transcription start sites (Ulitsky, 2016), producing RNA from previously untranscribed regions.

1.7.2 *De novo* ORFs

In contrast to processes involving existing genes, the *de novo* emergence of novel protein coding sequences from non-coding sequences is poorly understood. However, this process represents a significant source of evolutionary innovation; in *Drosophila*, ~12% of species or lineage specific new genes originated *de novo* (Zhou et al., 2008). Generally, *de novo* genes can be identified by comparisons with sister lineages, as a lack of homologs can indicate that the gene is unique to a single lineage. However this is not conclusive, as the gene may have been lost in the sister species, or may just be missing due to an annotation error.

To produce a peptide, a *de novo* gene must be transcriptionally and translationally active, include all of the associated sequence features required to regulate these processes, and not produce a product with strong deleterious effects. It is not clear in which order these features are gained, and “ORF first” (Begun et al., 2006) and “transcription first” (Levine et al., 2006) models have been proposed (Figure 1.7). In the ORF first model, a smORF occurs at random in non-coding DNA, gains a promoter, and becomes part of a transcriptional unit. The smORF then has the potential to be translated, producing a small peptide. Over time, this new gene could be subject to natural selection, gaining traits associated with canonical protein coding genes, such as increasing expression levels of its RNA and protein products, increasing length, displaying codon usage bias, and gaining a function at the peptide level. In the transcription first model, active transcription is present before the gain of a smORF. These models represent a continuum, where sequences can exist in the genome at each intermediate step, and also proceed "backwards" through these states, by truncation, and loss of function, translation, or transcription.

LncRNAs have been postulated to represent steps along this evolutionary continuum, in part due to their similarity to novel genes (Xie et al., 2012; Ruiz-Orera, Messeguer, et al., 2014). Novel genes tend to be shorter, contain fewer exons, evolve more rapidly, and be lowly expressed compared to protein coding genes; as are lncRNAs. In particular, new genes are commonly found be highly or even specifically expressed in the testes in mammals and flies, regardless of their mechanism of genesis (Levine et al., 2006; Wu, Irwin, and Zhang, 2011). The "out-of-testes" hypothesis has been proposed, which suggests that as the male reproductive organs tend toward rapid evolution due to the range of selective pressures acting upon them, they provide an environment conducive to the emergence of new genes (Kaessmann, 2010). The testes are known to have a chromatin environment conducive to promiscuous transcription, with a high abundance of RNA polymerase II (Schmidt and Schibler, 1995). LncRNAs expression is also enriched in the testes of humans and *Drosophila*, with the second highest levels of enrichment in the brain (Jandura and Krause, 2017) In humans, the transcription of many *de novo* genes have also been found to originate in the brain (Xie et al., 2012), in particular in early brain development (Zhang, Landback, et al., 2011). Interestingly, ~40% of human lncRNAs are specifically expressed in the brain (Derrien et al., 2012), and their misregulation has been linked to diseases including Alzheimer’s disease (Mus, Hof, and Tiedge, 2007; Faghihi, Modarresi, et al., 2008), Parkinson’s disease (Carrieri, Cimatti, et al., 2012)

Importantly, if lncRNAs are indeed a step in the generation of *de novo* coding sequences, they

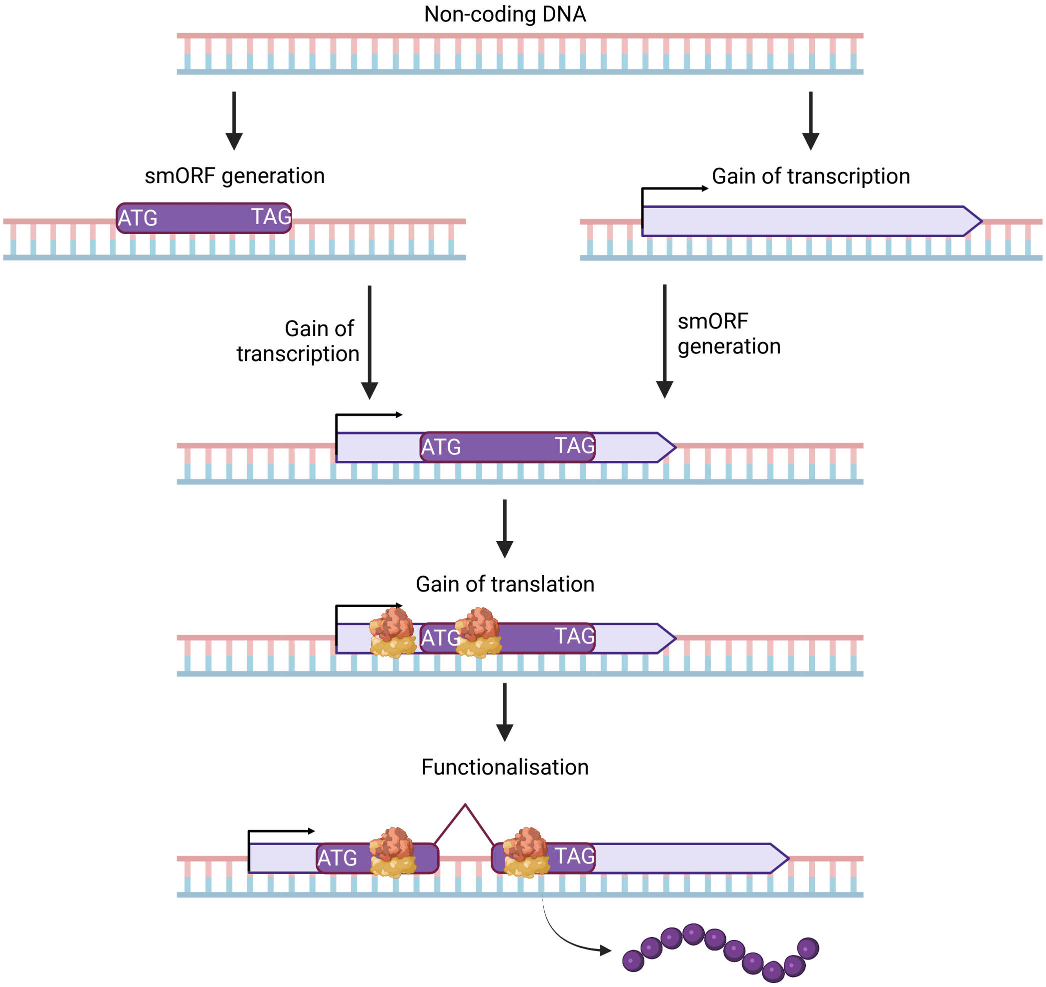


Figure 1.7: **Possible models for the emergence of novel protein coding genes from non-coding sequence.** Non coding sequence may follow the left hand "ORF first" or right hand "transcription first" model, where a random smORF occurs in the genome then gains transcription, or a portion of the genome becomes transcribed and subsequently gains a smORF. This smORF may then become translated, and over time gain function and features associated with protein coding genes. Dark purple boxes represent smORFs, and light purple arrows transcripts. Created with BioRender.com.

are a functional step. As discussed in Section 1.4, lncRNAs carry out many key functions as non-coding transcripts. These functional lncRNAs may or may not contain multiple untranslated smORFs. Ribosome profiling has opened up a world of lncRNAs undergoing active translation, producing both functional and non-functional peptides and representing further points in this continuum. Further, there are examples of transcripts which could represent the intermediate state of functioning as a lncRNA transcript, and producing a functional peptide. *SRA1* is a lncRNA which acts as a molecular scaffold, enhancing the activity of steroid receptors (Lanz et al., 1999). Via alternative splicing, *SRA1* also produces an mRNA that codes for a protein called *SRAP*. *SRAP* is less well studied than its lncRNA counterpart, but has been found to interact with transcription factors, and is thought be involved in transcriptional repression or modulating splicing (Ulveling, Francastel, and Hubé, 2011; Sheng et al., 2018). It is therefore interesting to consider at which point a lncRNA is considered "protein coding" enough to be reclassified, and if a specific subset of the oxymorons which are "translated lncRNAs" should be clearly defined.

1.8 LncRNAs in the human brain

LncRNAs are highly expressed in the human central nervous system, including many brain-specific lncRNAs (Derrien et al., 2012), and their dysregulation has been associated with a wide range of neurological diseases, including Alzheimer's disease (*BACE1-AS*) (Faghihi, Modarresi, et al., 2008), Fragile X syndrome (*FMR4*) (Khalil et al., 2008), and Parkinson's disease (*MALAT1*) (Cai et al., 2020; Zhang, Wang, et al., 2016). A subset of these lncRNAs regulate neuronal differentiation, a tightly controlled process which requires temporally and spatially specific expression profiles. For example, the conserved lincRNA *TUNAR* is transcribed in embryonic stem cells, and depletion of *TUNAR* expression using short hairpin RNA (shRNA) caused a reduction in cell proliferation in mouse (Lin, Chang, et al., 2014). Recent work identified a 48 aa transmembrane protein (pTUNAR) translated from *TUNAR* which impacts neural differentiation and neurite formation, and a longer 65 aa isoform which was more lowly translated (Senís et al., 2021).

1.8.1 SH-SY5Y cells as a model for human neuronal differentiation

Development of the human central nervous system begins in week 3 of embroygenesis, as the neural plate is formed and folds to create the neural tube, from which the brain and spinal cord will eventually be derived. Neural progenitor cells (NPCs) develop and neurogenesis begins as they divide and differentiate to produce neuronal and glial cells, eventually giving rise to the intricately arranged network which forms the human brain (Tao and Zhang, 2016). The study of lncRNA expression and translation during this process is particularly challenging, as techniques such as Poly-Ribo-Seq require a lot of primary tissue. Once neurons are fully matured they also cannot be propagated, so a small sample cannot provide sufficient material. As LncRNAs are generally poorly conserved, study of another model species will not provide in depth insight into those lncRNAs important in human neurons. Therefore, SH-SY5Y cells are often used as a ideal model to study expression and translation in the human brain and in neuronal differentiation.

SH-SY5Y cells are a human neuroblastoma cell line which originates from the SK-N-SH cell line, taken from a metastatic bone tumour biopsy of a 4-year old female neuroblastoma patient (Biedler, Helson, and Spengler, 1973). The SK-N-SH line contains two phenotypes; neuroblast-like cells and epithelial-like cells (Ross, Spengler, and Biedler, 1983), and three successive sub-clones selecting for cells with neuron-like characteristics resulted in the SH-SY5Y line (Biedler, Roffler-Tarlov, et al., 1978). Undifferentiated SH-SY5Y express immature neuronal markers and are phenotypically similar to neuroblasts, the nondividing cells eventually produced from NPCs which differentiate to form neurons. Upon treatment with retinoic acid (RA), SH-SY5Y cells can be differentiated to more mature neuron-like cells, with elongated neurites and decreased proliferation, providing a model for neuronal differentiation (Kovalevich, Santerre, and Langford, 2021).

1.9 Project objectives

1) To establish which lncRNAs are actively translated in human neuronal cells (Chapter 2).

LncRNAs are known to play key roles in human neurogenesis, but work thus far has focused on nuclear lncRNAs, and very little is known about cytoplasmic lncRNAs and their coding potential throughout this process. Here, we aim to detect lncRNAs undergoing active translation during human neuronal differentiation, despite the pervading view that these transcripts do not have coding potential. To achieve this, Poly-Ribo-Seq data from undifferentiated (Control) and differentiated (RA) SH-SY5Y cells was analysed to identify subpopulations of human cytoplasmic neuronal lncRNAs, based on their association with the translational machinery.

2) To identify distinguishing characteristics of translated lncRNAs compared to non-coding and canonical coding sequences (Chapter 3).

The expression and structure of lncRNAs is known to be similar to *de novo* coding sequences, and it has been postulated that lncRNA represent a step on the evolutionary continuum between non-coding and coding sequence. By focusing on translated lncRNAs we aimed to determine where they fall on this spectrum. Here the translated lncRNAs, their smORFs and resulting peptides are compared to canonical and non-canonical ORFs to characterise this sub-population of lncRNAs.

3) To determine the extent of sequence conservation of translated lncRNA smORFs in non-human species (Chapter 4).

Unlike canonical protein coding genes, lncRNAs generally lack sequence conservation, and approximately 30% of human lncRNAs are thought to be primate specific. By focusing on the sequence of the smORFs rather than the entire lncRNA transcript, I aimed to identify modular areas of sequence conservation, and potentially biologically important lncRNA peptides. To do so, the transcriptomes and proteomes of Primates and other reference quality genomes were searched for evidence of sequence conservation in the translated lncRNA smORFs.

Chapter 2

Identification of subpopulations of cytoplasmic lncRNAs expressed in human neuronal cells

2.1 Introduction

2.1.1 Cytoplasmic lncRNA populations

Initial studies of lncRNA function focused on lncRNAs which localise to the nucleus, with many found to have roles as epigenetic regulators of gene expression. Later work widened the field to include the many lncRNAs enriched in the cytoplasm. A study in the leukaemia cell line (K562) detected 54% of expressed lncRNAs in the cytoplasm, while work in HepG2 hepatocellular carcinoma cells found ∼75% present in the cytoplasm (Carlevaro-Fita et al., 2016; Benoit Bouvrette et al., 2018). One possible classification of cytoplasmic lncRNAs is based on their association with the translation machinery, creating three populations; A) Cytosolic lncRNAs, which do not associate with the translational machinery; B) Untranslated polysome-associated lncRNAs found to associate with other, translated mRNAs, i.e., present in polysome fractions, but not translated themselves; C) Translated lncRNAs (Figure 2.1) (Aspden et al., 2014). LncRNAs in category A) can perform a number of functions, such as *lincRNA-RoR*, which acts as an miRNA sponge and is thought to be involved in the maintenance of embryonic stem cell pluripotency (Wang, Xu, et al., 2013). LncRNAs in category B) can affect the translation of mRNAs, such as *BC200/BCYRN1*, which acts as a localised translational repressor in dendrites and is implicated in Alzheimer's disease (Mus, Hof, and Tiedge, 2007). The lncRNA *Uchl1-AS1* is downregulated in models of Parkinson's disease, and dynamically expressed in the cytoplasm where it associates with the protein coding transcript *UCHL1*, increasing translation via a SINEB2 domain (Carrieri, Cimatti, et al., 2012; Carrieri, Forrest, et al., 2015). Recent community efforts have aimed to standardise documentation of category C) lncRNAs, identifying 2,208 translated lncRNA ORFs from human ribosome footprinting data (Mudge et al., 2022).

Previous analysis of Poly-Ribo-Seq data identified these three populations of cytoplasmic lncRNAs in *D.melanogaster* S2 cells (Aspden et al., 2014). As Poly-Ribo-Seq includes a poly(A) selection step, it can be assumed that all lncRNAs identified in these populations have a poly(A) tail. As well as contributing to transcript stability, the poly(A) tail aids in the export of RNA to the cytoplasm (Fuke and Ohno, 2008), and it's length can affect translational efficiency (Subtelny et al., 2014). This does exclude a subset of lncRNAs with alternative

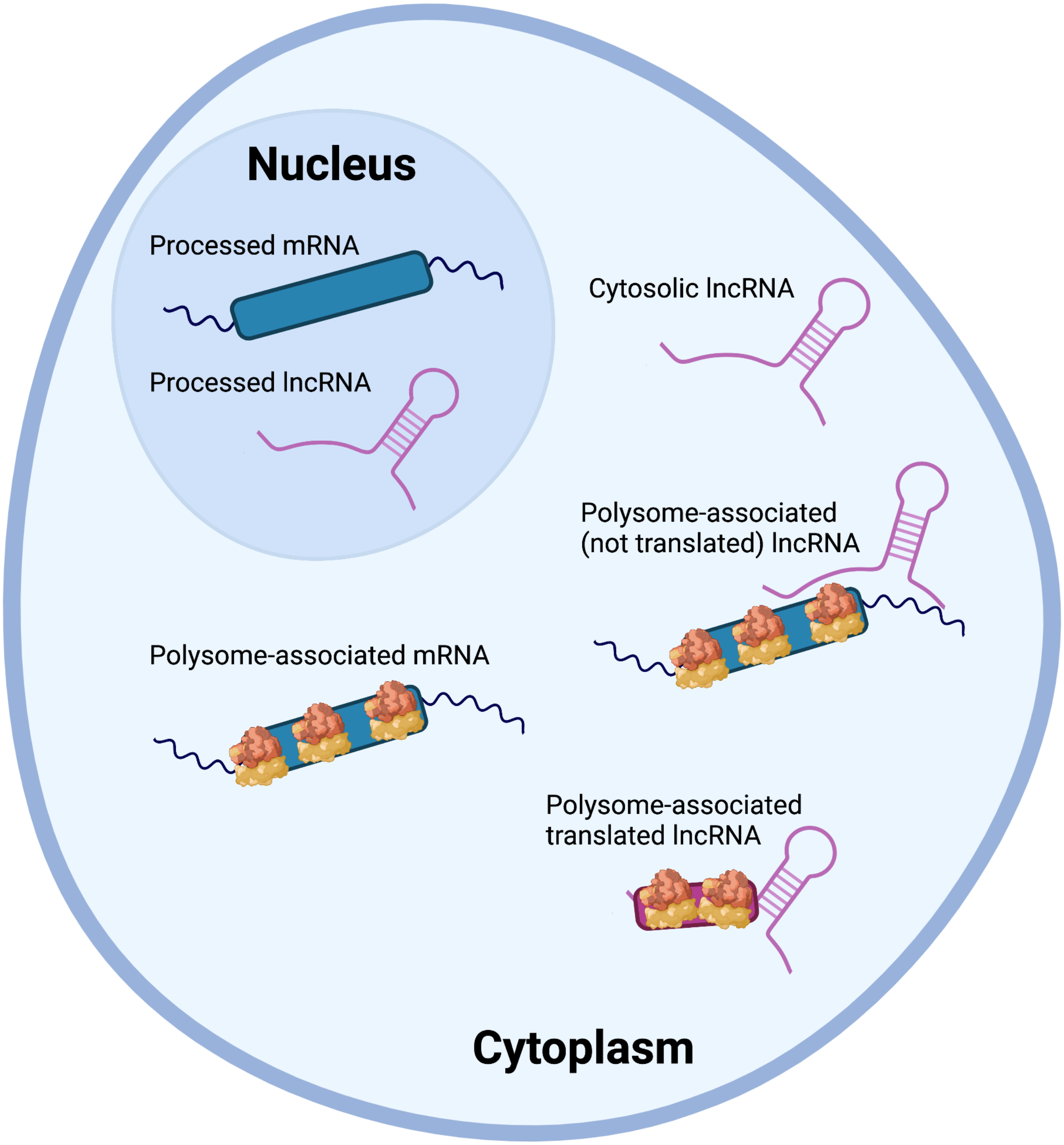


Figure 2.1: **Subpopulations of cytoplasmic lncRNAs.** Based on their interactions with the translational machinery there are four subpopulations of cytoplasmic RNAs: i) cytosolic lncRNAs, which do not associate with the translational machinery, ii) polysome-associated (not translated) lncRNAs, iii) polysome-associated translated lncRNAs and iv) polysome-associated translated mRNAs. Processed mRNA and lncRNA are shown in the nucleus. Created with BioRender.com.

stabilisation structures such as triple helical structures, which can also undergo nuclear export (Wilusz et al., 2012). The aim of this chapter is to identify these three populations of cytoplasmic lncRNAs in human neuronal cells using Poly-Ribo-Seq data, to understand the interaction of lncRNAs with the translation machinery and identify actively translated lncRNA smORFs.

2.1.2 Quantifying RNA expression from RNA-Seq

To effectively quantify RNA levels from RNA-Seq, a suitable normalisation method must be selected to adjust the raw count values. Most methods account for sequencing depth, to allow comparison between samples, and gene length, to allow comparison between different genes within the same sample. RNA composition may also be considered, as a few genes with extremely divergent expression levels will skew differential expression analyses. Reads per kilobase of exon per million reads mapped (RPKM) and fragments per kilobase of exon per million reads mapped (FPKM) are extensively used metrics used for comparisons within samples (Table 2.1), but they are not suitable for comparisons between replicates. This is because RPKM and FPKM normalise by the total number of sequencing reads, which is not a true measure of the number of transcripts in the sample. For example, if one replicate contains longer transcripts on average, the same number of reads can represent fewer transcripts (Wagner, Kin, and Lynch, 2012). Transcripts per million (TPM) is suitable metric for comparisons between replicates of the same sample type because the sum of all TPMs in a sample is constant, meaning they represent relative expression levels. For between sample comparisons, methods include trimmed mean of M-values (TMM) and the median of ratios from DESeq2, as they account for variance in the composition of RNA populations.

Table 2.1: **RNA-Seq count normalisation methods.** Various methods used to normalise raw RNA-Seq read counts, a brief description the factors these methods normalise by, and examples of software which use the method.

Measure	**Description**	**Example Software**	**References**
CPM (Counts Per Million)	Transcript counts scaled by total number of reads.		
RPKM (Reads Per Kilobase of exon per Million reads mapped)	Transcript counts are normalised by RNA length and total number of reads. Used in single-end RNA-Seq.	Sailfish	Mortazavi et al., 2008 Patro, Mount, and Kingsford, 2014
FPKM (Fragments Per Kilobase of exon per Million fragments mapped)	Fragments normalised by RNA length and total number of reads. Used in paired-end RNA-Seq.	Cufflinks, eXpress	Trapnell et al., 2012 Roberts and Pachter, 2013
TPM (Transcripts Per Million)	Estimate of relative transcripts abundance, normalised by RNA length.	Kallisto, Sailfish, Salmon	Li, Ruotti, et al., 2010 Wagner, Kin, and Lynch, 2012 Patro, Mount, and Kingsford, 2014
TMM (Trimmed Mean of M values)	Estimates scale factors between samples.	edgeR	Robinson and Oshlack, 2010 Robinson, McCarthy, and Smyth, 2010
Median of Ratios	Accounts for sequencing depth and RNA composition.	DESeq2	Anders and Huber, 2010 Love, Huber, and Anders, 2014

2.2 Materials and Methods

2.2.1 RNA-Seq and Ribo-Seq Data

2.2.1.1 Poly-Ribo-Seq of human neuroblastoma cell line (SH-SY5Y)

Poly-Ribo-Seq was carried out on Human Neuroblastoma SH-SY5Y cells by Dr Katerina Douka within the Aspden group, with single end 50bp Illumina sequencing. Detailed methods are outlined in Douka, Birds, et al., 2021. Cycloheximide treatment was used to immobilise ribosomes, and ribosome footprinting was performed using RNaseI, with conditions specifically optimised for these samples, including gradient loading and treatment temperatures, buffer composition and the duration of the treatment (Douka, Agapiou, et al., 2022). Poly-A selection was used to enrich the cytoplasmic and polysome-associated RNA samples for poly-A tailed mRNAs and lncRNAs, and reduce contamination from rRNA, the most abundant type of RNA in the cell. This was performed on undifferentiated (Control) cells and RA-treated differentiated (RA) cells, with three biological replicates for each condition. This produced the following sample types: A) Total cytoplasmic polyadenylated RNA-Seq, B) Polysome-associated polyadenylated RNA-Seq, and C) Ribo-Seq (Figure 1.5).

2.2.1.2 Data availability

All analysed Poly-Ribo-Seq datasets are deposited in GEO with ID GSE166214. Fasta files of human rRNA and high confidence hg38 tRNA were obtained from RiboGalaxy (Michel et al., 2016) and GtRNAdb Release 17 (Quek et al., 2015) respectively. The human reference genome and transcriptome (GRCh38.p12) and accompanying annotations were retrieved from GENCODE release 30 (Frankish et al., 2021).

2.2.2 Poly-Ribo-Seq analysis

2.2.2.1 Read trimming and alignment

Quality reports of RNA-Seq and Ribo-Seq data were made using Fastqc (v.0.11.9) (Andrews, 2010). Adapter sequences were trimmed using Cutadapt (v2.10) (Martin, 2011), with a minimum read length of 25bp. Untrimmed outputs were retained for RNA-seq reads (Figure 2.2). Low-quality reads with a score of <20 for 10% or more of the read were discarded using the FASTQ Quality Filter within the FASTX-Toolkit (v0.0.14) (Hannon, 2009). One was removed from the 3' end of the trimmed reads in order to improve alignment quality, and reads originating from rRNA and tRNA were aligned and removed using Bowtie2 (v2.3.4.3) (Langmead and Salzberg, 2012) (E-appendix; bowtie_index.sh, preprocessing_RNAseq.sh and preprocessing_Riboseq.sh). Remaining reads were mapped to the human reference genome using STAR (v2.7.5c) (Dobin et al., 2013). The STAR genome index was built with an sjdbOverhang of 73 (E-appendix; STAR_index.sh). Samtools (v1.10) (Li, Handsaker, et al., 2009) was used to create sorted, indexed bam files of the resulting alignments (E-appendix; STAR_alignment.sh).

2.2.2.2 Transcript quantification (RNA-Seq)

Trimmed RNA-Seq reads were mapped to the human reference transcriptome using Salmon v1.6.0 (Patro, Duggal, et al., 2017) (E-appendix; Salmon_quant.sh). Salmon calculates Transcripts per Million (TPM), an estimate of the relative abundance of each transcript in a sample, which normalises for transcript length and sequencing depth. TPM gives the proportion of reads in a sample that map to a given transcript, a more accurate way to compare samples than other measures. A gentrome and decoy file were created from the human transcriptome and genome (GENCODE release 30), using the generateDecoyTranscriptome.sh script from Salmon Tools (Patro, Srivastava, and Sarkar, 2022) (E-appendix; Decoy_transcriptome.sh).

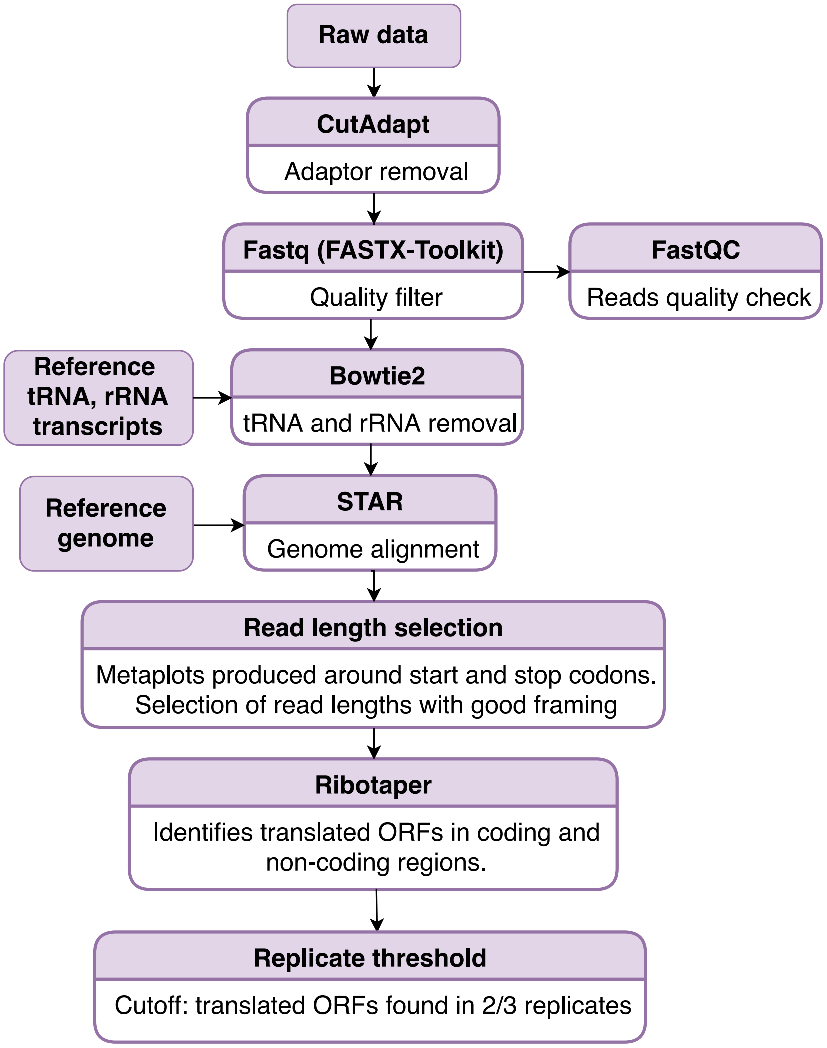

Raw data
CutAdapt
Adaptor removal
Fastq (FASTX-Toolkit)
Quality filter
FastQC
Reads quality check
Reference tRNA, rRNA transcripts
Bowtie2
tRNA and rRNA removal
Reference genome
STAR
Genome alignment
Read length selection
Metaplots produced around start and stop codons. Selection of read lengths with good framing
Ribotaper
Identifies translated ORFs in coding and non-coding regions.
Replicate threshold
Cutoff: translated ORFs found in 2/3 replicates

Figure 2.2: **Workflow for identification of translated ORFs from Ribo-Seq and RNA-Seq using Ribotaper.** 3' Adaptor sequences were trimmed from Polysome-associated RNA-seq and Ribo-seq data using Cutadapt v2.10 (Martin, 2011) with a minimum read length of 25bp, and untrimmed outputs retained for RNA-seq data. Low-quality reads with a score of >20 for 10% or more of the read were discarded using the Fastq quality filter in FASTX-Toolkit v0.0.14 (Hannon, 2009). FastQC v0.11.9 (Andrews, 2010) was used to produce quality reports on the data. One base was removed from the 3' ends of reads to improve alignment quality, and reads originating from rRNA and tRNA were aligned and removed using Bowtie2 v2.3.4.3 (Langmead and Salzberg, 2012). The splice aware aligner STAR v2.7.5c (Dobin et al., 2013) was used to align reads to the human genome. The STAR genome index was built with a sjdbOverhang of 73. Samtools v1.10 (Li, Handsaker, et al., 2009) was used to convert sam files to sorted, indexed bam files. Metaplots of aligned Ribo-seq data were generated using the metaplots.bash script from Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016), and used to select Ribo-seq read lengths exhibiting the best triplet periodicity, along with appropriate offsets. Ribotaper was used to identify actively translated smORFs. The final filter that was applied was the requirement for smORFs to be detected as translated in at least two of the three biological replicates for the condition (Control or RA).

This step prevents the mapping of reads which originate from an unannotated genomic locus to annotated transcripts with similar sequences. A salmon index was generated from this gentrome using a k-mer (minimum acceptable length for a valid match) of 27 (E-appendix; Salmon_index.sh). Tapestation readouts were used to estimate the mean and standard deviations of fragment lengths in the sequencing library (Appendix A Table A.1). Expression levels of the RNA-seq were quantified using Salmon (v.1.6.0) (Patro, Duggal, et al., 2017), with a threshold of $\geq$1 TPM to consider a transcript expressed, as in the Human Protein Atlas (Uhlén et al., 2015). Transcripts were considered lncRNAs if they were from the following GENCODE release 30 (Frankish et al., 2021) biotypes: lincRNA, 3prime overlapping ncrna, antisense, sense intronic, sense overlapping, macro lncRNA, bidirectional promoter lncrna.

Salmon outputs were read into edgeR v3.28.1 (Robinson, McCarthy, and Smyth, 2010) and normalised to allow comparison between Total and Polysome-associated RNA-Seq samples. Transcripts with very low counts across all samples were filtered, and counts were normalised for sequencing depth and effective library sizes (E-appendix; edgeR_norm.Rmd).

2.2.3 Identification of translated lncRNAs

2.2.3.1 Ribotaper

Metaplots of the aligned Ribo-Seq data were generated using the create_metaplots.sh bash script from Ribotaper (v1.3) (Calviello, Mukherjee, et al., 2016). These show the distance between the 5' ends of Ribo-Seq and the annotated start and stop codons from CCDS ORFs, allowing the locations of P-sites to be inferred. The equivalent "RNA-site" used by Ribotaper is at the 26th nt, arbitrarily chosen as a consistent position for each RNA-Seq read along the transcript. Read lengths exhibiting the best triplet periodicity were selected for each replicate, along with appropriate offsets (Appendix A, Table A.2). Actively translated smORFs were identified from the Ribo-Seq and Polysome-associated RNA-Seq, using Ribotaper (v1.3) (Calviello, Mukherjee, et al., 2016) (E-appendix; Ribotaper.sh). This method was also tested on the Total RNA-Seq. Ribotaper requires a given exon to contain more than 5 P-sites, exhibit significant 3-nt periodicity, and have 50% or more of the P-sites in frame with a start codon. If multiple in frame start codons are present, the most upstream start codon with a minimum of five P-sites in between it and the next ATG is selected. ORFs for which >30% of the Ribo-Seq coverage was only supported by multi-mapping reads were subsequently removed. For this portion of the analysis, I also required that a smORF be identified in at least two of the three biological replicates of a given condition (Control or RA) to consider it robustly translated.

ORFs are categorised by Ribotaper (v1.3) (Calviello, Mukherjee, et al., 2016) according to the gene biotype they originate from in the genome annotation; in this case, Gencode release 30 (Frankish et al., 2021). Consensus coding DNA sequence ORFs (CCDS ORFs) are found overlapping an annotated coding exon in a CCDS gene, whilst non-consensus coding DNA sequence coding ORFs (non CCDS coding ORFs) overlap an annotated coding exon in a non-CCDS protein coding gene. Upstream ORFs (uORFs) and downstream ORFs (dORFs), are defined as an ORF upstream or downstream of an annotated start codon or stop codon in a CCDS gene. non-coding ORFs (ncORFs) are any ORF in a non-CCDS gene, which are not overlapping a coding exon. All ORFs identified in other gene biotypes were classified as ncORFs.

2.2.4 General statistics and plots

Data were analysed and plotted in R (R Core Team, 2021), using packages including the tidyverse (Wickham et al., 2019), ggplot2 (Wickham, 2016), gghighlight (Yutani, 2022), VennDiagram (Chen, 2022), and viridis (Garnier et al., 2021).

ORFs identified using Total and Polysome-associated RNA-Seq were compared using Jaccard

Similarity indices, calculated as the number of obervations in the intersection of two sets, divided by the union of the sets.

2.3 Results

2.3.1 Poly-Ribo-Seq data quality and alignment

To identify subpopulations of human neuronal cytoplasmic lncRNAs based on their association with the translation machinery, high quality RNA-Seq and Ribo-Seq reads were selected and aligned to the human genome. Quality checks were performed on the data, then adapter sequences, low quality reads, and contaminants were removed, and reads were aligned to the human genome (Figure 2.3). Using Fastqc v0.11.9 (Andrews, 2010) to produce reports on the raw and processed data, it was confirmed that all low quality sequences and adapter sequences had been removed, and that the sequence length distribution was as expected. A higher proportion of reads were removed by Cutadapt v2.10 (Martin, 2011) from the Ribo-Seq samples than the RNA-Seq samples due to the length of the reads (Figure 2.3). Ribosome footprints of 28-34 nt and mRNA fragments of 50-80 nt were selected and gel purified for the Ribo-Seq. This meant that when the reads were subjected to 75 bp single end RNA sequencing, all of the raw Ribo-Seq sequences contained adapter sequence. When this was trimmed by Cutadapt, some of the Ribo-Seq sequences were shorter than my minimum read length of 25 bp. There was also a higher proportion of tRNA reads removed from the Ribo-Seq samples, as these reads represent the footprints of actively translating ribosomes and are therefore likely to have tRNA associated. The reads which passed these processing steps were of a high quality, and 77-99% of the usable reads were aligned to the human genome using STAR v2.7.5c (Dobin et al., 2013) (Figure 2.4, Appendix A Table A.4).

2.3.2 2,305 cytoplasmic lncRNAs are detected in SH-SY5Y cells

To establish which lncRNA transcripts were present in the cytoplasm in SH-SY5Y cells, expression levels were calculated from Total and Polysome-associated RNA-Seq using Salmon v1.6.0 (Patro, Duggal, et al., 2017). Salmon is a mapping software which finds the likely strand and position a query sequence originated from, as opposed to creating a base by base alignment. The recommended k-mer (minimum acceptable length for a value match) for reads of 75bp or longer is 31, but a smaller value may improve sensitivity. As the RNA-Seq read lengths range between 25 - 76 bp, with most between 46 - 72 bp (Figure 2.5), a range of smaller k-mers were tested (Table 2.2). The selected k of 27 gave a high mapping rate, and a high level of consensus between the three biological replicates in each condition and sample type, without being too short and increasing the risk of mapping to the wrong transcript.

A total of 42,593 transcripts were detected in one or more replicates or conditions, 2,305 of which were lncRNAs; 8.2% of the lncRNAs included in the Gencode v30 transcriptome (Frankish et al., 2021). There was a large amount of variation in lncRNA levels between replicates (Figure 2.6), in particular replicate 2 has a small overlap with the other replicates, likely due to variation in the concentrations of the cDNA libraries that produced these data. The concentrations of the Polysome and Total RNA-Seq libraries in Replicate 2 were all $<.8$ ng/μL, so these samples were less deeply sequenced than in other replicates. The concentrations of the Control Polysome and Total RNA-Seq libraries were highest in Replicate 1, and the RA Polysome and Total RNA-Seq libraries were most concentrated in Replicate 3, explaining why these replicates have the highest number of expressed lncRNAs unique to them when compared to the other replicates in the same condition and sample type.

In order to identify cytoplasmic lncRNAs that are not associated with the translational machinery, comparisons will need to be made between those present in the Total and Polysome-associated RNA-seq. However, Salmon quantifies the relative abundance of transcripts in TPM, allowing the proportion of reads to be compared between replicates of the same sample type. Therefore, further normalisation is needed before transcript abundance may be compared between Polysome and Total samples or Control and RA conditions.

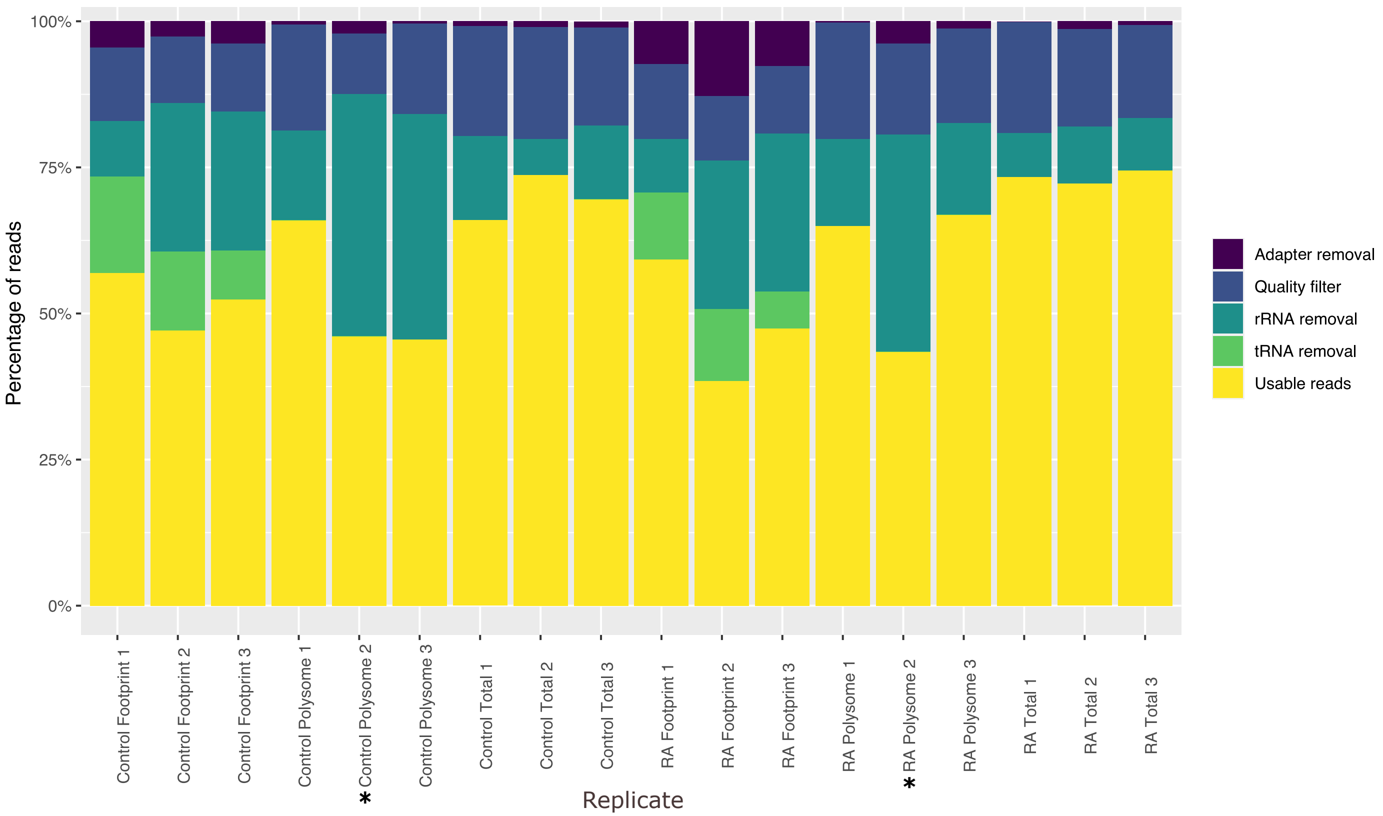


Figure 2.3: **Percentage of usable Ribo-Seq and RNA-Seq reads remaining from pre-processing steps.** The percentage of reads removed at each preprocessing step or remaining for alignment is given on the y-axis. The x-axis shows each RNA and Ribo-Seq sample. The starred replicates are Control Polysome 2, which was sequenced particularly deeply so has a higher number of raw reads than other polysome samples, and RA Polysome 2, which had the lowest percentage of usable reads. These are also highlighted in Appendix A, Table A.3. Adapter sequences were trimmed from raw reads using Cutadapt v2.10 (Martin, 2011). Low quality reads were removed using the Fastq quality filter in FASTX-Toolket v0.0.14 (Hannon, 2009). rRNA and tRNA contaminants were aligned and removed using Bowtie2 v2.3.4.3 (Langmead and Salzberg, 2012).

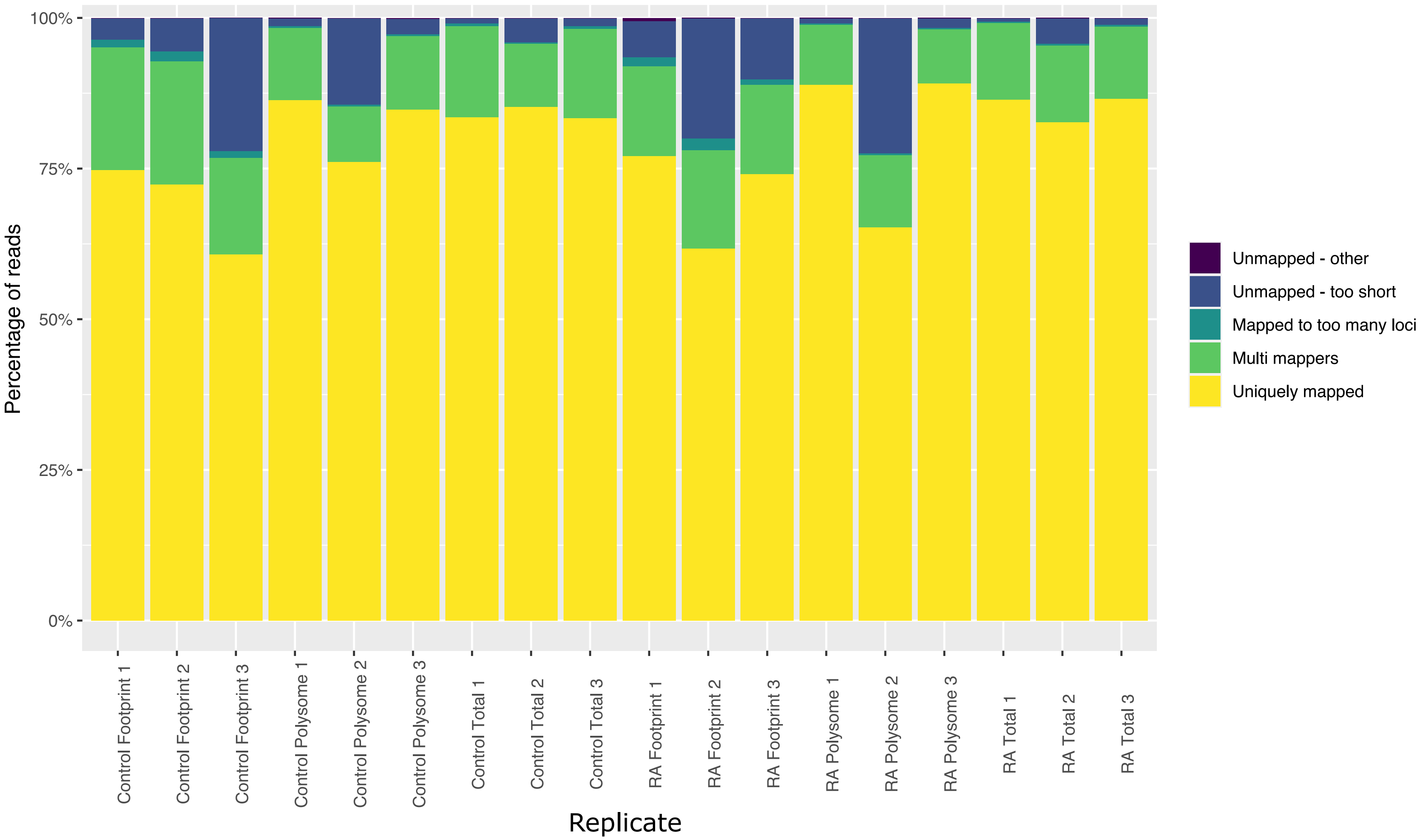


Figure 2.4: **Percentage of reads aligned to the human genome in each replicate and sample.** The y-axis shows percentage of reads in each alignment category. Uniquely mapped and multi mapped reads were used for translation analysis. The x-axis shows each RNA and Ribo-Seq sample.

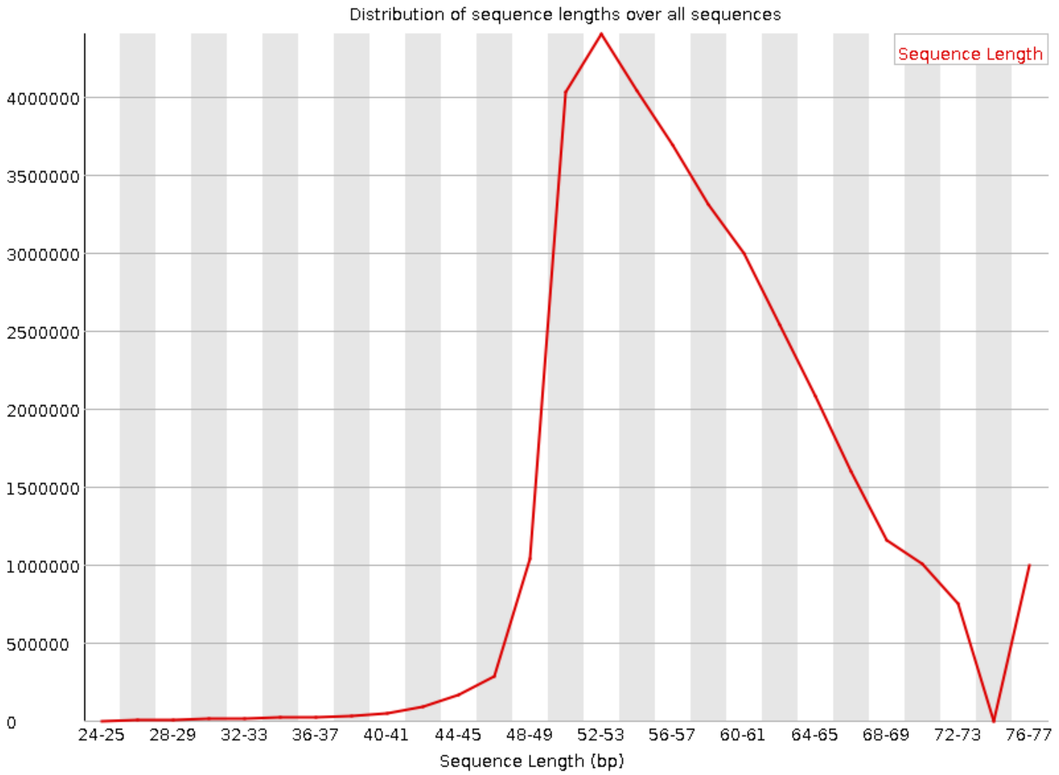


Figure 2.5: **RA Total RNA-Seq Replicate 1 sequence length distribution, post processing.** Example sequence length distribution plot produced by Fastqc v 0.11.9 (Andrews, 2010). The x-axis shows the sequence length in base pairs, and the y-axis shows the number of sequences.

Table 2.2: **Mapping rates of RNA-Seq to the human transcriptome using Salmon v1.6.0 (Patro, Duggal, et al., 2017)**. Rates for all Total and Polysome-associated RNA-Seq, in Control and RA conditions, for replicates 1-3. Indexes were generated using k-mers ranging from k=19 to k=31.

Replicate 1	**Mapping rate**			
Sample	**k-mer 19**	**k-mer 23**	**k-mer 27**	**k-mer 31**
Control Total	89.68%	89.44%	88.50%	86.39%
Control Polysome	90.06%	89.81%	89.00%	86.99%
RA Total	91.73%	91.59%	90.99%	89.33%
RA Polysome	90.81%	90.68%	90.17%	88.60%
Average mapping rate	**90.86%**	**90.69%**	**90.05%**	**88.31%**

Replicate 2	**Mapping rate**			
Sample	**k-mer 19**	**k-mer 23**	**k-mer 27**	**k-mer 31**
Control Total	81.64%	81.63%	81.48%	81.07%
Control Polysome	75.12%	75.09%	74.88%	74.37%
RA Total	81.92%	81.87%	81.58%	80.65%
RA Polysome	66.84%	66.76%	66.43%	65.50%
Average mapping rate	**74.63%**	**74.57%**	**74.29%**	**73.51%**

Replicate 3	**Mapping rate**			
Sample	**k-mer 19**	**k-mer 23**	**k-mer 27**	**k-mer 31**
Control Total	86.54%	86.40%	85.80%	84.27%
Control Polysome	87.34%	87.16%	86.50%	84.92%
RA Total	89.35%	89.20%	88.64%	87.16%
RA Polysome	87.48%	87.34%	86.77%	85.29%
Average mapping rate	**88.06%**	**87.90%**	**87.30%**	**85.79%**

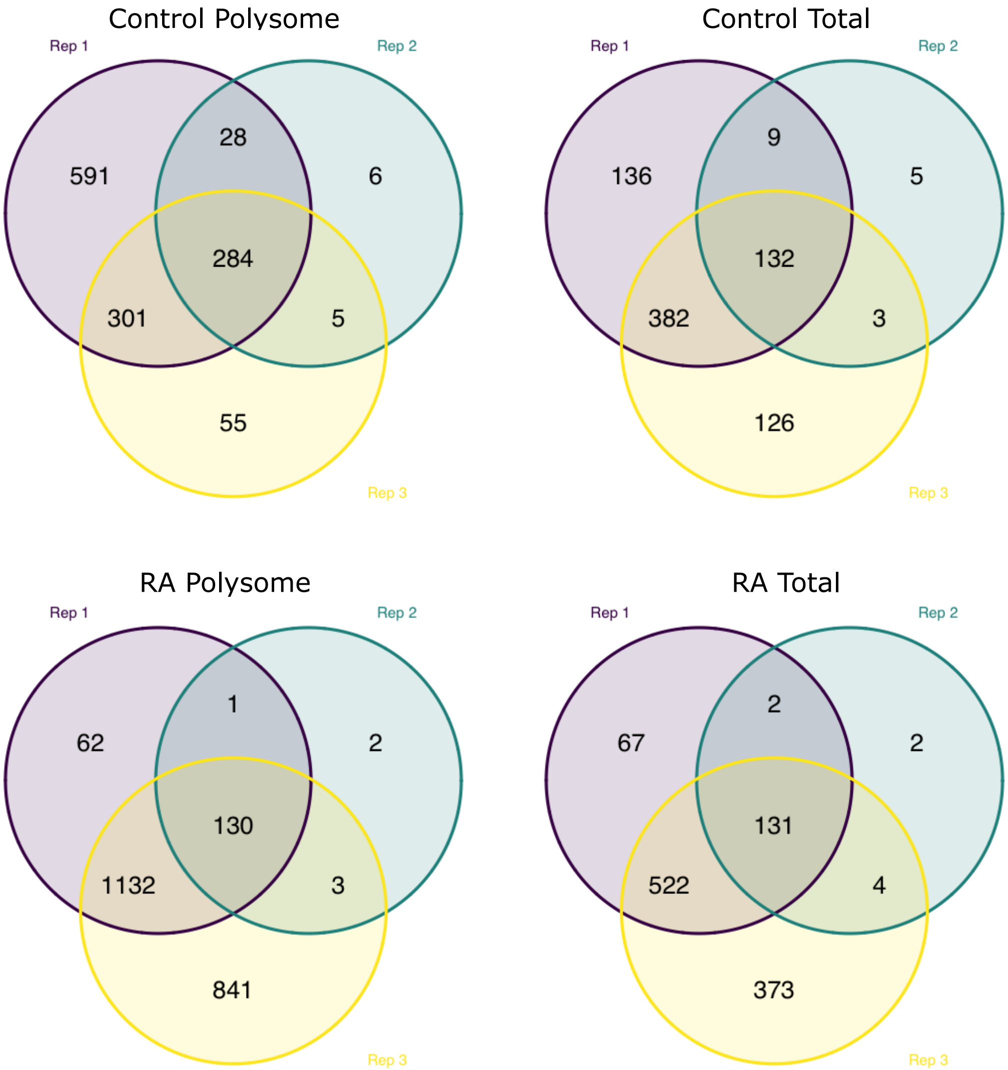


Figure 2.6: **Number of lncRNA transcripts detected in each replicate, in the Total and Polysome-associated RNA-Seq.** Each venn diagram shows the overlap between the three replicates for each condition and RNA-Seq type. A threshold of $\geq$1 TPM was used to detect presence of a transcript.

2.3.3 LncRNAs are translated in SH-SY5Y cells

2.3.3.1 Ribo-Seq data exhibits high levels of triplet periodicity

To identify ORFs undergoing active translation, the highest quality read lengths were selected for analysis from the Ribo-Seq datasets (Appendix A, Table A.2) using metaplots generated by Ribotaper (Calviello, Mukherjee, et al., 2016). Using high quality footprints is key to ensure that analysis is based on the footprints of actively translating ribosomes, and not scanning ribosomes, or other proteins bound to the RNA. High quality read lengths were selected based on their triplet periodicity, with a clear bias toward a single reading frame (Figure 2.7A), and if they exhibited a peak before the start codon which allows the position of the P-site to be inferred. As cycloheximide was used to inhibit translation, a build up of Ribo-Seq reads around the start codon and over the first 15 nt of the ORF in human was also expected (Douka, Agapiou, et al., 2022), and a length of approximately 28 - 32 nt (Aspden et al., 2014; Ingolia, Brar, et al., 2013). Read lengths which exhibited a bias toward multiple reading frames (Figure 2.7B) were discounted.

2.3.3.2 Comparison of detecting translation using Total or Polysome-associated RNA-Seq

To establish whether using Polysome-associated or Total RNA-Seq affected the results of translational analysis, runs of the Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016) pipeline were performed using both of these datasets. In translational analysis, RNA-Seq is used to establish which transcripts are present in a sample, then Ribo-Seq footprints which aligned to these transcripts are analysed to identify actively translated ORFs. Although we would expect broadly the same ORFs to be detected by both analyses, using the Polysome-associated RNA-Seq is likely to be more accurate, as in cases of multiple transcript isoforms harbouring the same ORF, Ribotaper selects the transcript with most RNA-Seq reads. If a transcript is highly expressed, but lowly translated, this isoform may be selected by Ribotaper instead of a more highly translated isoform if all transcripts present in the cytoplasm are included as in the Total RNA-Seq.

A total of 19,214 ORFs were detected as translated in 2/3 replicates in Control conditions, RA conditions, or in both conditions using the Polysome-associated RNA-Seq (Table 2.3). The Total RNA-Seq produced very similar results, and comparison of Total and Polysome ORFs gave Jaccard similarity indices of 93.4% for ORFs in Control conditions, 92.9% for RA conditions, and 90.9% for ORFs translated in both. The Polysome-associated RNA-Seq was therefore used in the translation analysis.

2.3.3.3 Poly-Ribo-Seq detects translational regulation during neuronal differentiation

To identify high level changes in translation between Control and RA conditions, the number of translated ORFs detected in each condition were summarised (Table 2.3). Translation was reduced following differentiation by RA treatment, with a 22% reduction in the number of ORFs detected. This was expected as previous work in the Aspden lab (Douka, Birds, et al., 2021) showed a reduction in the level of polysomes, and an increase in monosomes upon differentiation.

2.3.3.4 45 lncRNAs smORFs are translated in SH-SY5Y cells

To gain an understanding of the non-canonical ORFs translated in human neuronal cells, in particular lncRNA smORFs, ORFs identified in each condition were summarised (Table 2.4). The majority of translated ORFs in all conditions were canonical CCDS or nonCDDS ORFs in protein coding genes (Table 2.4). 71 uORFs and 5 dORFs were also identified in CCDS genes, and 80 other ncORFs were identified in other gene biotypes, including ORFs which do not overlap any coding exons in protein coding, non-CDDS genes. Across Control and RA

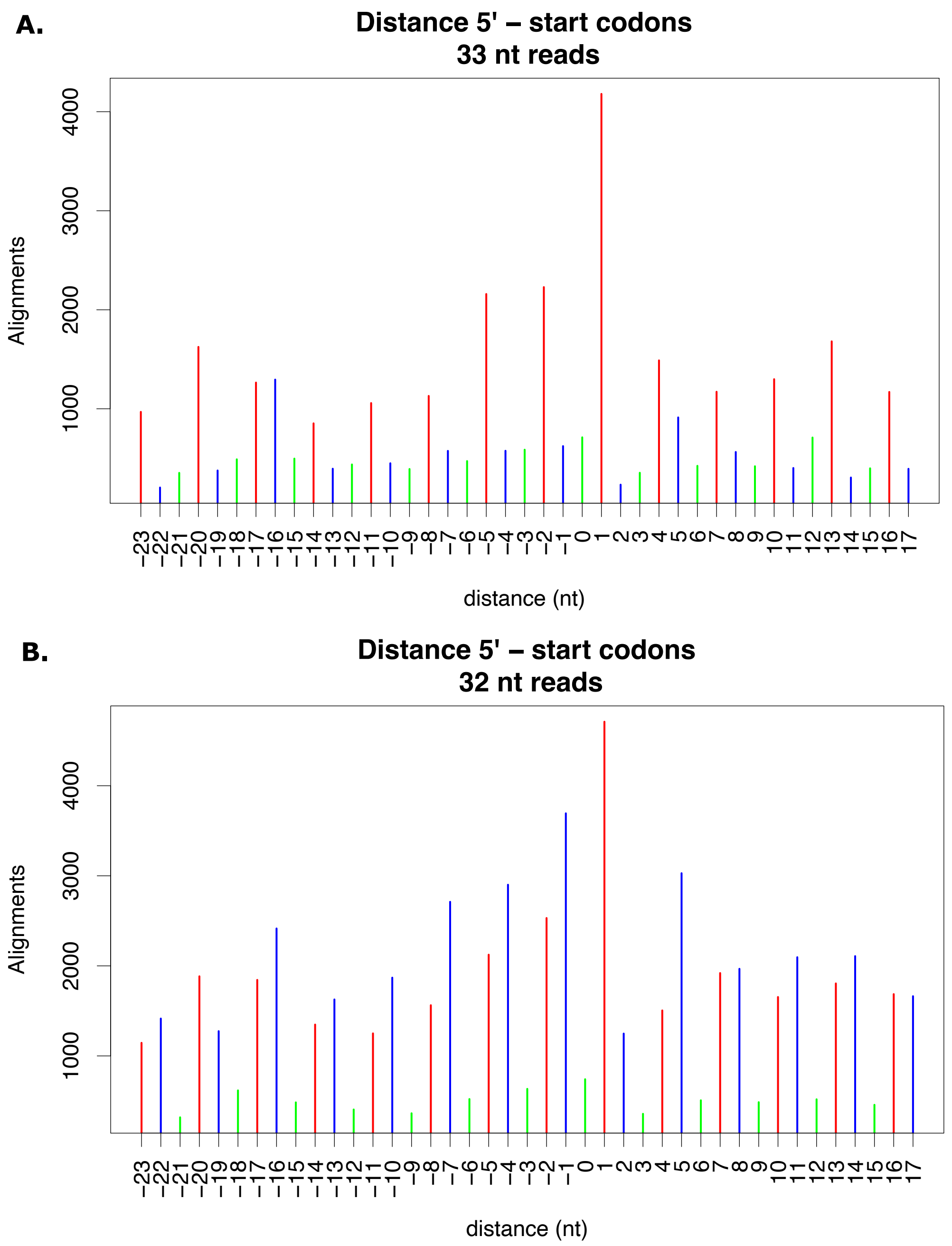


Figure 2.7: **Example metaplots of Ribo-Seq footprints.** Metagene plots of **A)** 33 nt and **B)** 32 nt Ribo-Seq reads from Control conditions, replicate 2. **A)** An example of good quality reads exhibiting framing bias to a single reading frame. **B)** An example of poor quality reads, without a clear bias to a single reading frame. The Ribo-Seq alignments have been down sampled to 10%, and aggregated. The x-axis shows the distance of the 5' end of the reads from annotated start codons in nt, where 0 is the position of the first nucleotide of the start codon. The y-axis shows the number of Ribo-Seq reads starting at that position. Bars are colour coded red, blue and green to indicate the three possible frames.

Table 2.3: **Actively translated ORFs identified in SH-SY5Y cells.** Gene Type denotes the biotype of gene the ORFs originate from, in Control conditions only, RA conditions only, and those found in both conditions (Overlap). LncRNA categories are highlighted in pink. ORFs were identified using Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016), from Polysome-associated RNA-Seq and Ribo-Seq. ORFs are included which were found to be translated in at least 2/3 replicates in a given condition, and any ORFs for which 30% of the Ribo-Seq coverage was supported by multimapping reads only were filtered out.

Gene Type	**Control Only**	**RA Only**	**Overlap**
Protein coding	6,318	2,747	10,026
Pseudogene	34	9	9
Antisense	9	7	5
lincRNA	10	9	1
Processed transcript	12	7	6
Sense intronic	0	1	0
Bidirectional promoter lncRNA	3	0	0
Total ORFs	6,387	2,780	10,047

conditions 45 lncRNA smORFs were identified (Figure 2.8), originating from 45 transcripts, from 25 genes. LncRNA smORFs were identified in the following gene biotypes; antisense, lincRNA, sense intronic lncRNA, and bidirectional promoter lncRNA.

2.3.3.5 Translated lncRNAs are lowly abundant in the cytoplasm

To investigate the levels of the translated lncRNA transcripts in the cytoplasm, their abundance in all replicates was quantified using the Total RNA-seq. Of the 45 lncRNA transcripts, 24 had a TPM $\geq$1 in one or more replicates, and for 22 of the transcripts this corresponded to one or more of the replicates they were translated in. This low level of expression means it is difficult to draw conclusions regarding the specificity of these translated lncRNAs expression, as they may be expressed in a number of tissues at levels we are unable to quantify without particularly deep sequencing.

Two of the translated lncRNA transcripts had a TPM of 0 in Control and RA conditions. However, as the transcript levels were quantified using Salmon (Fuke and Ohno, 2008), and the translational analysis was based on STAR (Dobin et al., 2013) alignments, some variation should be expected. This is because the methods use different mapping algorithms, meaning reads may be assigned to different isoforms or even different transcripts entirely. Further, the translational analysis uses the Polysome-associated RNA-Seq which includes a smaller population of transcripts, meaning the relative abundace of the translated lncRNAs will be greater.

2.3.4 Identifying the polysome-associated lncRNA subpopulation

Having identified which polyadenylated lncRNAs are present in the cytoplasm of SH-SY5Y cells, and which of these lncRNAs are actively translated, the next aim was to identify the subpopulation of untranslated, polysome-associated lncRNAs (Figure 2.1). This has previously been performed in *Drosophila melanogaster* (Aspden et al., 2014) by the straightforward comparison of Total cytoplasmic RNA-Seq and Polysome-associated RNA-Seq. Any lncRNAs found in the Polysome-associated sample and not in the Total sample were deemed Polysome-associated, and could be further split into untranslated and translated populations.

The TMM method as implemented in the edgeR package v3.28.1 (Robinson and Oshlack, 2010) was used allow comparisons between Total and Polysome-associated RNA-Seq results. TMM estimates scale factors between samples by assuming that the majority of genes (more than half) are not differentially expressed, therefore accounting for the fact that one or more genes may vary significantly between samples and affect the proportion of reads attributed to other genes. Because of this assumption, TMM normalised counts are not suitable to compare differences in RNA levels between Control and RA conditions, as we expect substantial changes in gene expression during neuronal differentiation. TMM normalised counts are expressed as CPM, which does not account for transcript length as TPM does. However, as the aim was to compare the expression level of the same transcripts between Total and Polysome-associated RNA-Seq, the transcript lengths were invariant. After genes with very low counts across all libraries were removed and the counts normalised, 41,021 transcripts remained, 3,710 of which were lncRNAs. This normalisation accounted for a lot of the variation previously observed between replicates, increasing the number of lncRNAs present in all three replicates in each condition and sample type (Figure 2.6, Figure 2.9). However, this normalisation did not allow for a straightforward comparison of Total and Polysome-associated RNA-Seq, as very similar numbers of transcripts were detected in the two samples in each condition (Figure 2.10).

This result may be because more than half transcripts are actually "differentially expressed" between Total and Polysome-associated samples, violating the assumptions of the TMM method. As the polysome-associated transcripts in the cell are a subset of all cytoplasmic

Table 2.4: **ORF types identified in SH-SY5Y cells.** The number of ORFs translated in Control conditions only, RA conditions only, both conditions (overlap), and the total number of ORFs. Type annotations are according to Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016). Coding ORFs consists of CCDS ORFs, which overlap known coding sequence regions in CCDS genes, and nonCCDS coding ORFs, which overlap an annotated coding exon in a non CCDS protein coding gene. uORFs and dORFs are ORFs in CCDS genes which do not overlap with any CDS exon, and are upstream and downstream of the annotated ORF respectively. ncORFs are ORFs are ORFs in non-CCDS genes, not overlapping any coding exon. An extra lncRNA smORF category was added to separate ORFs originating from the antisense, lincRNA, sense intronic lncRNA, or bidirectional promoter lncRNA gene biotypes.

Translated ORF Type	Control only	RA only	Overlap	Total
Coding ORF	6,268	2,731	10,014	19,013
uORF	44	15	12	71
dORF	4	1	0	5
ncORF	49	16	15	80
lncRNA smORF	22	17	6	45

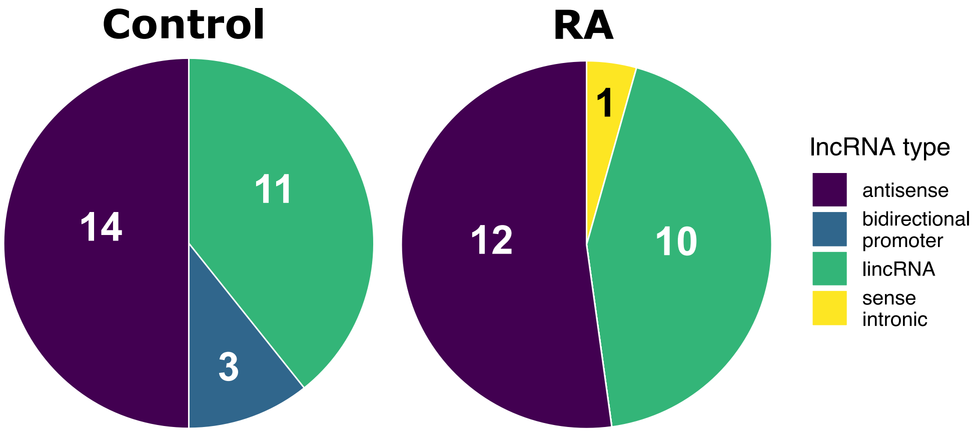
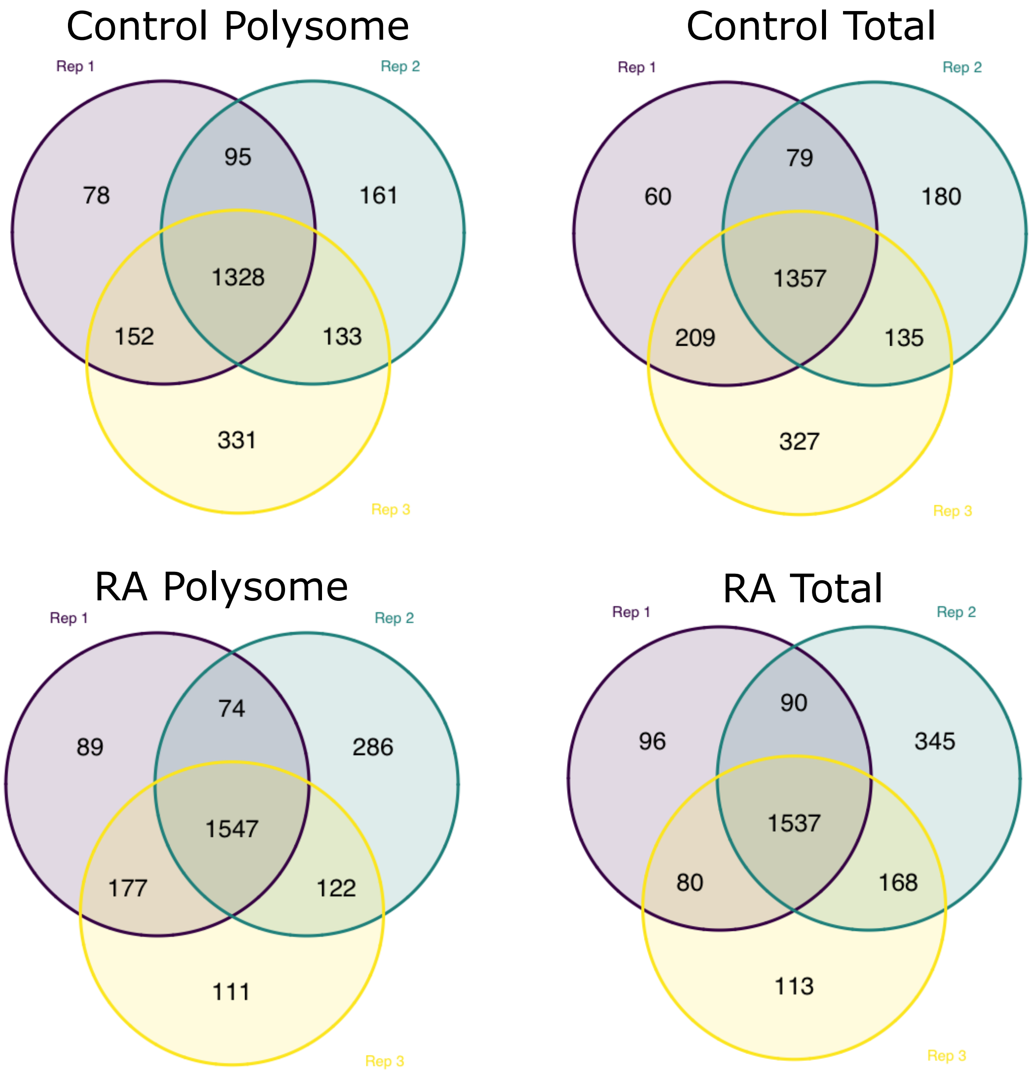


Figure 2.8: **Biotypes of translated lncRNAs found in Control and RA conditions.** A total of 45 lncRNA ORFs were identified using Polysome-associated RNA-Seq and Ribo-Seq using Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016).



Control Polysome

Control Total

RA Polysome

RA Total

Figure 2.9: **Normalised number of lncRNA transcripts present in each replicate, in the Total and Polysome-associated RNA-Seq.** Each venn diagram shows the overlap between the three replicates for each condition and RNA-Seq type. A threshold of ≥1 CPM was used to detect presence of a transcript.

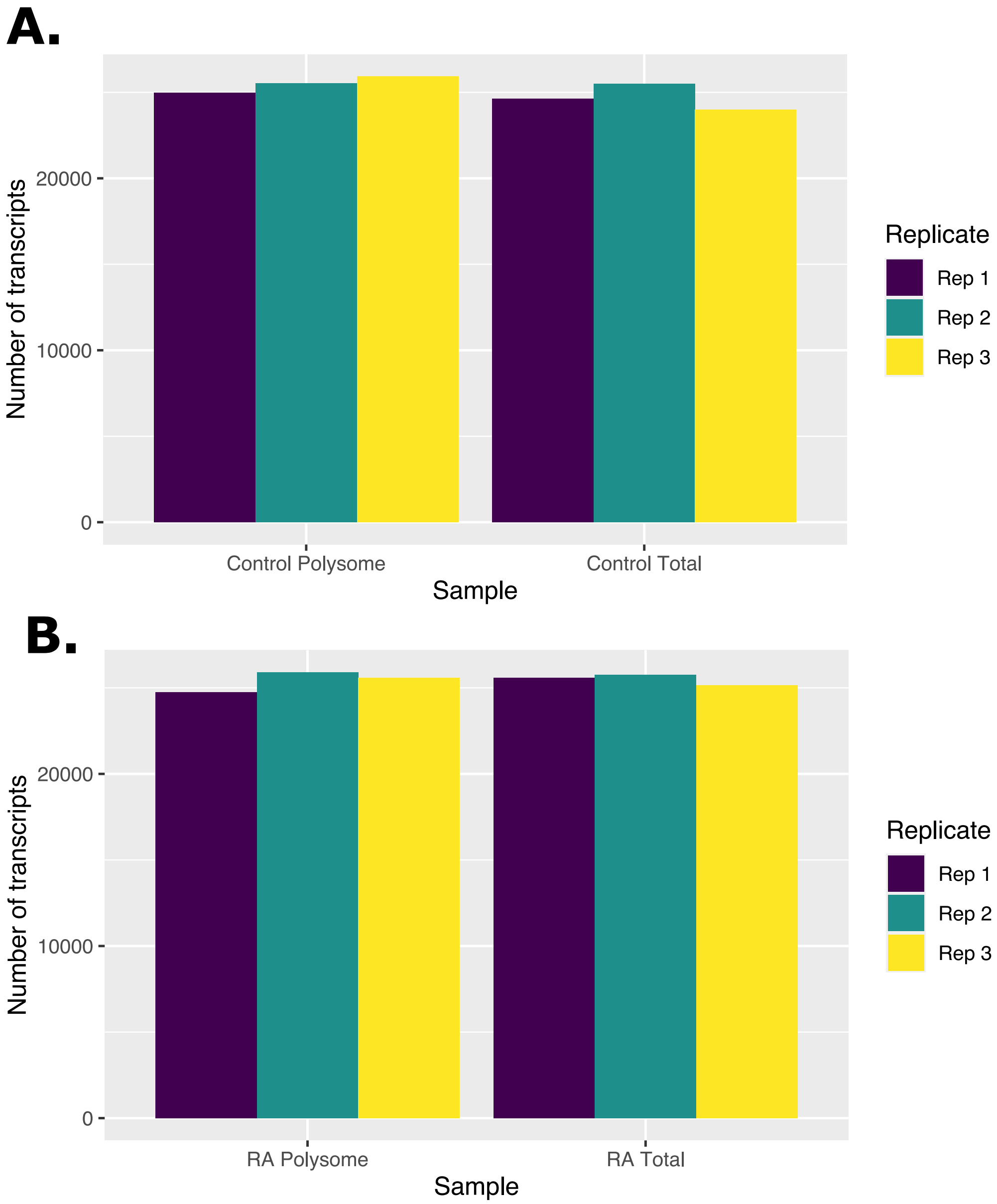


Figure 2.10: **Normalised number of transcripts present in the cytoplasm in each replicate, in A) Control and B) RA conditions, in Total and Polysome-associated RNA-Seq.** The x-axis shows the samples, and the y-axis shows the number of expressed transcripts. A threshold of ≥1 CPM was used to detect the presence of a transcript.

transcripts, a higher proportion of reads are likely to map to transcripts in this smaller population than the same transcripts in the Total RNA-Seq. Therefore all transcripts will exhibit different abundances between these samples, unless a transcript is very highly translated and all RNA molecules in the sample are polysome-associated.

In order to distinguish the untranslated, polysome-associated lncRNA population the CPMs were transformed and visualised. After removing any transcripts with a CPM of zero in the Total and the Polysome-associated RNA-Seq for a given replicate and condition, a small value $\alpha = 0.01$ was added to each CPM, to allow remaining zero values to be plotted. The following transformation was then applied, where CP represents the CPM of a given transcript in the Control Polysome-associated sample, and CT represents the CPM of the same transcript in the Control Total sample;

$$\log_2 \frac{CP+\alpha}{CT+\alpha}$$

This allowed the ratio of CPMs to be effectively visualised. All translated transcripts in each replicate and condition were also highlighted, although a small subset had a CPM of zero due to differences in the STAR (Dobin et al., 2013) and Salmon (Fuke and Ohno, 2008) mapping algorithms. In Figures 2.11 and 2.12, transcripts to the left of the origin are more highly abundant in the Total RNA-Seq, and transcripts to the right are more highly abundant in the Polysome-associated RNA-Seq. There is no clear enrichment of transcripts in either sample type as the mean ratio is close to the origin in all replicates, meaning that the majority of transcripts had near identical CPMs in the Total and Polysome-associated RNA-Seq. Therefore the untranslated, polysome-associated lncRNA population could not be identified from this analysis.

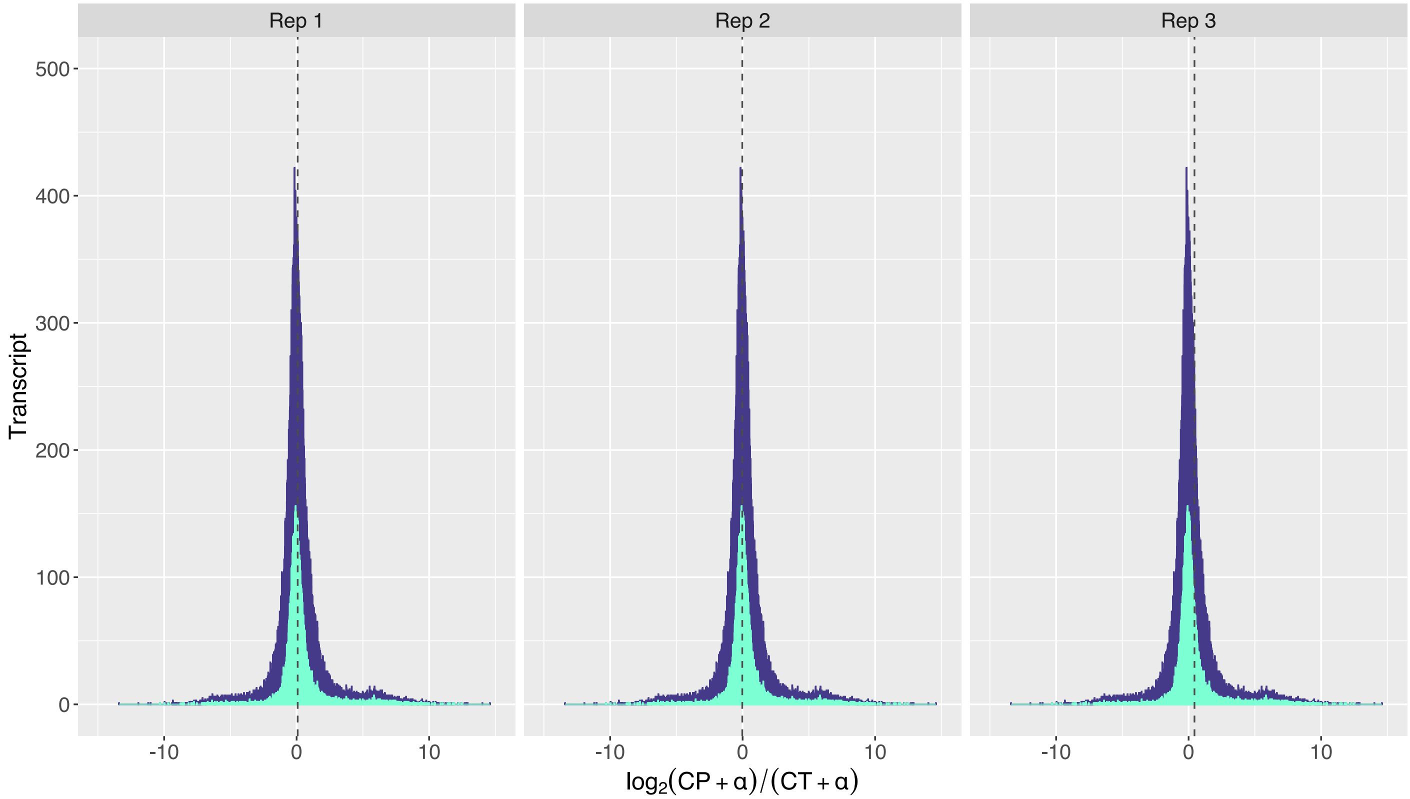


Figure 2.11: Log_2 **Total vs Polysome-associated RNA-Seq CPM, in Control conditions.** CPMs are visualised separately for Repliciates 1 to 3. The x-axis shows the ratio $\log_2 \frac{CP+\alpha}{CT+\alpha}$ for each transcript, where CP is the CPM of the transcript in Control Polysome-associated RNA-Seq, CT is the CPM of the transcript in Control Total RNA-Seq, and $\alpha = 0.01$. The y-axis shows the number of transcripts. Translated transcripts are highlighted in blue, and the mean ratio is highlighted by the dotted line.

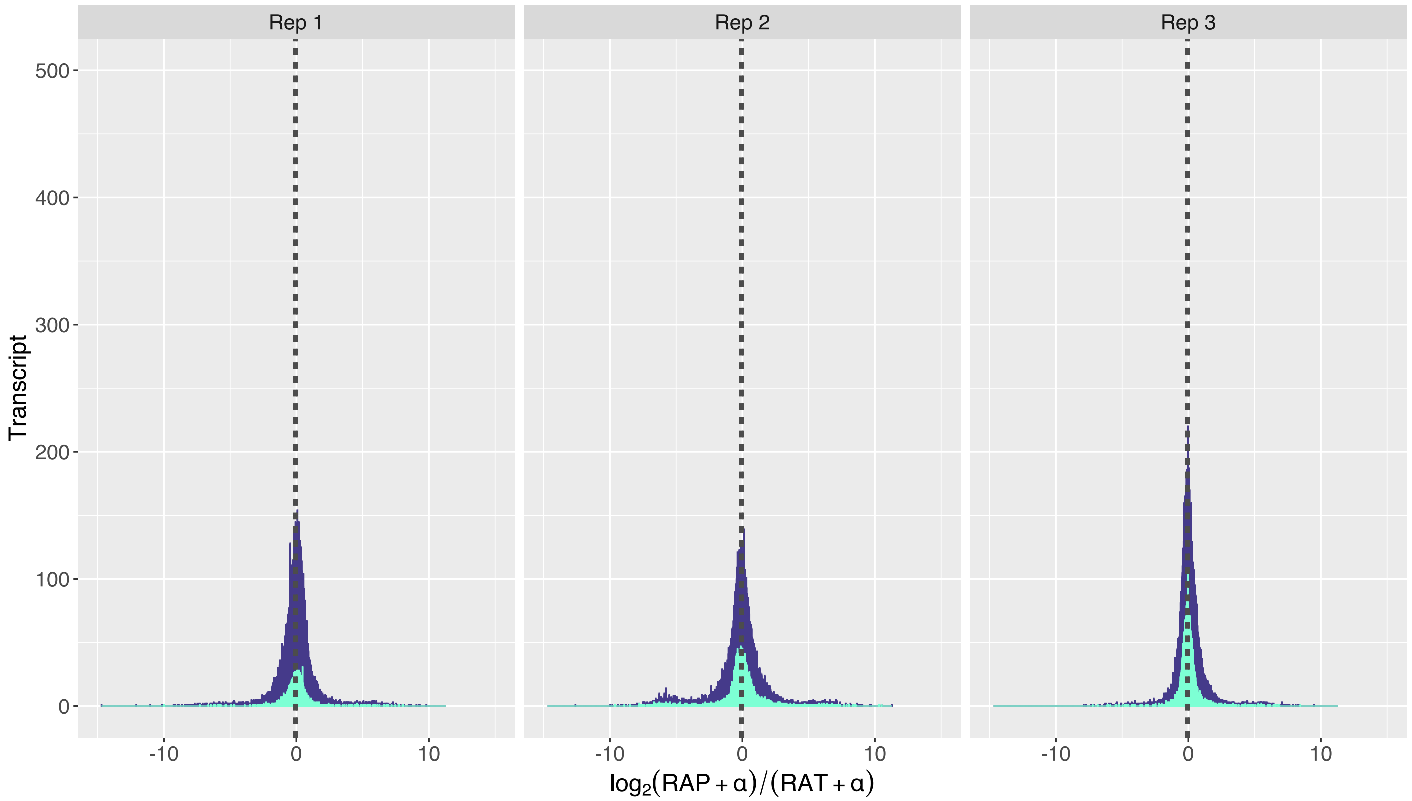


Figure 2.12: Log_2 **Total vs Polysome-associated RNA-Seq CPM, in RA acid treated conditions.** CPMs are visualised separately for Repliciates 1 to 3. The x-axis shows the ratio $\log_2 \frac{RAP+\alpha}{RAT+\alpha}$ for each transcript, where RAP is the CPM of the transcript in RA Polysome-associated RNA-Seq, RAT is the CPM of the transcript in RA Total RNA-Seq, and $\alpha = 0.01$. The y-axis shows the number of transcripts. Translated transcripts are highlighted in blue, and the mean ratio is highlighted by the dotted line.

2.4 Discussion

In summary, this chapter established that lncRNAs are both present in the cytoplasm and actively translated in SH-SH5Y cells, a model of human neuronal differentiation. Across undifferentiated Control and differentiated RA conditions, 45 lncRNA smORFs were identified (Figure 2.8), originating from 45 transcripts, from 25 genes. These translated transcripts represent ~2% of the total cytoplasmic lncRNA population.

The majority (41/45) of the lncRNA smORFs were in lncRNAs of the antisense or lincRNA biotypes (Figure 1.1), which corresponds to the proportions of lncRNAs in the Gencode v30 annotation (Derrien et al., 2012) (Table 1.2). No translated smORFs were identified in sense overlapping lncRNAs, both due to the poor annotation levels of this class of lncRNAs, and the difficultly in determining whether sequencing reads originate from the lncRNA, or the coding RNA they overlap.

Due to variation in the Total and Polysome-associated RNA-Seq data, it was not possible to separate out the populations of cytosolic and polysome-associated untranslated lncRNAs. This reflects previous work which found that the vast majority of cytoplasmic lncRNAs in SH-SH5Y cells were neither enriched in nor depleted from the polysomes (Aspden et al., 2014). Further, early work using Ribo-Seq data found evidence of translation in large proportions of lncRNA transcripts, including the majority of lincRNAs in mouse (Ingolia, Lareau, and Weissman, 2011), and up to 45% of lncRNAs in zebrafish (Chew et al., 2013). Newer methods which employ analysis of triplet periodicity to distinguish ribosome association from active translation have reduced the number of lncRNAs found to be translated (Guttman, Russell, et al., 2013), but these early works are still indicative that a large proportion of lncRNAs are associated with the translational machinery. However, having identified the subset of translated lncRNAs, comparisons can be made with the larger population of untranslated lncRNAs which are expressed and exported to the cytoplasm, allowing features unique to translated lncRNAs to be elucidated.

Our inability to separate out the populations of cytosolic and polysome-associated lncRNAs highlights the importance of experimental design, and the need to plan based on your research question. The data this work is based on were originally produced to investigate the coding potential of cytoplasmic lncRNAs, and to characterise their role in the early stages of human neuronal differentiation (Douka, Birds, et al., 2021). Poly-Ribo-Seq (Aspden et al., 2014) is therefore the ideal method for this question, as the presence of multiple ribosomes lends credence to the translation of novel ORFs from the "non-coding" transcriptome. However, to distinguish polysome-associated non-translated lncRNAs from cytosolic lncRNAs, slight changes to the protocol would be needed. Namely, sequencing the other fractions in the sucrose gradient. This would allow a fairer comparison of the enrichment of lncRNAs in the non-ribosome, monosome, and polysome fractions. An alternative approach would be use of a spike in control, which provides a known quantity which can be used to normalise transcript abundance across samples(Choe et al., 2005; Hoerth, Reitter, and Schott, 2022).

Further, the analysis methods used in this chapter highlight a barrier to research into lncRNA subpopulations; cost. Although there is now an abundance of publically available Ribo-Seq data in a wide range of species and sample types, much of these data only include one replicate per condition, due to the large funding and time requirements of the Ribo-Seq protocol. It would therefore be difficult to apply this analysis to another cell type or species using publically available data, as multiple replicates are required to account for biological variability, and to mitigate the low abundance of lncRNA transcripts.

Ribotaper (Calviello, Mukherjee, et al., 2016) was selected to perform this translational analysis was selected because the focus of the methodology on the periodic distribution of ribosomes along the ORF, and its ability to identify novel ORFs. However, a downside of using Ribotaper to measure translation is the lack of ability to identify ORFs on different splice variants in the same sample. Once an ORF has been reported on a transcript, the possibility that it originates from another isoform is not considered. Given that lncRNAs are often poorly annotated, or in some cases have a very large number of splice variants, there is a large chance that ORFs may be erroneously attributed to a given splice variant. This will also miss any cases where multiple splice variants with the same or very similar ORFs are in fact both being translated. Since this work was carried out, the authors of Ribotaper have released a new tool to address this issue; ORFquant (Calviello, Hirsekorn, and Ohler, 2020), which detects ORF translation across multiple transcript isoforms.

2.4.1 Conclusions

Overall, this chapter establishes that a small population of lncRNAs are translated in undifferentiated and differentiated human neuronal cells, despite the pervading view that these transcripts do not have coding potential. The following chapters will examine the properties of these lncRNAs and their potential significance in the evolution of new protein coding genes.

Chapter 3

Characterisation of translated lncRNAs

3.1 Introduction

3.1.1 Features of translated lncRNA smORFs

Although the translation of lncRNA smORFs and other non-canonical ORFs is becoming more widely recognised, efforts to standardise their annotation are still preliminary (Mudge et al., 2022; Chothani, Adami, Widjaja, et al., 2022). Therefore, relatively little is known about the features of translated lncRNAs and their smORFs, and what we do know is often based on analysis of translated lncRNAs in a single species or tissue/cell type. Further, the wider lncRNA population is a particularly heterogenous class of ncRNAs due to the relatively arbitrary definition of "ncRNAs of ≤ 200 nt", creating the largest category of human ncRNAs (Frankish et al., 2021). Given this context, we may expect to observe a wide range of characteristics in terms of the composition and structure of translated lncRNAs.

The "novel ORF consortium" reported 2,208 human lncRNA ORFs, ranging in length from 51 to 1050 nt, with a median of 153 nt (Mudge et al., 2022). Analysis of the full dataset of 7,264 ORFs from the consortium (also including uORFs, dORFs, and internal out-of-frame ORFs) found that 89.6% were evolutionarily young, lacking significant protein homology outside of primates, and 3% were human specific (Sandmann et al., 2023). In *Drosophila melanogaster*, translated lncRNA ORFs displayed significantly higher levels of conservation than ORFs which were only associated with the translational machinery (Patraquim et al., 2022). The translational profiles of lncRNA ORFs are also comparable to those of protein coding ORFs, in particular their footprint coverage and framing across the ORF (Ruiz-Orera, Messeguer, et al., 2014; Patraquim et al., 2022).

3.1.2 Features of *de novo* ORFs

LncRNAs have been found to be a source of *de novo* protein coding genes in a range of species, including human (Vakirlis, Vance, et al., 2022; An et al., 2023), primates (Chen, Shen, et al., 2015; Xie et al., 2012), rodents (Petrzilek et al., 2022), and *Drosophila* (Aspden et al., 2014; Reinhardt et al., 2013). It has been hypothesised that *de novo* coding genes exist on a continuum between canonical protein coding genes and non-coding sequences, which is reflected in their features (Vakirlis, Acar, et al., 2020). These include shorter genes and ORFs than annotated protein coding genes, and fewer exons and known domains, in mouse, human, zebrafish, and stickleback (Neme and Tautz, 2013). A large proportion of *de novo* genes are single exon in Drosophila (57%) (Zhao et al., 2014) and human (Wu, Irwin, and Zhang, 2011). In rodents, *de novo* ORF sequences were often found to be contained in one exon (Murphy and McLysaght, 2012).

The expression of *de novo* genes has also been found to be similar to that of lncRNAs; low expression (Palmieri, Kosiol, and Schlötterer, 2014; Zhao et al., 2014), in a highly tissue specific manner (Levine et al., 2006; Toll-Riera et al., 2009). In the the yeast *Saccharomyces cerevisiae* higher transcript abundance levels correlated with well categorised, strongly conserved ORFs (Carvunis et al., 2012), as did increased ORF length, closer proximity to transcription factor binding sites, and altered codon usage bias.

In new ORFs in budding yeast, a propensity to encode transmembrane domains is associated with beneficial fitness effects (Vakirlis, Acar, et al., 2020). However, other work on *de novo* coding sequences in human found only 2% of ORFs encoded a transmembrane domain (Vakirlis and McLysaght, 2019), and a study of rodent *de novo* coding sequences found no known domains (Murphy and McLysaght, 2012).

There are also conflicting reports in the literature regarding *de novo* protein coding genes and GC content, with some reporting that new human ORFs emerged from G/C-rich lncRNAs(Chen, Shen, et al., 2015), while others suggest *de novo* genes are A/T-rich (Ruiz-Orera, Messeguer, et al., 2014). This quality is important, as A/T-richness mediates the number of start (ATG) and stop (TAA,TAG,TGA) codons which are present in a transcript, therefore affecting the length and density of randomly occurring ORFs (McLysaght and Hurst, 2016). For example, in human CDS sequences, approximately 2.2 codons per 100 nt are ATGs (Athey et al., 2017), and they are found throughout the CDS, not only at the translation initiation site. Indeed, ATG codons are often found in frame and downstream from start codons with weak a Kozak context (Benitez-Cantos et al., 2020), and over half of human mRNAs have at least one upstream ATG (Andreev et al., 2022)

Given these mixed reports, this chapter aims to classify the sub-population of translated lncRNAs in terms of their translation, transcript sequences, peptide sequences, and their expression. By comparing these characteristics from translated lncRNAs with those of protein coding and non-canonical ORFs we will establish where their properties fall on the spectrum of non-coding to coding sequences.

3.2 Materials and Methods

3.2.1 Metaplots

To create ribosome footprint metaplots, 100 translated ORFs were randomly selected from the set of expressed, translated, protein coding transcripts, and compared to the translated lncRNA ORFs. P-site counts were computed for each position along the transcripts in a 75 nt window around the start and stop codons of the ORF. These counts were scaled by the total number of reads in the two windows for each transcript, and the mean normalised counts were plotted for each position in the two windows.

3.2.2 Translational Efficiency

TE was estimated for all translated ORFs in each condition, where TE was equal to the mean number of P sites per ORF, normalised by the median P sites per ORF per condition, divided by the mean number of RNA sites per ORF, normalised by the median RNA sites per ORF per condition.

3.2.3 Differential translation efficiency

Differential translation efficiency analysis was performed using the deltaTE (Chothani, Adami, Ouyang, et al., 2019) alternate protocol (E-appdendix; TE_Total_DESEQ2.Rmd). This analysis was based on transcript-level read counts from all identified translated ORFs from the Ribo-seq and Total RNA-seq, not gene-level counts as described in the protocol. To create these read counts, Browser Extensible Data (BED) files produced by Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016) (Section 2.2.3.1) describing the ORFs translated in each replicate and condition were combined and filtered to only include ORFs passing the criteria of being present in 2/3 replicates in a given condition. The chromStart and chromEnd fields were altered as follows to ensure they spanned from the first nucleotide of the start codon to the first nucleotide of the stop codon for each ORF:

- Positive stranded ORF, first region; chromStart - 1
- Positive stranded ORF, last region; chromEnd - 1
- Negative stranded ORF, first region; chromStart - 1
- Negative stranded ORF, last region; chromEnd + 1

The resulting BED file was used to filter files describing P and RNA sites produced by Ribotaper during translational analysis (Section 2.2.3.1), using the create_tracks.bash script from Ribotaper (E-appdendix; Bespoke_bed.sh). The resulting datatracks were summed to give read counts of RNA-seq and Ribo-seq reads present on each ORF.

3.2.4 Peptide composition analysis

The compositions of peptides from the protein coding ORFs, lncRNA smORFs, uORFs, and dORFs were calculated using extractACC from the protr package v1.6-2 (Xiao et al., 2015). The average amino acid composition of each group was calculated as the median. Random control expected frequencies were taken from King and Jukes (King and Jukes, 1969).

3.2.5 Mass spectrometry analysis

Analysis of the mass spectrometry data was performed by Dr Elton Vasconcelos, bioinformatics technical support officer at LeedsOmics. Two published SH-SY5Y cell mass proteomics datasets were analysed: PXD010776 (Murillo et al., 2018) and PXD014381 (Brenig et al., 2020). Binary raw files (.raw) were downloaded from PRIDE (Perez-Riverol et al., 2022) then converted to human-readable MGF format using ThemoRawFileParser (Hulstaert et al., 2020). The amino acid sequences of translated lncRNA smORFs were added to the whole *Homo sapiens* proteome dataset (20,379 entries) downloaded from UniProtKB (The UniProt Consortium, 2021) on Nov/2019.

The new FASTA file was then used as a custom database on Comet v2019.01.2 (Eng, Jahan, and Hoopmann, 2013) search engine runs that scanned all MS/MS files (.mgf) against it. Default settings were used in Comet with the following exceptions according to the MS/MS data type. iTRAQ-4plex (PXD010776): decoy_search = 1, peptide_mass_tolerance= 10.00, fragment_bin_tol = 0.1, fragment_bin_offset = 0.0, theoretical_fragment_ions = 0, spectrum_batch_size = 15000, clear_mz_range = 113.5-117.5, add_Nterm_peptide = 144.10253, add_K_lysine = 144.10253, minimum_peaks = 8. Label-free (PXD014381): decoy_search = 1, peptide_mass_ tolerance = 10.00, fragment_bin_tol = 0.02, fragment_bin_offset = 0.0, theoretical_fragment _ions = 0, spectrum_batch_size = 15000. CometUI (Eng, Jahan, and Hoopmann, 2013) was employed for analysing MS/MS data and setting a false discovery rate (FDR) threshold of 10% per peptide identification. This FDR threshold was selected due to expected low abundance levels of the target smORFs.

3.2.6 Domain analysis

HMMER v3.3.1 (Mistry, Finn, et al., 2013) was used to scan lncRNA peptides against the Pfam-A Hidden Markov Model (HMM) library v34.0 in a local installation of PfamScan v1.6 (Mistry, Bateman, and Finn, 2007). PfamScanner (Moore, 2022) was used to parse these outputs, with an e-value cut-off of 0.001 (E-appendix; pfam.sh).

3.2.7 "Non-stringent" translated lncRNA dataset

An expanded, "non-stringent" dataset of translated lncRNA smORFs was identified using the methods described in Section 2.2.3.1, without the requirement that a smORF be identified in at least two of the three biological replicates of a given condition (Control or RA).

3.2.8 PyschENCODE data analysis

Gene expression in RPKM from human and rhesus macaque brains was downloaded from PyschENCODE (Zhu, Sousa, et al., 2018; Akbarian et al., 2015). These data were filtered to lncRNAs using the GENCODE v25 annotation (Frankish et al., 2021), and the 105 genes in the annotation which corresponded to translated lncRNA smORFs from the "non-stringent" set were selected. For each sample, a further 25 random sets of 105 lncRNAs were selected, and the Kruskal-Wallis test applied to establish whether set had a significant effect on RPKM. A two-tailed multiple comparison test was applied to test whether there was a significant difference between the translated set and the 25 randomly selected tests. This non-parametric test was selected as the expression of the lncRNA sets did not fulfil the assumption of a normal distribution.

3.2.9 General plots and statistics

Data were analysed and plotted in R (R Core Team, 2021), using packages including the tidyverse (Wickham et al., 2019), tibble (Müller and Wickham, 2021), seqinr (Charif and Lobry, 2007), protr (Xiao et al., 2015), ggplot2 (Wickham, 2016), gggenes (Wilkins, 2023), ggpubr (Kassambara, 2023), gghighlight (Yutani, 2022), VennDiagram (Chen, 2022), ComplexHeatmap (Gu, Eils, and Schlesner, 2016), RColorBrewer (Neuwirth, 2022), and viridis (Garnier et al., 2021).

Boxviolin plots of ORF length, TE, exon number, start and stop codon content, and GC content were created using ggstatsplot v0.10.0 (Patil, 2021). As the data did not follow a normal distribution, a non-parametric statistical approach was used; a Kruskal wallis test, followed by Dunn's Tests for pairwise comparisons with the bonferroni correction.

3.3 Results

Analysis of Poly-Ribo-Seq data (Chapter 2) identified 45 actively translated lncRNA smORFs in SH-SY5Y cells, originating from 21 antisense, 20 intergenic, 1 sense intronic, and 3 bidirectional lncRNA transcripts. This chapter aims to understand the features, expression, and potential function of these lncRNAs to gain insight into their importance, and to determine whether they represent a distinct population of lncRNAs based on these properties.

3.3.1 Characterisation of the translational landscape of lncRNA smORFs

To understand the nature of the non-canonical translation of the 45 lncRNA smORFs, this section compares their translational activity to that of protein coding ORFs and other non-canonical ORFs.

3.3.1.1 LncRNA smORF footprint distributions are comparable to protein coding ORFs

The translated lncRNA smORFs clearly display triplet periodicity, with an average of ~75% in frame Ribo-Seq reads in Control and ~74% in RA conditions. To compare the pattern of ribosome footprints to protein coding ORFs, 100 ORFs were randomly selected from the set of expressed, translated, CCDS protein coding ORFs. P-sites around the start and stop codons of the translated lncRNAs and protein coding ORFs were plotted (Figures 3.1 and 3.2). The distribution of footprints is similar between lncRNA and protein coding ORFs in Control and RA conditions, with a clear bias toward the reading frame corresponding to the start codon, and a drop-off of footprints at the stop codon. This is indicative that ribosome footprinting across the lncRNA smORFs is that of genuine translation events.

The translated lncRNA smORF LINC01116-201 (ENST00000295549.8_609_822) exemplifies this strong periodicity, encoding a 216 nt smORF which is translated in Control replicates 1 and 2, and RA replicate 3 (Figure 3.3). LINC01116 is induced upon RA-induced differentiation, and exhibits increased ribosome footprinting (Douka, Birds, et al., 2021). This increased transcript abundance means LINC01116-201 is an ideal example to illustrate the triplet periodicity observed during active translation, with 80% of the Ribo-seq reads mapped to frame 2. The few Ribo-seq reads aligned outside the smORF display a more equal distribution, with no clear bias to any reading frame. This smORF also demonstrates how lncRNAs can contain many canonical start and stop codons within a smORF. However, as they are all in reading frames 0 and 1, they do not affect the translation of the LINC01116-201 smORF.

3.3.1.2 LncRNA smORF translational efficiency is comparable to protein coding ORFs

Previous work in two isogenic human cancer cell models found the TE of translated cytoplasmic lncRNAs to be slightly lower than mRNAs when expression levels are accounted for (Ji, Song, et al., 2015). To assess the level at which the 45 lncRNA smORFs are translated, the TE of lncRNA smORFs, protein coding ORFs, uORFs and dORFs were determined (Figure 3.4). This was calculated as the normalised number of ribosome footprints relative to number RNA-seq reads. Kruskal-Wallis tests reveal a significant effect of ORF type on TE in Control ($H(3) = 34.13, p = 1.86e^{-}07)$) and RA conditions ($H(3) = 31.71, p = 6.01e^{-}07)$). In Control conditions, there was no significant difference between lncRNA and protein coding ORFs, and while in RA conditions there was a significant difference (p=0.02), and lncRNA smORFs had a higher median TE of 0.77 compared to 0.02 for protein coding ORFs. This provides further evidence that these lncRNAs are undergoing genuine translation events.

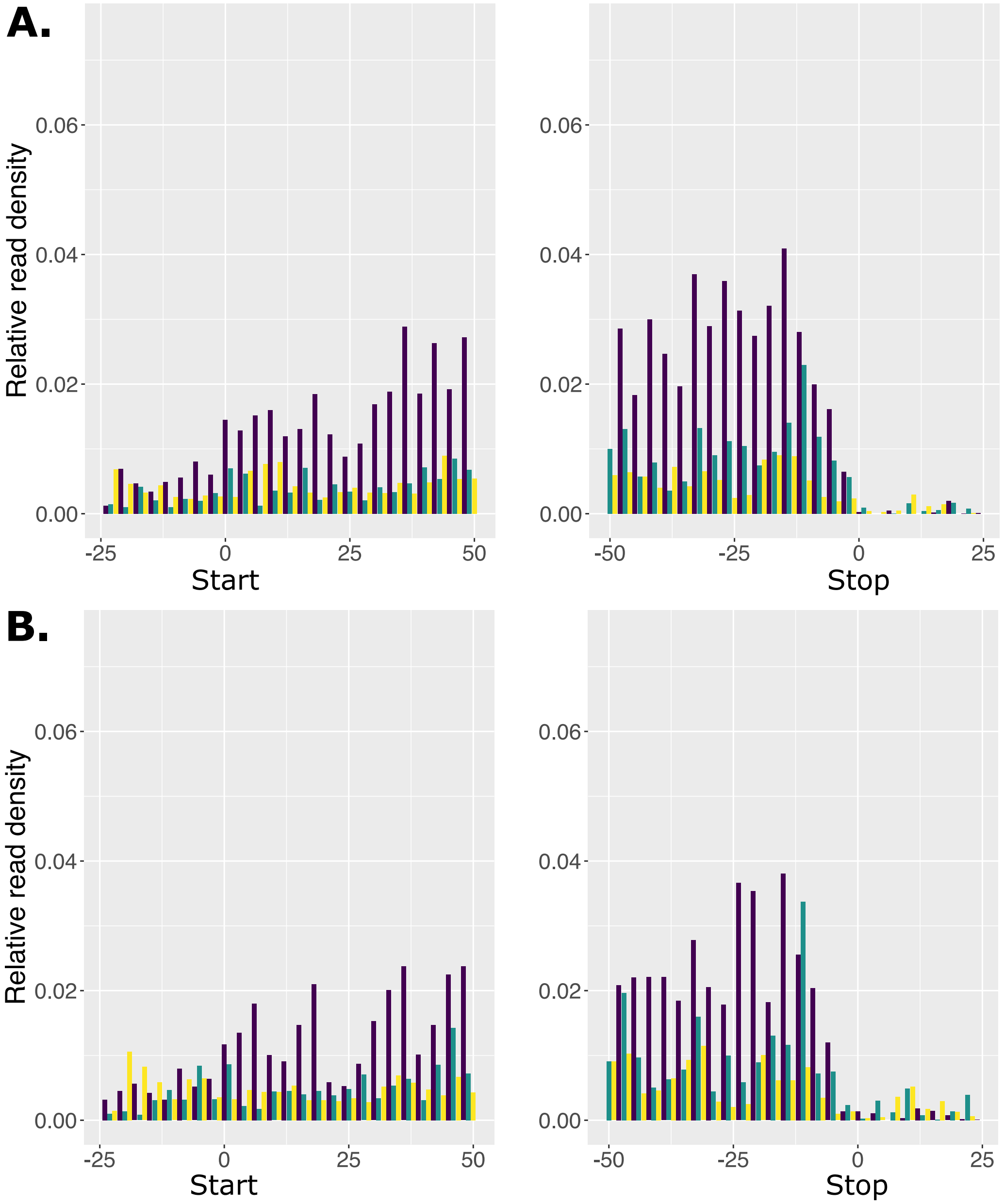


Figure 3.1: **Distribution of protein coding ORF P-sites.** Metagene plots showing the distribution the P-sites of ribosome reads around the start and stop codons of 100 randomly selected CCDS protein coding ORFs in A) Control and B) RA conditions. The x-axis shows positions along the ORF, where position 0 denotes the start or stop codon. The y axis shows the relative Ribo-seq read density. The plot is colour coded, corresponding to the three possible reading frames.

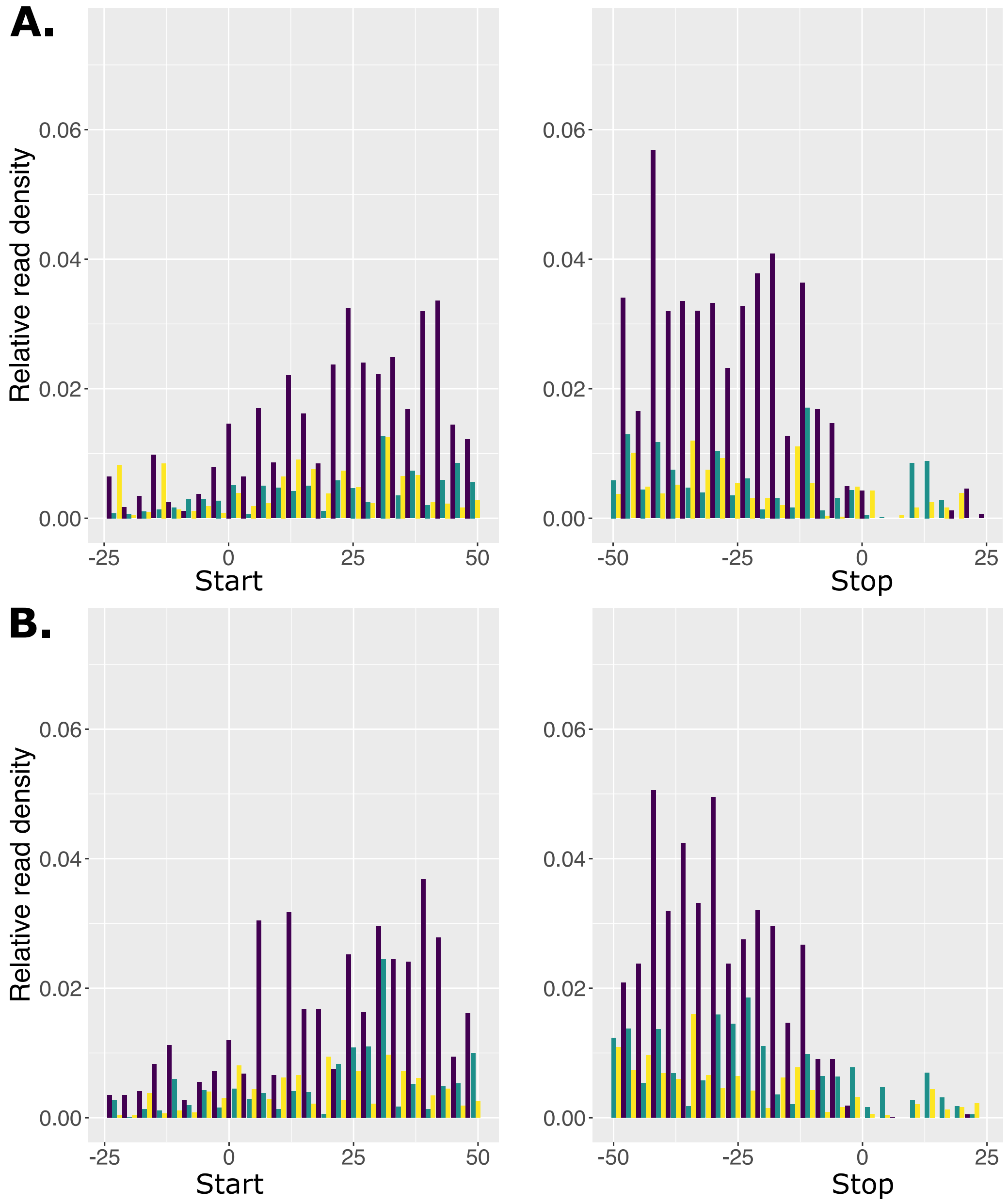


Figure 3.2: **Distribution of translated lncRNA smORF P-sites.** Metagene plots showing the distribution the P-sites of ribosome reads around the start and stop codons of the 45 translated lncRNA smORFs in A) Control and B) RA conditions. The x-axis shows positions along the ORF, where position 0 denotes the start or stop codon. The y axis shows the relative Ribo-seq read density. The plot is colour coded, corresponding to the three possible reading frames.

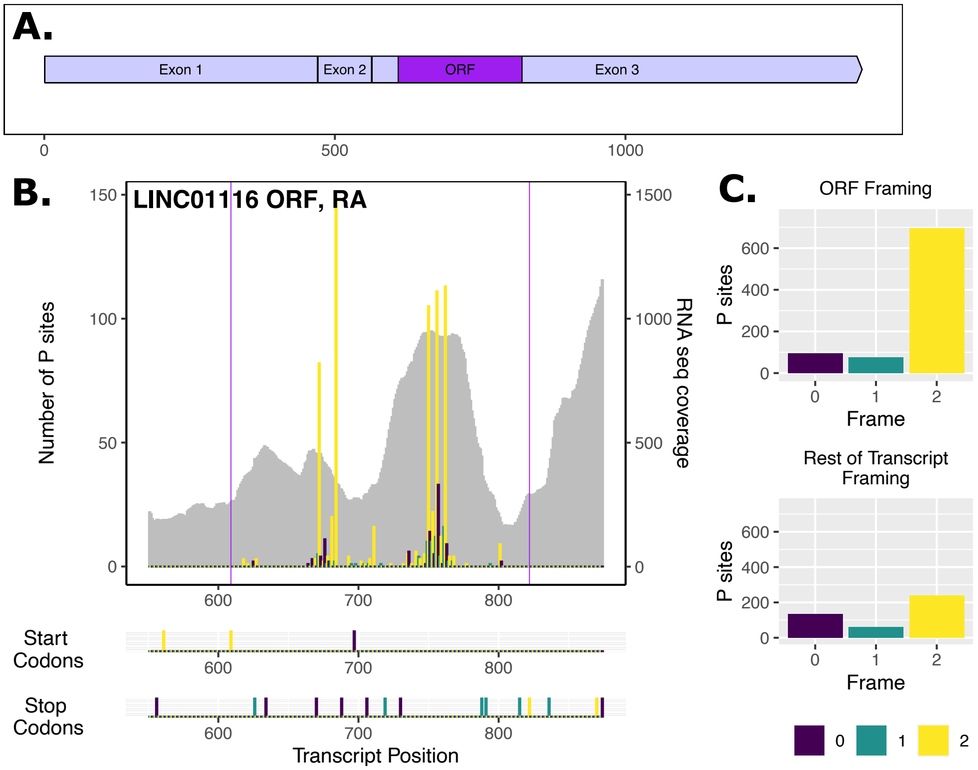


Figure 3.3: **Ribosome footpinting across LINC01116 smORF. A)** The transcript LINC01116, which contains a 216nt smORF in the 3rd exon, highlighted in dark purple. The x-axis shows distance along the transcript in nucleotides. **B)** The x axis shows distance along the LINC01116 transcript in nucleotides. The start and end of the smORF is marked by purple lines, corresponding to the portion of the transcript highlighted in A). The positions of all canonical start and stop codons, colour coded by frame, are shown below the main plot. The left y-axis shows the number of P-sites, which are colour coded by frame, arbitrarily designated as frame 0; dark purple, frame 1; turquoise, and frame 2; yellow. The right y-axis shows RNA-seq coverage, which is plotted in grey. The Ribo-seq and RNA-seq results are from RA conditions, in replicate 3. **C)** P-site framing within and without the smORF. The x-axises show the three possible frames, colour coded as in B). The y-axises show counts of Ribo-seq read P-sites. 80% of the Ribo-seq reads within the smORF are in the same frame (frame 2), whilst reads across the rest of the transcript are more evenly distributed.

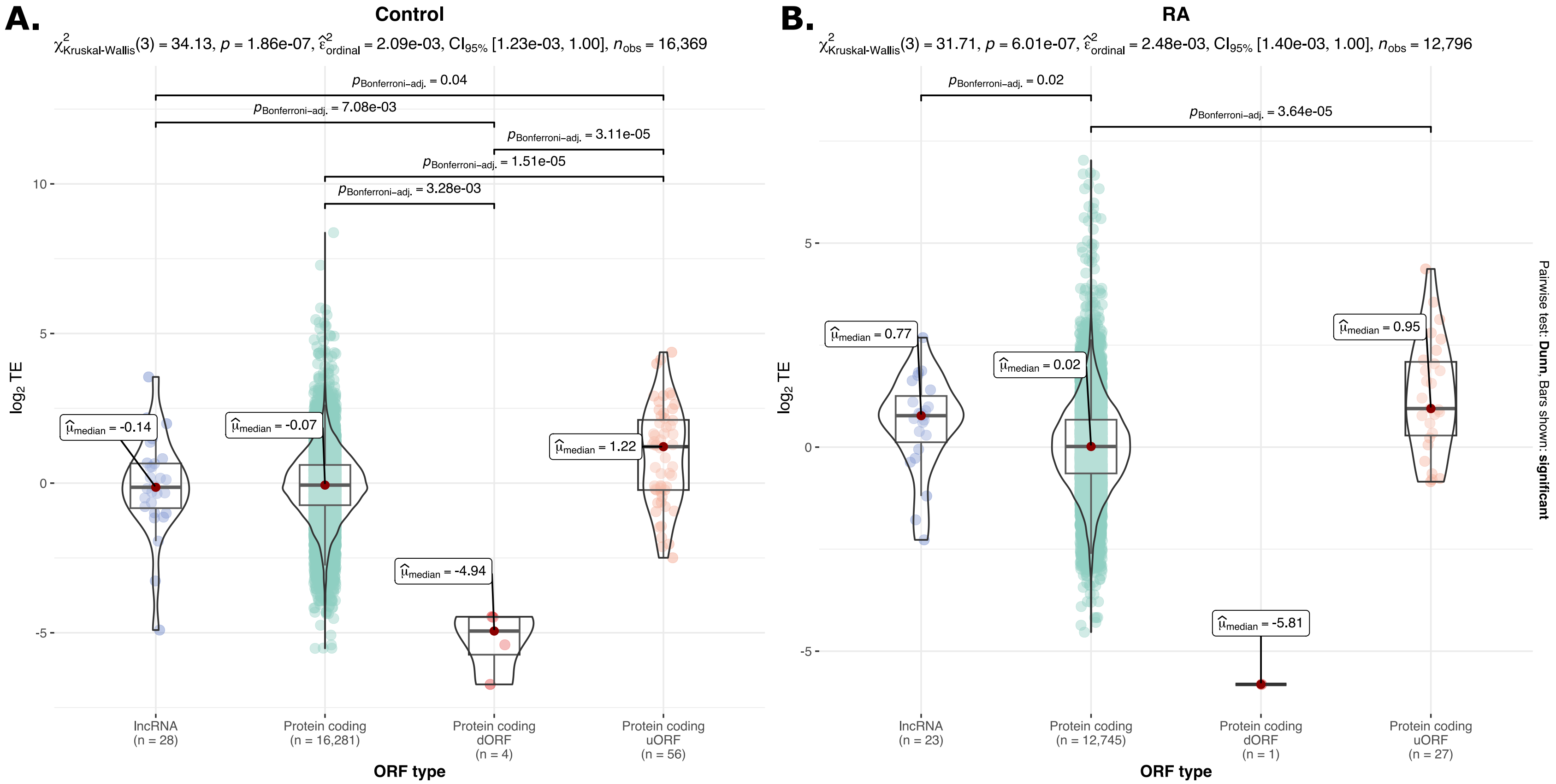


Figure 3.4: **Translational efficiency of ORF classes.** TE of lncRNA smORFs, protein coding ORFs, uORFs and dORFs in A) Control and B) RA conditions. The x-axis shows ORF type, where n is the number of ORFs in each category and condition. The y-axis shows log_2TE, where TE = Normalised P sites per ORF/RNA sites per ORF. The violin plots display the distribution of TEs, overlaid with box plots to display the interquartile ranges. The red dots denote the median values. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test. Significant pairwise comparisons are displayed.

A possible explanation for the lncRNA smORFs higher median TE in RA conditions may be the global reduction in translation levels upon RA-induced differentiation in SH-SY5Y cells (Douka, Birds, et al., 2021). If the lncRNA smORFs represent very new coding sequences, they may not be regulated as effectively during neuronal differentiation as established protein coding genes.

The dORFs could only be compared to the other populations in Control conditions, as only one dORF was translated in RA conditions. However, it is clear that dORFs are translated at much lower efficiencies, due to the need for ribosomes to reinitiate after translating or scanning over the main ORF. uORFs were significantly different from protein coding ORFs in both Control ($p = 1.51e^{-}05$) and RA ($p = 3.64e^{-}05$) conditions, with a higher median TE.

A limitation of this analysis is that it is based on outputs from Ribotaper (Calviello, Mukherjee, et al., 2016) which only reports the Ribo-seq and RNA-seq coverage of ORFs which passed the various thresholds to be considered translated. Therefore the TEs are likely to be skewed higher than reality, as ORFs just below the threshold to be considered translated in a given replicate are not represented in the data.

3.3.1.3 No differential translation was detected in lncRNAs

To investigate if the lncRNA smORFs were undergoing translational changes between Control and RA conditions, differential translation analysis was performed using the Total RNA-seq, and Ribo-seq (Chothani, Adami, Ouyang, et al., 2019). A principle component analysis (PCA) (Appendix B, Figure B.1) was used to check for batch effects, confirming that read counts cluster by sequencing type and replicate. Changes in Ribosome footprints (ΔRPF), RNA-seq reads (ΔRNA), and TE (ΔTE) were calculated (Figure 3.5). Of the 19,214 ORFs translated in either condition, only 2 were differential translation efficiency ORFs (DTEO) ($p_a dj < 0.05$), both of which were CCDS protein coding ORFs. DTEOs are defined as ORFs with changes in Ribo-seq footprinting independent of transcription levels, indicative of a change in translational efficiency. 2,642 translated ORFs were differentially transcribed ORFs (DTO) ($p_a dj < 0.05$), 3 of which were lncRNA smORFs; these are ORFs under significant transcriptional control as their RNA-seq and Ribo-seq read counts are changing at the same rate. No lncRNA smORFs were differentially translated, and only ∼7% underwent signicant transcriptional changes; as hypothesised in the previous section, this lack of significant regulation may be due to the lncRNA smORFs being new coding sequences.

3.3.2 Characterisation of lncRNA smORFs

3.3.2.1 LncRNA smORFs are shorter than canonical coding sequences

De novo coding genes and their ORFs are shorter than well characterised, conserved protein coding genes in a range of species (Carvunis et al., 2012; Neme and Tautz, 2013). LncRNA ORFs are also shorter, as if they contained large ORFs they would likely also be classified as protein coding genes. To establish how the 45 translated lncRNA smORFs compare to protein coding and *de novo* ORFs, the lengths of translated ORFs were visualised (Figure 3.6). The lncRNA smORFs ranged in length from 51 to 816 nt, with a median of 180 nt; consistent with the literature (Mudge et al., 2022). There was a signifcant difference between the protein coding ORF (median of 1413 nt) and lncRNA smORF lengths ($p = 1.26e^{-}27$).

There was no significant difference between lncRNA ORFs and uORFs (median of 102 nt), although uORFs did have the longest transcript population, with a median of 3,832 nt compared to the protein coding median of 3,040 nt. This likely due to the fact that there needs to be sufficient distance between a uPRF and the main ORF for ribosomes to reinitiate following translation of the uORF. As this dORF population is particularly small with only 5 translated dORFs identified, conclusions cannot be drawn from their length distribution. This small population is to be expected, as for a dORF to be translated the translational

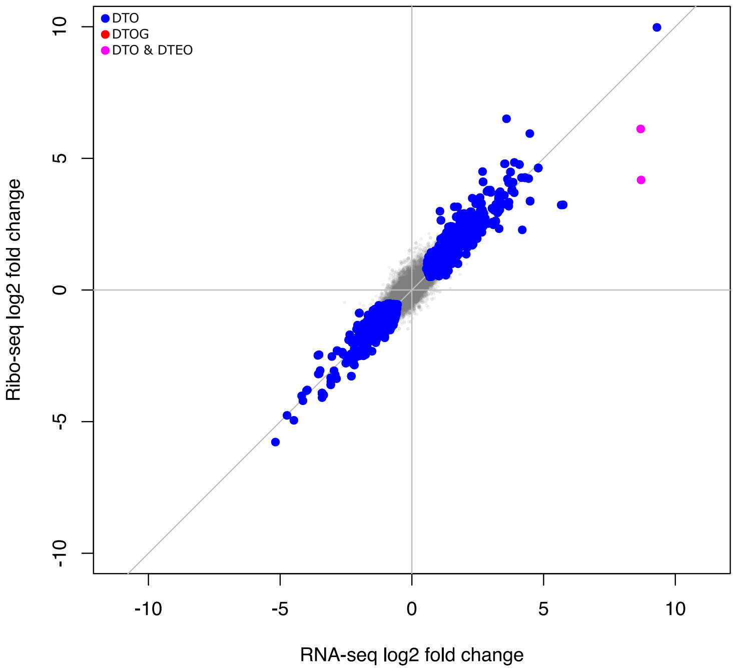


Figure 3.5: **Plot of log fold change for each ORF in RNA (RNA-Seq) versus ribosome footprints (Ribo-Seq).** The x-axis shows changes in RNA-seq counts, and the y-axis shows changes in Ribo-Seq counts. ORFs highlighted in blue are DTOs, driven by transcriptional regulation with significant ΔRPF and ΔRNA but no significant ΔTE. Red ORFs are DTEOs, driven by translational regulation with significant ΔRPF and ΔTE, but no significant ΔRNA. Pink ORFs are both DTO and DTEO; undergoing differential transcription and translation efficiency.

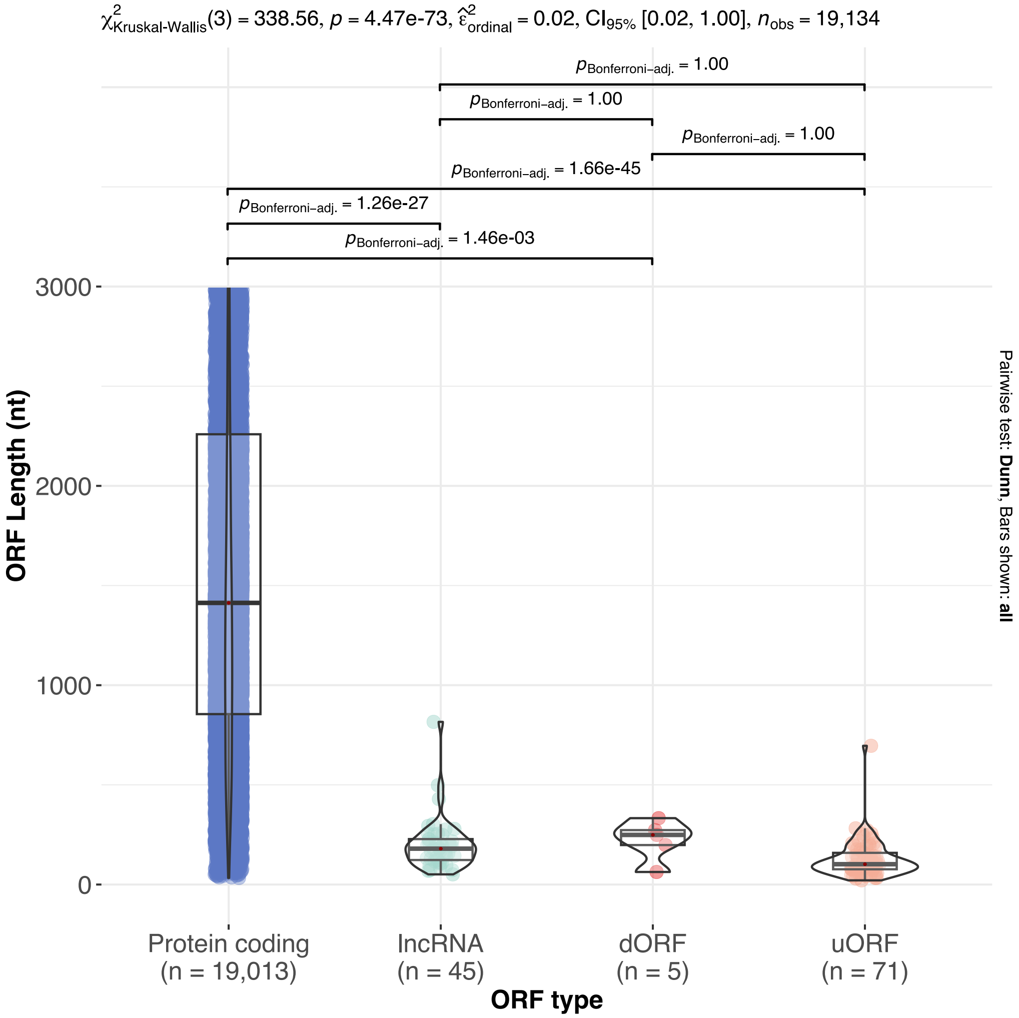


Figure 3.6: **Translated ORF lengths (nt).** The x axis shows categories of translated ORFs, n denotes the number of ORFs in each category. The y axis shows the length in nucleotides. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test. A ylim of 3,000 nt was used to crop the figure due to a high number of protein coding outliers; the full figure is in Appendix B, Figure B.2

machinery must remain on a transcript beyond the main ORF and reinitiate, which occurs with low efficiency.

3.3.3 Characterisation of lncRNA smORF peptides

Here the properties of the peptides produced from these smORFs are examined. This contextualises the peptides in comparison to characterised proteins, and provides supporting evidence for their synthesis.

3.3.3.1 LncRNA peptide composition is similar to canonical proteins

Previous analysis of *Drosophila melanogaster* translated lncRNA smORF peptides found specific amino acid usage indicative of genuine protein products (Aspden et al., 2014). To establish if the translated human lncRNA smORFs display specific amino acid usage, the average amino acid composition of protein coding, lncRNA, uORF, and dORF peptides was calculated, and compared to the expected amino acid frequencies by chance (King and Jukes, 1969) (Figure 3.7). LncRNA peptides cluster most closely with uORF peptides, but in general exhibit a similar composition to canonical proteins. Specifically, all of the ORF types displayed a lower use of arginine than by chance, a feature of translated proteins (King and Jukes, 1969), although lncRNA peptides contained a higher proportion than protein coding peptides. No enrichment of hydropobic amino acids was observed, unlike in previous studies of non canonical ORFs (Aspden et al., 2014; Wacholder et al., 2023).

3.3.3.2 LncRNA peptide synthesis is supported by mass spectrometry evidence

To further validate the translation of the neuronal lncRNA smORFs, analysis of publicly available mass spectrometry data was performed by Dr Elton Vasconcelos, as described in Section 3.2.5. Two proteomics datasets were used from undifferentiated SH-SY5Y cells (PXD010776, Murillo et al., 2018) and RA treated SH-SY5Y cells (PXD014381, Brenig et al., 2020) to search for evidence of peptides from the 45 lncRNA smORFs, 5 dORFs and 71 uORFs. Evidence was found for 18% of lncRNA smORF peptides, 8% of uORFs and 40% of dORFs (Table 3.1). This relatively low level of support is to be expected, as the small size of these peptides can hinder their detection via mass spectrometry. Interestingly, some lncRNA smORF peptides which were only translated in Control conditions were found in RA-treated SH-SY5Y cells, and vice versa. This suggests that lncRNAs are more actively translated that our analysis reveals, perhaps due to the stringent framing cut-offs used in Ribotaper (v1.3) (Calviello, Mukherjee, et al., 2016).

Since this analysis was completed, a community effort to standardise ORFs detected in Ribo-seq studies has been released which pooled ORFs from 7 Ribo-seq studies, and examined 16 mass spectrometry datasets from a wide range of human cell lines, tissue types and disease states for supporting evidence (Mudge et al., 2022). In this pooled Ribo-Seq dataset, 35/45 lncRNAs smORFs were present either as an identical sequence, or as a "smORF isoform", originating from the same gene and identical over the majority of their sequences. Evidence for 19/35 of these smORFs was found in mass spectrometry data (Table 3.2), bringing the total percentage of lncRNA smORF peptides with mass spectrometry evidence to 66%. This is a strong indication that these smORFs are indeed undergoing active translation and the peptide remains in the cell, as opposed to undergoing immediate degradation. Further, this indicates that the translation of the majority of the lncRNA smORFs is not specific to human neuronal cells.

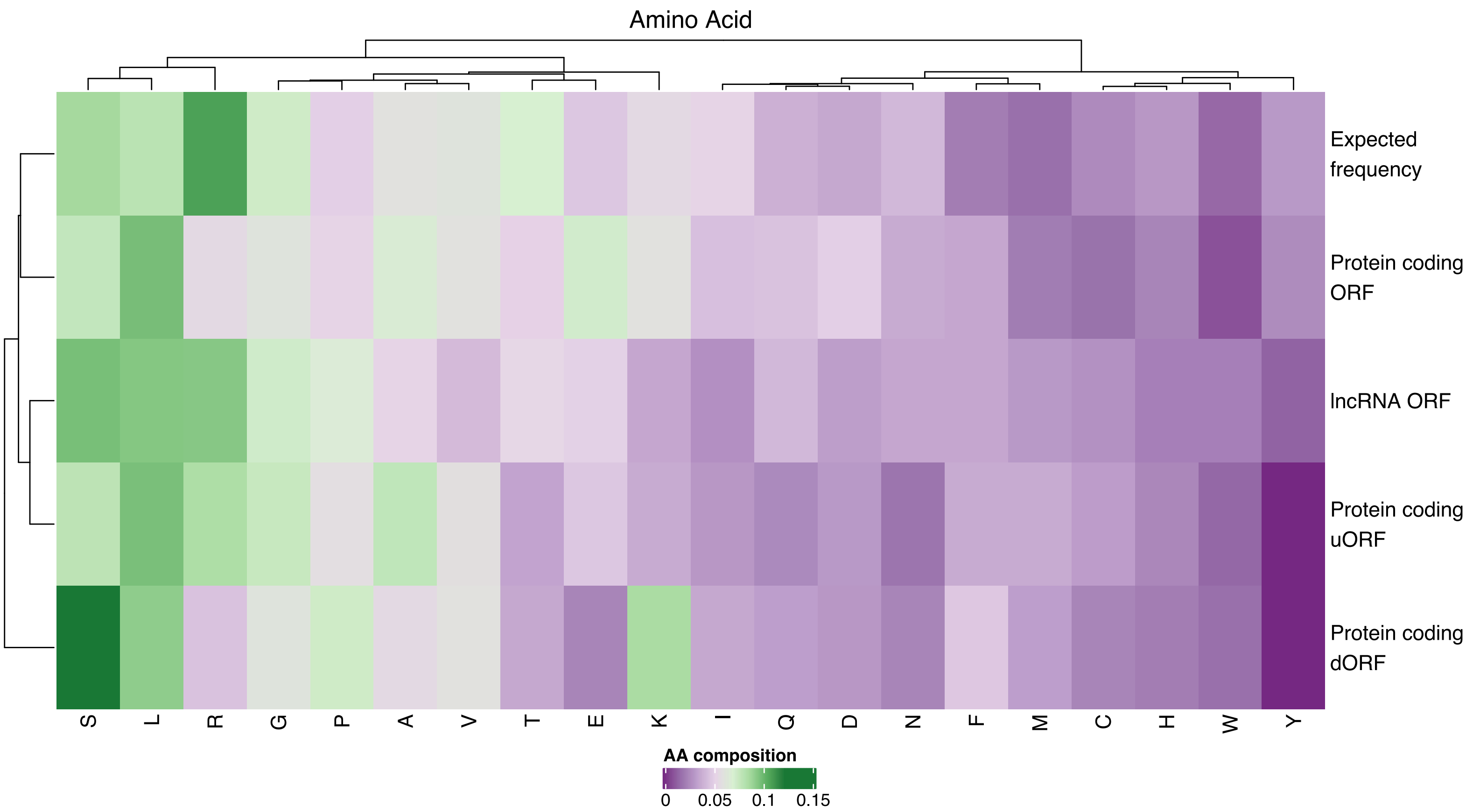


Figure 3.7: **Heatmap of average amino acid usage.** Amino acid usage of lncRNA smORF peptides compared to canonical, dORF, and uORF proteins, and the expected amino acid frequency by chance. Rows show types of ORFs and the expected frequency by chance. Columns show amino acids. The trees show the clustering of the rows and columns. Average frequencies of amino acids in each category are indicated on a scale from purple to green, where purple indicates a low frequency.

Table 3.1: **Non canonical peptides found in mass spectrometry datasets from SH-SY5Y cells.** ORF ID, ORF Type, and Transcript name denote the ORF, type of non-coding ORF, and transcript that each peptide originates from. Datasets denote the proteomic dataset each peptide was detected in (Murillo et al., 2018; Brenig et al., 2020).

ORF ID	ORF Type	Transcript Name	Replicate	Datasets
ENST00000429940.6_97_595	lncRNA	LINC00839-202	RA replicate 2,3	Undifferentiated SH-SY5Y cells, RA treated SH-SY5Y cells
ENST00000609803.2_330_426	lncRNA	AC008124.1-201	RA replicate 1,3	Undifferentiated SH-SY5Y cells
ENST00000526036.1_1226_2042	lncRNA	AP001372.2-201	Control replicate 1,2,3; RA replicate 3	RA treated SH-SY5Y cells
ENST00000518942.1_154_277	lncRNA	AC064807.1-20	Control replicate 1,2	RA treated SH-SY5Y cells
ENST00000603633.2_174_258	lncRNA	LINC00221-201	Control replicate 2,3	RA treated SH-SY5Y cells
ENST00000651711.1_64_313	lncRNA	AC068616.1-201	RA replicate 2,3	RA treated SH-SY5Y cells
ENST00000602414.5_453_561	lncRNA	SNHG8-201	Control replicate 2; RA replicate 2, 3	RA treated SH-SY5Y cells
ENST00000449419.5_342_636	lncRNA	ENTPD1-AS1-207	Control replicate 1,2,3	RA treated SH-SY5Y cells
ENST00000377694.2_904_967	dORF	IGFBPL1-201	Control replicate 2,3	Undifferentiated SH-SY5Y cells
ENST00000597781.5_258_507	dORF	SMIM7-215	Control replicate 1,2; RA replicate 3	Undifferentiated SH-SY5Y cells
ENST00000003912.7_13_217	uORF	NIPAL3-201	RA replicate 2,3	Undifferentiated SH-SY5Y cells
ENST00000334746.10_220_316	uORF	BTBD7-202	RA replicate 2,3	Undifferentiated SH-SY5Y cells
ENST00000370766.8_29_107	uORF	ZNF75D-202	Control replicate 1,2, 3; RA replicate 3	Undifferentiated SH-SY5Y cells
ENST00000371103.8_22_163	uORF	LCOR-204	RA replicate 2,3	Undifferentiated SH-SY5Y cells
ENST00000261558.8_238_319	uORF	AP5M1-201	Control replicate 1,2; RA replicate 2, 3	RA treated SH-SY5Y cells
ENST00000368194.7_18_144	uORF	ARHGEF11-202	RA replicate 2,3	RA treated SH-SY5Y cells

Table 3.2: **LncRNA smORF peptides found in mass spectrometry datasets.** ORF ID and Transcript name denote the ORF and transcript that each peptide originates from. Datasets denotes the cell or tissue types each peptide was detected in, in the GENCODE community study (Mudge et al., 2022; Prensner, Enache, et al., 2021; van Heesch, Witte, et al., 2019; Chong et al., 2020; Calviello, Mukherjee, et al., 2016).

ORF ID	Transcript name	Datasets
ENST00000419422.1_379_634	AC006329.1-201	Re-analysis of multiple human datasets
ENST00000609803.2_330_426	AC008124.1-201	Heart tissue/cells; Re-analysis of multiple human datasets
ENST00000499459.2_92_287	AC008966.1-203	Re-analysis of multiple human datasets
ENST00000590750.1_25_283	AC020928.2-201	Re-analysis of multiple human datasets
ENST00000526036.1_1226_2042	AP001372.2-201	HEK293 cells; Heart tissue/cells; Melanoma cell lines, lung cancer samples
ENST00000440088.5_140_317	APTR-204	Melanoma cell lines, lung cancer samples
ENST00000501177.7_136_388	CRNDE-201	Heart tissue/cells
ENST00000625445.1_752_881	EBLN3P-202	Heart tissue/cells
ENST00000424349.1_116_296	FGD5-AS1-202	Heart tissue/cells; Re-analysis of multiple human datasets
ENST00000515376.5_746_851	HAND2-AS1-235	Heart tissue/cells
ENST00000603633.2_174_258	LINC00221-201	Melanoma cell lines, lung cancer samples
ENST00000295549.8_609_822	LINC01116-201	Melanoma cell lines, lung cancer samples
ENST00000453910.5_151_262	MIR99AHG-209	Heart tissue/cells
ENST00000454935.1_477_633	OLMALINC-201	Re-analysis of multiple human datasets
ENST00000641571.1_548_773	OLMALINC-204	Re-analysis of multiple human datasets
ENST00000437621.6_279_426	PSMG3-AS1-201	Heart tissue/cells
ENST00000602414.5_453_561	SNHG8-201	Heart tissue/cells; Re-analysis of multiple human datasets
ENST00000504792.6_74_230	THAP9-AS1-204	Heart tissue/cells
ENST00000566220.2_141_570	TUG1-205	Melanoma cell lines, lung cancer samples

3.3.3.3 Few lncRNA peptides contain known domains

Although some studies have found a high abundance of transmembrane domains in translated smORFs and *de novo* coding genes (Carvunis et al., 2012; Aspden et al., 2014; Vakirlis, Acar, et al., 2020), others didn't identify any known domains in *de novo* coding sequences (Murphy and McLysaght, 2012). To examine the annotated domain content of the translated lncRNA peptides, they were compared to the Pfam-A HMM library (Mistry, Bateman, and Finn, 2007), where each HMM represents a protein family or domain. Of the 45 lncRNA smORFs, only 1 returned a Pfam HMM (Mistry, Bateman, and Finn, 2007). 2/71 uORFs also returned HMMs, and none were found in the translated dORFs. The domain returned by the lncRNA smORF peptide was a helix-turn-helix (HtH) domain, a common DNA-binding motif (Alberts et al., 2002). That so few lncRNA smORFs contained known domains is to be expected, given that the median size of the lncRNA peptides is 54 aa, while the median length of the 19,179 HMM models in Pfam-A v34 (Mistry, Bateman, and Finn, 2007) is 128 aa.

3.3.4 Characterisation of cytoplasmic translated lncRNA transcripts

3.3.4.1 Translated lncRNAs contain more exons than de novo coding genes

Compared to annotated protein coding genes, a larger proportion of *de novo* genes are single exon (Wu, Irwin, and Zhang, 2011; Neme and Tautz, 2013; Zhao et al., 2014). In order to compare the translated lncRNA transcripts to translated protein coding transcripts and untranslated lncRNA transcripts, their exon distributions were visualised (Figure 3.8). The translated lncRNA population had a range of 1-6 exons, with a median of 3, more than the average for *de novo* genes. There was a significant difference between the translated lncRNA transcripts and the translated protein coding transcripts (median = 11) ($p = 1.76e^{-}19$), and no signifcant difference between the translated lncRNAs and other annotated lncRNA transcripts (Gencode v30, Frankish et al., 2021). However, the untranslated lncRNA population likely contains lncRNAs which are translated in other cell types, tissues, or developmental periods, so firm conclusions cannot be drawn from this particular comparison.

3.3.4.2 LncRNAs contain significantly more stop codons than protein coding transcripts

The basic requirement to form an ORF is an in frame start and stop codon pair. To establish whether the 45 translated lncRNAs were enriched for canonical start (ATG) or stop codons (TAA, TAG, TGA), the codon content of protein coding, untranslated, and translated lncRNA transcripts was compared. In human CDS sequences approximately 2.2 codons per 100 nt are ATGs (Athey et al., 2017), and given that many protein coding ORFs contain multiple downstream ATGs, we may expect the full transcript sequence to contain a similar proportion of ATGs or fewer. All categories of transcript had a median of 1.7 ATG start codons per 100 nucleotides (Figure 3.9), and a Kruskal-Wallis test revealed no significant effect of transcript type on number of start codons per nucleotide (H(2) = 0.07, p = .97).

As stop codons are selected against across the ORFs of protein coding transcripts to prevent truncated protein products, we expect protein coding transcripts will contain fewer stop codons than lncRNA transcripts. Translated lncRNA smORFs may also contain fewer stop codons than the wider lncRNA population if the peptide product is functional. Comparison of the transcript categories revealed that protein coding transcripts had the lowest median stop codons per 100 nt (3.59) (Figure 3.10), and transcript type had a significant effect on number of stop codons per 100 nucleotides (H(2) = 5844.49 p = .00). The protein coding transcript population was significantly different from both untranslated (p > 0.00) and translated lncRNA ($p = 1.11e^{-0.6}$). This indicates a greater selection pressure against stop codons in protein coding transcripts than lncRNAs. The lack of significant difference between the untranslated and translated lncRNA transcripts (p = 0.36) suggests that this

Table 3.3: **Protein domains identified in non canonical peptides.** ORF ID and Transcript name denote the ORF and transcript that each peptide originates from. ORF Type is the category of each non-coding ORF. HMM name is the hidden markov model returned from the Pfam-A database (Mistry, Bateman, and Finn, 2007). Type is the type of HMM returned, where a domain is structural unit and a family is a collection of related protein regions.

ORF ID	Transcript name	ORF Type	HMM name	Type
ENST00000526036.1_1226_2042	AP001372.2-201	lncRNA	Helix-turn-helix Tc5 transposase DNA-binding domain	Domain
ENST00000470557.2_1570_2266	PTRH2-203	uORF	CENP-B N-terminal DNA-binding domain	Domain
ENST00000470557.2_1570_2266	PTRH2-203	uORF	Helix-turn-helix Tc5 transposase DNA-binding domain	Domain
ENST00000282869.10_1106_1187	ZNF117-201	uORF	KRAB box	Family

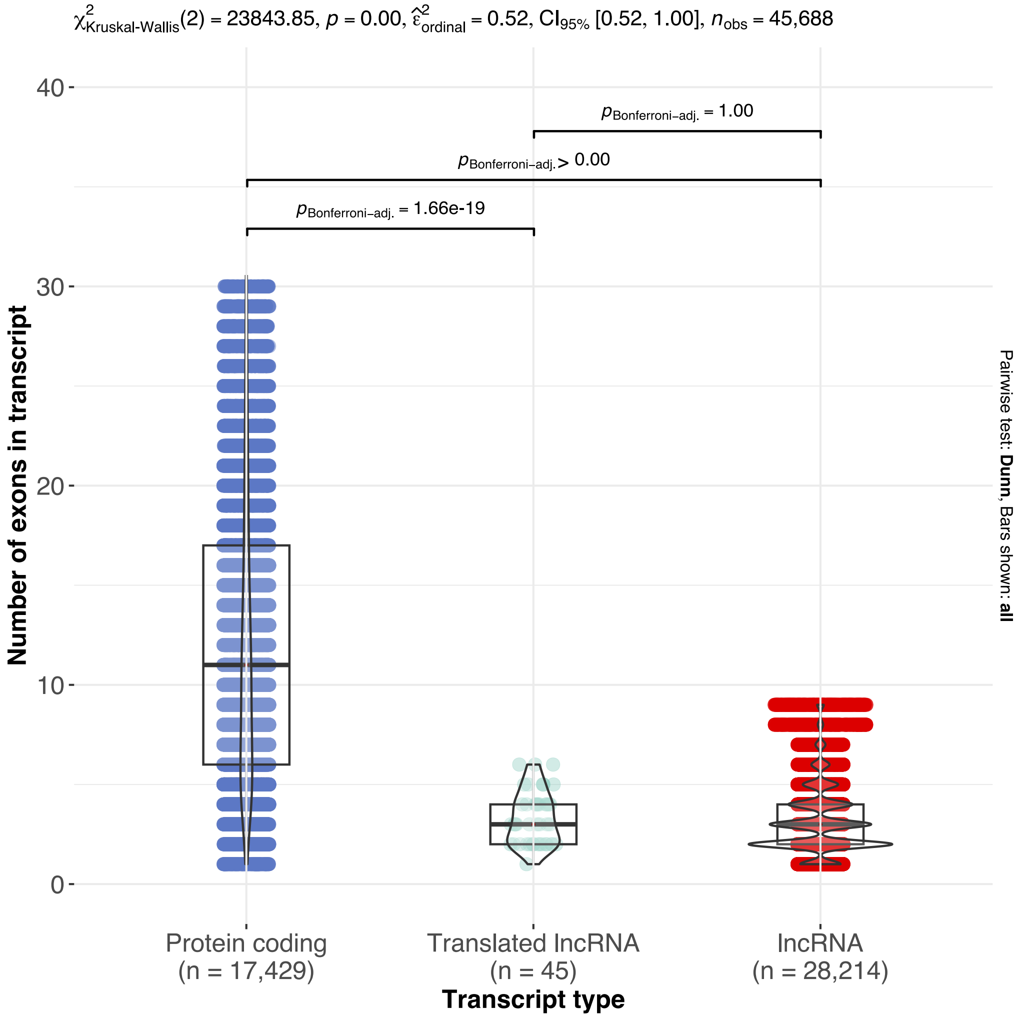


Figure 3.8: **Number of exons in transcripts.** The x axis shows categories of transcripts, and n denotes the number of transcripts. The y axis shows the number of exons in the transcripts. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test. A ylim of 30 was used to crop the figure due to a high number of protein coding outliers; the full figure is in Appendix B, Figure B.3

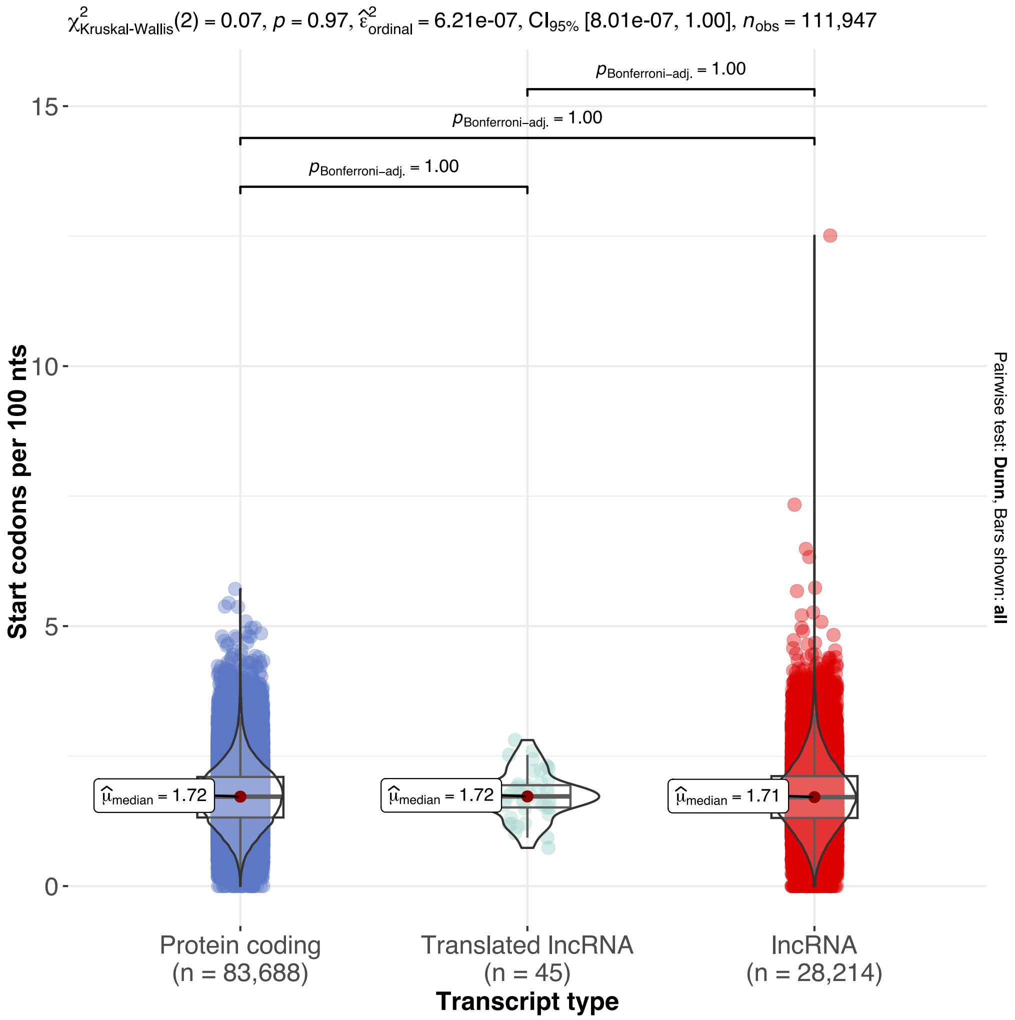


Figure 3.9: **Canonical start codons (ATG) per 100 nucleotides in protein coding, untranslated lncRNA, and translated lncRNA transcripts.** The x axis shows transcript types, and the y axis shows start codons per nucleotide. The violin plots display the distribution of start codons per 100nt, and are overlaid with box plots to display the interquartile ranges. The red dots denote the median values. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test. Transcript sequences used from human reference genome Release 30 (GRCh38.p12), Gencode (Frankish et al., 2019)

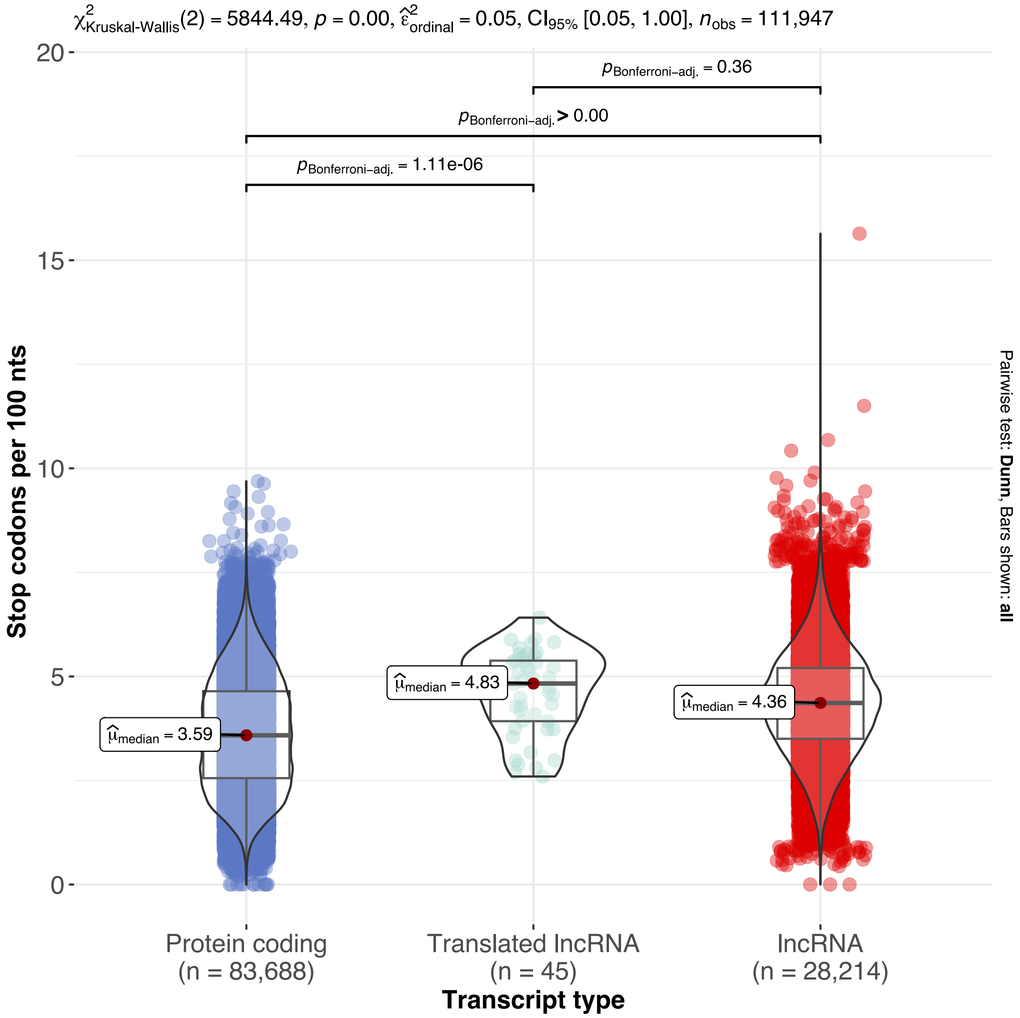


Figure 3.10: **Canonical stop codons (TAA,TAG,TGA) per 100 nucleotides in protein coding, untranslated lncRNA, and translated lncRNA transcripts.** The x axis shows transcript types, and the y axis shows start codons per nucleotide. The violin plots display the distribution of stop codons per 100nt, and are overlaid with box plots to display the interquartile ranges. The red dots denote the median values. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test. Transcript sequences used from human reference genome Release 30 (GRCh38.p12), Gencode (Frankish et al., 2019)

selection pressure is not acting on the translated lncRNA population. However, as in the previous exon number comparison, the untranslated lncRNA population likely contains lncRNAs which are translated in other conditions, so firm conclusions cannot be drawn from this particular comparison.

3.3.4.3 Translated lncRNAs are A/T-rich

To investigate GC content in the translated lncRNA, their distribution was compared to 100 randomly selected translated protein coding transcripts, and 100 randomly selected untranslated lncRNAs. The GC-richness of the genome regions these transcripts are in was not accounted for by this analysis. The translated protein coding transcripts had a median GC content of 51%, compared to 45% in translated and untranslated lncRNA (Figure 3.11). A Kruskal-Wallis test found that transcript type had a significant effect on GC content ($p = 4.54e^{-}0.4$). If the translated lncRNAs do indeed represent *de novo* genes, this aligns with hypotheses that *de novo* genes are AT-rich (Ruiz-Orera, Messeguer, et al., 2014).

3.3.5 Characterisation of the transcriptional landscape of translated lncRNAs

Compared to mRNA, lncRNAs are generally expressed at lower levels, but in a more tissue, cell type and developmental stage specific manner (Jandura and Krause, 2017). This low, specific expression has also been observed in *de novo* human coding sequences (Vakirlis, Vance, et al., 2022). This section therefore aims to establish if the transcription of the translated lncRNAs differs from the wider human lncRNA population.

Following informal discussions with leaders in the field, we decided to expand the following analysis to include all translated smORFs in all replicates, thus removing the requirement that a smORF be identified in at least two of the three biological replicates of a given condition. The larger "non-stringent" dataset contains 355 translated lncRNA smORFs, from 288 lncRNA transcripts, in 208 genes, across Control and RA conditions. This increases the likelihood of detecting transcription of a translated lncRNA in other samples, given the specificity of lncRNA expression.

3.3.5.1 Translated lncRNAs are significantly upregulated in human brain tissue

To establish whether the expression of the translated lncRNA transcripts differs from the wider human lncRNA population in the brain, data from the human brain evolution focused arm of the PsychENCODE consortium (Zhu, Sousa, et al., 2018; Akbarian et al., 2015) was analysed. These data included RNA-seq spanning 16 brain regions from prenatal and postnatal rhesus macaque (*Macaca mulatta*) brains, matched to equivalent regions and developmental stages in existing human brain data (Li, Santpere, et al., 2018). In particular, samples from the frontal cortex, hippocampus, and striatum were included, areas of the brain which have dopaminergic innervation, corresponding to the dopaminergic markers expressed by the SH-SY5Y cell line (Forster et al., 2016). The XSAnno framework (Zhu, Li, et al., 2014) had been applied to these data to create common annotation sets of homologous genes between human and macaque.

Of the 208 translated lncRNA genes, 105 were in the dataset, corresponding to 166 lncRNA smORFs. To compare the expression profiles of translated neuronal lncRNAs to untranslated lncRNAs, 25 random sets of untranslated 105 lncRNA genes were selected in each sample and tissue type. In all tissue types in human (Table 3.4) and macaque (Appendix B, Table B.1), gene set had a significant effect on RPKM. The expression of the translated gene set (Set 1) was significantly different from the different from the randomly selected gene sets, with a higher expression level in all tissue types. For example in a representative sample from the human hippocampus (Figure 3.12), the translated lncRNA set had a median RPKM of 1.69,

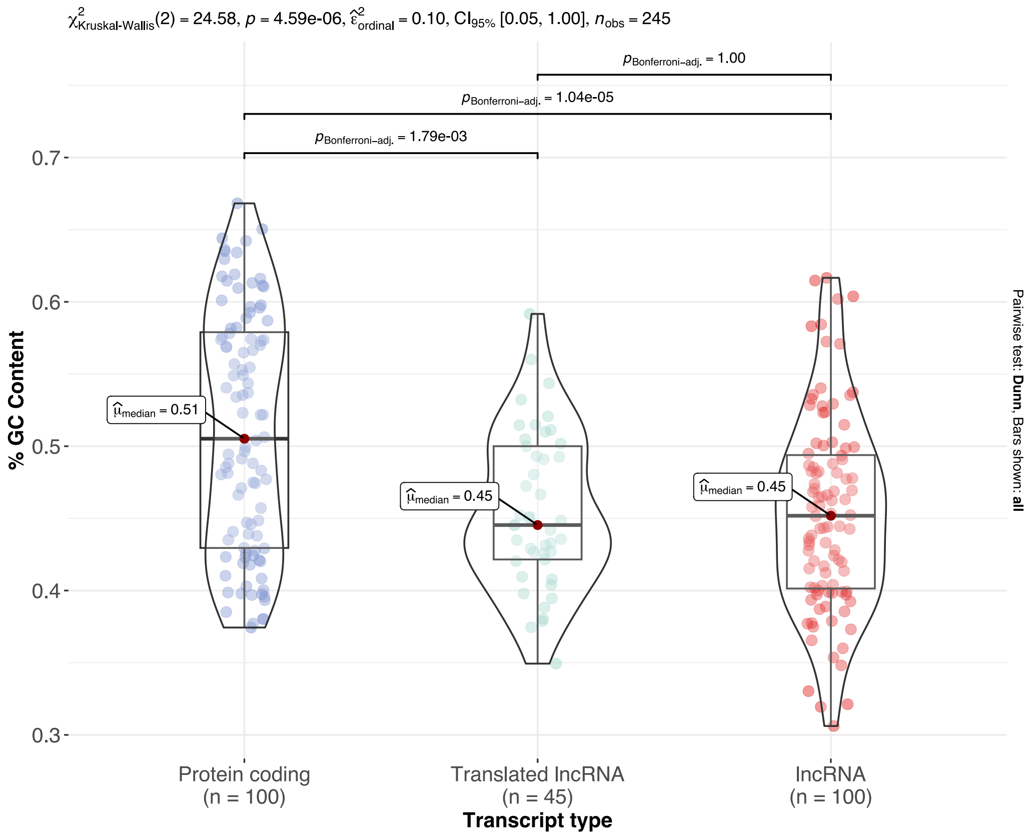


Figure 3.11: **GC content in protein coding, untranslated lncRNA, and translated lncRNA transcripts.** Comparison of 100 random translated protein coding transcripts, translated lncRNA, and 100 random untranslated lncRNAs. The x axis shows transcript types, and the y axis shows percentage GC content. The violin plots display the distribution of percentage GC content, and are overlaid with box plots to display the interquartile ranges. The red dots denote the median values. A Kruskal-Wallis test was used to compare the groups. Transcript sequences used from human reference genome Release 30 (GRCh38.p12), Gencode (Frankish et al., 2019)

Table 3.4: **Expression of translated lncRNA genes in the human brain.** Tissue denotes the area of the brain analysed, and number of samples denotes the number of individuals samples were taken from. Kruskal-Wallis indicates if a Kruskal-Wallis test found that the lncRNA gene set had a significant effect on RPKM in each sample, and p-values summarises the range of p-values from each sample. Multiple comparison describes whether a multiple comparison test found that the levels of translated lncRNAs (RPKM) were significantly different from randomly selected lncRNAs, and in what direction. RPKM values from the PsychENCODE consortium (Zhu, Sousa, et al., 2018; Akbarian et al., 2015).

Tissue	No. of samples	Kruskal-Walls	P-values	Multiple Comparison
Hippocampus (HIP)	28	Significant	<2.2e-16	True, higher
Medial prefrontal cortex (MFC)	29	Significant	<2.2e-16	True, higher
Dorsolateral prefrontal cortex (DFC)	29	Significant	e-13 or smaller	True, higher
Orbital prefrontal cortex (OFC)	29	Significant	e-13 or smaller	True, higher
Ventrolateral prefrontal cortex (VFC)	31	Significant	e-12 or smaller	True, higher
Amygdala (AMY)	25	Significant	<2.2e-16	True, higher
Striatum (STR)	23	Significant	<2.2e-16	True, higher
Primary motor cortex (M1C)	25	Significant	e-13 or smaller	True, higher
Primary somatosensory cortex (S1C)	24	Significant	e-14 or smaller	True, higher
Inferior posterior parietal cortex (IPC)	30	Significant	e-14 or smaller	True, higher
Primary auditory cortex (A1C)	29	Significant	e-15 or smaller	True, higher
Superior temporal cortex (STC)	30	Significant	e-16 or smaller	True, higher
Inferior temporal cortex (ITC)	26	Significant	<2.2e-16	True, higher
Primary visual cortex (V1C)	29	Significant	<2.2e-16	True, higher
Mediodorsal nucleus of thalamus (MD)	23	Significant	<2.2e-16	True, higher
Cerebellar cortex (CBC)	29	Significant	e-13 or smaller	True, higher

compared to medians of 0 to 0.25 in the randomly selected untranslated sets. This suggests that the translated lncRNA set are transcriptionally upregulated in the human and macaque brain, more so than the known enrichment of lncRNA expression in the brain (Jandura and Krause, 2017).

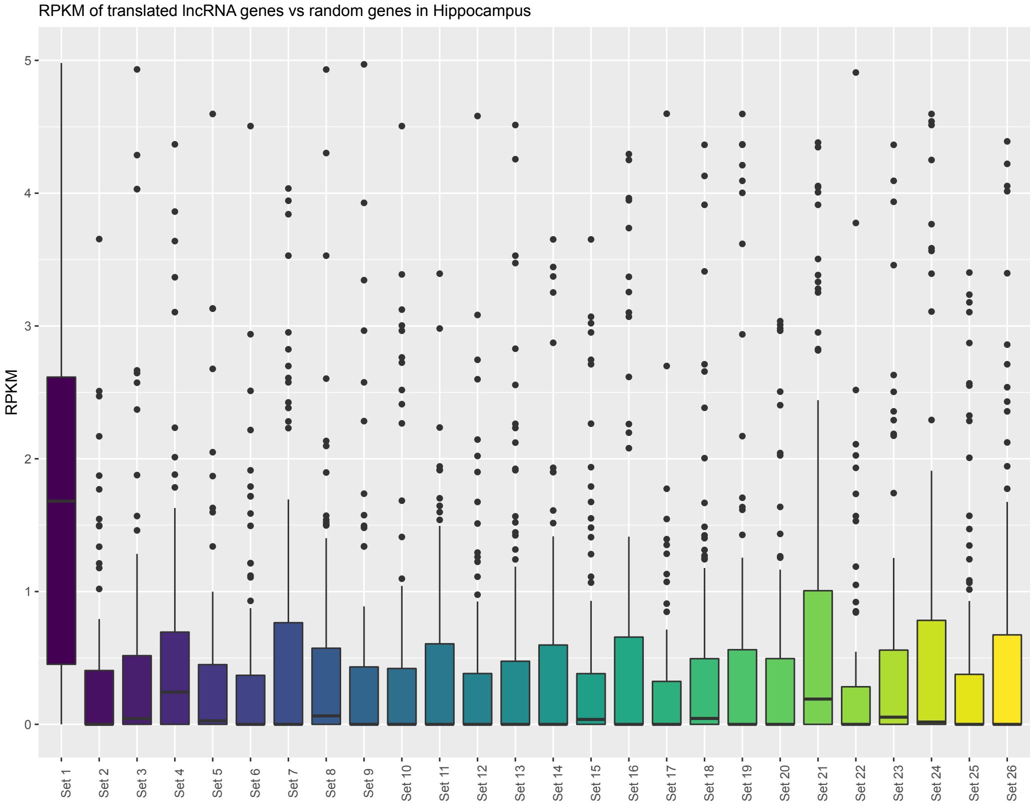


Figure 3.12: **RPKM of lncRNA genes in human hippocampus sample.** The x-axis shows sets of human lncRNA genes, where Set 1 corresponds to the 105 translated lncRNA genes, and Sets 2 - 26 correspond to 105 randomly selected lncRNA genes. The y-axis shows the distribution of RPKM for genes in this set. RPKM values from the PsychENCODE consortium (Zhu, Sousa, et al., 2018; Akbarian et al., 2015), human sample HSB98.

3.4 Discussion

This chapter has classified the sub-population of translated human neuronal lncRNAs, establishing that their smORFs have ribosome footprint distributions and translation efficiencies comparable to that of protein coding ORFs, in line with the literature (Ruiz-Orera, Messeguer, et al., 2014; Patraquim et al., 2022). From the "stringent" set of lnRNA smORFs, 35/45 were present in other human Ribo-Seq datasets, either as an identical sequence, or as a "smORF isoform". No differential lncRNA translation was observed, however this is likely an artefact of the data as only two protein coding ORFs were found to have significantly different TE between Control and RA conditions.

Translated lncRNA smORFs are shorter than canonical protein coding ORFs, and the peptides synthesised from these smORFs exhibit amino acid compositions similar to that of canonical proteins, but tend not to contain known domains. Unlike previous studies, no enrichment for hydrophobic amino acids or transmembrane domains was observed (Aspden et al., 2014; Wacholder et al., 2023). Mass spectrometry evidence was found for 66% of the "stringent" set of 45 translated lncRNA smORFs, from a range of human tissues including heart tissue/cells, melanoma cell lines and lung cancer samples.

The translated lncRNA transcripts contained more exons than *de novo* genes, and significantly fewer than protein coding genes. They also contained similar proportions of start codons as in protein coding transcripts, with the same median of 1.72 ATG codons per 100 nts. Translated lncRNAs did however contain significantly more stop codons per 100 nts than protein coding transcript, indicative of a greater selection pressure against stop codons in the protein coding transcript population. Given that protein coding and translated lncRNA transcripts contained similar proportions of start codons, this enrichment for stop codons is likely responsible for a large portion of the AT-richness of translated lncRNAs compared to protein coding transcripts.

The expression of the wider "non stringent" set translated lncRNAs was significantly enriched in human and macaque brain tissue, throughout development time points from pre to postnatal, although only 105/208 translated transcripts were present in the dataset.

Although models have been proposed for the *de novo* emergence of protein coding genes from non coding sequence (Carvunis et al., 2012), the process remains poorly understood. A growing number of studies have suggested that lncRNAs are a source of *de novo* protein coding sequence (Ruiz-Orera, Messeguer, et al., 2014; Chen, Shen, et al., 2015; Couso and Patraquim, 2017; Sandmann et al., 2023). The characteristics of the translated human neuronal smORFs explored in this chapter correspond to many of those observed in *de novo* protein coding sequences including low expression, and a lack of known protein domains in the resulting peptide. Some features of the smORFs were "further" along the evolutionary continuum (Figure 1.7) of non-coding to coding sequence, containing more exons than *de novo* genes, similar proportions of start codons, and producing stable peptides that could be detected using mass spectrometry. However, the majority of these analyses were on a small subset of translated lncRNAs, identified in one cell line in humans. Continued work to understand the wider translatome in a wide range of tissues, developmental time points and species will be required to build a clearer picture of how lncRNAs relate to *de novo* coding sequences.

A missing aspect of the analysis of start codon enrichment, and of ORF identification throughout this thesis, is the consideration of non AUG start codons. Translation in eukaryotes can also initiate at start codons which differ by 1nt from AUG, with lower efficiency (Kozak, 1989; Peabody, 1989). Translation from non AUG start codons not only expands the non-canonical proteome, but also regulates the translation of canonical protein codings ORFs in a similar manner to uORFs (Andreev et al., 2022). Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016) and the subsequently released software ORFquant (Calviello, Hirsekorn, and Ohler, 2020) strictly focus on ORFs using the canonical AUG start codon. Future methods may look to define possible start sites by considering the full context of the translation initiation site, to further expand our understanding of lncRNA smORF translation.

Although work is ongoing to expand our knowledge of the human translatome by Ribo-Seq analysis of primary biological material (Chothani, Adami, Widjaja, et al., 2022), the translational events described in this thesis were identified in SH-SY5Y cells. Though a useful model for human neuronal differentiation, SH-SY5Y cells share limitations with a large portion of Ribo-Seq studies performed using cell lines; the biology altering mutations required to create and maintain a cell line, genetic changes over multiple passages, and potential for contamination (Marx, 2014). All of these changes can create a transcriptome and translatome which do not reflect the biological reality of primary material. It is therefore reassuring that the translated lncRNAs are significantly enriched compared to untranslated lncRNAs based on analysis of primary human brain tissue, and that Ribo-Seq and mass spectrometry evidence from samples which include primary human tissues supports their active translation.

Alongside other publically available RNA-Seq, the expression data used in this chapter could be further analysed by examining the expression of individual translated lncRNAs across the developmental timepoints in the brain and other tissues. Although the lncRNA smORFs were identified in SH-SH5Y cells, this does not mean that their translational activity is specific to neuronal differentiation, partially given that lncRNA expression is enriched in the brain, second only to the testes in human (Jandura and Krause, 2017). Combined with Ribo-seq data from other tissues, it may be possible to begin to understand whether the lncRNAs are functioning purely at the RNA level, at the peptide level, or both, and whether this varies across time and tissue type.

3.4.1 Conclusions

In summary, this chapter has established that the translated lncRNAs and their smORFs exhibit characteristics between those of coding and non-coding sequence, much like *de novo* ORFs and other translated lncRNAs identified in the literature. The next chapter will look to test our hypothesis that lncRNAs are sources of functional peptides by investigating their sequence conservation in and beyond primates, using a sequence based approach which focuses on the translated smORF.

Chapter 4

Translated lncRNA smORFs exhibit sequence conservation across species

4.1 Introduction

Evolutionary sequence conservation has long been considered indicative of coding potential and even peptide function. This chapter investigates the levels of sequence conservation across a sample of species for the translated lncRNA smORFs and their corresponding peptide products described in Chapter 3.

In order to identify lncRNAs that are conserved across species, measures beyond direct sequence conservation are often employed. These approaches include the analysis of syntenic regions, secondary and tertiary RNA structures, splicing patterns, and function. In lncRNAs that exhibit sequence conservation across species, it is generally modular, and shorter than we observe in mRNA (Johnsson et al., 2014); for example as in the lncRNAs Xist and HOTAIR (Johnsson et al., 2014; Brockdorff, 2018). Nevertheless, the majority of lncRNAs are considered to be lineage specific, with less than 3% of human lincRNAs conserved in one or more non-primate mammals (Perez-Riverol et al., 2022).

Further, although lncRNAs are often found to contain potential ORFs, they are generally smORFs of 100 codons or fewer. This length is at the limit of many sequence similarity search tools. Despite this limitation, an increasing number of conserved, translated lncRNAs have been described in the literature. One such example is LINC00961, a polyadenylated lncRNA encoding a 90 aa peptide; small regulatory polypeptide of amino acid response (SPAR/SPAAR) (Matsumoto, Pasut, et al., 2017). SPAR is conserved in primates, and is syntenous in mouse with 65% aa sequence identity between the mouse and human peptides, including a conserved N-terminal transmembrane domain (Matsumoto, Pasut, et al., 2017; Spiroski et al., 2021). Functional analysis of SPAR has revealed a role in suppressing activation of mammalian target of rapamycin complex 1 (mTORC1) in response to amino acids in mouse. Another example of a lncRNA containing a translated smORF is LINC00948, which encodes the 46 aa peptide myoregulin (MLN) (Anderson et al., 2015). Initially discovered in human, mouse and rat, MLN is conserved in a wide range of mammalian species (Lu et al., 2020). MLN is structurally similar to phospholamban (PLN) and sarcolipin (SLN), type II single-pass transmembrane proteins involved in regulation of muscle performance and cardiovascular disease, and shares their function as a regulator of SERCA (Nelson et al., 2016).

Despite these difficulties, the majority of known functional small peptides in human are conserved in other species (Choi, Kim, and Nam, 2019). This reflects a bias in research efforts, based on the aforementioned assumption that conservation indicates function. Therefore, although any sequentially conserved lncRNA smORFs we identify from those described in Chapter 2 may represent strong targets for *in vitro* study, all of the translated lncRNA smORFs should be considered as potentially biologically important.

In this Chapter, we build upon previous analyses where we found evidence of sequence conservation in *Hominidea* in 17 of the 45 smORFs in the stringent set (Douka, Birds, et al., 2021). Here, the analysis is extended to include all 355 translated lncRNA smORFs, a greater range of species with broader divergence times, and more stringent filtering and analysis of results. Using a combination of BLAST (Altschul et al., 1990) search strategies which focus on the smORF sequence of the lncRNA, the aim of this Chapter was to find small regions of sequence conservation, in line with the previously observed modular lncRNA sequence conservation. Historically, sequence conservation has often been identified after the functions of candidate lncRNAs had been elucidated. For example Xist, one of the first lncRNAs to be discovered in 1991 (Brown, Ballabio, et al., 1991), was identified in human as part of the X-chromosome inactivation centre, and subsequently also in mouse due to syntenic conservation (Brockdorff et al., 1991). Upon alignment of the human and mouse Xist, conserved sequences were identified in 5' tandem repeats (Brown, Hendrich, et al., 1992). Xist has since been found in all placental mammals, with varying levels of sequence conservation.

This sequence based approach is not often used for lncRNAs due to their perceived lack of protein coding potential, and is not effective for particularly small smORFs as it is unlikely to return results which pass the necessarily stringent e-value cutoff. However, using a dataset of translated smORFs allows conservation to be investigated at the amino acid level, thus eliminating the noise of synonymous substitutions. For example, the translated smORFs in *Drosophila sarcolamban* were found to be conserved from flies to vertebrates using a combination of tBLASTn searches, phylogenetically informed consensus sequence analysis, and structural homolog searches (Magny et al., 2013)

Given the low levels of conservation expected in lncRNAs, closely related species from *Hominoidea* and the wider Primates, as well as a small number of other mammals with reference quality genomes were selected for this analysis (Figure 4.1). *Gallus gallus* (chicken) was included as a non-mammal outgroup, and *Monodelphis domestica* (opossum) as a representative of *Marsupialia*. The phylogenetic relationships of these species are robust and well resolved (Perelman et al., 2011; Meredith et al., 2011), meaning the presence and absence of conserved translated smORFs throughout the phylogeny can be used to build strong hypotheses about their evolutionary origins.

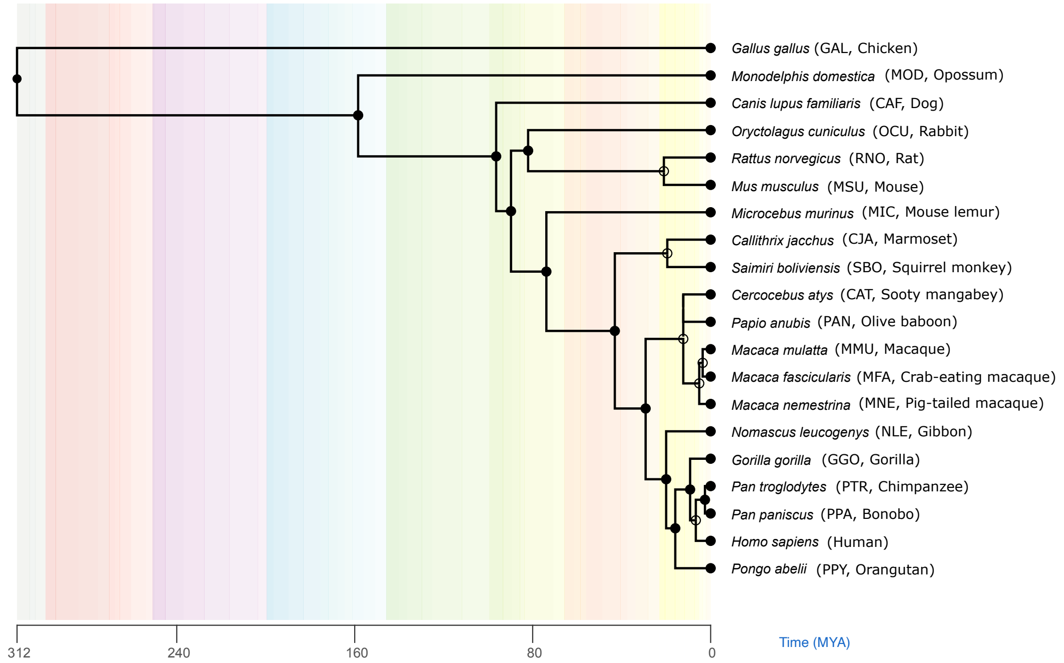

Gallus gallus (GAL, Chicken)
Monodelphis domestica (MOD, Opossum)
Canis lupus familiaris (CAF, Dog)
Oryctolagus cuniculus (OCU, Rabbit)
Rattus norvegicus (RNO, Rat)
Mus musculus (MSU, Mouse)
Microcebus murinus (MIC, Mouse lemur)
Callithrix jacchus (CJA, Marmoset)
Saimiri boliviensis (SBO, Squirrel monkey)
Cercocebus atys (CAT, Sooty mangabey)
Papio anubis (PAN, Olive baboon)
Macaca mulatta (MMU, Macaque)
Macaca fascicularis (MFA, Crab-eating macaque)
Macaca nemestrina (MNE, Pig-tailed macaque)
Nomascus leucogenys (NLE, Gibbon)
Gorilla gorilla (GGO, Gorilla)
Pan troglodytes (PTR, Chimpanzee)
Pan paniscus (PPA, Bonobo)
Homo sapiens (Human)
Pongo abelii (PPY, Orangutan)
312
240
160
80
0
Time (MYA)

Figure 4.1: **Phylogeny of species included in conservation analysis.** 19 species with reference quality genome sequences were selected from Ensembl 104 (Cunningham et al., 2022). These species included *Hominoidea* and Primates, as well as a small number of other mammals and *Gallus gallus* as outgroup. The species names, a three letter abbreviation, and the common names are given. The abbreviations will be used throughout this chapter. Estimated divergence times are shown on the X axis, as millions of years ago (MYA), the colour banding is merely to assist in differentiating different bands of time. Solid circles indicate nodes that map directly to the NCBI Taxonomy (Schoch et al., 2020), and open circles are nodes created during the polytomy resolution process described in Hedges et al., 2015. *Canis lupus familiaris* date is as per *Canis lupus*; according to Freeman et al. (Freedman and Wayne, 2017) there is less than 35 KYA difference in divergence time. Figure created in TimeTree of Life v4, using phylogeny and divergence times from the TimeTree database, which contains divergence times and timetrees from 4,075 articles (Kumar et al., 2017; Hedges et al., 2015).

4.2 Materials and Methods

4.2.1 Data sources

Translated lncRNA smORFs were identified as described in 2.2.3, without the requirement that a smORF be identified in at least two of the three biological replicates of a given condition. Translated smORFs are classified as 'smORF isoforms' if they originate from the same gene and at least 50% of the smORF sequences overlap in the same frame.

4.2.2 Species sampling

Genomes, transcriptomes and proteomes were downloaded from Ensembl 104 (Cunningham et al., 2022) to span *Hominoidea* and primates (Appendix C, Table C.1). As LncRNAs are generally poorly conserved, the sampling strategy was to deeply sample within closely related species of primates and include a small number of other species at different divergence depths within mammals and vertebrates. The estimated divergence times of all included species are shown in Figure 4.1.

In addition, to allow us to assess the depth at which conservation levels drop off, we included a small number of other mammal species with high quality genomes and annotations: *Mus musculus* (mouse) and *Canis lupus familiaris* (dog), and *Monodelphis domestica* (opossum) as a representative of marsupials. We also included the outgroup *Gallus gallus* (chicken). To account for irregularities in mouse genetics due to laboratory breeding, genomes, transcriptomes and proteomes of 16 laboratory and wild mouse strains from the Mouse Genome Project (Lilue et al., 2018) were included in the analysis. Selected species, along with their genome identifier and assembly quality, are detailed in Table 4.1.

As the translated lncRNAs were originally identified using the Gencode v30 (Frankish et al., 2021) annotation, Gencode v38 (Frankish et al., 2021) was included to confirm that no major annotation changes had occurred in the human lncRNA transcripts of interest (Appendix **??**). Gencode v30 (Frankish et al., 2021) was the final annotation to split lncRNAs into the following biotypes: 3prime_overlapping_ncRNA, antisense, bidirectional_promoter_lncRNA, lincRNA, macro_lncRNA, non_coding, processed_transcript, sense_intronic and sense_overlapping. These biotypes have been replaced by a generic lncRNA category in future annotations. The number of genes and transcripts annotated in each species transcriptome and proteome are detailed in Appendix C, Table C.1.

4.2.3 Sequence similarity searching

Three levels of sequence similarity search were performed on the set of 355 translated lncRNA smoRFs, which originate from 288 lncRNA transcripts (E-appendix; peptide_blast.sh). The BLAST (Altschul et al., 1990) suite of programs were employed as follows; i) a peptide-centric search using BLASTp v2.9.0+ (Figure 4.2), ii) a transcript-centric search using BLASTn v2.9.0+ (Figure 4.3), and iii), a smORF-centric search using tBLASTn v2.9.0+ (Figure 4.4). Details of the searches are provided in the next sections.

i) For the peptide-centric search (Figure 4.2), the amino acid sequence of each lncRNA smORF peptide was compared to a database created using the proteomes of the species listed in Table 4.1, using default settings and an e-value of 0.00001. As some query sequences returned a large number of subject hits, the pool of potential homologous peptide sequences were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. This measure is an estimate, as the BLAST alignment length includes gap characters, but was effective for removing low quality results. If no subject hits passed these filters for a given query in each searched species, the hit with the smallest e-value was retained.

Table 4.1: **Publicly available genomes used in conservation analysis.** Species sampled are provided with common and latin names. ’Source’ refers to the dataset the annotations were sourced from. ’Genome’ refers to the version of the annotations included in this analysis. ’Updated’ refers to the date the annotation was last updated. ’N50 contig length’ refers to the N50 statistic of each genome assembly, which is a measure of quality. ’Notes’ refers to any extra relevant information on the annotations. Genome sources: Gencode (Frankish et al., 2021), Ensembl 104 (Cunningham et al., 2022), and The Mouse Genome Project (Lilue et al., 2018).

Species	**Source**	**Genome**	**Updated**	**N50 contig length**	**Notes**
Human (*Homo sapiens*)	Gencode	Release 30 (GRCh38.p12)	04/2019	57,879,411	Final Gencode version to include detail on lncRNA transcript types. Version used to identify translated lncRNA smORFs.
Human (*Homo sapiens*)	Gencode	Release 38 (GRCh38.p13)	05/2021	57,879,411	Newest version at time of analysis.
Bonobo (*Pan paniscus*)	Ensembl	panpan1.1	03/2020	66,676	
Chimpanzee (*Pan troglodytes*)	Ensembl	Pan_tro_3.0	03/2020	384,816	
Gorilla (*Gorilla gorilla gorilla*)	Ensembl	gorGor4	03/2020	52,934	
Sumatran orangutan (*Pongo abelii*)	Ensembl	PPYG2	08/2012	-	
Gibbon (*Nomascus leucogenys*)	Ensembl	Nleu_3.0	12/2017	35,148	
Olive baboon (*Papio anubis*)	Ensembl	Panu_3.0	03/2020	149,817	
Sooty mangabey (*Cercocebus atys*)	Ensembl	Caty_1.0	01/2018	112,942	

Table 4.1 continued from previous page

Species	Source	Genome	Updated	N50 contig length	Notes
Pig-tailed macaque (*Macaca nemestrina*)	Ensembl	Mnem_1.0	01/2018	106,897	
Crab-eating macaque (*Macaca fascicularis*)	Ensembl	Macaca_ fascicularis_6.0	08/2020	21,344,639	
Macaque (*Macaca mulatta*)	Ensembl	Mmul_10	12/2019	46,608,966	
Bolivian squirrel monkey (*Saimiri boliviensis boliviensis*)	Ensembl	SaiBol_1.0	03/2020	38,823	
Marmoset (*Callithrix jacchus*)	Ensembl	ASM275486v1	05/2019	155,284	
Mouse lemur (*Microcebus murinus*)	Ensembl	Mmur_3.0	03/2020	210,702	
Rabbit (*Oryctolagus cuniculus*)	Ensembl	OryCun2.0	05/2019	64,648	
Mouse (*Mus musculus*)	Ensembl	GRCm39	03/2021	32,273,079	
Mouse (*Mus musculus*)	Mouse Genomes Project	129S1_SvImJ_v1	01/2018	236,538	Strain: high incidence of spontaneous testicular teratomas.
Mouse (*Mus musculus*)	Mouse Genomes Project	A_J_v1	01/2018	472,329	Strain: inbred mice widely used to model cancer and for carcinogen testing.
Mouse (*Mus musculus*)	Mouse Genomes Project	AKR_J_v1	01/2018	383,916	Strain: useful in cancer, immunology, and metabolism research.

Table 4.1 continued from previous page

Species	Source	Genome	Updated	N50 contig length	Notes
Mouse (*Mus musculus*)	Mouse Genomes Project	BALB_cJ_v1	01/2018	414,263	Strain: susceptibility to developing the demyelinating disease upon infection with Theiler’s murine encephalomyelitis virus.
Mouse (*Mus musculus*)	Mouse Genomes Project	C3H_HeJ_v1	01/2018	443,242	Strain: research areas including cancer, infectious disease, sensorineural, and cardiovascular biology research.
Mouse (*Mus musculus*)	Mouse Genomes Project	C57BL_6NJ_v1	01/2018	438,357	Strain: homozygous for Crb1rd8, the retinal degeneration 8 mutation.
Mouse (*Mus musculus castaneus*)	Mouse Genomes Project	CAST_EiJ_v1	01/2018	379,379	Strain: derived from wild mice trapped in Thailand.
Mouse (*Mus musculus*)	Mouse Genomes Project	CBA_J_v1	01/2018	468,391	Strain: research includes immunology and inflammation, metabolism, hearing and cochlear function, infectious disease, and fetal development.
Mouse (*Mus musculus*)	Mouse Genomes Project	DBA_2J_v1	01/2018	385,934	Strain: widely used inbred strain. Characteristics include low susceptibility to developing atherosclerotic aortic lesions, high-frequency hearing loss, susceptibility to audiogenic seizures.
Mouse (*Mus musculus*)	Mouse Genomes Project	FVB_NJ_v1	01/2018	161,206	Strain: multipurpose inbred strain. Commonly used for transgenic injection.

Table 4.1 continued from previous page

Species	Source	Genome	Updated	N50 contig length	Notes
Mouse (*Mus musculus*)	Mouse Genomes Project	LP_J_v1	01/2018	483,943	Strain: high susceptibility to audiogenic seizures, and have a fairly high incidence of tumours that develop later in life.
Mouse (*Mus musculus*)	Mouse Genomes Project	NOD_ShiLtJ_v1	01/2018	353,461	Strain: polygenic model for autoimmune type 1 diabetes.
Mouse (*Mus musculus*)	Mouse Genomes Project	NZO_HlLtJ_v1	01/2018	256,342	Strain: develop severe obesity.
Mouse (*Mus musculus musculus*)	Mouse Genomes Project	PWK_PhJ_v1	01/2018	866,061	Strain: wild-derived inbred strain.
Mouse (*Mus musculus*)	Mouse Genomes Project	SPRET_EiJ_R	01/2018	742,570	Strain: wild-derived inbred strain.
Mouse (*Mus musculus domesticus*)	Mouse Genomes Project	WSB_EiJ_v1	01/2018	194,719	Strain: derived from wild mice trapped in Eastern Shore, Maryland.
Rat (*Rattus norvegicus*)	Ensembl	Rnor_6.0	01/2017	100,500	
Dog (*Canis lupus familiaris*)	Ensembl	CanFam3.1	10/2020	12,024,593	
Opossum (*Monodelphis domestica*)	Ensembl	ASM229v1	05/2019	107,990	

Table 4.1 continued from previous page

Species	Source	Genome	Updated	N50 contig length	Notes
Chicken (Red jungle fowl) (*Gallus gallus*)	Ensembl	GRCg6a	03/2021	17,655,422	

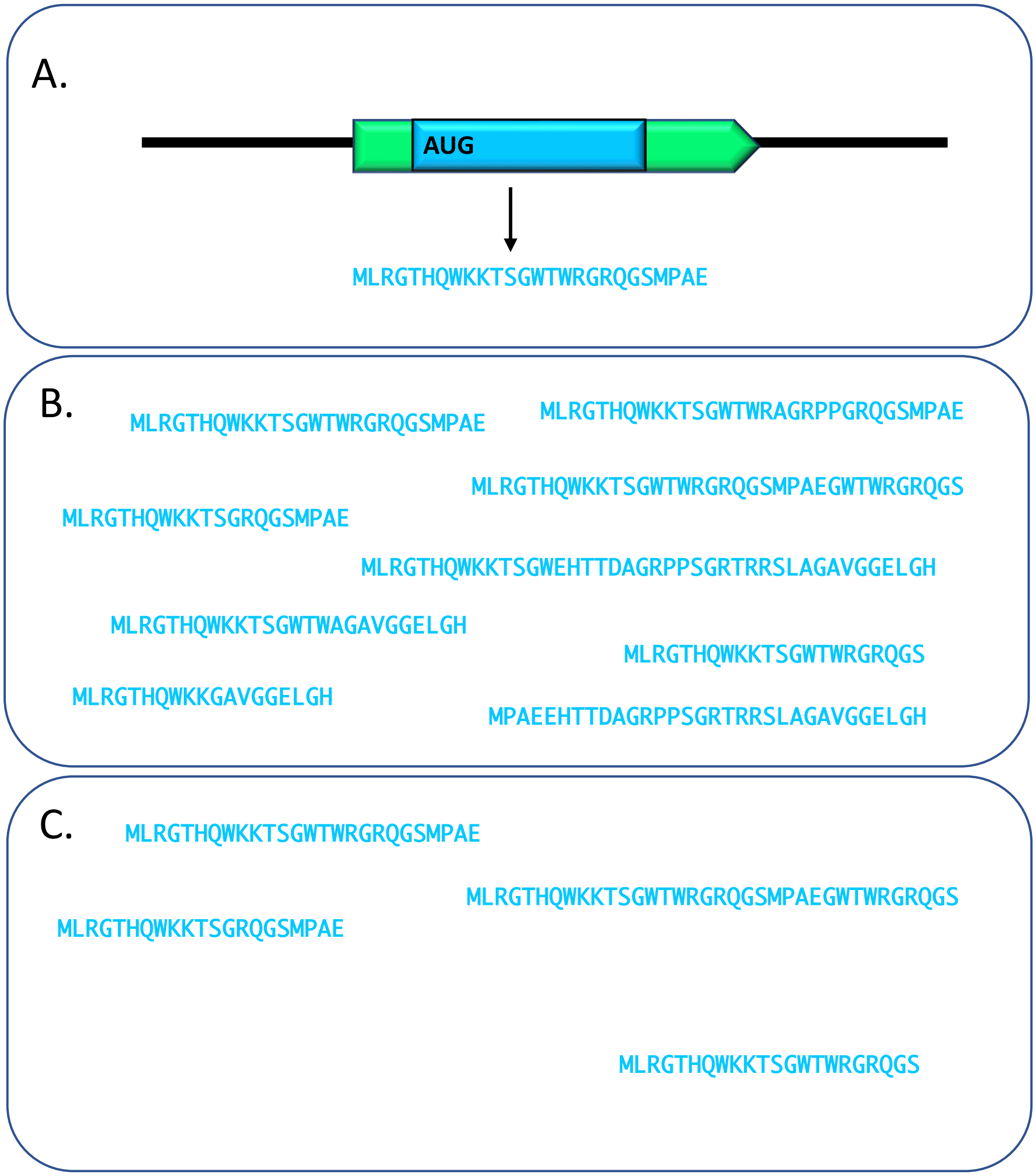

A.
AUG
MLRGTHQWKKTSGWTWRGRQGSMPAE
B.
MLRGTHQWKKTSGWTWRGRQGSMPAE
MLRGTHQWKKTSGWTWRAGRPPGRQGSMPAE
MLRGTHQWKKTSGWTWRGRQGSMPAEGWTWRGRQGS
MLRGTHQWKKTSGRQGSMPAE
MLRGTHQWKKTSGWEHTTDAGRPPSGRTRRSLAGAVGGELGH
MLRGTHQWKKTSGWTWAGAVGGELGH
MLRGTHQWKKTSGWTWRGRQGS
MLRGTHQWKKGAVGGELGH
MPAEEHTTDAGRPPSGRTRRSLAGAVGGELGH
C.
MLRGTHQWKKTSGWTWRGRQGSMPAE
MLRGTHQWKKTSGWTWRGRQGSMPAEGWTWRGRQGS
MLRGTHQWKKTSGRQGSMPAE
MLRGTHQWKKTSGWTWRGRQGS

Figure 4.2: **Graphical illustration of the peptide-centric search of lncRNA smORFs using BLASTp. A.** The amino acid sequence of each lncRNA smORF peptide was compared to databases of annotated proteins from the selected species using BLASTp v2.9.0+ (Altschul et al., 1990). **B.** A pool of potential hits were returned for each peptide in each species (e-value of 0.0001). **C.** Subject hits with percentage identity below 75% and coverage below 50% were filtered. Coverage was calculated as alignment length divided by query sequence length. If no subject hits passed these filters for a given query in each searched species, the hit with the smallest e-value was retained. The green arrow represents a transcript, and the blue box an ORF. The sequences of letters represent peptide sequences.

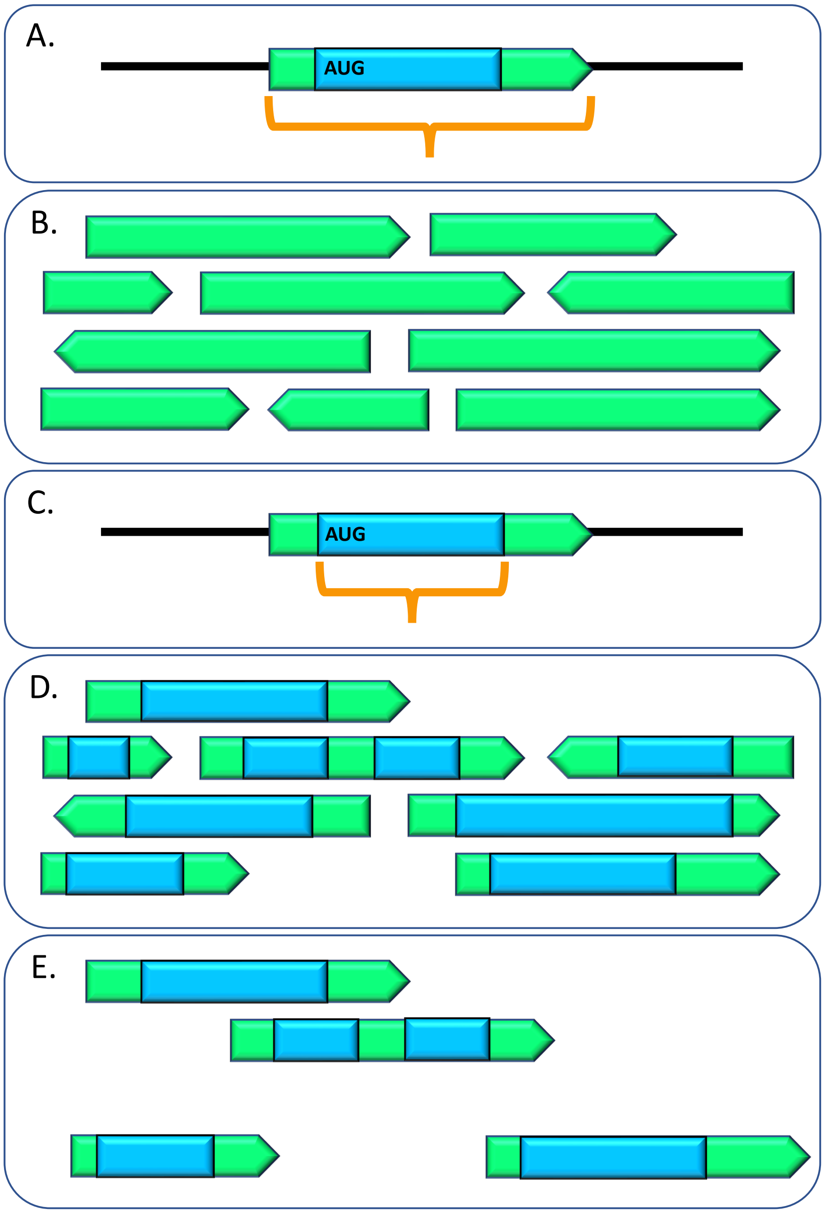


Figure 4.3: **Graphical illustration of the transcript-centric search of lncRNA smORFs using BLASTn. A.** The nucleotide sequence of each translated lncRNA transcript was compared to databases of annotated coding and non-coding transcripts from the selected species using BLASTn v2.9.0+ (Altschul et al., 1990). **B.** A pool of potential hits were returned for each transcript in each species (e-value of 0.0001). These transcripts were then pooled to create a new, curated BLAST database for each species. **C.** The nucleotide sequence of each translated lncRNA smORF was compared to the new BLAST database of whole transcript hits from the selected species using BLASTn v2.9.0+ (Altschul et al., 1990). **D.** A pool of potential hits were returned for each smORF in each species, (e-value of 0.0001). **E.** Subject hits were filtered by removing all hits with percentage identity below 75%, and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. If no subject hits passed these filters for a given query in each searched species, the hit with the smallest e-value was retained. The green arrows represent a transcripts, and the blue boxes ORFs. The orange brackets highlight the portion of the query used in the BLASTn search.

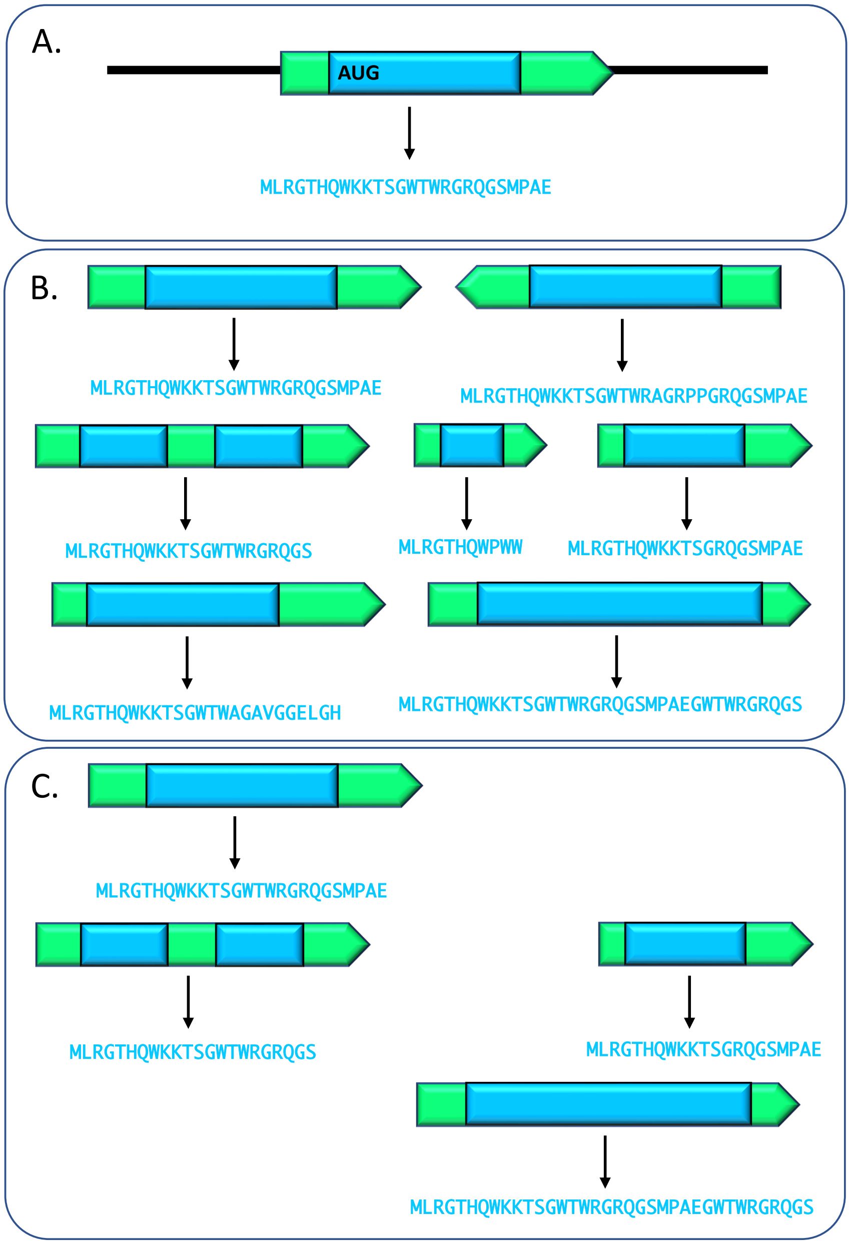

A.
AUG
MLRGTHQWKKTSGWTWRGRQGSMPAE
B.
MLRGTHQWKKTSGWTWRGRQGSMPAE
MLRGTHQWKKTSGWTWRAGRPPGRQGSMPAE
MLRGTHQWKKTSGWTWRGRQGS
MLRGTHQWPWW
MLRGTHQWKKTSGRQGSMPAE
MLRGTHQWKKTSGWTWAGAVGGELGH
MLRGTHQWKKTSGWTWRGRQGSMPAEGWTWRGRQGS
C.
MLRGTHQWKKTSGWTWRGRQGSMPAE
MLRGTHQWKKTSGWTWRGRQGS
MLRGTHQWKKTSGRQGSMPAE
MLRGTHQWKKTSGWTWRGRQGSMPAEGWTWRGRQGS

Figure 4.4: **Graphical illustration of the ORF-centric search of lncRNA smORFs using tBLASTn. A.** The amino acid sequence of each lncRNA smORF peptide was compared to databases of coding and non coding transcripts from the selected species, translated in all six possible reading frames, using tBLASTn v2.9.0+ (Altschul et al., 1990). **B.** A pool of potential hits were returned for each peptide in each species (e-value of 0.0001). **C.** Subject hits were filtered by removing all hits with percentage identity below 75%, and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. If no subject hits passed these filters for a given query in each searched species, the hit with the smallest e-value was retained. The green arrows represent a transcripts, and the blue boxes ORFs. The sequences of letters represent peptide sequences.

ii) For the transcript-centric search (Figure 4.3), the nucleotide sequence of each translated lncRNA transcript was compared to a database created using the combined coding and non-coding transcriptomes of the species listed in Table 4.1, using default settings and an e-value of 0.00001. The pool of potential homologous transcript sequences were used to create a new, curated BLAST database for each species. The nucleotide sequence of each translated lncRNA smORF was compared to the curated database using BLASTn v2.9.0+ (Altschul et al., 1990), with default settings and an e-value of 0.00001. The remaining subject hits were then filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. If no subject hits passed these filters for a given query in each searched species, the hit with the smallest e-value was retained.

iii) For the smORF-centric search (Figure 4.4), the amino acid sequence of each lncRNA smORF peptide was compared to a database created using the combined coding and non-coding transcriptomes of the species listed in Table 4.1, translated in all six possible reading frames. Default settings and an e-value of 0.00001 were used. The pool of potential homologous sequences were filtered by removing all subject hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. If no subject hits passed these filters for a given query in each searched species, the hit with the smallest e-value was retained.

4.2.4 Sequence alignment

Multiple sequence alignments (MSAs) were generated with hits that satisfied the criteria described above from the peptide, transcript, and ORF-centric searches for each translated lncRNA query sequence. As each set of potentially homologous sequences could contain a different combination of insertions, deletions, rearrangements, and mutations, three different algorithms were tested; (Figure 4.5) MUltiple Sequence Comparison by Log-Expectation (MUSCLE) v5.0.1428 (Edgar, 2004) (E-appendix; Muscle5_peptide.sh), MAFFT v7.487 (Katoh and Standley, 2013) (E-appendix; MAFFT_peptide.sh), and ClustalW v2.1 (Larkin et al., 2007) (E-appendix; CLUSTAL_peptide.sh), using default settings. These were selected as popular, well benchmarked similarity based methods, as phylogeny aware methods were not suitable for these data.

These alignments were then compared using MetAl v1.1.0 (Blackburne and Whelan, 2012) (E-appendix; Metal_peptide.sh) which calculates the d_{pos} alignment distance metric, which incorporates information on the position of gaps in MSAs. If MetAl found $\leq 5\%$ difference between these alignments, the MAFFT result was carried forward, as the MAFFT algorithm performed the most efficiently. Else, noRMD v1.2 (Thompson et al., 2001) was used to select the optimal alignment, based on the mean pairwise distance between sequences in continuous sequence space, summed over the full length of the alignment (E-appendix; normd_peptide.sh).

The ORF-centric results (tBLASTn) were aligned as nucleotide sequences, and also converted to aa sequences and aligned to eliminate the noise of synonymous substitutions.

4.2.5 Assessment of alignments

Alignments from the peptide, transcript, and smORF-centric searches were assessed using the Heads-or-Tails (HoT) score (Landan and Graur, 2007) (Figure 4.5), within Guidance v2.02 (Sela et al., 2015) (E-appendix; hot_peptide.sh). By comparing a set of co-optimal alignments to the standard alignment, the uncertainty of each residue, residue-pair, column and sequence in an alignment is calculated and scored (0-1). As Guidance does not accept MUSCLE v5.0.1428 (Edgar, 2004) alignments, the MSAs that were generated using this method were visually inspected using Jalview v2.10.2b2 (Waterhouse et al., 2009). For amino

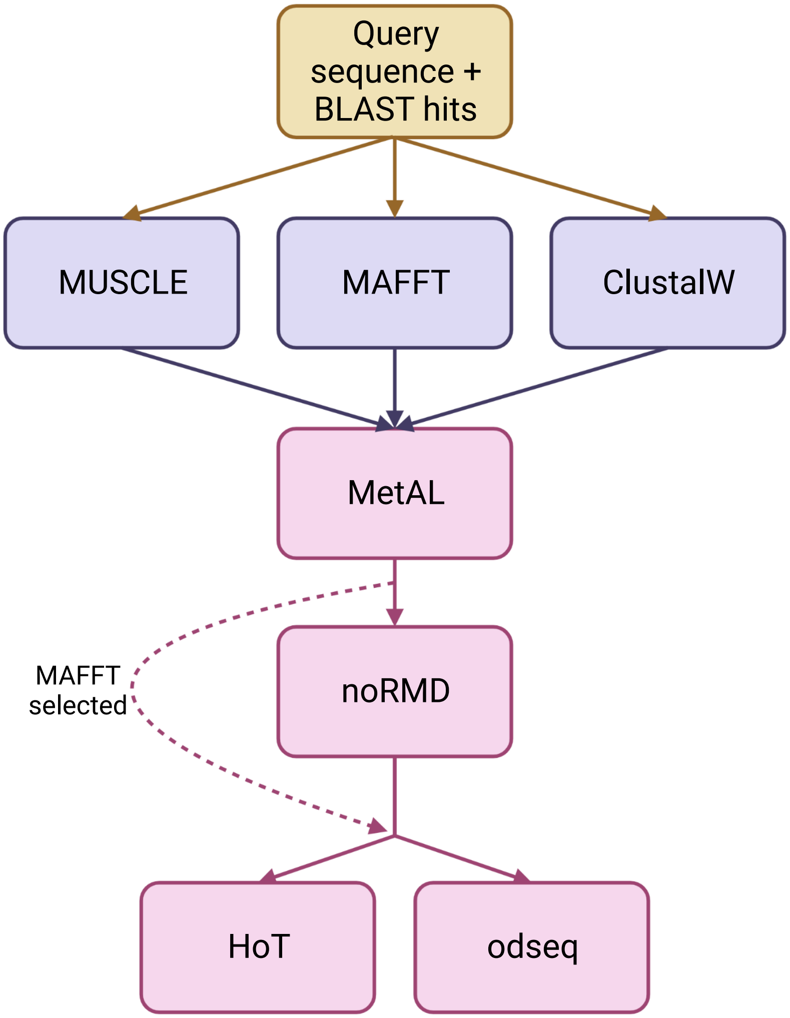


Figure 4.5: **Overview of pipeline for the alignment and assessment of quality of BLAST results.** Each query sequence was aligned with the BLAST subject hits that met our thresholds, using three different algorithms - MUSCLE v5.0.1428 (Edgar, 2004), MAFFT v7.487 (Katoh and Standley, 2013), and ClustalW v2.1 (Larkin et al., 2007). These alignments were compared using MetAl v1.1.0 (Blackburne and Whelan, 2012). If MetAl found $\leq 5\%$ difference between these alignments, the MAFFT result was carried forward. Else, noRMD v1.2 (Thompson et al., 2001) was used to select the "best" alignment, based on the mean pairwise distance between sequences in continuous sequence space, summed over the full length of the alignment. Alignments were then assessed using the HoT (Landan and Graur, 2007) score, a measure of alignment uncertainty based on co-optimal alignments, within Guidance v2.02 (Sela et al., 2015), and odseq v1.22.0 (Jiménez, 2022), which detects outliers using the average distance between sequences. Figure created with BioRender.com.

acid tBLASTn results, wildcard characters (representing stop codons) and gaps introduced by tBLASTn v2.9.0+ (Altschul et al., 1990) were removed in order to analyse them using Guidance. Alignments were also assessed using odseq v1.22.0 (Jiménez, 2022), which detects outliers using the average distance between sequences (Figure 4.5).

The evidence for sequence conservation was manually evaluated for each translated lncRNA smORF, considering the peptide, transcript, and smORF-centric searches. Alongside visual examination of the alignments to check for clear misalignment and modular conservation, the following metrics were considered:

i) The HoT score (Landan and Graur, 2007) for the entire alignment; this was classed as high ($x > 0.9$), medium ($0.9 \geq x > 0.5$), or low ($0.5 \geq x$). Modular areas, such as around the smORF in a longer transcript alignment, may still be well conserved even if the whole sequence alignment does not score highly, so this metric was more relevant to the shorter peptide alignments.

ii) The results of odseq v1.22.0 (Jiménez, 2022) analysis allow us to determine which sequences are outliers. A number of scenarios can be tested in this way, e.g if the query sequence is the only sequence to be deemed an outlier, it is likely that a group of related subject sequences have been returned that are not homologous to the query, but are homologous to one another. The alignment can be pruned and realigned to investigate this further.

iii) Other considerations include how many non-human species results were returned, and whether the pattern of subject hits followed the species phylogeny. For example, if results are only found in one distantly related species, this may be an artefact of how the annotations were produced. The number of potential homologs in the BLAST search was also informative as a large number of poorly aligned results may suggest that the lncRNA contains a repetitive sequence. In all cases the upstream and downstream ATGs were taken into account, as a smORF could be partially conserved between species if different start codons were used but the frame was maintained.

Using these metrics, the translated lncRNA smORFs were grouped into the following categories based on their results in all BLAST searches; A) No BLAST results. These smORFs did not return any filter passing results from any BLAST search in the queried non-human species. B) No convincing evidence. Results were returned for these lncRNA smORFs, but they aligned very poorly with the query smORF sequence. C) Results in 1 or 2 species. These smORFs returned well aligned results, but only in one or two of the queried species. D) Evidence of sequence conservation. For these smORFs, well aligned results were returned for multiple species in one or more of the BLAST searches, and this evidence was convincing when considered alongside the HoT score (Landan and Graur, 2007), conserved start codons, and if query sequences were marked as outliers.

The number of results filtered at each step in the BLASTp, BLASTn and tBLASTn sequence searching, alignment and assessment pipeline are summarised in Figures 4.6, 4.7, and 4.8.

4.2.6 General statistics and plots

Data were analysed and plotted in R (R Core Team, 2021), using packages including the tidyverse (Wickham et al., 2019), tibble (Müller and Wickham, 2021), plot.matrix (Klinke, n.d.), reshape (Wickham, 2007), matrixStats (Bengtsson et al., 2022), seqinr (Charif and Lobry, 2007), protr (Xiao et al., 2015), ggplot2 (Wickham, 2016), gggenes (Wilkins, 2023), ggpubr (Kassambara, 2023), VennDiagram (Chen, 2022) ggtext (Wilke and Wiernik, 2022), ggtree (Yu et al., 2017), and viridis (Garnier et al., 2021).

To assess whether there was a significant association between the "stringent" and "non-stringent" lncRNA smORF groups and conservation category, a Chi-square test was performed.

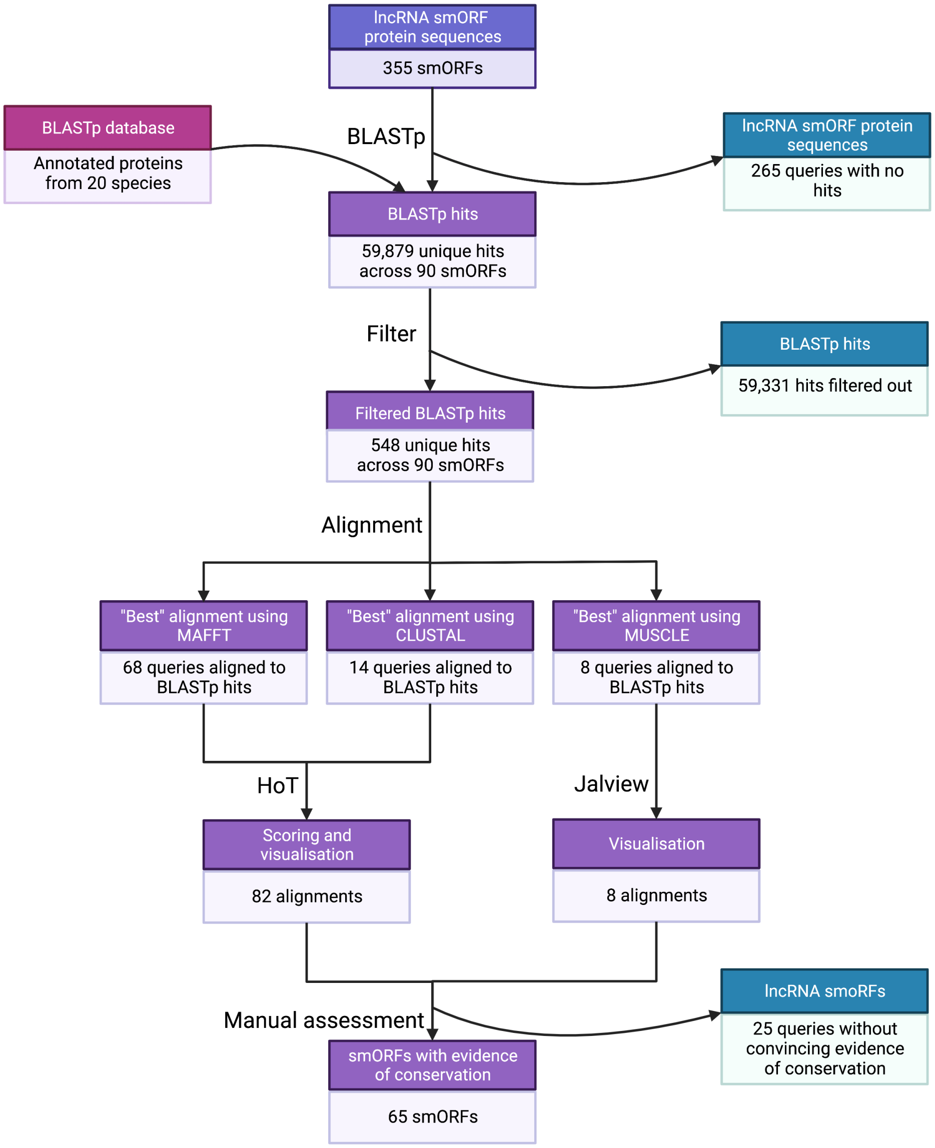

lncRNA smORF protein sequences
355 smORFs
BLASTp database
Annotated proteins from 20 species
BLASTp
lncRNA smORF protein sequences
265 queries with no hits
BLASTp hits
59,879 unique hits across 90 smORFs
Filter
BLASTp hits
59,331 hits filtered out
Filtered BLASTp hits
548 unique hits across 90 smORFs
Alignment
"Best" alignment using MAFFT
68 queries aligned to BLASTp hits
"Best" alignment using CLUSTAL
14 queries aligned to BLASTp hits
"Best" alignment using MUSCLE
8 queries aligned to BLASTp hits
HoT
Jalview
Scoring and visualisation
82 alignments
Visualisation
8 alignments
Manual assessment
lncRNA smoRFs
25 queries without convincing evidence of conservation
smORFs with evidence of conservation
65 smORFs

Figure 4.6: **BLASTp alignment assessment pipeline.** Translated lncRNA smORF peptide sequences were compared to the proteomes of 20 species including human using BLASTp v2.9.0+ (Altschul et al., 1990), and any queries with no hits removed. The BLASTp results were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. Else, the hit with the smallest e-value was retained. The remaining hits were then aligned to their query lncRNA smORF peptide using MAFFT v7.487 (Katoh and Standley, 2013), ClustalW v2.1 (Larkin et al., 2007) and MUSCLE v 5.0.1428 (Edgar, 2004) and the "best" method selected. MAFFT and CLUSTAL alignments were scored and visualised using Guidance v2.02 (Sela et al., 2015). MUSCLE alignments were visualised using Jalview V2 (Waterhouse et al., 2009). Following manual assessment 65/355 smORFs with evidence of sequence conservation remained. Figure created with BioRender.com.

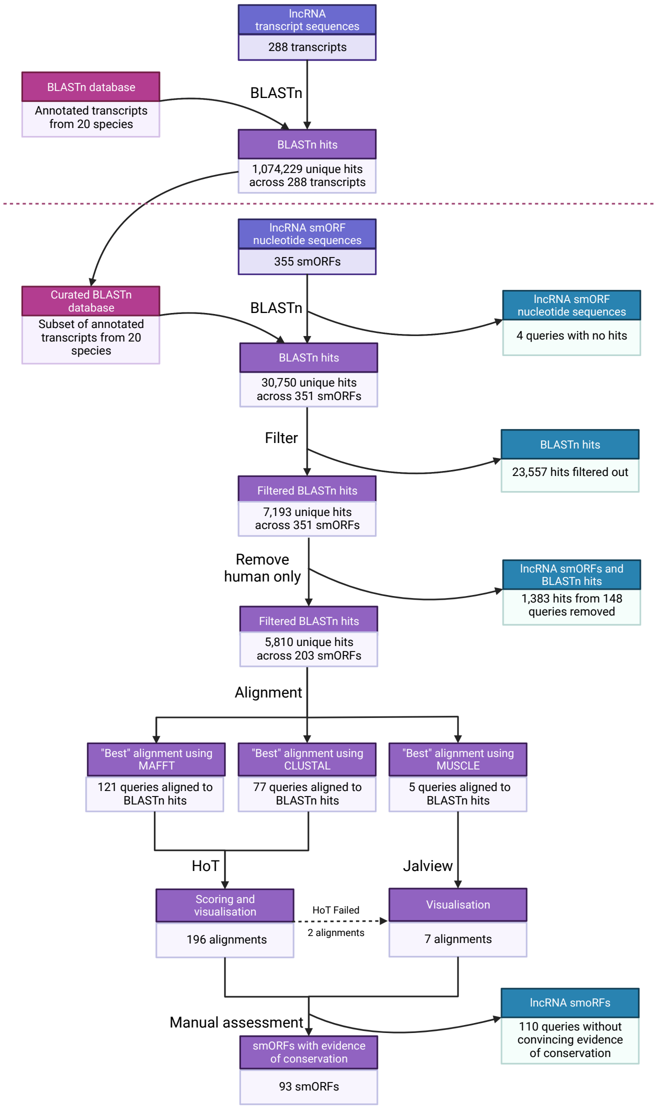

lncRNA transcript sequences
288 transcripts
BLASTn database
Annotated transcripts from 20 species
BLASTn
BLASTn hits
1,074,229 unique hits across 288 transcripts
lncRNA smORF nucleotide sequences
355 smORFs
Curated BLASTn database
Subset of annotated transcripts from 20 species
BLASTn
lncRNA smORF nucleotide sequences
4 queries with no hits
BLASTn hits
30,750 unique hits across 351 smORFs
Filter
BLASTn hits
23,557 hits filtered out
Filtered BLASTn hits
7,193 unique hits across 351 smORFs
Remove human only
lncRNA smORFs and BLASTn hits
1,383 hits from 148 queries removed
Filtered BLASTn hits
5,810 unique hits across 203 smORFs
Alignment
"Best" alignment using MAFFT
121 queries aligned to BLASTn hits
"Best" alignment using CLUSTAL
77 queries aligned to BLASTn hits
"Best" alignment using MUSCLE
5 queries aligned to BLASTn hits
HoT
Jalview
Scoring and visualisation
196 alignments
HoT Failed
2 alignments
Visualisation
7 alignments
Manual assessment
lncRNA smoRFs
110 queries without convincing evidence of conservation
smORFs with evidence of conservation
93 smORFs

Figure 4.7: **BLASTn alignment assessment pipeline.** Translated lncRNA smORF transcript sequences were compared to the transcriptomes of 20 species including human using BLASTn v2.9.0+ (Altschul et al., 1990). From these hits a new, curated BLASTn database was created, and the sequences of the smORFs only were compared to this new database. Any queries with no hits were removed. The BLASTn smORF results were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. Else, the hit with the smallest e-value was retained. The remaining hits were then aligned to their query lncRNA smORF and transcript sequences using MAFFT v7.487 (Katoh and Standley, 2013), ClustalW v2.1 (Larkin et al., 2007) and MUSCLE v 5.0.1428 (Edgar, 2004) and the "best" method selected. MAFFT and CLUSTAL alignments were scored and visualised using Guidance v2.02 (Sela et al., 2015). MUSCLE alignments were visualised using Jalview V2 (Waterhouse et al., 2009). Following manual assessment 93/355 smORFs with evidence of sequence conservation remained. Figure created with BioRender.com.

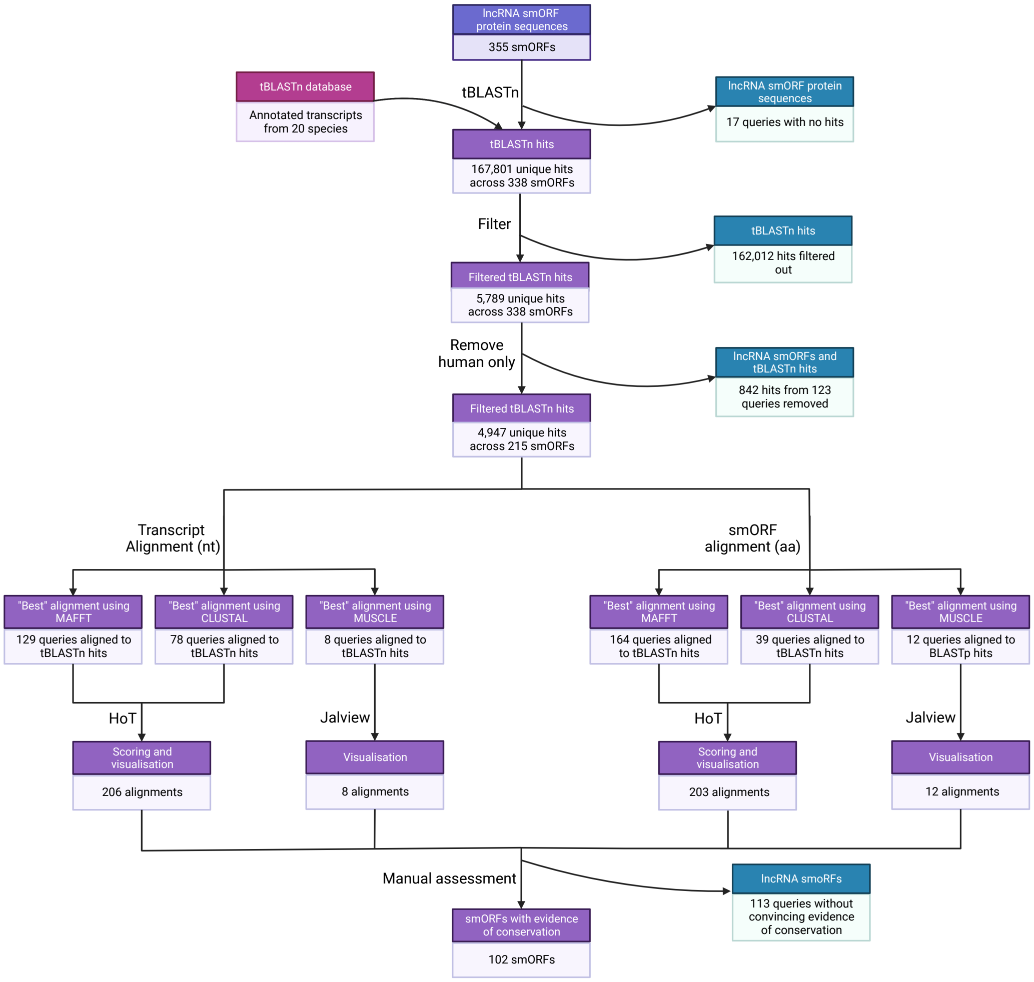

lncRNA smORF protein sequences
355 smORFs
tBLASTn database
Annotated transcripts from 20 species
tBLASTn
lncRNA smORF protein sequences
17 queries with no hits
tBLASTn hits
167,801 unique hits across 338 smORFs
Filter
tBLASTn hits
162,012 hits filtered out
Filtered tBLASTn hits
5,789 unique hits across 338 smORFs
Remove human only
lncRNA smORFs and tBLASTn hits
842 hits from 123 queries removed
Filtered tBLASTn hits
4,947 unique hits across 215 smORFs
Transcript Alignment (nt)
smORF alignment (aa)
"Best" alignment using MAFFT
129 queries aligned to tBLASTn hits
"Best" alignment using CLUSTAL
78 queries aligned to tBLASTn hits
"Best" alignment using MUSCLE
8 queries aligned to tBLASTn hits
"Best" alignment using MAFFT
164 queries aligned to tBLASTn hits
"Best" alignment using CLUSTAL
39 queries aligned to tBLASTn hits
"Best" alignment using MUSCLE
12 queries aligned to BLASTp hits
HoT
Jalview
HoT
Jalview
Scoring and visualisation
206 alignments
Visualisation
8 alignments
Scoring and visualisation
203 alignments
Visualisation
12 alignments
Manual assessment
lncRNA smoRFs
113 queries without convincing evidence of conservation
smORFs with evidence of conservation
102 smORFs

Figure 4.8: **tBLASTn alignment assessment pipeline.** Translated lncRNA smORF peptide sequences were compared to the transcriptomes of 20 species including human using tBLASTn v2.9.0+ (Altschul et al., 1990), and any queries with no hits removed. The tBLASTn results were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. Else, the hit with the smallest e-value was retained. Both the nucleotide sequence and the translated amino acid sequence of each query lncRNA smORF were aligned to the remaining hits using MAFFT v7.487 (Katoh and Standley, 2013), ClustalW v2.1 (Larkin et al., 2007) and MUSCLE v 5.0.1428 (Edgar, 2004). and the "best" method selected. MAFFT and CLUSTAL alignments were scored and visualised using Guidance v2.02 (Sela et al., 2015). MUSCLE alignment were visualised using Jalview V2 (Waterhouse et al., 2009). Following manual assessment 102/355 smORFs with evidence of sequence conservation remained. Figure created with BioRender.com.

This test was selected as it is suitable for categorical data, and the data meets the assumptions of mutual exclusivity, independence, and the assumption that at least 80% or more of the expected values equal to or greater than 5.

4.3 Results

4.3.1 Generation of inclusive list of translated lncRNA smoRFs

To investigate the sequence conservation of translated lncRNA smORFs and their peptide products within mammals, a suitable dataset of lncRNA smORFs was required. The majority of Chapter 3 focused on a the properties of a stringent set of 45 lncRNA smORFs, selected due to our confidence in their active translation in SH-SY5Y cells. As in Section 3.3.5, a larger dataset is more suitable to for this question to increase the likelihood of detecting sequence conservation, as lncRNAs are generally considered to be under little selective pressure at the sequence level. The requirement that a smORF be identified in at least two of the three biological replicates of a given condition to be considered translated was therefore removed (Section 2.2.3.1). This produced a dataset of 355 lncRNA smORFs actively translated in SH-SY5Y cells, in control or RA conditions.

This wider dataset of 355 lncRNA smORFs also includes ‘smORF isoforms’. These occur when genes with multiple transcript splice variants are expressed and translated. In some cases the smORFs from these splice variants are identical or overlap significantly, producing the same or very similar peptides; smORF isoforms (Figure 4.9). Here smORFs have been considered isoforms if they originate from the same gene and at least 50% of the smORF overlaps in the same frame. For example, many smORF isoforms occur when different in-frame start codons are identified across different replicates for what is likely the same smORF, due to the position of Ribo-seq reads. When smORF isoforms are grouped, the 335 smORFs collapse into 242 smORFs.

4.3.2 Putative homologs returned in all 18 sampled species

To assess whether the sequences of the translated lncRNA smORFs were conserved within vertebrates, three levels of sequence similarity searches were performed; i) a peptide-centric search, ii) a transcript-centric search and iii) an ORF-centric search.

4.3.2.1 Peptide-centric search

To search for homologous sequences in currently annotated peptides in the selected species, a BLASTp search (Altschul et al., 1990) was performed. Of the 355 lncRNA smORF peptide queries, 90 returned putative homologs in one or more of the non-human species in the dataset (Appendix C, Figure C.1). The percentage identity and coverage filters described in Section 4.2.3 removed a large number of low quality hits, reducing the mean number of hits per query from 665.3 to 6.089 hits (Table 4.2). This significantly reduced the amount of noise for later analysis, and made alignment of these results both possible and helpful.

As expected, as the evolutionary distance from human increased, fewer lncRNA smORF peptides returned hits (Ulitsky, Shkumatava, et al., 2011). Only three of the smORF peptides were also found in the outgroup *Gallus gallus*. These three queries also returned hits in all other selected species (Figure 4.10). On closer inspection two of the three peptides originate from overlapping, in frame smORFs on the same transcript, and are therefore classified as ’smORF isoforms’ (Figure 4.9). One of these query peptides from smORF ENST00000526036.1_1226_2042, referred to as smORF A throughout, is discussed in more detail later in this chapter.

Only 18/355 of the peptide queries returned one or more hits in human (Figure C.1, Table 4.3), which was to be expected as this was a search of novel/unannotated peptides against a database of canonical, annotated peptides. The results returned in human are likely due to a shared domain or region within the peptide, although a shared evolutionary history is also possible. To identify shared domains, the query and hit peptides were searched against

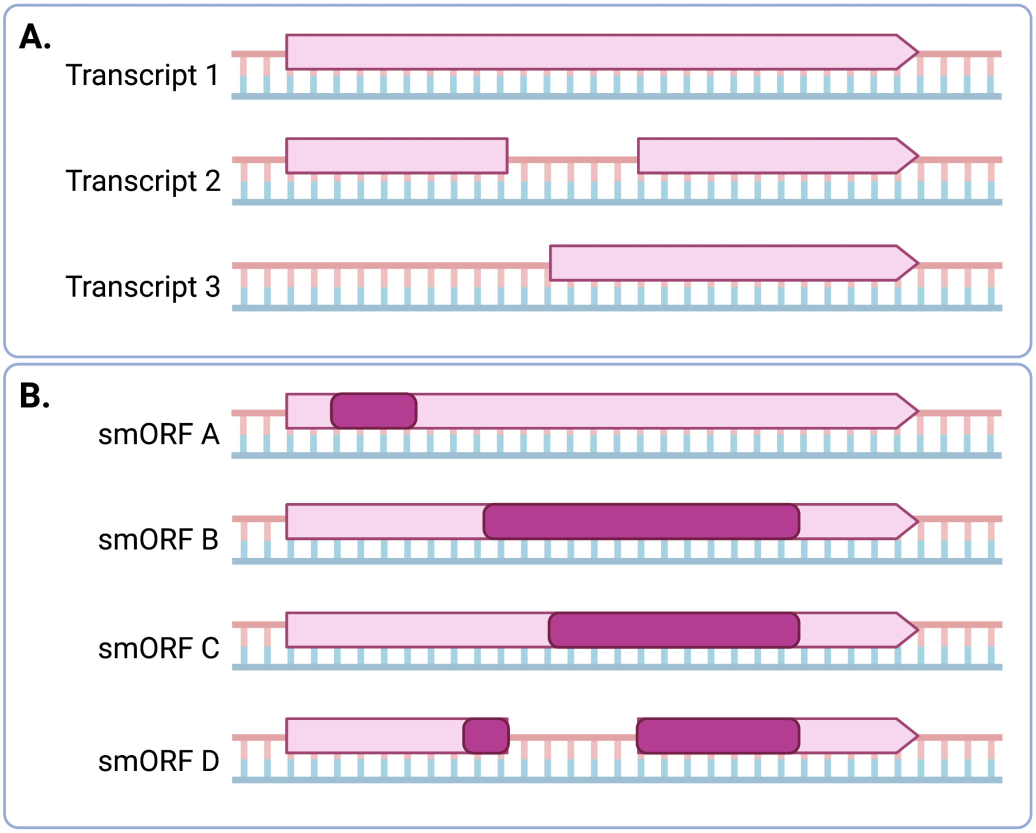


Figure 4.9: **Example of translated lncRNA smORF isoforms.** The arrows represent transcripts, and dark pink boxes represent smORFs. **A.** Three splice variants transcribed from the same gene. **B.** Four potential smORFs identified from the splice variants. smORF A does not overlap with any other smORFs from this gene, so will be investigated independently. smORFs B to D overlap significantly. If these smORFs are translated in the same frame, and therefore share ≥ 50 of their peptide sequence, they will be considered together as smORF isoforms.

Table 4.2: **Summary of BLASTp hits, pre and post-filtering.** The pool of potential homologous peptide sequences were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. Else, the hit with the smallest e-value was retained.

	No. of hits per query	No. of filtered hits per query
Range	1-28,078	1-36
Median	4	3.5
Mean	665.3	6.089

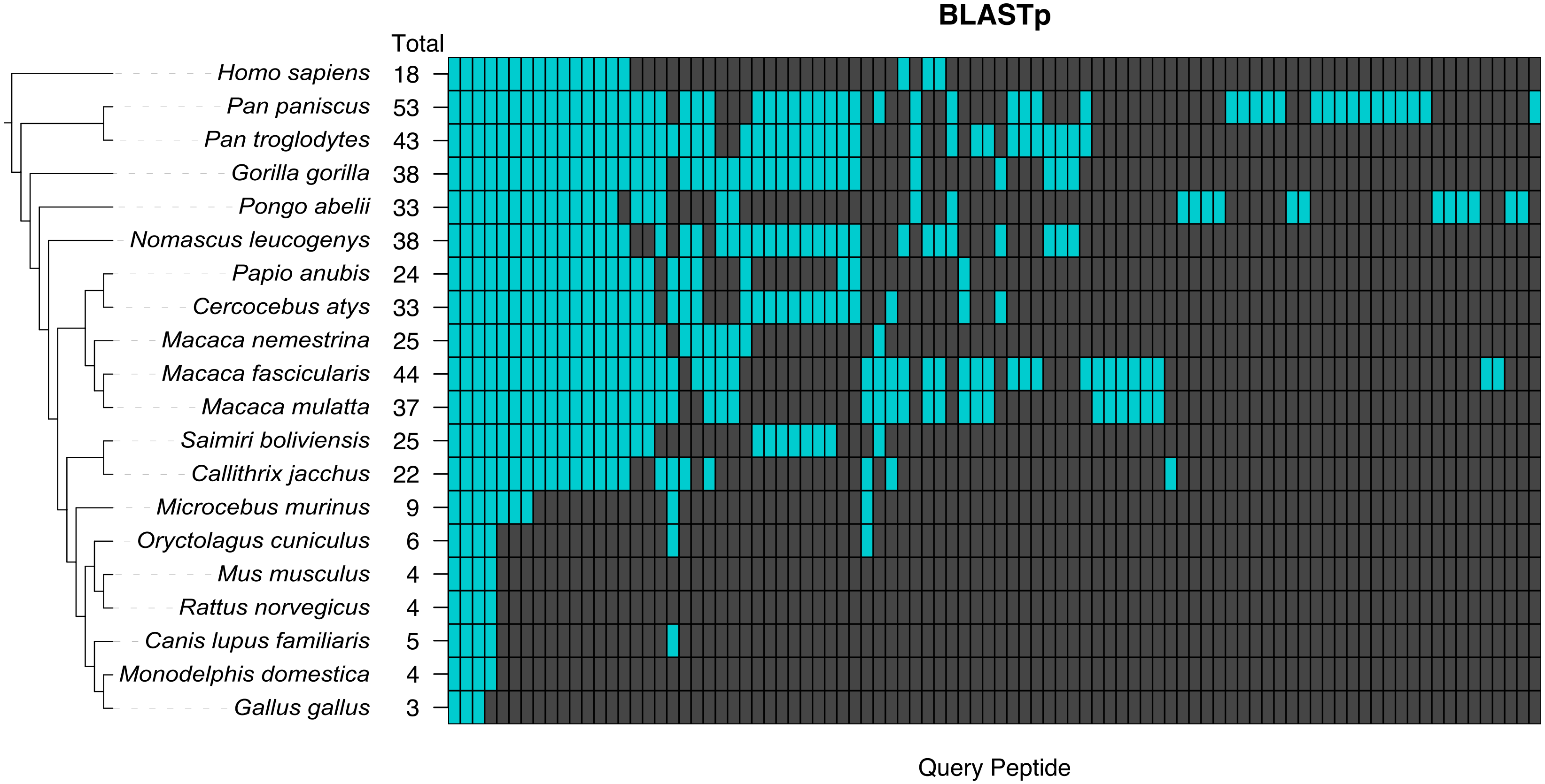


Figure 4.10: **Distribution of BLASTp hits returned for lncRNA smORF peptides.** Initial results from BLASTp v2.9.0+ (Altschul et al., 1990), as described in Figure 4.2B. The phylogeny shows the species included in the conservation analysis, and the total number of lncRNA smORF queries which returned hits in each species is given at each branch end. Each column represents a lncRNA smORF peptide query, and the blue cells represent one or more hits that were returned for that query in a given species. Dark grey cells indicate that no hits were returned.

Table 4.3: **Human peptides returned by BLASTp search.** Query peptide denotes the lncRNA smORF peptide query entered into the BLASTp search. Hit peptide denotes the human peptide hit returned by BLASTp. Orthologous region describes any domains shared between the query and hit peptides, identified using PRINTS BLAST. Coverage is calculated as the length of the alignment from the BLAST result, divided by the length of the query peptide. Percent identity is the percentage of identical matches between the aligned portion of the query and hit peptides. PRINTS BLAST is an interface to a BLAST v2.2.4 (Altschul et al., 1990) search of all protein sequences contained within the PRINTS database release 42.0 (Attwood et al., 1994).

Query Peptide	Hit Peptide	Orthologous region	Coverage	Percent Identity
ENST00000424349.1_491_656	ENSP00000369673.3	Query and hit contain a HTH_48 domain.	82%	76%
ENST00000424349.1_491_656	ENSP00000373354.3	Query and hit contain a HTH_48 domain.	82%	76%
ENST00000424349.1_491_656	ENSP00000403145.1	Query and hit contain a HTH_48 domain.	82%	76%
ENST00000295549.8_609_822	ENSP00000403769.2	Hit contains three F138DOMAINs. Query contains F138DOMAIN like region.	62%	64%
ENST00000295549.8_609_822	ENSP00000502572.1	Query and hit contain F138DOMAIN like region.	68%	56%
ENST00000344893.4_298_883	ENSP00000392024.3	Query and hit contain KRAB domains.	100%	63%
ENST00000344893.4_409_883	ENSP00000498410.1	Hit overlaps with hit from query smORF isoform (ENST00000344893.4_298_883).	100%	60%
ENST00000371417.3_542_1070	ENSP00000354971.4	Query contains a F138DOMAIN. Hit contains a F138DOMAIN like region.	20%	74%
ENST00000424989.1_284_479	ENSP00000419129.1	No annotated domain. Short alignment to propionyl-CoA carboxylase subunit beta (PCCB) hit protein.	40%	81%
ENST00000506795.1_100_544	ENSP00000402917.1	No annotated domain. Short alignment to gamma-glutamyltransferase 5 (GGT5) hit peptide.	30%	66%
ENST00000507387.1_303_633	ENSP00000427662.1	No annotated domain. Well aligned to start of follistatin like 4 (FSTL4) hit peptide.	39%	86%
ENST00000507759.5_202_508	ENSP00000384436.2	Hit contains F138DOMAIN. Query contains F138DOMAIN like region.	32%	70%

ENST00000507759.5_202_508	ENSP00000505475.1	Hit contains F138DOMAIN. Query contains F138DOMAIN like region.	32%	70%
ENST00000526036.1_1226_2042	ENSP00000461610.1	Query and hit contain a DNA binding HTH_Tnp_Tc5 domain.	38%	41%
ENST00000526036.1_1226_2042	ENSP00000483808.1	Query and hit contain a DNA binding HTH_Tnp_Tc5 domain.	38%	41%
ENST00000526036.1_1340_2042	ENSP00000461610.1	As in ENST00000526036.1_1226_2042.	28%	37%
ENST00000526036.1_1340_2042	ENSP00000483808.1	As in ENST00000526036.1_1226_2042.	28%	37%
ENST00000534914.5_131_503	ENSP00000431299.1	Hit contains a F138DOMAIN. Query contains F138DOMAIN like region.	43%	60%
ENST00000548329.1_212_563	ENSP00000472170.1	No annotated domain. Short alignment with zinc finger protein 675 (ZNF657) hit peptide.	36%	86%
ENST00000549357.1_66_315	ENSP00000492329.1	No annotated domain. Well aligned to start of gamma-aminobutyric acid type A receptor subunit gamma2 (GABRG2) hit peptide.	64%	74%
ENST00000549357.1_90_315	ENSP00000492329.1	As in ENST00000549357.1_66_315.	60%	69%
ENST00000551597.6_54_252	ENSP00000338107.4	No annotated domain. End of query well aligned to end of cancer antigen 1 (CAGE1) hit peptide.	33%	86%
ENST00000608396.2_548_869	ENSP00000499080.1	Query and hit contain F138DOMAIN like region.	41%	71%
ENST00000608396.2_548_869	ENSP00000501710.1	Query and hit contain F138DOMAIN like region.	33%	83%
ENST00000632111.1_163_529	ENSP00000490329.1	Query and hit contain F138DOMAIN like region.	23%	82%

PRINTS BLAST (Attwood et al., 1994; Altschul et al., 1990), a database containing 2156 protein family 'fingerprints', encoding 12,444 individual motifs, and the Ensembl annotation (Cunningham et al., 2022) of the hits was examined. Of the 18 peptides, 11 contained a known protein domain which overlapped with the same domain in the hit peptide (Table 4.3). Helix-turn-helix (HtH) and F138DOMAIN domains were the most common.

More query peptides (≥ 22) returned hits in the non-human Primates in the dataset (with the exception of *Microcebus murinus*) than in human, which may be indicative of differences in annotation pipelines; for example higher stringency during annotation of human protein coding genes.

4.3.2.2 Transcript-centric search

To search for sequence conservation of the translated lncRNA smORFs at the nucleotide level, the sequence of each translated lncRNA transcript was compared to databases of annotated coding and non-coding transcripts from the selected species using BLASTn (Altschul et al., 1990). The hits were pooled to create a custom BLAST database for each species, and a second BLASTn search performed using the smORF nucleotides sequence only (Figure 4.3). This filtered the hits to those with sequence similarity across the lncRNA smORF in particular, instead of other parts of the transcript.

All 288 lncRNA transcripts (ranging in length from 180 to 19,245 nt) returned themselves from the human transcriptome in the intial BLASTn (accounting for annotation versions) (Appendix C, Figure C.2), confirming that the selected e-value cut off was reasonable. In total, 260/288 translated lncRNA transcript queries returned putative homologs in one or more of the non-human species in the dataset in the initial BLASTn (Altschul et al., 1990) search. There was a drop off in the number of lncRNA smORF transcripts returning results outside of the Primates, with 114 queries returning results in *Microcebus murinus* (mouse lemur) compared to 37 in *Oryctolagus cuniculus* (rabbit) (Appendix C, Figure C.2).

The smORF nucleotide sequences (ranging in length from 9 to 816 nt) were then used for a the second BLASTn search (Altschul et al., 1990) against the curated BLAST database. This was effective in filtering out lower quality hits from the pooled database, reducing the mean number of hits per query from 3,729.96 to 87.61 hits (Table 4.4). Following this initial filter, the percentage identity and coverage filters (Section 4.2.3) did not remove as many low quality hits as in the previous section, but did further reduce the mean number to 20.49 hits. A small number of queries remained with a particularly large number of results, with the highest number of hits for a single query being 1,515.

203/355 translated lncRNA smORFs returned sequences in the non-human species in the dataset. Very few results (1 to 8) were returned in queried non-primate species (Figure 4.11). When human results are included, BLASTn hits were returned for 351/355 translated lncRNA smORF sequences (Figure 4.11, Appendix C Figure C.3) from 287/288 lncRNA transcripts. All 351 of these smORFS returned their own transcript from the human transcriptome from Gencode v30 (Frankish et al., 2021), and 350/351 in Gencode v38 (accounting for annotation versions). The transcript which was not returned by its smORF has been reannotated as a retained intron in Gencode v38, and was approximately half its previous length. The four smORFs which did not return any hits from the human transcriptome are all < 24 nucleotides in length, so were likely unable to pass the stringent e-value requirement.

In both the transcript and smORF BLASTn (Altschul et al., 1990) searches, there was some variation in the hits returned from the 16 mouse transcriptomes. If a hit was returned in at least half (8/16) of the queried mouse transcriptomes, then this was classified as a mouse hit.

Table 4.4: **Summary of BLASTn hits at transcript search, ORF search, and filtering steps.** The pool of potential homologous transcript sequences were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. Else, the hit with the smallest e-value was retained.

	No. of hits per transcript query	No. of hits per ORF query	No. filtered of hits per ORF query
Range	1 - 34,943	1 - 8,240	1 - 1,515
Median	51	10	9
Mean	3,729.96	87.61	20.49

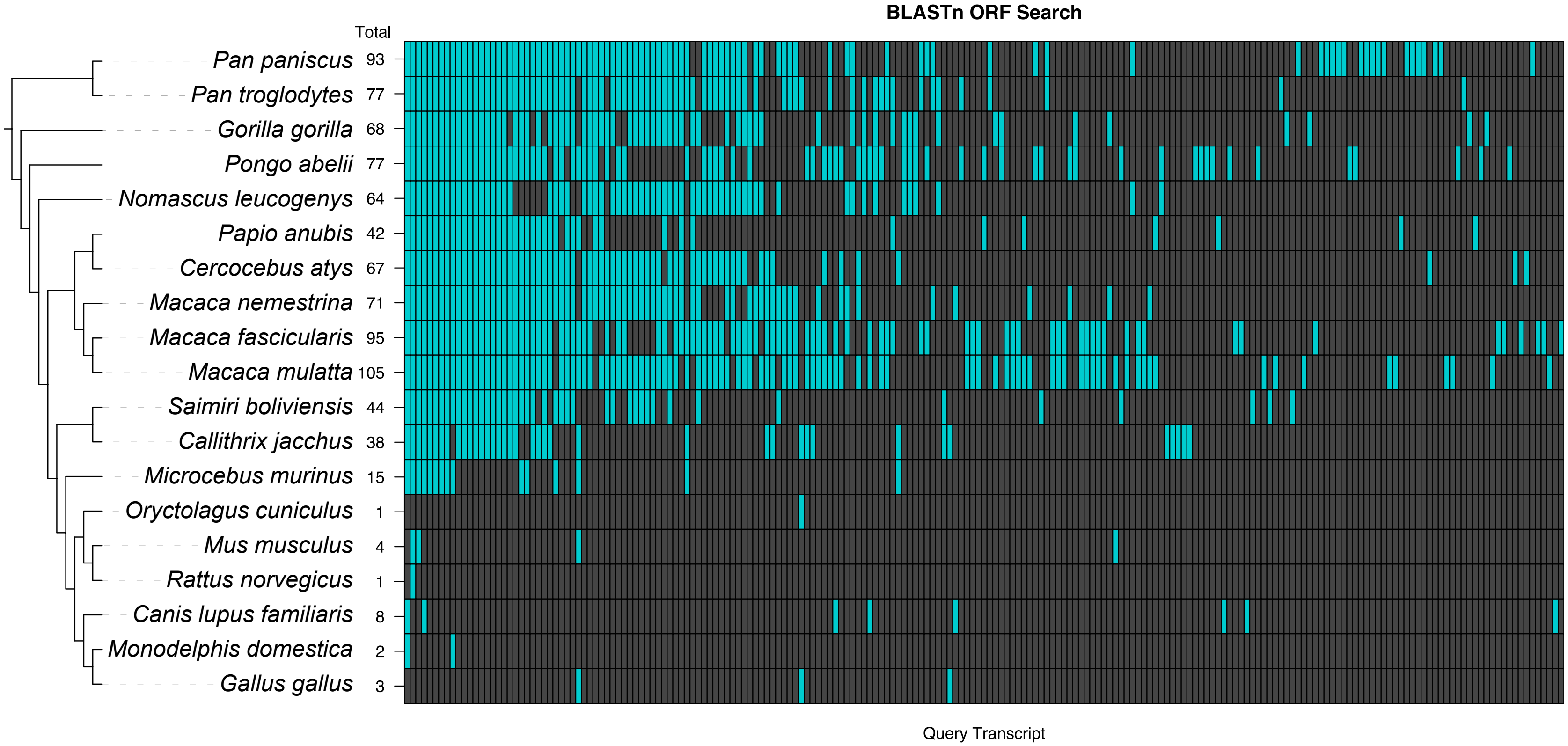


Figure 4.11: **Distribution of BLASTn hits returned for translated lncRNA smORFs in non-human species.** Initial results of smORF BLASTn, as described in Figure 4.3D. Human results have not been included as all but four query smORFs returned their transcript of origin in human. The phylogeny shows the species included in the conservation analysis, and the total number of lncRNA smORF queries which returned hits in each species is given at each branch end. Each column represents a lncRNA smORF query, and the blue cells represent one or more hits that were returned for that query in a given species. Dark grey cells indicate that no hits were returned.

4.3.2.3 ORF-centric search

To identify homologous sequences in the entire annotated transcriptomes of the selected species, a tBLASTn search (Altschul et al., 1990) was performed using the lncRNA smORF peptide sequences (ranging in length from 3 to 272 aa). 215/355 lncRNA smORF peptide queries in the ORF-centric tBLASTn returned one or more putative homologs in non-human species in the dataset (Appendix C Figure C.4). The percentage identity and coverage filters described in Section 4.2.3 reduced the mean number of hits per query from 496.4 to 17.13 hits (Table 4.5). A small number of queries remained with a large number of results, with the highest number of hits for a single query being 449, but this was a large reduction from the initial range of 1 to 33,039 hits.

There were a small number (17) of lncRNA smORF peptide query sequences that did not return any hits in human; ranging in length from 3 - 19 aa. The their short length is likely to have prevented results from passing the stringent e-value threshold, although one 18 aa query peptide did return a hit in human. Peptides of <24 aa are considered particularly small, and difficult to identify and characterise (Ruiz-Orera and Albà, 2018). Of the 338/355 lncRNA smORF peptides which did return one or more hits in human, 337 peptides returned their transcript of origin from Gencode v30, and 334 in Gencode v38 (Frankish et al., 2021). Interestingly, a higher number of queries returned one or more hits in *Macaca mulatta* (macaque) and *Macaca fascicularis* (crab-eating macaque) (105 and 95 queries) than in the more closely related *Hominidae* (Figure 4.12).

4.3.2.4 Putative homologs returned for ∼64% of translated smORFs

When the peptide-centric, transcript-centric and ORF-centric sequence searches are combined, 227/355 translated lncRNA smORF queries returned possible evidence of conservation in one or more of the selected non-human species. This was a surprising result given the low levels of sequence conservation generally expected in lncRNA. Figure 4.13 shows a high level summary summary of the results; it does not describe whether the same results were returned from each search, or in how many species each hit was found. The 90 queries returning hits from BLASTp were a subset of the 215 queries returning hits in tBLASTn, and 86 of these queries returned hits from all three BLAST searches. As the BLASTp search is of annotated proteins, this overlap is to be expected; if a query aligns well with an annotated protein from a given species, tBLASTn is likely to return the same or a similar result.

12 of the queries only returned non-human hits in the BLASTn search (Figure 4.13), and upon closer examination they only returned hits from one or two species. These 12 smORFs produced particularly short peptides, with a median length of 34.5aa, compared to 60aa for all 227 smORFs which returned BLAST hits, and 42aa for the 128 smORFs which returned no BLAST hits (Table 4.6). This short sequence length means any results were likely filtered out by our stringent e-value cut-off in the BLASTp and tBLASTn searches, as short sequences are more likely to align by chance rather than due to a true biological relationship. A further 20 queries only returned non-human hits in the tBLASTn search (Figure 4.13). This is likely due to synonymous substitutions, which alter the nucleotide sequence of the smORF without affecting the the amino acid sequence of the peptide.

The majority (191/227) of the translated lncRNA smORF queries which returned hits in one or more of the non-human species in the dataset did so from both BLASTn and tBLASTn searches (Figure 4.13). As the BLASTn focused on the smORF nucleotide sequence, this may indicate that the smORFs are well conserved at the nucleotide and aa level, but this depends on the hits returned for each individual query.

A proportion of the translated lncRNA smORFs (36%) did not return any evidence of sequence conservation. 11/128 of these smORFs were from the stringent set of 45 lncRNA smORFs (24%), translated in two or more replicates in the same condition. This suggests

Table 4.5: **Summary of tBLASTn hits, pre and post-filtering.** The pool of potential homologous sequences were filtered by removing all hits with percentage identity below 75% and coverage below 50%. Coverage was calculated as alignment length divided by query sequence length. Else, the hit with the smallest e-value was retained.

	No. of hits per query	No. of filtered hits per query
Range	1-33,039	1-449
Median	10	9
Mean	496.4	17.13

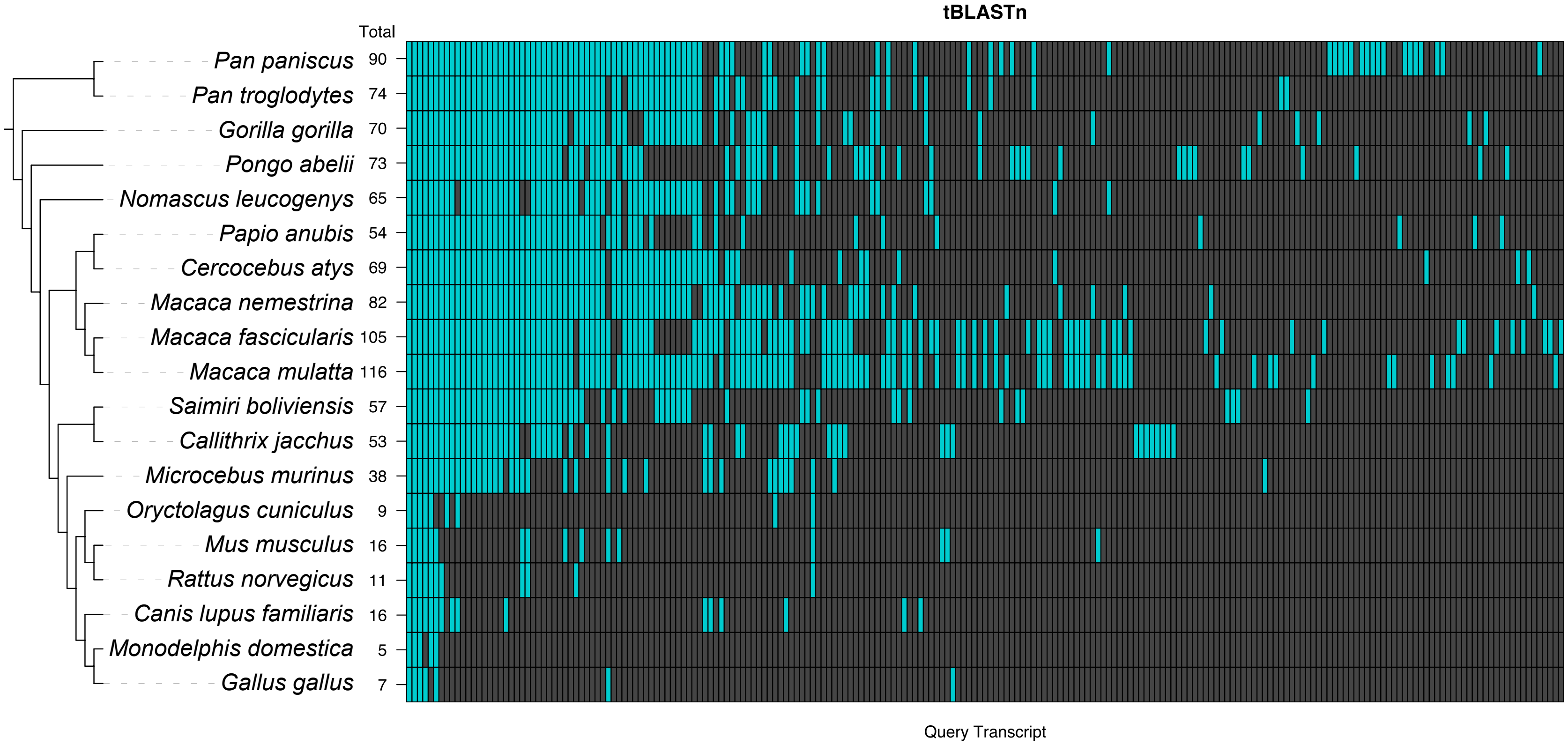


Figure 4.12: **Distribution of tBLASTn hits returned for lncRNA smORF peptides in queried non-human species.** Initial results from tBLASTn v2.9.0+ (Altschul et al., 1990), as described in Figure 4.4B. Human results have not been included as all but 17 query smORFs returned one or more hits in human. The phylogeny shows the species included in the conservation analysis, and the total number of lncRNA smORF peptide queries which returned hits in each species is given at each branch end. Each column represents a lncRNA smORF peptide query, and the blue cells represent one or more hits that were returned for that query in a given species. Dark grey cells indicate that no hits were returned.

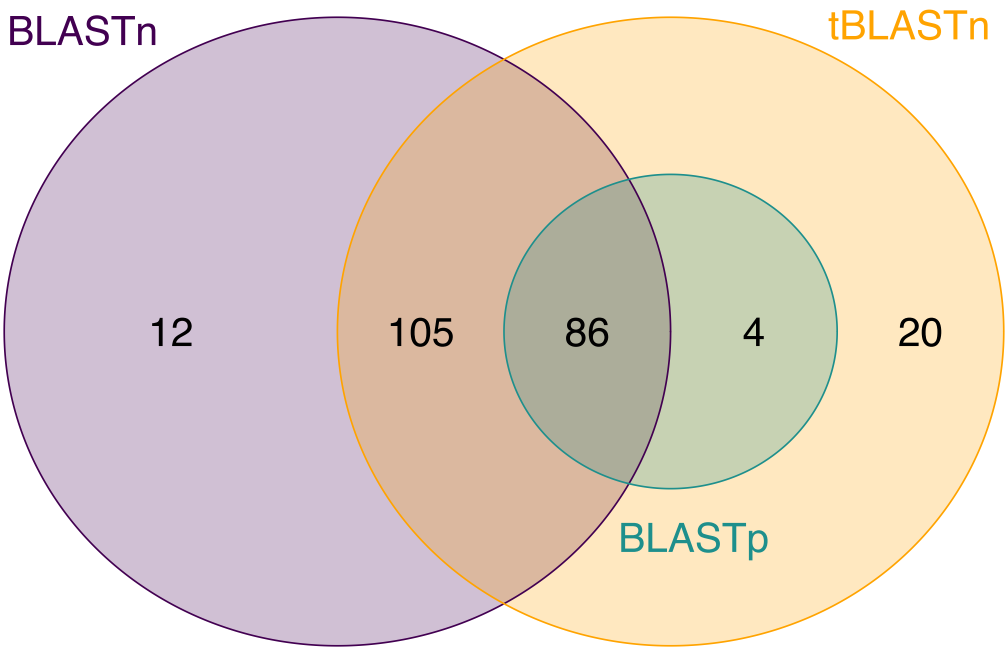


Figure 4.13: **LncRNA smORF queries returning one or more BLAST hits in non-human species.** Intersection between the number of queries returning one or more BLAST hits in non-human species in BLASTp, BLASTn, and tBLASTn (Altschul et al., 1990) searches.

Table 4.6: **Summary of translated lncRNA peptide lengths.** Average measures of the peptide lengths of all translated lncRNA peptide queries, peptides returning possible evidence of conservation from one or more queried non-human species using BLASTp, BLASTn or tBLASTn, peptides which did not return any results from non-human species, and peptides which only returned hits from non-human species using BLASTn.

	Mean	**Mode**	**Median**	**Range**
All peptide queries	64.3 aa	26 aa	54 aa	3 - 272 aa
BLAST hits in one or more non-human species	72.6 aa	37 aa	60 aa	17 - 272 aa
No BLAST hits in non-human species	49.4 aa	26 aa	42 aa	3 - 193 aa
Hits in non-human species from BLASTn only	33.7 aa	18 aa	34.5 aa	17 - 56 aa

that it was reasonable to expand our dataset to include all 355 smORFs for this analysis, given that a comparable proportion of the smORFs returned possible evidence of sequence conservation.

4.3.3 Assessing sequence conservation in translated lncRNA smORFs

To investigate if each set of putative homologs could represent true homologs of the lncRNA smORF queries, they were aligned to the lncRNA smORF sequences. The quality of these alignments were then assessed as described in Section 4.2.5.

4.3.3.1 Alignment of putative homologs

To allow the levels of smORF sequence conservation to be assessed, each translated lncRNA smORF query was aligned with all of their remaining subject hits from each search strategy using three different algorithms, as described in 4.2.4. MetAl v1.1.0 (Blackburne and Whelan, 2012) found no significant difference between alignment methods for 84 of the 215 ORF-centric aa queries and their hits, likely due to their short length limiting the amount of potential high confidence alignments in alignment space. However, 90/215 of these queries' alignments were unable to be compared by MetAl as MUSCLE v5.0.1428 (Edgar, 2004) handled gap characters differently to the other algorithms, creating alignments of differing length (Table 4.7). These queries were therefore only compared using noRMD v1.2 (Thompson et al., 2001). One of the transcript-centric queries (ENST00000442007.1_154_283) and its hits were aligned by MAFFT v7.487 (Katoh and Standley, 2013) alone due to a high number of remaining hits; therefore this alignment was not assessed using MetAl v1.1.0 (Blackburne and Whelan, 2012).

MAFFT v7.487 (Katoh and Standley, 2013) produced the "best" alignment for the majority of queries in all search strategies ($\sim$53%), followed by ClustalW v2.1 ($\sim$29%) (Larkin et al., 2007) (Table 4.7). MUSCLE v5.0.1428 (Edgar, 2004) was selected for less than 5% of the alignments.

4.3.3.2 Assessment of alignments

To analyse the conservation of each smORF, the alignments created from all three BLAST v2.9.0+ (Altschul et al., 1990) strategies were assesed using the metrics described in Section 4.2.5; the HoT score (Landan and Graur, 2007), odseq v1.22.0 (Jiménez, 2022) results, and the species hits were returned in, alongside visual examinations. The translated lncRNA smORFs were categorised as A) No BLAST results, B) No convincing evidence, C) Results in 1 or 2 species, or D) Evidence of sequence conservation, based on these results.

$\sim$6% of lncRNA smORF alignments revealed no convincing evidence of conservation

Only 15/227 smORFs with BLAST v2.9.0+ (Altschul et al., 1990) hits exhibited no convincing evidence (Category B). SmORFs in this category had low to medium HoT scores (Landan and Graur, 2007) for the whole alignment for the majority of the BLAST (Altschul et al., 1990) search strategies, the human sequences were outliers compared to the other species according to odseq v1.22.0 (Jiménez, 2022), and visual examination of the alignment did not reveal any well aligned modules of conservation, particularly over the smORF. However, it is likely that if some of these smORFs were examined in more depth, for example using 1-to-1 alignments of the smORF and each result, evidence to support re-categorisation of this smORF into Category C) or even D) may be found. A limitation of this study is the time required to carry out this depth of analysis on the 227 translated lncRNA smORFs which returned BLAST results.

Table 4.7: **Alignment methods selected for BLAST results using MetAl and noRMD.** Search strategy refers to the different sequence homology searches performed on the translated lncRNA smORFs using the BLAST (Altschul et al., 1990) suite of programs. No significant difference; MAFFT refers to the number of queries which were found to have no significant difference between each alignment algorithm by MetAl v1.1.0 (Blackburne and Whelan, 2012), and therefore aligned using MAFFT 7.487 (Katoh and Standley, 2013). MAFFT, MUSCLE and ClustalW refer to the number of queries for which MAFFT v7.487 (Katoh and Standley, 2013), MUSCLE v 5.0.1428 (Edgar, 2004), or ClustalW v2.1 (Larkin et al., 2007) were found to be the "best" aligner by noRMD v1.2 (Thompson et al., 2001). MetAl failed refers to the number of alignments for which MetAl v1.1.0 (Blackburne and Whelan, 2012) was unable to compute comparisons. In these cases the alignments were only assessed using noRMD v1.2 (Thompson et al., 2001).

Search strategy	No significant difference; MAFFT	MAFFT	MUSCLE	ClustalW	MetAl failed
Peptide-centric	13	55	8	14	0
Transcript-centric	2	119	5	77	1
smORF-centric (nt)	2	127	8	78	0
smORF-centric (aa)	84	80	12	39	90

Putative homologs found in 1 or 2 species only for ~35% of translated lncRNA smORFs

Category C) smORFs only returned hits in one or two species, so even if the hits aligned well with the smORF, they could not be taken as evidence of sequence conservation. This is because the hits could be an artefact of the methods used to create the human genome annotation, or the annotations of the other species. If a smORF is under selective pressure and well conserved, we would expect to see some evidence in other closely related species. This evidence may be missing because the transcripts these smORFs are in are not annotated yet. Alternatively they may not exist, and the conservation we see between a small number of species is due to mis-annotation. A further, unlikely option is that the smORFs are truly under strong selective pressure in these species, and have been lost in the other queried species. Depending on the phylogenetic position of the species the smORF was found in, this may require several independent loss events.

A total of 120 smORFs returned results in 2 queried non-human species or fewer within all three BLAST (Altschul et al., 1990) strategies (Figure 4.14); 110 of which were classified as Category C), although 10 are in B) (No convincing evidence). Only 22 of these smORFs returned BLASTp results, with the majority in *Pan paniscus* (bonobo); the queried species most closely related to human. Approximately 39% of the 120 the smORFs returned BLAST hits from *Macaca fascicularis* (crab-eating macaque) or *Macaca mulatta* (macaque) (Figure 4.14), likely due to differences in their annotation methods (Ensembl, 2019; Ensembl, 2021a) compared to the other queried primates (Ensembl, 2018; Ensembl, 2021b), which were annotated using the same pipeline. This difference is reflected in the number of transcripts in the annotations (Table C.1), as *Macaca fascicularis* (crab-eating macaque) and *Macaca mulatta* (macaque) have 6,628 and 4,773 annotated lncRNAs respectively, compared to only 640 annotated lncRNAs in *Macaca nemestrina* (pig-tailed macaque).

Evidence of sequence conservation found for ~27% of translated lncRNA smORFs

The original set of 45 ‘stringent’ lncRNA smORFs approximately split into thirds, with 31% smORFs falling into categories A) and B) with no results, or no convincing evidence of sequence conservation. 35% of the smORFs were in category C)(results in 1 or 2 species), and 33% were classified as D) with convincing evidence of sequence conservation.The larger, ‘non-stringent’ dataset of 355 translated lncRNA smORFs had a similar split; 40% were categorised as A) or B), 31% as C), and 29% as D. The majority of smORFs in all categories were from antisense or intergenic lncRNA genes, reflecting the proportions of lncRNAs annotated in the human genome, as these are more straightforward to identify.

Some of the translated lncRNA smORFs were not unique, as for genes with multiple transcripts, or transcripts with multiple splice variants, slight variations on the same smORF could be found to be translated in different replicates by Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016). In cases where these smORF isoforms (Figure 4.9) are present, their results have been grouped, taking the "best" category result for each group. If two smORFs were similar but not identical, they were grouped if at least half of the peptide sequence was shared. Using this strategy the 355 lncRNA smORF isoforms collapse into 242 smORFs. This provides a more accurate estimate of the number of smORFs found with evidence of sequence conservation, with 65/242 translated lncRNA smORFs exhibiting convincing evidence of conservation (category D) (Figure 4.15). To test if there was a significant association between being in the grouped ‘stringent’ (37 smORFs) or ‘non-stringent’ datasets (242 smORFs) and the level of conservation found a Chi-square goodness of fit test was performed. The proportions of smORFs in categories A to D did not differ by stringency, X^2(3,242)=5.82, p=.1209. This supports the inclusion of the wider pool of lncRNA smORFs, given that a similar proportion returned convincing evidence of sequence conservation as in the ‘stringent’ set.

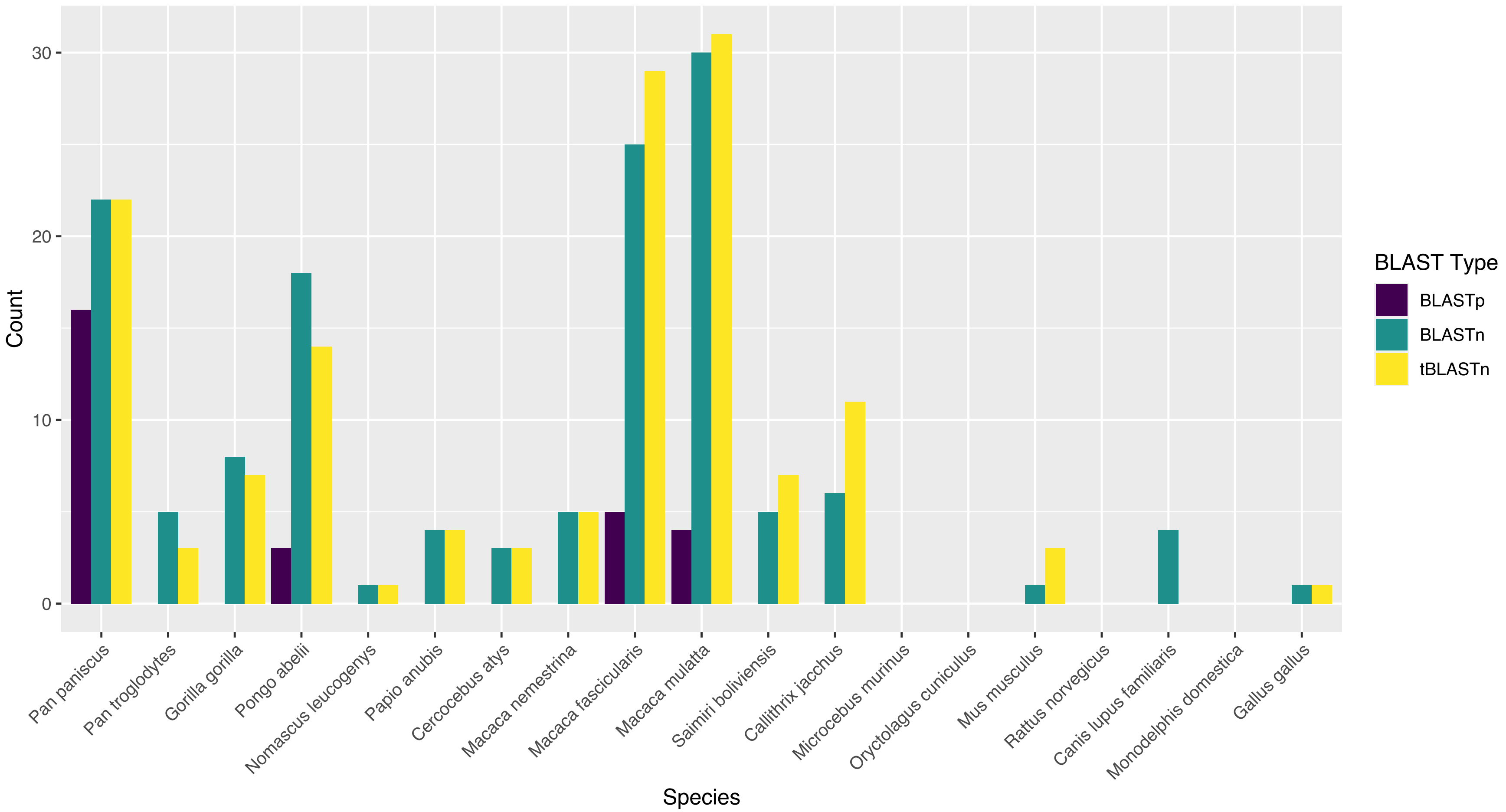


Figure 4.14: **Translated lncRNA smORFs returning results in 2 queried non-human species or fewer.** Count of translated lncRNA smORFs which returned results in 2 queried non-human species or fewer within all three BLAST (Altschul et al., 1990) strategies. The plot shows colour coded counts for the number of these smORFs returned by each search strategy, in each species.

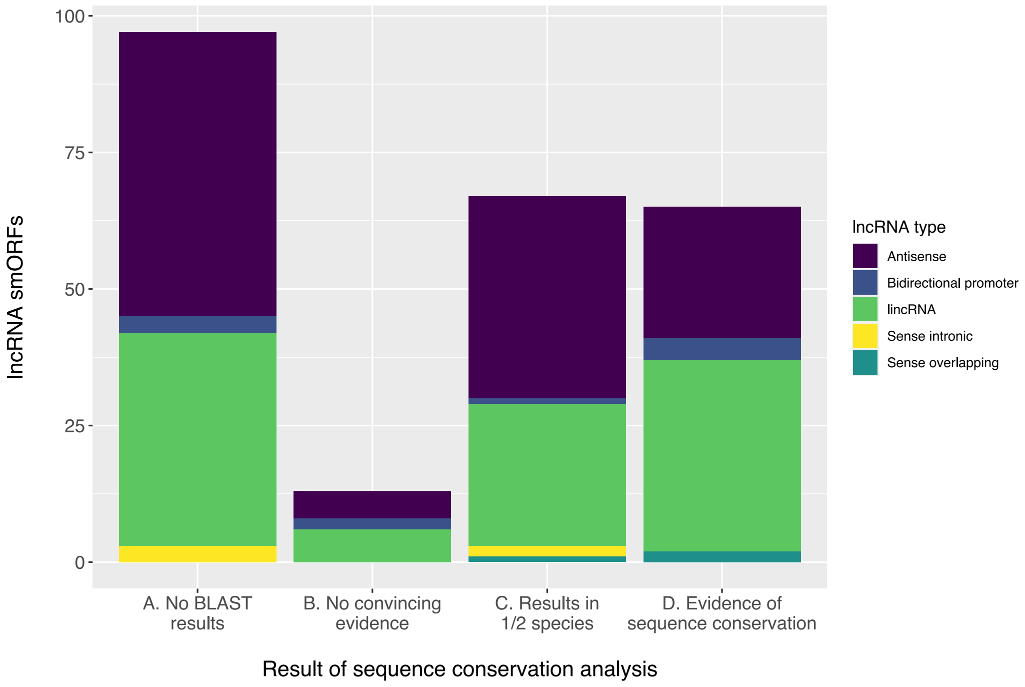


Figure 4.15: **Evidence of sequence conservation in grouped lncRNA smORF isoforms.** The number of smORFs categorised as A) No BLAST results, B) No convincing evidence, C) Results in 1/2 species, and D) Evidence of sequence conservation, when smORFs isoforms were combined. Bars are colour coded according to the type of lncRNA gene the smORFs originate from; antisense, bidirectional promoter, lincRNA, sense intronic, or sense overlapping lncRNA genes.

4.3.4 Characterisation of the conservation of individual translated lncRNA smORFs

In the above section, all of the results from the BLAST v2.9.0+ (Altschul et al., 1990) analysis have been summarised. Here, these results are intepreted in the context of specific translated lncRNA smORFs, using selected examples from categories B) No convincing evidence, C) Results in 1 or 2 species, and D) Evidence of sequence conservation. To understand the biological implications of these results they are discussed in the context of evidence of functionality from this work and the literature.

4.3.4.1 LIPT2-AS1-201: smORFs A and B

LIPT2-AS1-201 (ENST00000526036.1) is a 2,493 nt lncRNA transcript transcribed from LIPT2-AS1 (ENSG00000254837.2) which contains two translated smORF isoforms (Table 4.16A., Figure 4.16B.). LIPT2-AS1-201 is expressed in 43% of the human tissues included in the SyntDB database (Bryzghalov, Szcześniak, and Makałowska, 2020), is highly expressed in neural cell and neural progenitor cells, and is down regulated in large-cell medulloblastoma (Birks et al., 2013), the second most common paediatric brain tumour. The two smORF isoforms are ENST00000526036.1_1226_2024, referred to as 'smORF A', and ENST0000052603-6.1_1340 _2024, referred to as 'smORF B'. Both of these smORFs are examples from conservation category D) (evidence of sequence conservation). SmORF A (272 aa) represents an N-terminal extension of smORF B, as smORF A starts slightly upstream of and in the same frame as smORF B (234 aa), and they end at the samnde stop codon (Figure 4.16.B).

Putative homologs of smORF A and B returned in all 18 sampled species

SmORFs A and B are particularly interesting as they returned BLAST v2.9.0+ (Altschul et al., 1990) hits from all species in the dataset, including the outgroup *Gallus gallus* (chicken) for smORF A (Figure 4.17). The majority of lncRNAs are thought to be lineage-specific (Hezroni, Koppstein, et al., 2015), so the possibility of human lncRNA sequence conservation outside of mammals or even primates is intriguing. SmORF A was identified as conserved in our preliminary analysis (Douka, Birds, et al., 2021), although as a smaller set of species were used this was only observed in *Pongo abelii* (orangutan).

Hits were found in all queried non human species using BLASTp and tBLASTn for smORF A. SmORF B also returned hits in all non human species using BLASTp and tBLASTn except the outgroup *Gallus gallus*, as the *Gallus gallus* peptide aligned to the N terminus of the smORF A peptide which is which is not found in smORF B. The majority of hits were the same for the two smORFs, aside from the aforementioned *Gallus gallus* result, and different splice variants were returned from the same *Papio anubis* gene (Table 4.8). Results were only returned for four species for both smORFs using BLASTn. This is likely due to BLASTn searches being performed using nucleotide sequences, as opposed to the amino acid sequences used in the other BLAST searches. At the nucleotide level, synonymous substitutions can occur which produce the same aa sequence, meaning that the nucleotide sequence of a transcript can diverge even when under selective pressure. Although we have attempted to account for this by initially searching with the full lncRNA transcript nucleotide sequence, followed by the smORF nucleotide sequence, the non smORF portions of the transcript may also have diverged further, meaning possible transcript hits were unable to pass the e-value cut-off filter.

A.

smORF	Replicate	Frame	Start	Stop	P-sites	P-sites in frame
smORF A	Rep 1 Control	2	1226	2042	43	67%
smORF A	Rep 2 Control	2	1226	2042	58	79%
smORF A	Rep 3 Control	2	1226	2042	36	94.4%
smORF A	Rep 3 RA	2	1226	2042	99	79%
smORF B	Rep 2 RA	2	1340	2042	91	70%

B.

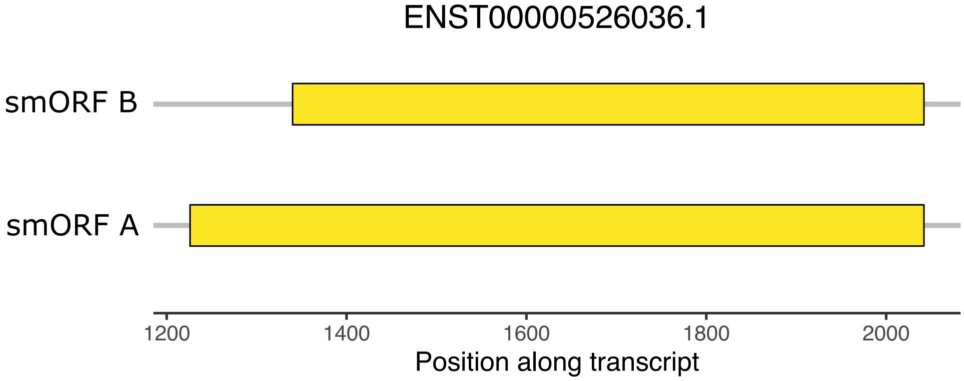


Figure 4.16: **Summary of translated smORFs identified in LIPT2-AS1-201.** A) smORF refers to the identifier used for each smORF in this chapter. SmORF ID refers to the identifier given to the smORF, based on the transcript ID and the position of the smORF along the transcript. Replicate refers to the Poly-Ribo-Seq replicate the smORF was identified in, and under which condition (Control or Retinoic Acid). Start and stop refer to the positions of the start and stop codons of the smORF along the transcript. P sites refer to the number of ribosome footprints identified along the smORF, and P sites in frame gives the number of those footprints that are in frame with the identified smORF. B) Two potential translated lncRNA smORFs were identified in LIPT2-AS1-201; smORF A (ENST00000526036.1_1226_2024) and smORF B (ENST00000526036.1_1340_2024). The plot shows their relative positions along the transcript.

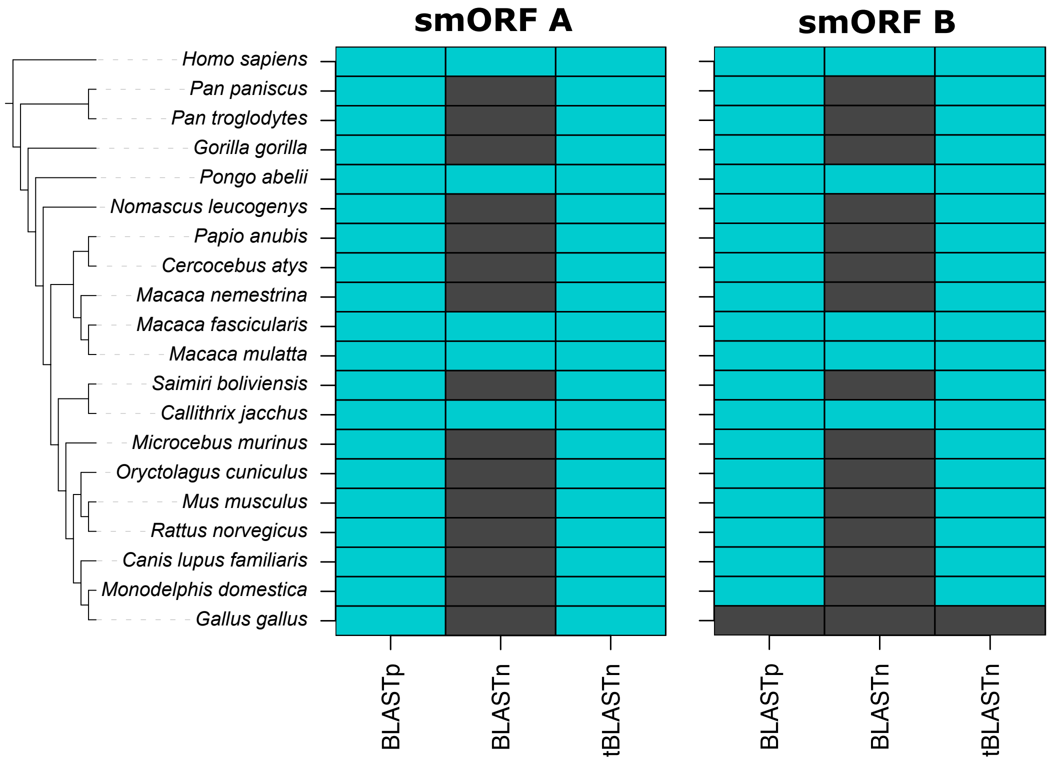


Figure 4.17: **Summary of BLAST search results for translated lncRNA smORFs identified in LIPT2-AS1-201.** The phylogeny shows the species included in the conservation analysis. The columns summarise the BLAST v2.9.0+ (Altschul et al., 1990) hits returned for LIPT2-AS1-201 smORF A (ENST00000526036.1_1226_2024) and smORF B (ENST00000526036.1_1340_2024). Each column represents a BLAST search, and the blue cells represent a one or more hits that were returned for that BLAST search in a given species. Dark grey cells indicate that no hits were returned.

Table 4.8: **BLAST results for translated lncRNA smORFs identified in LIPT2-AS1-201.** 'Species' refers to the included species, by their common and scientific name. 'Hit transcript' refers to transcripts from the included species, and 'BLASTn' and 'tBLASTn' denote whether LIPT2-AS1-201 smORF A or smORF B returned these transcripts from BLASTn and tBLASTn searches. 'Hit peptide' refers to peptides from the included species, and 'BLASTp' denotes which of the smORFs returned these peptides from BLASTp searches. 'Syntenous?' gives a brief description of the position of the gene, in comparison to LIPT2-AS1 which is bidirectional to the human LIPT2-201 gene.

Species	Hit transcript	BLASTn	tBLASTn	Hit peptide	BLASTp	Syntenous?
Marmoset (*Callithrix jacchus*)	ENSCJAT00000096635.1	smORF A, B	smORF A, B	ENSCJAP00000069628.1	smORF A, B	Yes, bidirectional to LIPT2-201
Macaque (*Macaca mulatta*)	ENSMMUT00000018364.4	smORF A, B	smORF A, B	ENSMMUP00000036837.3	smORF A, B	Yes, bidirectional to LIPT2-201
Crab-eating macaque (*Macaca fascicularis*)	ENSMFAT00000084317.1	smORF A, B	-	ENSMFAP00000047715.1	-	No, STX11-204
Crab-eating macaque (*Macaca fascicularis*)	ENSMFAT00000085873.1	smORF A, B	smORF A, B	ENSMFAP00000061433.1	smORF A, B	Yes, bidirectional to LIPT2-201
Orangutan (*Pongo abelii*)	ENSPPYT00000034012.1	smORF A, B	smORF A, B	ENSPPYP00000024748.1	smORF A, B	Yes, bidirectional to LIPT2-201
Chicken (*Gallus gallus*)	ENSGALT00000062186.3	-	smORF A	ENSGALP00000048513.3	smORF A	No
Opossum (*Monodelphis domestica*)	ENSMODT00000035523.2	-	smORF A, B	ENSMODP00000058528.1	smORF A, B	No, JRK-201
Dog (*Canis lupus familiaris*)	ENSCAFT00000089198.1	-	smORF A, B	ENSCAFP00000075353.1	-	No, JRK
Dog (*Canis lupus familiaris*)	ENSCAFT00000015636.2	-	-	ENSCAFP00000014470.2	smORF A, B	No, TIGD2-201. Archived gene.
Rabbit (*Oryctolagus cuniculus*)	ENSOCUT00000054082.1	-	smORF A, B	ENSOCUP00000029923.1	smORF A, B	No
Rat (*Rattus norvegicus*)	ENSRNOT00000000199.4	-	smORF A, B	ENSRNOP00000007697.2	smORF A, B	No, Jrk-202
Mouse (*Mus musculus*)	ENSMUST00000050234.4	-	smORF A, B	ENSMUSP00000051842.3	smORF A, B	No, Jrk-201
Mouse Lemur (*Microcebus murinus*)	ENSMICT00000063729.1	-	smORF A, B	ENSMICP00000048026.1	smORF A, B	No, JRK-201

Table 4.8 continued from previous page

Species	Hit transcript	BLASTn	tBLASTn	Hit peptide	BLASTp	Syntenous?
Bolivian squirrel monkey (*Saimiri boliviensis boliviensis*)	ENSSBOT00000032997.1	-	smORF A, B	ENSSBOP00000016192.1	smORF A, B	No, JRK-201
Sooty mangabey (*Cercocebus atys*)	ENSCATT00000069085.1	-	smORF A, B	ENSCATP00000044643.1	smORF A, B	No, JRK-201
Olive baboon (*Papio anubis*)	ENSPANT00000008710.2	-	smORF B	ENSPANP00000002239.1	-	No, JRK-203
Olive baboon (*Papio anubi*)	ENSPANT00000041574.1	-	smORF A	ENSPANP00000037301.1	smORF A, B	No, JRK-202
Pig-tailed macaque (*Macaca nemestrina*)	ENSMNET00000066965.1	-	smORF A, B	ENSMNEP00000042462.1	smORF A, B	No, JRK-201
Gibbon (*Nomascus leucogenys*)	ENSNLET00000001586.2	-	smORF A, B	ENSNLEP00000001501.1	smORF A, B	No, JRK-201
Gorilla (*Gorilla gorilla gorilla*)	ENSGGOT00000030873.2	-	smORF A, B	ENSGGOP00000018106.2	smORF A, B	No, JRK-201
Chimpanzee (*Pan troglodytes*)	ENSPTRT00000093865.1	-	smORF A, B	ENSPTRP00000069879.1	smORF A, B	No, JRK-201
Bonobo (*Pan paniscus*)	ENSPPAT00000052462.1	-	smORF A, B	ENSPPAP00000029612.1	smORF A, B	No, JRK-201

Subset of smORF A and B hits returned due to a shared Helix-turn-Helix motif

Two clear categories of hits emerged from the smORF A and B BLAST results. The first were peptides originating from various Jrk HtH proteins (Table 4.8), including two annotated proteins from the human Jrk HtH gene; ENSP00000483808.1 and ENSP00000461610.1. The HtH motif is an abundant DNA binding motif, often found in proteins involved in the regulation of gene expression (Roy and Kundu, 2021). To determine if the smORF A and B peptides shared any motifs with these proteins, a domain analysis was performed using PfamScan v1.6 (Mistry, Bateman, and Finn, 2007) (as described in 3.2.6). This identified a HTH_Tnp_Tc5 domain at the N-terminus of both peptides.

To understand how well the HtH motif was conserved across the BLAST hits, the tBLASTn results for LIPT2-AS1-201 (Figure 4.18) were compared to the HMM logo for the HTH_Tnp_Tc5 family (Figure 4.19). The positions most conserved within the model of the motif are also well conserved across the aligned results from the smORF A tBLASTn search. An exception is the *Gallus gallus* (chicken) result (ENSGALT00000062186.3), which does not originate from a JRK gene, although it does appear to share some of the motif. The HtH motif could be indicative of smORF A having a function in DNA binding, particularly given its high levels of conservation. SmORF B, however, begins halfway through the motif (highlighted in pink in Figure 4.18), so is unlikely to function in this capacity.

Syntenous conservation of LIPT2-AS1 in Primates

The second category of BLAST hits for smORF A and B were sytenous to LIPT2-AS1. As lncRNA sequences are often poorly conserved, it can be important to consider their position on the chromosome relative to canonical protein coding genes to find evidence of conservation, so it is encouraging that this result has arisen from a sequence based approach.

LIPT2-AS1 is a bidirectional lncRNA gene which overlaps with the promoter of the protein coding LIPT2-201 gene. The BLAST hits returned from the monkeys *Callithrix jacchus, Macaca mulatta*, and *Macaca fascicularis* (Marmoset, macaque and crab-earing macaque), and the great ape *Pongo abelii* (orangutan) were all protein coding genes, bidirectional to LIPT2-201 genes in these species (Table 4.8). Aside from an extra *Macaca fascicularis* transcript, these were the only non human results returned from the BLASTn searches for both smORFs. When aligned, the ATG of smORF B is conserved in all four syntenous transcripts, whilst the ATG of smORF A has been replaced by TTG in both *Macaca* results (Figure 4.20). However an upstream, in frame ATG is conserved in all four species and in human, which would enable the translation of a longer peptide. Indeed, the canonical peptides translated from the *Macaca mulatta, Macaca fascicularis* and *Pongo abelii* results begin from this start codon, highlighted in green in Figure 4.20. Of all the hits returned from tBLASTn, the peptides identified in these four syntenous transcripts aligned most closely with the human query peptides (Figure 4.18; highlighted in orange), with pairwise identity distances from smORF A ranging from 19% - 26%.

Adjacent to this group, but not syntenous, was the peptide returned from *Oryctolagus cuniculus* (rabbit) (ENSOCUP00000029923.1) which had a pairwise identity distance of 52%. When the full canonical peptides returned from these species are aligned with the other BLASTp results (Figure 4.20), the *Oryctolagus cuniculus* peptide aligned well with the start of smORF A, followed by a large insertion and lower levels of similarity across the rest of the peptide.

Although the human LIPT2-AS1 gene currently has no annotated orthologs, using the syntenous *Callithrix jacchus* gene as a query in Ensembl Release 107 (Cunningham et al., 2022) returned results for four of our species of interest (Table 4.9). The *Macaca fascicularis*, *Macaca mulatta*, and *Oryctolagus cuniculus* genes match the results identified by BLAST. Both of the macaque results had a Gene Order Conservation (GOC) score of 100, indicative of strong

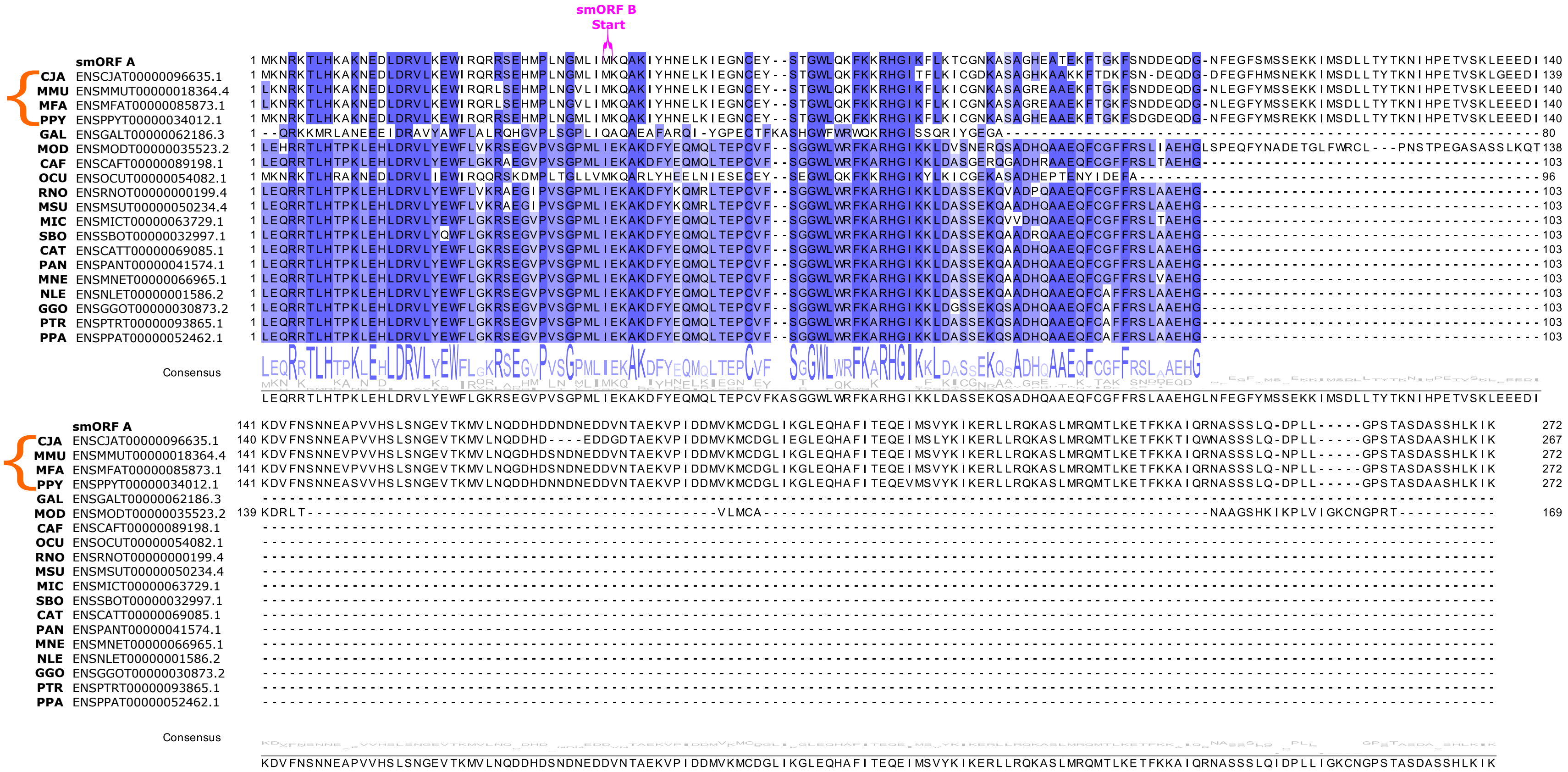


Figure 4.18: **Alignment of results from LIPT2-AS1-201 smORF A tBLASTn search.** Results returned from queried non-human species using tBLASTn v2.9.0+ (Altschul et al., 1990). The alignment is coloured by percentage identity, and the consensus sequence shown below. The four syntenous results are highlighted in orange, and start of the smORF B peptide is highlighted in pink. Created in Jalview v2.10.2b2 (Waterhouse et al., 2009).

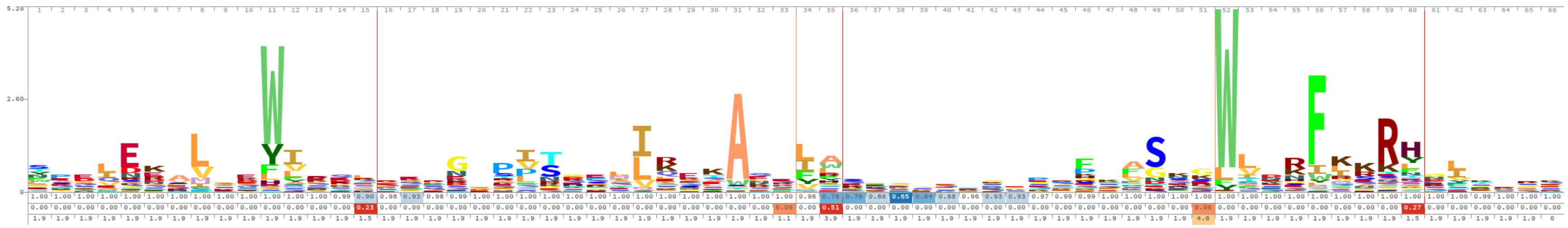


Figure 4.19: **Hidden Markov Model logo for the HTH_Tnp_Tc5 family.** The height of the each letter stack corresponds to the level of conservation at that position, and the height of each letter within a stack corresponds to the frequency of that letter at that position. The table at the bottom of the figure shows occupancy - the probability of observing a letter at that position, as opposed to a gap - in the top row, where a darker blue background indicates lower occupancy. Insert probability is in the middle row, giving the probability of an insertion of one or more letters between that position and the next letter in the sequence. The bottom row shows expected insert length, the expected length of an insertion between that postion and the next letter in the sequence, if one is present. For both insert rows, a darker red background indicates higher values When a cell in the insert probability row is highlighted, a vertical bar of the same colour indicates where the insertion would be in the sequence. From Pfam (Schuster-Böckler, Schultz, and Rahmann, 2004).

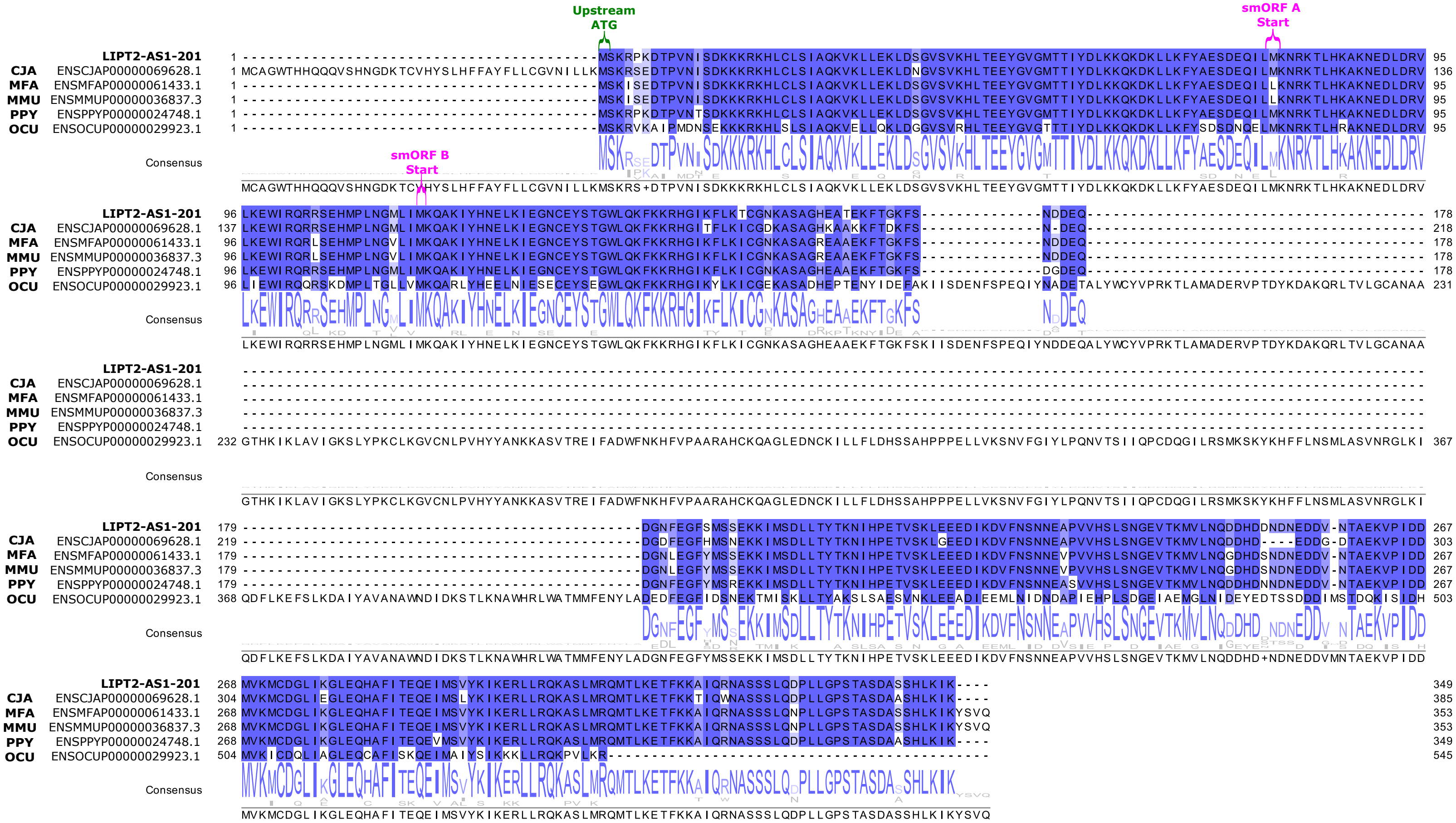


Figure 4.20: **Alignment of selected results from LIPT2-AS1-201 smORF A BLASTp search.** Syntenous results returned from queried non-human species using BLASTp v2.9.0+ (Altschul et al., 1990). The alignment is coloured by percentage identity, and the consensus sequence shown below. The start of the smORF A and B peptides are highlighted in pink. The position of an upstream, conserved methionine is highlighted in green. Created in Jalview v2.10.2b2 (Waterhouse et al., 2009).

Table 4.9: **Ortholog pairs to the *Callithrix jacchus* gene ENSCJAG00000053703.** Results according to Ensembl Release 107 (Cunningham et al., 2022). ‘Species’ refers to the included species, by their common and scientific name. ‘Type’ refers to the type of orthology found between a species pair. ‘Ortholog’ refers to the id of any orthologous genes found. ‘Target %id’ refers to the percentage of the orthologous sequence matching the query sequence. ‘Query %id’ refers to the percentages of the query sequence matching the orthologous sequence. ‘GOC score’ refers to the Gene Order Conservation score, which indicates how many of the four closest neighbours of a gene match between orthologous pairs. ‘WGA coverage’ refers to the Whole Genome Alignment Coverage, a calculation of the coverage of the alignment over the ortholog pair, ranging from 0-100. ‘High confidence’ is a yes/no score of the confidence of the ortholog, based on percentage identity, GOC score, and WGA coverage.

Species	**Type**	**Ortholog**	**Target %id**	**Query %id**	**GOC Score**	**WGA Coverage**	**High Confidence**
Crab-eating macaque (*Macaca fascicularis*)	1-to-1	ENSMFAG00000049105	76.20%	91.81%	100	n/a	Yes
Macaque (*Macaca mulatta*)	1-to-1	ENSMMUG00000013093	76.20%	91.81%	100	n/a	Yes
Olive baboon (*Papio anubis*)	1-to-many	ENSPANG00000036988	75.64%	91.13%	50	n/a	No
Olive baboon (*Papio anubis*)	1-to-many	ENSPANG00000037152	75.64%	91.13%	50	n/a	No
Rabbit (*Oryctolagus cuniculus*)	1-to-1	ENSOCUG00000031325	29.17%	54.27%	0	n/a	No

syntenous conservation around the gene. No results were returned for *Pongo abelii*, as the transcript and associated gene identified by BLAST have been retired from the latest Ensembl annotation. Two more orthologs were also identified in *Papio anubis* (olive baboon), which align well to smORF A (Figure 4.21). These orthologs were not identified by BLAST as they were not in the *Papio anubis* annotation (Panu_3.0) at the time of analysis, and a new annotation (Panubis1.0) has now been released. *Papio anubis* actually has three annotated LIPT2 paralogs with syntenous antisense genes, but smORF A is not conserved in the third antisense paralog. These syntenous results are summarised in Figure 4.22.

smORF A is conserved in euarchontoglires

Based on the species studied, it is likely that smORF A was present in the last common ancestor of *Homo sapiens* and *Oryctolagus cuniculus*, represented by a green star (Figure 4.22). SmORF A may have originated earlier in evolutionary time, but more species would need to be included in the analysis to investigate this. The species in which smORF A is conserved are highlighted by green circles, and the high levels of conservation between their sequences suggests functionality. The *Oryctolagus cuniculus* smORF differs most from smORF A, which is to be expected as it is the most distantly related to *Homo sapiens*. However, the N-terminus of the *Oryctolagus cuniculus* smORF which contains the HtH DNA binding domain is most conserved, suggesting that function may also be conserved.

It is most likely that smORF A is also conserved in the majority of the other queried species which diverged from the last common ancester of *Homo sapiens* and *Oryctolagus cuniculus*, but these orthologs are currently unannotated. Over the course of this analysis orthologous transcripts have been added to and removed from annotations, and we lack the quality of genome annotation to confirm the absence of smORF A in these species. If smORF A truly is absent in any of these species, there may have been a gene loss event, or a change of the function of the gene causing the loss of the smORF sequence.

Interestingly, although all of the orthologs identified are annotated as protein coding, they were all annotated automatically by Ensembl's (Cunningham et al., 2022) pipeline, and have no function assigned to their peptides aside from likely DNA binding due to the HtH motif. Has this difference in annotation simply arisen from the higher levels of scrutiny applied to the human genome?

LIPT2-AS1-201 is translated in Control and RA conditions

SmORF A was was detected as translated in all three Control replicates (Figure 4.16.A), and is therefore part of the stringent lncRNA smORF set. SmORF A exhibits high triplet periodicity (Control Replicate 2; Figure 4.23.B), with 79% of the Ribo-seq reads mapped to frame 1 across the smORF. This bias is indicative of active translation. Very few Ribo-seq reads mapped to the rest of the transcript, outside of the smORF (Figure 4.23C.) SmORF A was also translated in RA conditions, replicate 3, and smORF B was only detected in RA conditions, replicate 2. Given that smORF A was detected in 4 of the 6 samples, it is possible that only smORF A is truely translated, and smORF B was detected due to a low number of reads mapping between the smORF A and B start codons.

To examine if the conserved upstream ATG could be a potential smORF start in LIPT2-AS1-201, footprints upstream of the smORFs were examined. Only a very small number of Ribo-seq reads mapped at this position (Figure 4.24A., postion 995 on the x-axis), demostrating why no translated smORFs were detected starting from this position by Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016)

The non-canonical translation of the lncRNA LIPT2-AS1-201 is also supported by the literature. Analysis of previously published mass spectrometry results in RA-induced SH-SY5Y

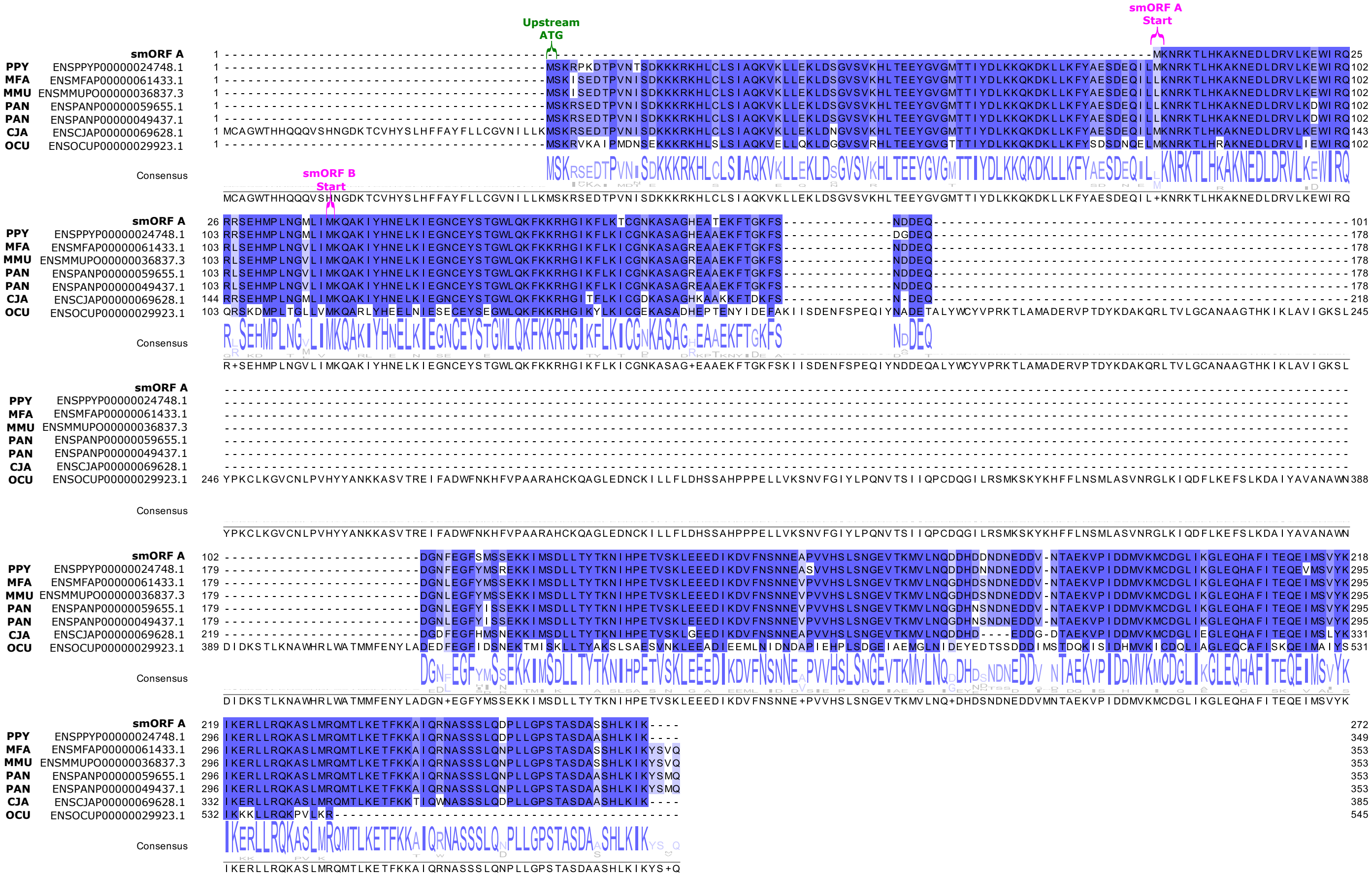


Figure 4.21: **Alignment of conserved LIPT2-AS1-2011 smORF A peptides.** Selected results returned from queried non-human species using BLASTp v2.9.0+ (Altschul et al., 1990) and Ensembl release 107 (Cunningham et al., 2022) queries. The alignment is coloured by percentage identity, and the consensus sequence shown below. The start of the smORF A and B peptides are highlighted in pink. The position of an upstream, conserved methionine is highlighted in green. Created in Jalview v2.10.2b2 (Waterhouse et al., 2009).

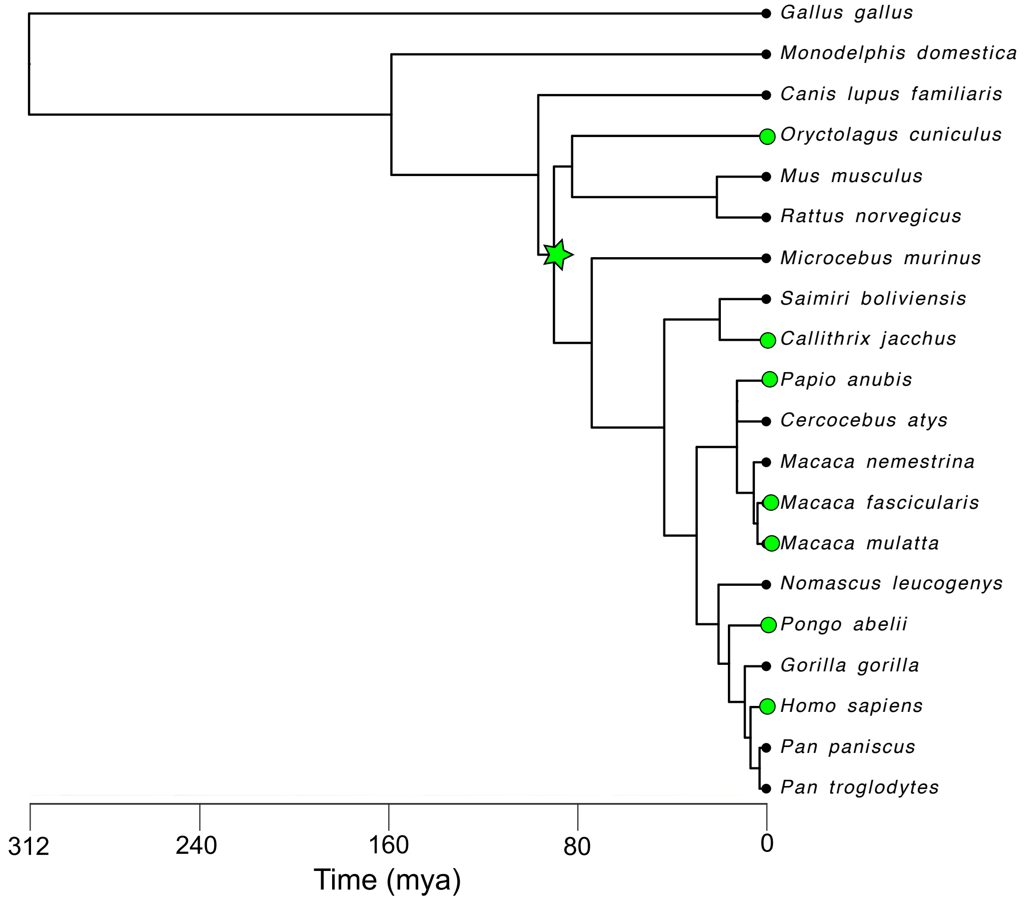


Figure 4.22: **Phylogenetic distribution of smorf A conservation across euarchontoglires.** The phylogeny shows the species included in the conservation analysis, with estimated divergence times in millions of years ago (Kumar et al., 2017). Species in which smORF A is conserved are highlighted in green, and the green star represents the ancestral gene shared by these species.

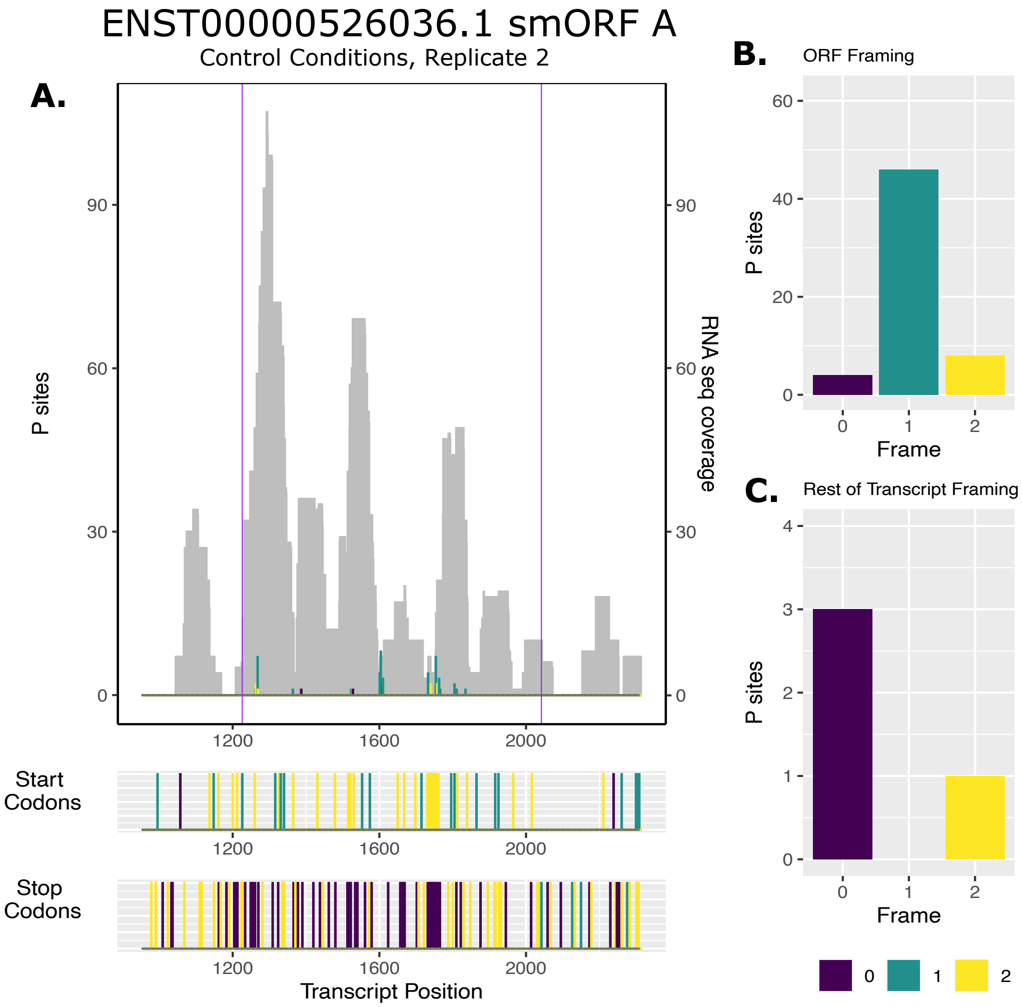


Figure 4.23: **Detail of the Ribo-seq P-sites and framing within LIPT2-AS1-201 smORF A (ENST00000526036.1_1226_2042) in Control Replicate 2.** A) Number of Ribo-seq P-sites in colour, and RNA-seq coverage in grey, across the LIPT2-AS1-201 transcript in Control conditions, replicate 2. Ribo-seq reads are colour coded by frame, arbitrarily designated as frame 0; dark purple, frame 1; turquoise, and frame 2; yellow. The start and end of smORF A is marked by purple lines. The positions of all canonical start and stop codons, colour coded by frame, are shown below the plot. B) P-site framing within the smORF. C) P-site framing without the smORF. 79% of the Ribo-seq reads within the smORF are in the same frame (frame 1), whilst very few Ribo-seq reads were found across the rest of the transcript.

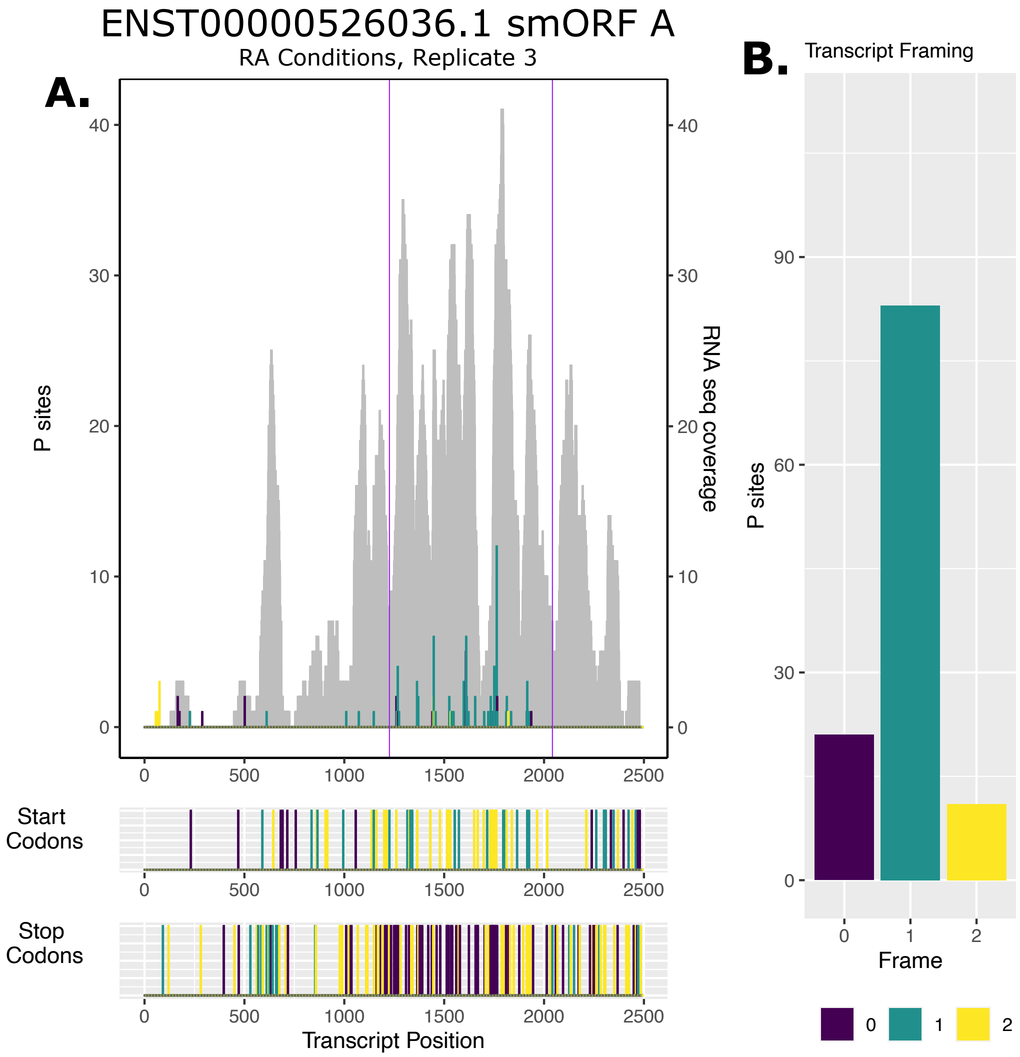


Figure 4.24: **Detail of the Ribo-seq P-sites and framing within LIPT2-AS1-201 in RA replicate 3.** A) Number of Ribo-seq P-sites in colour, and RNA-seq coverage in grey, across the LIPT2-AS1-201 transcript in RA conditions, replicate 3. Ribo-seq reads are colour coded by frame, arbitrarily designated as frame 0; dark purple, frame 1; turquoise, and frame 2; yellow. The start and end of smORF B is marked by purple lines. The positions of all canonical start and stop codons, colour coded by frame, are shown below the plot. B) P-site framing across the transcript.

cells (Brenig et al., 2020) supported the production of a peptide from smORF A. Further, translation from the conserved upstream ATG identified in the conservation analysis performed here (Figure 4.21) has been identified by Ribo-seq in three studies; translated in human embryonic stem cells, HEK293 cells (a human embryonic kidney cell line), and human heart tissue (Calviello, Mukherjee, et al., 2016; van Heesch, Witte, et al., 2019; Gaertner et al., 2020; Mudge et al., 2022). This longer smORF was also found in mass spectrometry data from the HEK293 cells, human heart tissue, and in a further study using human leukocyte antigen bound peptides (Calviello, Mukherjee, et al., 2016; van Heesch, Witte, et al., 2019; Chong et al., 2020; Mudge et al., 2022). Given this evidence and the conservation of the upstream ATG, it seems likely that the upstream ATG represents the "true" start of the smORF, and smORFs A and B represent translation from alternative start sites (Kochetov, 2008).

4.3.4.2 ZEB1-AS1-203: smORF C

From the 67 grouped smORFs that were found to have 1-2 hits in other species (category C), an example was selected to illustrate the potential meaning of such a result. One such translated lncRNA smORF was identified in ZEB1-AS1-203 (ENST00000441257.1), transcribed from ZEB1-AS1 (ENSG00000237036), a gene which overlaps the promoter of the transcriptional repressor ZEB1. ENST00000441257.1_75_243 produces a 56 aa peptide, and going forward will be referred to as smORF C.

SmORF C was translated in RA replicates 2 and 3, and is also described in the literature, identified as translated in two isogenic human cancer cell models - a Src-inducible mammary epithelial model, and a Ras-dependent fibroblast model - and in HEK293 (a human embryonic kidney cell line), HeLa-S3 (a clonal derivative of the HeLa adenocarcinoma cell line), and K562 cells (a myelogenous leukemia cell line) (Ji, Song, et al., 2015; Martinez et al., 2020).

ZEB1-AS1-203 is particularly interesting as the ZEB1-AS1 lncRNA is well established as oncogenic. ZEB1-AS1 is overexpressed in a wide range of cancers, including osteosarcoma (Liu and Lin, 2016), prostate cancer (Su et al., 2017), and gastric cancer, (Ma et al., 2019; Chai et al., 2019), and in pulmonary fibrosis (Qian et al., 2019). A number of mechanisms of action have been proposed for this association, including acting as a molecular sponge or as a ceRNA, but the current consensus is that ZEB1-AS1 upregulates the expression of ZEB1. However, there is not yet any examination of the possible translation of ZEB1-AS1 in the context of cancer or other disease in the literature.

smORF C exhibits modular sequence conservation

SmORF C was previously identified as conserved in our earlier, less stringent work (Douka, Birds, et al., 2021), in which we found evidence of sequence conservation in *Gorilla gorilla*. Hits were found in *Gorilla gorilla* and *Macaca mulatta* by BLASTn (Altschul et al., 1990) and in *Gorilla gorilla* only by tBLASTn (Figure 4.25, Table 4.10). ZEB1-AS1 splice variants were returned in human from BLASTn and tBLASTn. Two other smORFs were identifed in ZEB1-AS1-208, a splice variant of ZEB1-AS1, producing peptides of 21 and 26 aa. The 21 aa peptide has an identical peptide sequence to the C-terminus of smORF C. However, neither returned any BLAST hits, like due to their particularly short lengths.

The N-terminus of the smORF C peptide and the *Gorilla gorilla* aa sequences align well (Figure 4.26), as do the start of the transcripts and the corresponding half of the smORF (Figure 4.27). The *Gorilla gorilla* gene is also syntenous to ZEB1-AS1, as it is antisense to the *Gorilla gorilla* ZEB1-202 gene (Table 4.10). The *Macaca mulatta* transcript, however, originated from *Macaca mulatta* ZEB1-201 (ENSMMUT00000043036.1), which does not currently have any annotated antisense genes. This transcript was returned by BLASTn due to a short, reverse alignment to smORF C, so to properly visualise this alignment the reverse

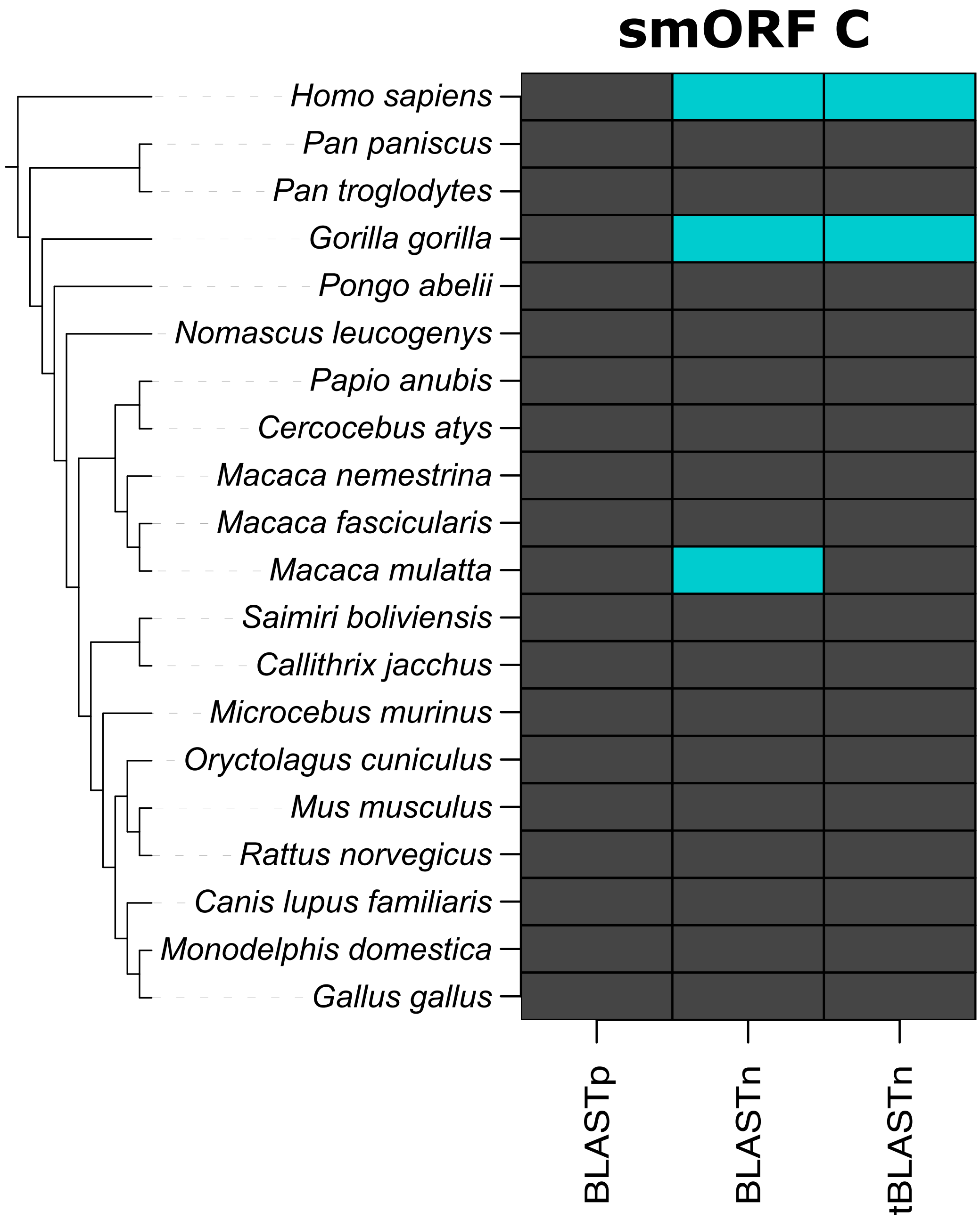


Figure 4.25: **Summary of BLAST search results for translated lncRNA smORFs identified in ZEB1-AS1-203.** The phylogeny shows the species included in the conservation analysis. The columns summarise the BLAST v2.9.0+ (Altschul et al., 1990) hits returned for smORF C (ENST00000526036.1_75_243). Each column represents a BLAST search, and the blue cells represent a one or more hits that were returned for that BLAST search in a given species. Dark grey cells indicate that no hits were returned.

Table 4.10: **BLAST results for the translated lncRNA smORF identified in ZEB1-AS1-203.** 'Species' refers to the included species, by their common and scientific name. 'Hit transcript' refers to transcripts from the included species, and 'BLASTn' and 'tBLASTn' denote whether smORF C returned these transcripts from BLASTn and tBLASTn searches. 'Hit peptide' refers to peptides from the included species, and 'BLASTp' denotes which of the smORFs returned these peptides from BLASTp searches. 'Syntenous?' gives a brief description of the position of the gene, in comparison to ZEB1-AS1.

Species	**Hit transcript**	**BLASTn**	**tBLASTn**	**Hit peptide**	**BLASTp**	**Syntenous?**
Gorilla (*Gorilla gorilla gorilla*)	ENSGGOT00000068843.1	smORF C	smORF C	None; ncRNA	-	Yes, antisense to ZEB1-202
Macaque (*Macaca mulatta*)	ENSMMUT00000043036.3	smORF C	-	ENSMMUP00000036050.3	-	No, ZEB1-201

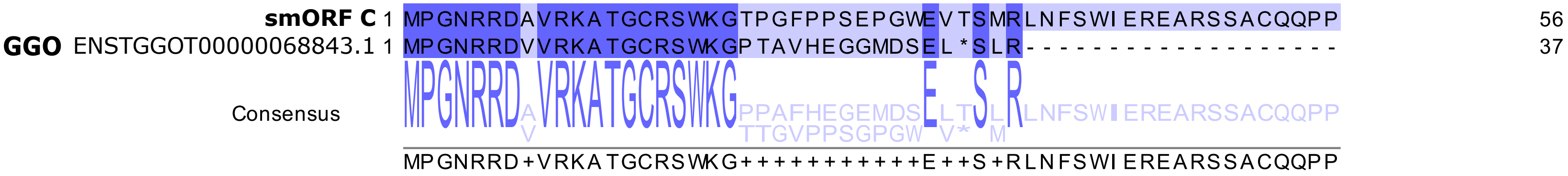

smORF C 1 MPGNRRDAVRKATGCRSWKGTPGFPPSEPGWEVTSMRLNFSWIEREARSSACQQPP 56
GGO ENSTGGOT00000068843.1 1 MPGNRRDVVRKATGCRSWKGPTAVHEGGMDSEL*SLR------------------- 37
Consensus
MPGNRRD+VRKATGCRSWKG+++++++++++E++S+RLNFSWIEREARSSACQQPP

Figure 4.26: **Alignment of results from smORF C tBLASTn search.** Results returned from queried non-human species using tBLASTn v2.9.0+ (Altschul et al., 1990). The alignment is coloured by percentage identity, and the consensus sequence shown below. Created in Jalview v2.10.2b2 (Waterhouse et al., 2009).

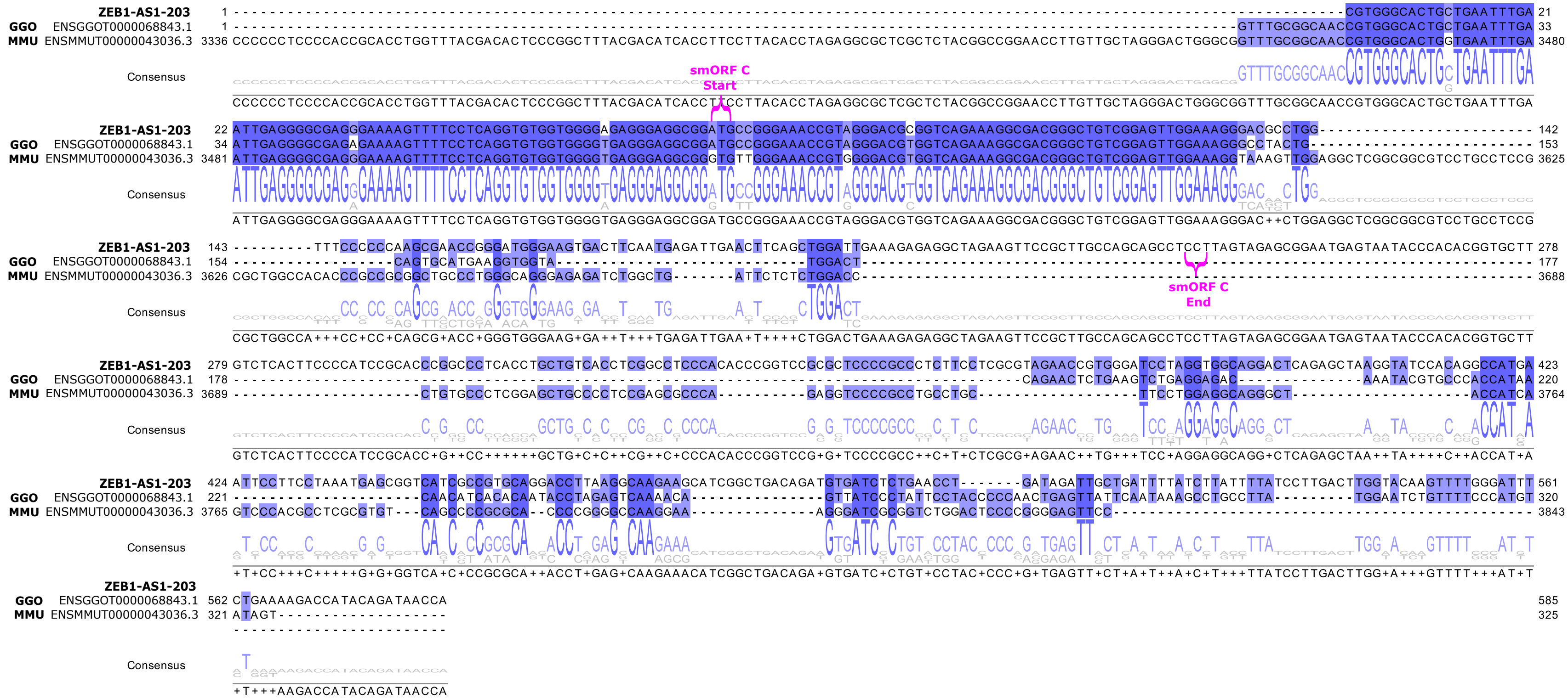


Figure 4.27: **Alignment of results from smORF C BLASTn search.** Results returned from queried non-human species using BLASTn v2.9.0+ (Altschul et al., 1990). The alignment has been cropped to show the area including the ZEB1-201-203 transcript only. The alignment is coloured by percentage identity, and the consensus sequence shown below. The start and end of smORF C are highlighted in pink. Created in Jalview v2.10.2b2 (Waterhouse et al., 2009).

compliment of ENSMMUT00000043036.1 was aligned to the other BLASTn results (Figure 4.27). ENSMMUT00000043036.1 is much longer than the other transcripts at 3843 nt, compared to ZEB1-201-203 at 585 nt, so the figure is also cropped to the aligned portion of the transcript. This small aligned area is at the start of ZEB1-AS1, where it overlaps the start of ZEB1 in human, so may be due to selective pressure on ZEB1.

These alignments show clear modules of conserved sequences in the transcripts, and in *Gorilla gorilla* partial conservation of the peptide sequence. However, these modules overlap the protein coding ZEB1 gene in the antisense direction. Therefore this conservation is likely due to selective pressure acting on the ZEB1 sequence, not on smORF C. Further, if smORF C was truly conserved we would expect to observe some conservation in other closely related species, such as another macaque or great ape. However, given the importance of ZEB1-201 in disease, this and other examples from Category C represent strong candidates for further investigation using more comprehensive genome annotations.

4.3.4.3 LINC03011-201: smORF D

ENST00000445681.1_322_376 is an example of the 13 grouped smORFs with no convincing evidence (category B), which will be refered to as smORF D. SmoRF D produced a 18 aa peptide, and was identified in LINC03011-201 (ENST00000445681.1), transcribed from LINC03011. SmoRF D was translated in RA replicate 3, and the same peptide was also identified in two Ribo-seq studies in the literature from another splice variant of LINC03011 (LINC03011-205), in human cancer cell models; a Src-inducible mammary epithelial model, and a Ras-dependent fibroblast model, and in human embryonic stem cells (Ji, Song, et al., 2015; Gaertner et al., 2020).

Poor sequence conservation over smORF D

SmORF D only returned results from the BLASTn (Altschul et al., 1990) search strategy in two macaque species. Three transcripts from *Macaca fascicularis* and two from *Macaca mulatta* passed the filter thresholds (Figure 4.28, Table 4.11). Alignment of these hits (Figure 4.29) does not reveal any large modules of conservation over the smORF; instead there are smaller areas of sequence similarity, and no conserved start codons.

All of the transcripts returned from non-human species are protein coding; however none of these proteins were returned by the other BLAST search strategies, and they aligned poorly to the short peptide produced from smORF D. Three of the five returned transcripts do exhibit some synteny with LINC03011, as they are upstream of the RABGEF1 gene (Table 4.11). In human, LINC03011 is upstream and antisense to RABGEF1 pseudogene 1.

This analysis finds no convincing evidence that smORF D is conserved in the queried species, although there may be some conservation of the whole lncRNA transcript sequence. There is a possibility that LINC03011 may have a shared evolutionary history with the identified *Macaca fascicularis* and *Macaca mulatta* transcripts given their syntenous genomic postitions and small modules of sequence similarity. For example, they may share a function at the sequence level, as all of the transcripts returned by BLASTn are novel transcripts with no elucidated function as yet. As potentially homologous sequences were only returned from *Macaca fascicularis*(crab-eating macaque) and *Macaca mulatta* (macaque), whose annotations have been already discussed (Section 4.3.3.2), any future investigations of this lncRNA and smORF should also consider other annotations.

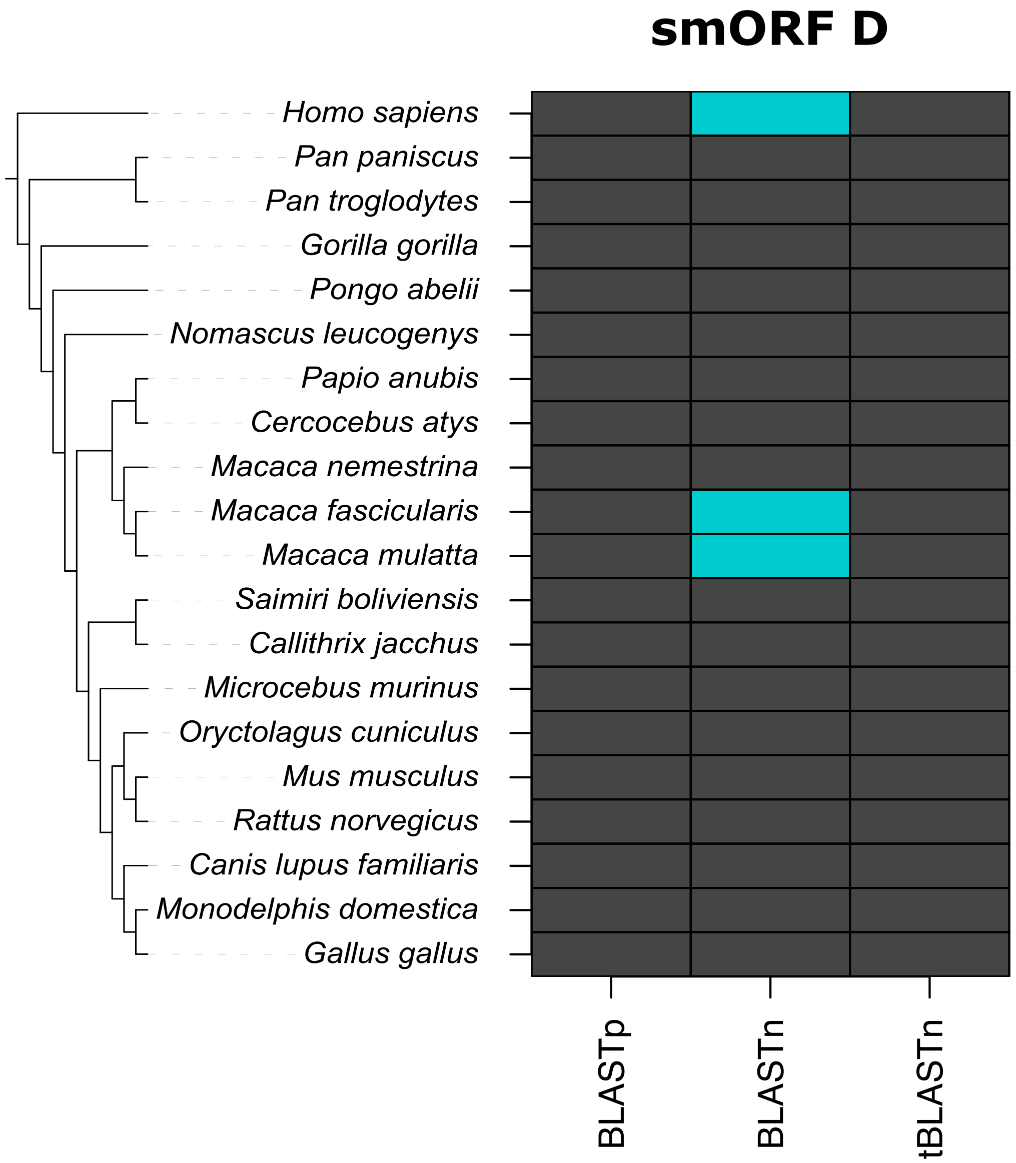


Figure 4.28: **Summary of BLAST search results for translated lncRNA smORFs identified in LINC03011-201.** The phylogeny shows the species included in the conservation analysis. The columns summarise the BLAST v2.9.0+ (Altschul et al., 1990) hits returned for smORF D (ENST00000445681.1_322_376). Each column represents a BLAST search, and the blue cells represent a one or more hits that were returned for that BLAST search in a given species. Dark grey cells indicate that no hits were returned.

Table 4.11: **BLAST results for translated lncRNA smORFs identified in LINC03011-201.** ‘Species’ refers to the included species, by their common and scientific name. ‘Hit transcript’ refers to transcripts from the included species, and ‘BLASTn’ and ‘tBLASTn’ denote whether smORF D returned these transcripts from BLASTn and tBLASTn searches. ‘Hit peptide’ refers to peptides from the included species, and ‘BLASTp’ denotes which of the smORF returned these peptides from BLASTp searches. ‘Syntenous?’ gives a brief description of the position of the gene, in comparison to LINC03011.

Species	**Hit transcript**	**BLASTn**	**tBLASTn**	**Hit peptide**	**BLASTp**	**Syntenous?**
Crab-eating macaque (*Macaca fascicularis*)	ENSMFAT00000089185.1	smORF D	-	ENSMFAP00000052836.1	-	Yes, upstream and antisense to RABGEF1
Crab-eating macaque (*Macaca fascicularis*)	ENSMFAT00000073877.1	smORF D	-	ENSMFAP00000057602.1	-	Yes, upstream of RABGEF1
Crab-eating macaque (*Macaca fascicularis*)	ENSMFAT00000074186.1	smORF D	-	ENSMFAP00000050139.1	-	No
Macaque (*Macaca mulatta*)	ENSMMUT00000107908.1	smORF D	-	ENSMMUP00000081027.1	-	Yes, upstream and antisense to RABGEF1
Macaque (*Macaca mulatta*)	ENSMMUT00000079644.1	smORF D	-	ENSMMUP00000066219.1	-	No

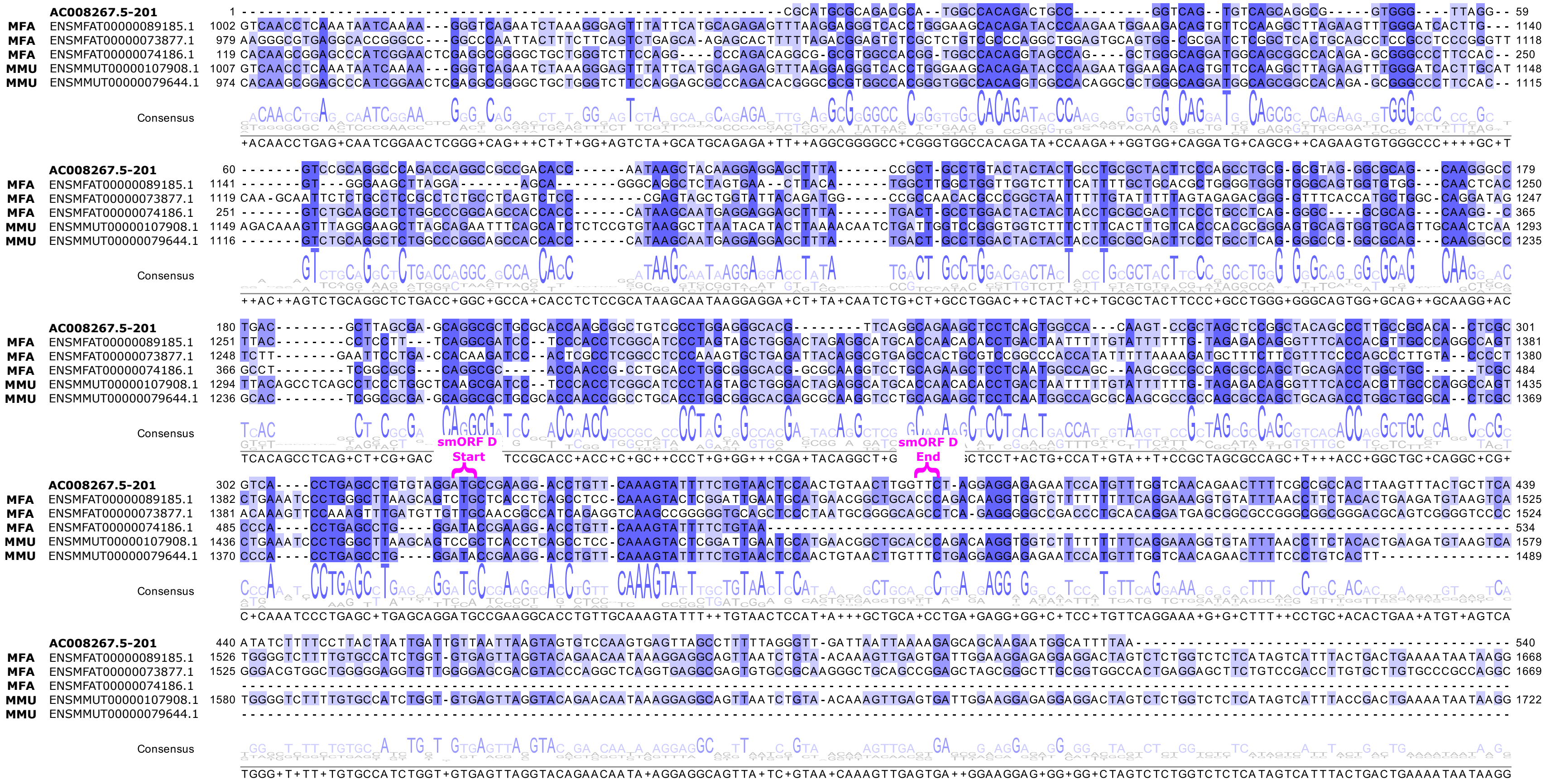


Figure 4.29: **Alignment of results from LINC03011-201 smORF A BLASTn search.** Results returned from queried non-human species using BLASTn v2.9.0+ (Altschul et al., 1990). The alignment is coloured by percentage identity, and the consensus sequence shown below. The figure is cropped to the portion of alignment showing LINC03011-201. The start and end the smORF D are highlighted in pink. Created in Jalview v2.10.2b2 (Waterhouse et al., 2009).

4.4 Discussion

In summary, the 355 translated lncRNA smORFs can be collapsed down into 242 unique smORFs when isoforms are grouped. There is convincing evidence of sequence conservation in non-human species for 65/242 smORFs. These results include both novel peptides, and some of the few conserved lncRNA smORF peptides already described in the literature. For example ATP synthase-associated peptide (ASAP), encoded by LINC00467. This analysis found evidence of sequence conservation across the ASAP smORF, and variation from the smORF isoform discovered in the original paper (Ge et al., 2021).

Three examples of individual translated lncRNA smORFs were described in detail, including smORF A from LIPT2-AS1-201, which was sequentially and syntenously conserved across euarchontoglires. Importantly, although the peptide sequence varied more with increasing evolutionary distance, a helix-turn-helix motif was well conserved. This DNA binding motif is found in proteins involved in regulation of gene expression (Roy and Kundu, 2021), suggesting a potential function for smORF A, perhaps with links to disease given the down regulation of LIPT2-AS1-201 in large-cell medulloblastoma (Birks et al., 2013). Further, inclusion of orthology analysis from Ensembl (Cunningham et al., 2022), and evidence from other Ribo-Seq and mass spectrometry analyses (Calviello, Mukherjee, et al., 2016; van Heesch, Witte, et al., 2019; Brenig et al., 2020; Chong et al., 2020; Mudge et al., 2022), revealed conservation and translation of a longer smORF than originally identified by our analysis. This highlights the importance of combining data from a wide range of sources. Although the basis of this work is a high quality Poly-Ribo-Seq dataset, publically available ’omics datasets and the literature are key to help build a picture of these lncRNA smORFs and their biological importance.

For those ORFs like smORF C, for which conservation was only observed in a few species, or smORF D for which no strong sequence conservation was found across the smORF, further study may still uncover biologically important functions. For example, it may be the act of translation of the smORF that is functional, not the resulting peptide. The lncRNA may be competing with other transcripts for ribosomes, or producing peptides which stall and sequester ribosomes, reducing the translation of other proteins. If so, we would not expect to see strong conservation across the whole smORF sequence.

We have found evidence of sequence conservation in 17/45 translated lncRNA smORFs from the stringent set, equivalent to 13/37 grouped smORF isoforms (Douka, Birds, et al., 2021). The work in this chapter both supports and builds on this initial study, finding evidence of sequence conservation for 15/37 of these smORFs (category D), and results in 1 or 2 species for 10/37 (category C). This was achieved by expanding our search to include a greater evolutionary range of species, and by introducing tBLASTn alignments at the aa level to produce better visualisations of smORF conservation.

This variation between studies does highlight the downside of the manual part of this work, where the strands of the sequence alignment analysis are pulled together to categorise each smORF. However, this analysis includes stringent filters on the initial BLAST (Altschul et al., 1990) results for all search strategies, using p-value, percentage identity and coverage to effectively remove a high number of low quality results. Adding further automated filters, for example to identify syntenous hits, or to prioritise alignments with conserved start codons, would improve the consistency of this work by removing the element of human judgement, but this would also cause the loss of key data for many lncRNA smORFs. By excluding these data that may not represent conservation, interesting results which provide deeper context would be lost. For example, smORF A and B include subject hits which do not represent conservation but a shared protein domain, giving clues to the potential function of these peptides.

Starting with a larger set of candidate smORFs than in our earlier work, including those for which we only had evidence of translation in a single replicate, was found to be a valid and fruitful method. When the categorisations of the smORFs with and without the stringent set (translated in two or more replicates in a given condition) are compared, found no significant association with stringency and conservation category. Therefore it is no more likely to find a smORF with evidence of sequence conservation (category D) in the stringent than in the non-stringent set.

Further, evidence for a "stringent" lncRNA smORF for which we can truly argue for translation does not need to come from a single dataset. The study of non-canonical ORFs is a fast growing field. Although Poly-Ribo-Seq is particularly suited to lncRNA smORF identification given the requirement of multiple bound ribosomes, Ribo-seq and mass spectrometry results are now available in a range of tissues, cell types, and conditions. This is due in no small part to a drive for open science, and many journal's requirement for data to be shared in a publically available format. These data can be used to further bolster our confidence in a non-stringent smORF's translation, and to provide further context on when and where these smORFs are translated. Community projects are working to bring these data together, and create a reference annotation of non-canonical ORFs identified using Ribo-seq. For example, the work currently supported by Gencode (Mudge et al., 2022), which also includes ORFs in UTRs, and internal out of frame ORFs found in canonical mRNA CDSs.

Genome annotations in general have been a key factor affecting the results in this chapter. Many of these lncRNA smORFs, particularly those with results in 1 or 2 species (category C), could in fact be conserved at the sequence level in more of the queried species than identified in this work. However, this analysis was limited by the contents of the annotations - if a transcript is not included in an annotation, it will not be found. Importantly for this work, there is bias in the kind of lncRNAs included in annotations, in particular those which are more straightforward to identify. For example, the Ensembl primate annotation pipeline (Ensembl, 2018; Ensembl, 2021b) focuses on lincRNAs, using a lack of overlap with protein coding genes and lack of Pfam domain to identify them. However, 2-exon lincRNA models are excluded by this pipeline.

Another issue particular to lncRNAs is the time and tissue specificity of their expression, and their low abundance (Ponting, Oliver, and Reik, 2009). Transcriptome annotations depend heavily on the depth of the sequencing used to build them, so depending on the tissues and developmental time points used, lncRNAs can be missed. Further, genome assemblies are often built using the another, well annotated model species annotation as a scaffold, introducing bias. This is a particular concern for lncRNA smORFs with results in 1 or 2 species (category C), as their sequence conservation in one or two closely related species could not be a reflection of reality, but an artefact from this scaffolding. The Ensembl annotation, for example, used the human assembly as a reference (Ensembl, 2018; Ensembl, 2021b).

As the genome annotations were acquired from Ensembl, many of the annotations were built using the same pipeline, in particular the primate genomes (Ensembl, 2018; Ensembl, 2021b). If the pipeline excludes a particular area of the genome or particular type of transcript in one species, it is likely to do so in another closely related species. Future analysis would include multiple genomes for the same species from different sources as a step towards combating this, although all annotation pipelines still face difficulties with lncRNAs.

Basic human error can also introduce issues. When the *Pongo abelii* (orangutan) genome was first published in 2011, ten re-sequenced genomes were also shared by the Orangutan Genome Consortium (Locke et al., 2011). A recent paper revealed that nine of these ten sequenced genomes were accidentally switched, leading to the mis-assignation of sex, and in one case species (Banes et al., 2022). There is always a none zero risk that errors such as this can occur, and perpetuate through the field.

Using the original, raw reads which were used to create the annotations could give greater confidence that a smORF truly isn't present in a species. However, as annotations continue to increase in quality and species coverage, this is likely to be required less and less. Consider Ensembl (Cunningham et al., 2022), which at launch in 2000 only featured a human annotation, with a promise of mouse and worm to come soon. At time of writing there are 314 annotated species on Ensembl, with a rapid release pipeline working to make new annotations available more quickly. Further, as Gencode looks to incorporate non-canonical ORFs in their human annotation, more researchers will be aware of the need to consider annotating beyond the canon of protein coding genes. This will lag behind in non-human species however, due to the work involved and differences in non-canonical ORFs between species. For example, the difference in effective population sizes between *Drosophila* and human contributes to differences in selective pressure upon lncRNA sequences (Haerty and Ponting, 2013).

4.4.1 Conclusions

This work has produced a pool of potentially biologically important translated lncRNA peptides, with evidence of sequence conservation within and in some cases beyond primates. Given the assumption that lncRNAs typically lack strong sequence conservation (Mudge et al., 2022), to find this in 27% of of these human neuronal lncRNA smORFs is a surprising and intriguing result.

Further, this does not exclude the smORFs without evidence of conservation from future research efforts. These may be limited by annotation, by specificity of expression, or may even represent particularly evolutionarily young smORFs. These in particular are important to study to elucidate their place in the *de novo* evolution of new protein coding genes.

Chapter 5

Discussion

5.1 General discussion

LncRNAs are a heterogeneous class of ncRNAs, forming ∼23% of the human transcriptome (Derrien et al., 2012). Initially considered to be "junk DNA", we now know that lncRNAs are key regulators of processes such as cell differentiation and development (Flynn and Chang, 2014; Statello et al., 2021), and their dysregulation has been implicated in a wide range of diseases (Bao et al., 2019). In particular, lncRNAs are enriched and in some cases specifically expressed in the human brain (Derrien et al., 2012; Jandura and Krause, 2017), and are misregulated in Alzheimer's disease (Mus, Hof, and Tiedge, 2007; Faghihi, Modarresi, et al., 2008), Parkinson's disease (Carrieri, Cimatti, et al., 2012), and other neurodegenerative disorders (Yang et al., 2021). However, the majority of well characterised lncRNAs thus far are nuclear, and little is known about the ∼54% of lncRNAs which are exported to the cytoplasm (Carlevaro-Fita et al., 2016), and their roles in neurogenesis.

Further, advances in 'omics technologies have expanded our understanding of the ORFeome and translatome to include non-canonical ORFs. Cytoplasmic lncRNA have been found to associate with the translational machinery (Carlevaro-Fita et al., 2016), and translated lncRNA smORFs are have been identified using ribosome profiling in a growing number of tissues, cell lines, and development time points (Mudge et al., 2022; Chothani, Adami, Widjaja, et al., 2022).

This PhD thesis sought to identify and characterise populations of cytoplasmic human neuronal lncRNAs based on their association with the translational machinery, using a model of neuronal differentiation; SH-SY5Y cells. Further, it aimed to gain an understanding of the role of translated lncRNAs in the context of the evolution of novel protein coding sequences, and to determine the extent of their conservation in non-human species.

5.2 Cytoplasmic lncRNAs are actively translated in human neuronal cells

Poly-Ribo-Seq of undifferentiated and RA-treated differentiated SH-SH5Y cells revealed the active translation of 45 lncRNAs, representing ∼2% of the detected cytoplasmic lncRNAs (E-appendix; lncRNA_summary.xlsx). Upon expansion to include a "less stringent" set of translated lncRNA smORFs, 242 grouped smORF isoforms were detected (Figure 4.9). These smORFs are mostly found in antisense and lincRNA transcripts, reflecting the current bias in current lncRNA annotations (Derrien et al., 2012). However, this could also be due to potential cytoplasmic enrichment of these types of lncRNAs. We were not able to separate out the sub-populations of cytosolic and polysome-associated lncRNA transcripts from these data, but based on the literature it is likely that a large proportion of the lncRNAs interacted with the translational machinery (Ingolia, Lareau, and Weissman, 2011; Chew et al., 2013; Aspden et al., 2014).

LncRNAs are lowly expressed compared to mRNA. This means that despite the high quality of the Poly-Ribo-Seq analysed in this thesis, there may only be a small number of ribosome footprints mapped to actively translated lncRNA smORFs, reducing the likelihood of the full smORF being detected as translated. Further, the sequence of a given footprint effects the likelihood of it being sucessfully sequenced. For example, sequencing coverage is significantly decreased for sequences with particularly high (>70%) or low (<40%) GC content (Amr and Funke, 2015). This difficulty is evidenced by the different smORF isoforms identified by Ribotaper between replicates. In some cases, such as between undifferentiated (Control) and differentiated (RA) samples, this may be indicative of translational regulation; use of an alternative start site. However it is more likely that there were insufficient reads to pass Ribotaper's (Calviello, Mukherjee, et al., 2016) requirements for in-frame ribosomal footprints between an ATG and the next upstream, in-frame ATG. For example, analysis of publicly available data revealed that the true start side of smORF A (Section 4.3.4.1, Figure 4.21) is likely upstream of the ATG identified in this translational analysis. To distinguish which isoforms are translated, future work would involve combining the replicates to improve read coverage across the smORFs, as performed in the literature (Ruiz-Orera, Messeguer, et al., 2014; Ruiz-Orera and Albà, 2019). This follow up analysis is currently underway in the Aspden lab using both Ribotaper (Calviello, Mukherjee, et al., 2016) and the newer software ORFquant (Calviello, Hirsekorn, and Ohler, 2020), and has led to the detection of 147 high confidence lncRNA-smORFs, including many which are exact matches or isoforms of those described in this work.

To further improve the detection of lncRNA translation from these data, work is also ongoing to use wider range of Ribo-Seq analysis tools, including RiboTISH (Zhang, He, et al., 2017), Riboflow (Ozadam, Geng, and Cenik, 2020), Riboviz2 (Cope et al., 2022), and PRICE (Erhard et al., 2018). As in the recent work to establish a comprehensive picture of human translation using a wide range of cell and tissue types, which identified 1,652 lncRNA ORFs (Chothani, Adami, Widjaja, et al., 2022), this combination of approaches allows a greater number of lncRNA smORFs to be detected. For example, PRICE aims to address variation in P-site positions within footprints, while RiboTISH accounts for non-canocial start site, factors which are not accounted for by the Ribotaper pipeline.

This analysis also focused on a very small window of human neuronal differentiation, as Poly-Ribo-Seq was only performed on undifferentiated SH-SY5Y cells, and differentiated SH-SY5Y cells following 3 days of treatment with RA. RA treatment does not result in the terminal differentiation of SH-SY5Y cells, meaning lncRNAs expressed and translated at intermediate and later stages of neuronal differentiation were not captured. Although costly, performing Poly-Ribo-Seq at more frequent time points and following treatment with other differentiation agents would build a fuller picture of lncRNA translation in human neurogenesis. It is also possible to culture neurons on porous membranes, allowing the main body of the cell to be analysed seperately from the neurites(**fuscoNeuronalRibosomesExhibit2021**). Performing Ribo-Seq on the cell body and neurites could provide information about the localisation and possible function of the lncRNA peptides, as localised translation is common in neurites given the long distance from the cell body andrelative lack of space.

Although Poly-Ribo-Seq data was used here in order to identify lncRNA smORFs actively translated by polysomes, rather than single scanning or spuriously bound ribosomes, this does exclude particularly small smORFs that are only large enough to be bound by a single ribosome at a time (Heyer and Moore, 2016). Recent analysis used a modified version of Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016) with no minimum length for ORF assignment, and more stringent requirements for number and framing of P-sites, to identify particularly small smORFs of 3 - 15 aa. 221 of these ORFs were detected in publically available Ribo-Seq data from human brain, testis, liver, kidney, and heart, (Sandmann et al., 2023). Future work should look to ensure that smORFs such as these are not excluded from our detection pipelines.

5.3 Translated lncRNAs share features with protein coding and *de novo* coding sequence

In order to identify distinguishing characteristics of translated lncRNA, in Chapter 3 the translated human neuronal lncRNA smORFs were compared to protein coding ORFs, uORFs, and dORFs. Many of the lncRNA smORFs exhibited similar characteristics to protein coding ORFs, such as start codon enrichment, amino acid composition and high levels of triplet periodicity and translational efficiency, in line with the literature (Ruiz-Orera, Messeguer, et al., 2014; Patraquim et al., 2022). They also exhibited features indicative of novel coding sequence, such as a lack of known protein domains (Murphy and McLysaght, 2012), and low RNA expression levels (Palmieri, Kosiol, and Schlötterer, 2014; Zhao et al., 2014). The translation of 77% of the 45 "stringent" smORFs was validated by other publically available Ribo-Seq datasets, and evidence for the production of stable peptides found for 66% of these smORFs in mass spectrometry data (Mudge et al., 2022). Given the challenges of detecting small peptides using mass spectrometry, this is a substantial proportion of the peptides. Future work would look to expand this work to validate members of the wider, non-stringent lncRNA smORF set using public Ribo-Seq and mass spectrometry data, in particular using mass spectrometry from more timepoints across neurogenesis.

These features were however investigated prior to establishing the extent of sequence conservation of the translated lncRNA smORFs. Further understanding of their characteristics can be gained using this context. Are the lncRNA smORFs with convincing evidence of sequence conservation more similar to protein coding ORFs compared to the rest of the translated lncRNA population, or do they exhibit novel features more like that of *de novo* coding sequence? However, this analysis should be performed with the caveat that the inclusion of more sequencing data from other tissues or annotations from other species could provide sufficient evidence to recategorise a gicen lncRNA smORF.

The expression of the translated lncRNA transcripts was significantly enriched in human and macaque brain tissue compared to untranslated lncRNAs. This is a key result, as the lncRNA smORFs were identified in a cell line, and this confirms their upregulation compared to the wider lncRNA population in neuronal tissues. In our earlier work (Douka, Birds, et al., 2021), analysis performed by Andreas Kosteletos found that the 45 "stringent" translated lncRNAs were more likely to have developmentally dynamic expression than untranslated lncRNAs, particularly in the gonads and brain. 68% of the 45 translated lncRNAs were associated with central nervous system cancers, and only 24% had no significant disease associations. Expanding this analysis to include the "non stringent" set of 242 grouped smORF isoforms would help to establish whether the conserved translated smORFs (Category D) have a distinct expression profile compared to the wider set of translated lncRNAs. Further work could also investigate and compare the changes in expression of nearby genes and overlapping genes to the translated lncRNA throughout neurogenesis. A recent study in yeast found that changes in antisense *de novo* transcripts expression were similar to changes in the genes they overlapped in stress conditions, and given the large-scale changes in transcriptional control throughout neurogenesis we may expect to see a similar result (Blevins et al., 2021).

5.4 Translated lncRNAs exhibit sequential and syntenic conservation

In chapter 4 the extent of sequence conservation of the translated lncRNA smORFs in non-human species was investigated. Convincing evidence of sequence conservation was found for 65/242 grouped smORF isoforms, and 67/242 returned potential conservation in 1 or 2 non-human species.

This conservation analysis has expanded upon our earlier work (Douka, Birds, et al., 2021), including the larger "non stringent" set of translated lncRNA smORFs, more stringent filters, a greater range of species, and alignment of the tBLASTn hits as amino acid sequences. Given the small number of non-human species included in our earlier work (6), it is unsurprising that of the 17 smORFs with evidence of sequence conservation, 13 only returned results in 1 or 2 species. Despite the size of this initial analysis, the majority of the results are supported by the larger, in depth analysis in this thesis. All 4 lncRNA smORFs which returned originally convincing hits in 3 or more species were classified as Category D (evidence of sequence conservation), and 12/13 which returned results in 1 or 2 species (Category C) remained in this category, or returned convincing evidence in a greater number of species.

On average, the conserved, translated lncRNA smORFs were longer than other lncRNA smORFs, with a median peptide length of 71 aa. Being short did not preclude the detection of sequence conservation, however; shorter peptides including 3 with a length of 25 aa were conserved. A higher proportion of conserved smORFs were also found in publicly available Ribo-Seq and mass spectrometry data, suggesting that these smORFs may be more widely translated than the other lncRNA smORFs.

The sequence conservation and biological context of three individual smORFs has been described in detail, including smORF A in LIPT-AS1 (Section 4.3.4.1), which is sequentially and syntenously conserved across euarchontoglires. To do so, information from a variety of sources was integrated, including the protein domains contained in the smORF peptide, the expression of LIPT2-AS1-201, analyses of Ribo-Seq and mass spectrometry data in the literature, and genome annotations. Follow up work in the Aspden Lab has aimed to understand the potential function of smORF A, by knockdown and overexpression of LIPT2-AS1-201 to investigate potential effects on neuronal differentiation, and by FLAG-tagging to investigate the localisation of the peptide. This has revealed specific subcellular localisation and functions of the LIPT2-AS1-201 peptide in neuronal differentiation.

For 110/242 grouped smORF isoforms, no BLAST results (Category A) or no convincing evidence (Category B) was returned. However, this lack of conservation does not preclude function. Recent work in *Saccharomyces cerevisiae* detected thousands of non canonical ORFs, only 14 of which were under purifying selection (Wacholder et al., 2023). Despite this, a proportion of the unconserved non canonical ORFs were found to provide fitness benefits, contributing to regulatory processes including DNA repair, stress response, and post-transcriptional regulation. 12 *de novo* lncRNA peptides, including two human specific peptides, had significant fitness effects upon disruption of the ORFs in human cell lines using CRISPR-Cas9 (Vakirlis, Vance, et al., 2022).

To understand the translation initiation of lncRNA smORFs, future work should include analysis of the Kozak context of the translated lncRNA smORF both in the individual smORFs, and across the BLAST hits from all conservation categories. For those lncRNAs with a TIS which is particularly strong (close to the consensus sequence), but with little sequence conservation across the smORF, this may indicate that it is the act of translation that is functional, rather than the resulting peptide. As with other lncRNA functions which sometimes only require very small modules of sequence conservation, e.g. molecular decoys, we may find evidence of conservation in other species by looking for syntenous transcripts, with similar expression profiles, and similarly strong TIS and ribosomal footprinting. Further, analysis of the Kozak sequences would be a useful tool to establish the likely "true" start codon of smORF isoform.

Further work could also seek to compare the sequences of the translated lncRNA transcripts, to identify shared motifs and possible modes of regulation. For example, a study of 74 human/hominoid-specific *de novo* genes, including a gene implicated in neuronal maturation, identified distincitive elements controlling their export from the nucleus (An et al., 2023).

5.5 Conclusions

The work presented in this thesis highlights the potential involvement of translated cytoplasmic lncRNA in human neuronal differentiation. Poly-Ribo-Seq data was analysed from SH-SH5Y cells, revealing 242 actively translated smORFs with triplet periodicity and translational efficiency comparable to protein coding ORFs. The sequences of 27% of these smORFs were conserved in non-human species, suggesting that the resulting peptides may contribute to neuronal development.

Appendix A

Chapter 2

Ensembl files type used, v104 (Cunningham et al., 2022):

- **species_name.cdna.all.fa**: The super-set of all transcripts resulting from Ensembl gene predictions.
- **species_name.ncrna.fa**: The super-set of all transcripts resulting from Ensembl short and long non-coding gene predictions. Combined with species_name.cdna.all.fa to create a transcript sequence file.
- **species_name.pep.all.fa**: The super-set of all translations resulting from Ensembl gene predictions.

Gencode files used, v30 and v38 (Frankish et al., 2021)

- **gencode.version.transcripts.fa**
- **gencode.version.pc_translations.fa**

Table A.1: **Mean and standard deviation of RNA and Ribo-Seq fragment lengths.** Mean, minimum, and maximum fragment lengths were estimated from tapestation readouts. Standard deviation was calculated as $SD_{Estimate} = \frac{max-min}{4}$.

Sample	Replicate	Mean	Min	Max	Standard Deviation
Control Total	1	188	160	240	20
Control Polysome	1	185	160	230	17.5
Control Footprint	1	168	140	190	12.5
RA Total	1	187	160	230	17.5
RA Polysome	1	188	160	230	17.5
RA Footprint	1	162	140	190	12.5
Control Total	2	198	175	240	16.25
Control Polysome	2	194	175	240	16.25
Control Footprint	2	161	135	190	13.75
RA Total	2	189	160	230	17.25
RA Polysome	2	182	160	225	16.25
RA Footprint	2	157	130	180	12.5
Control Total	3	184	160	220	15
Control Polysome	3	178	150	215	16.25
Control Footprint	3	159	137.5	180	10.6
RA Total	3	180	150	220	17.5
RA Polysome	3	178	155	220	16.25
RA Footprint	3	152	130	175	11.25

Table A.2: **Ribo-Seq read lengths and offsets.** High quality read lengths selected for each Ribo-Seq replicate based on metaplots, used in translational analysis using Ribotaper v1.3 (Calviello, Mukherjee, et al., 2016). The P-site offset is the estimated offset of the P-site from the start of each Ribo-Seq read.

Replicate	**Read Length**	**P-site Offset**
Control, Replicate 1	31	13
Control, Replicate 1	33	5
Control, Replicate 2	31	10
Control, Replicate 2	33	5
Control, Replicate 3	31	7
Control, Replicate 3	33	5
RA, Replicate 1	33	5
RA, Replicate 2	31	13
RA, Replicate 2	33	5
RA, Replicate 3	31	10
RA, Replicate 3	33	5

Table A.3: **Number of reads aligned to the human genome in each sample.** Supplementary to Figure 2.3. Samples are named by type of sequencing, and the conditions the cells were in. These were total cytoplasmic polyadenylated RNA-seq, Polysome-associated polyadenylated RNA-Seq, and Ribo-Seq, in Control and RA conditions. Rep refers to the replicate each sample was from. The starred samples are those highlighted in Figure 2.3. Adapter sequences were trimmed from raw reads using Cutadapt v2.10 (Martin, 2011). Low quality reads were removed using the Fastq quality filter in FASTX-Toolket v0.0.14 (Hannon, 2009). rRNA and tRNA contaminants were aligned and removed using Bowtie2 v2.3.4.3 (Langmead and Salzberg, 2012). Remaining reads were mapped to the human genome using STAR (v2.7.5c) (Dobin et al., 2013). The overall percentage of raw reads mapped to the genome is also given.

Sample	Rep	No. of raw reads	Reads after adapter trimming	Reads after quality filter	Reads after tRNA removal	Reads after rRNA removal	Reads aligned to genome	Percent aligned
Control Total	1	74580218	73987868	59925151	49196300	49185248	48522608	65%
Control Polysome	1	75815765	75394406	61610625	49981898	49977859	49147072	65%
Control FP	1	210729343	201306731	174680552	154701468	119932694	114046520	54%
RA Total	1	47091265	47011115	38052453	34540528	34533610	34242186	73%
RA Polysome	1	64468224	64317646	51443298	41870107	41868254	41396996	64%
RA FP	1	180333341	167105627	143917744	127508671	106887772	98311891	55%
Control Total	2	84553605	83708049	67496433	62296427	62284252	59617098	71%
***Control Polysome**	2	251494676	246274020	220113977	115912944	115870271	98844317	39%
Control FP	2	172613271	168133150	148443740	104648901	81336826	75497312	44%
RA Total	2	77512549	76500077	63512964	55983849	55967658	53401852	69%
***RA Polysome**	2	63950891	61494601	51557736	27775961	27759910	21436972	34%
RA FP	2	206766558	180372772	157443114	104910323	79532804	62060178	30%
Control Total	3	54946328	54337292	45155000	38201612	38195104	37524685	68%
Control Polysome	3	74521025	74230186	62637374	33944162	33938799	32907862	44%
Control FP	3	123263935	118550426	104255049	74876309	64561118	49588008	40%
RA Total	3	59056664	58680906	49252943	43978544	43973849	43362788	73%
RA Polysome	3	69816758	68974561	57634216	46683472	46680700	45808989	66%
RA FP	3	189414183	174948478	152968776	101740468	89880360	79892176	42%

Table A.4: **Number of reads aligned to the human genome.** Supplementary to Figure 2.4. Samples are named by type of sequencing, and the conditions the cells were in. These were total cytoplasmic polyadenylated RNA-seq, Polysome-associated polyadenylated RNA-Seq, and Ribo-Seq, in Control and RA conditions. Rep refers to the replicate each sample was from. Uniquely mapped and multi mapped reads were combined to give the total number of mapped reads.

Sample	Rep	Input	Uniquely mapped	Multi mappers	Mapped to too many loci	Unmapped; too short	Unmapped; other	Total mapped	Percent mapped
Control Total	1	49,185,248	41,078,769	7,443,839	194,789	440,530	27,321	48,522,608	99%
Control Polysome	1	49,977,859	43,182,199	5,964,873	152,027	634,680	44,080	49,147,072	98%
Control FP	1	119,932,694	89,662,264	24,384,256	1,577,906	4,221,959	86,309	114,046,520	95%
RA Total	1	34,533,610	29,865,627	4,376,559	81,834	192,733	16,857	34,242,186	99%
RA Polysome	1	41,868,254	37,228,791	4,168,205	92,030	346,335	32,893	41,396,996	99%
RA FP	1	106,887,772	82,419,111	15,892,780	1,579,965	6,459,205	536,711	98,311,891	92%
Control Total	2	62,284,252	53,126,051	6,491,047	133,285	2,500,744	33,125	59,617,098	96%
Control Polysome	2	115,870,271	88,258,244	10,586,073	336,626	16,619,940	69,388	98,844,317	85%
Control FP	2	81,336,826	58,868,737	16,628,575	1,302,064	4,498,797	38,653	75,497,312	93%
RA Total	2	55,967,658	46,288,901	7,112,951	178,482	2,342,605	44,719	53,401,852	95%
RA Polysome	2	27,759,910	18,119,568	3,317,404	89,589	6,213,304	20,045	21,436,972	77%
RA FP	2	79,532,804	49,106,533	12,953,645	1,574,700	15,822,873	75,053	62,060,178	78%
Control Total	3	38,195,104	31,836,976	5,687,709	147,140	498,734	24,545	37,524,685	98%
Control Polysome	3	33,938,799	28,778,073	4,129,789	121,007	869,959	39,971	32,907,862	97%
Control FP	3	64,561,118	39,218,990	10,369,018	699,043	14,247,669	26,398	49,588,008	77%
RA Total	3	43,973,849	38,082,036	5,280,752	113,647	471,056	26,358	43,362,788	99%
RA Polysome	3	46,680,700	41,602,278	4,206,711	125,101	703,277	43,333	45,808,989	98%
RA FP	3	89,880,360	66,612,048	13,280,128	845,857	9,095,295	47,032	79,892,176	89%

Appendix B

Chapter 3

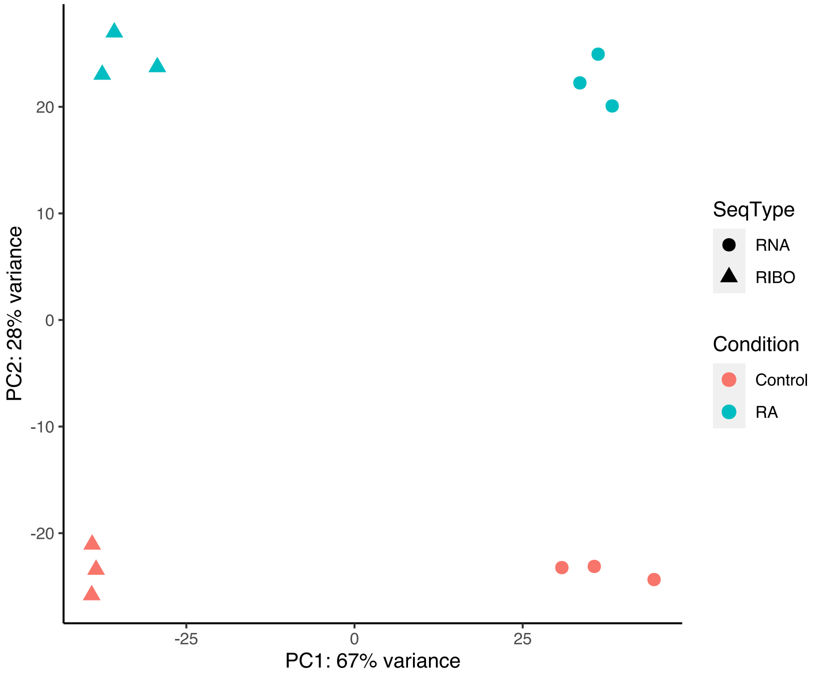


Figure B.1: **Principle component analysis of RNA-seq and Ribo-Seq datasets, from control and RA conditions.** PC1 is visualised on the x-axis, and PC2 is visualised on the y-axis, and together these components account for 95% of the variance in the data. PC1 separates the sample types, indicating that the largest sources of variance is between sequencing type. The second largest source of variance, PC2, separates the conditions.

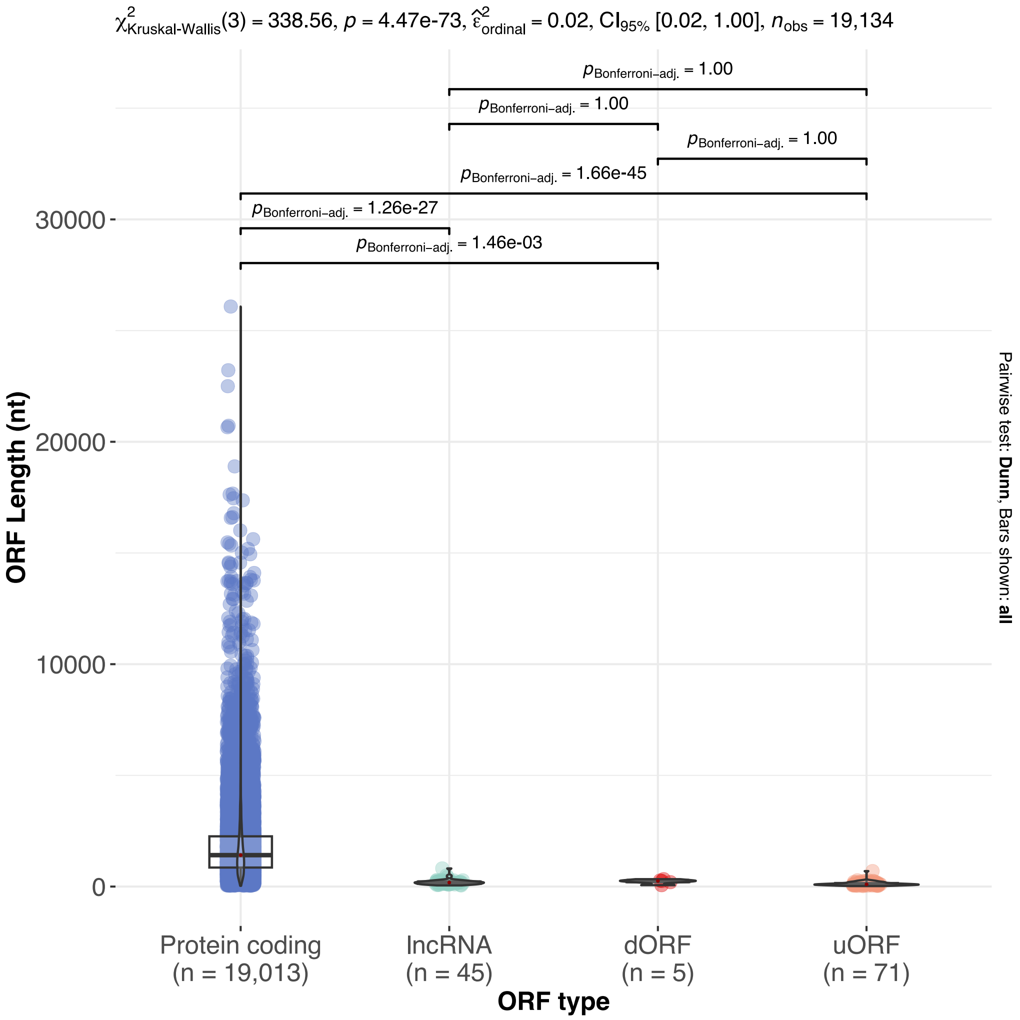


Figure B.2: **Translated ORF lengths (nt).** The x axis shows categories of translated ORFs, n denotes the number of ORFs in each category. The y axis shows the length in nucleotides. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test.

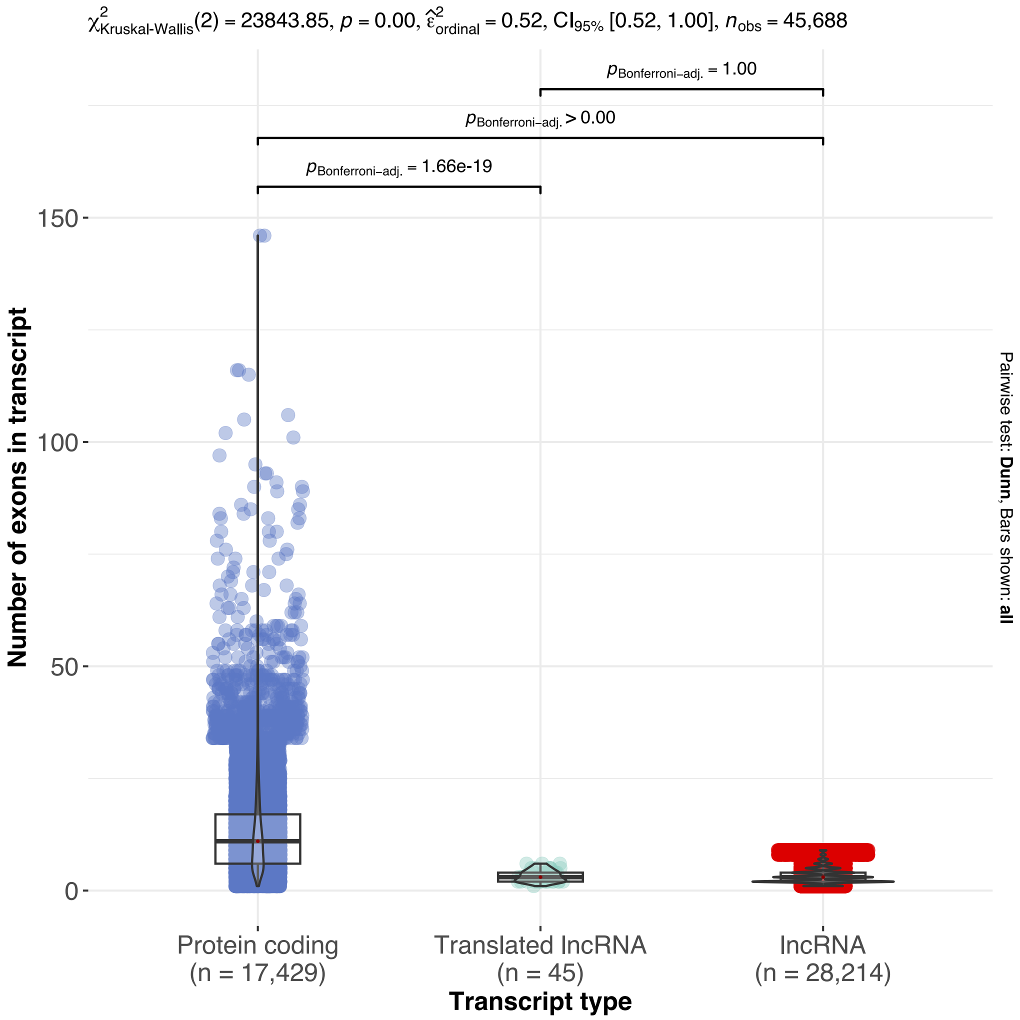


Figure B.3: **Number of exons in transcripts.** The x axis shows categories of transcripts, and n denotes the number of transcripts. The y axis shows the number of exons in the transcripts. A Kruskal-Wallis test was used to compare the groups, and pairwise comparisons were made using Dunn's test.

Table B.1: **Expression of translated lncRNA genes in the macaque brain.** Tissue denotes the area of the brain analysed, and number of samples denotes the number of individuals samples were taken from. Kruskal-Wallis indicates if a Kruskal-Wallis test found that the lncRNA gene set had a significant effect on RPKM in each sample, and p-values summarises the range of p-values from each sample. Multiple comparison describes whether a multiple comparison test found that the levels of translated lncRNAs (RPKM) were significantly different from randomly selected lncRNAs, and in what direction. RPKM values from the PsychENCODE consortium (Zhu, Sousa, et al., 2018; Akbarian et al., 2015).

Tissue	**No. of samples**	**Kruskal-Walls**	**P-values**	**Multiple Comparison**
Hippocampus (HIP)	22	Significant	0.00011 or smaller	True, higher (549/550)
Medial prefrontal cortex (MFC)	25	Significant	e-5 or smaller	True, higher (608/625)
Dorsolateral prefrontal cortex (DFC)	22	Significant	e-7 or smaller	True, higher
Orbital prefrontal cortex (OFC)	25	Significant	e-5 or smaller	True, higher (591/600)
Ventrolateral prefrontal cortex (VFC)	24	Significant	0.00042 or smaller	True, higher (591/600)
Amygdala (AMY)	21	Significant	e-8 or smaller	True, higher (521/525)
Striatum (STR)	22	Significant	e-8 or smaller	True, higher (545/550)
Primary motor cortex (M1C)	24	Significant	e-5 or smaller	True, higher (598/600)
Primary somatosensory cortex (S1C)	24	Significant	e-7 or smaller	True, higher (590/600)
Inferior posterior parietal cortex (IPC)	24	Significant	e-5 or smaller	True, higher (599/600)
Primary auditory cortex (A1C)	23	Significant	e-5 or smaller	True, higher (571/575)
Superior temporal cortex (STC)	21	Significant	0.00025 or smaller	True, higher (520/525)
Inferior temporal cortex (ITC)	22	Significant	e-5 or smaller	True, higher (546/550)
Primary visual cortex (V1C)	22	Significant	e-5 or smaller	True, higher (548/550)
Mediodorsal nucleus of thalamus (MD)	24	Significant	e-6 or smaller	True, higher
Cerebellar cortex (CBC)	20	Significant	e-6 or smaller	True, higher (499/500)

Appendix C

Chapter 4

Table C.1: **Number of annotated genes and transcripts in species used in conservation analysis.** Species sampled are provided with common and latin names. ’Number of genes’ refers to the number of genes defined by the annotation, split into protein coding, long non-coding, small non-coding, misc. non-coding, and pseudogenes, and a total. ’Number of transcripts’ refers to the number of transcripts defined by the annotation, split into protein coding, long non-coding, and a total.

Species	**Number of genes**						**Number of transcripts**		
	Protein coding	**Long non-coding**	**Small non-coding**	**Misc non-coding**	**Pseudo genes**	**Total**	**Protein coding**	**Long non-coding**	**Total**
Human v30 (*Homo sapiens*)	19,986	16,193	7,576	-	14,706	58,870	83,688	30,369	208,621
Human v38 (*Homo sapiens*)	19,955	17,944	7,567	-	14,773	60,649	86,757	48,752	237,012
Bonobo (*Pan paniscus*)	21,210	1,497	4,998	2,004	549	30,258	-	-	53,360
Chimpanzee (*Pan troglodytes*)	23,534	1,786	5,640	2,284	485	33,729	-	-	61,457
Gorilla (*Gorilla gorilla gorilla*)	21,794	490	5,207	2,071	522	30,084	-	-	53,705
Orangutan (*Pongo abelii*)	20,424	-	5,796	1,200	1,023	28,443	-	-	29,447
Gibbon (*Nomascus leucogenys*)	20,794	7	4,479	1,979	567	27,826	-	-	47,559
Olive baboon (*Papio anubis*)	21,647	714	4,686	2,006	423	29,476	-	-	53,493
Sooty mangabey (*Cercocebus atys*)	20,926	545	4,525	1,923	540	28,459	-	-	53,925
Pig-tailed macaque (*Macaca nemestrina*)	21,060	640	4,584	1,997	584	28,865	-	-	54,453
Crab-eating macaque (*Macaca fascicularis*)	22,504	6,628	4,687	3,274	1,457	38,550	-	-	66,619

Table C.1 continued from previous page

Species	Number of genes						Number of transcripts		
	Protein coding	Long non-coding	Small non-coding	Misc non-coding	Pseudo genes	Total	Protein coding	Long non-coding	Total
Macaque (*Macaca mulatta*)	21,761	4,773	4,712	3,419	767	35,432	-	-	64,228
Bolivian squirrel monkey (*Saimiri boliviensis boliviensis*)	19,380	384	4,146	3,153	439	27,502	-	-	49,071
Marmoset (*Callithrix jacchus*)	22,615	6,709	6,384	4,260	1,190	41,158	-	-	66,095
Mouse lemur (*Microcebus murinus*)	18,895	665	3,983	2,785	470	26,798	-	-	46,396
Rabbit (*Oryctolagus cuniculus*)	20,612	5,736	2,520	63	656	29,587	-	-	51,853
Mouse (*Mus musculus*)	22,213	11,310	5,526	562	13,649	53,260	-	-	149,478
Mouse (*Mus musculus*)	20,720	5,930	4,627	549	6,401	38,227	-	-	102,148
Mouse (*Mus musculus*)	20,698	5,952	4,631	550	6,397	38,228	-	-	102,254
Mouse (*Mus musculus*)	20,659	5,941	4,623	552	6,359	38,134	-	-	102,168
Mouse (*Mus musculus*)	20,709	5,943	4,622	553	6,371	38,198	-	-	102,138
Mouse (*Mus musculus*)	20,580	5,926	4,588	550	6,258	37,902	-	-	101,710
Mouse (*Mus musculus*)	20,901	5,970	4,725	558	6,826	38,980	-	-	103,002

Table C.1 continued from previous page

Species	**Number of genes**						**Number of transcripts**		
	Protein coding	**Long non-coding**	**Small non-coding**	**Misc non-coding**	**Pseudo genes**	**Total**	**Protein coding**	**Long non-coding**	**Total**
Mouse (*Mus musculus castaneus*)	20,319	5,853	4,448	548	5,912	37,080	-	-	102,099
Mouse (*Mus musculus*)	20,589	5,911	4,577	549	6,269	37,895	-	-	101,632
Mouse (*Mus musculus*)	20,658	5,938	4,588	551	6,346	38,081	-	-	101,908
Mouse (*Mus musculus*)	20,580	5,916	4,592	549	6,263	37,900	-	-	101,520
Mouse (*Mus musculus*)	20,717	5,942	4,602	549	6,376	38,186	-	-	101,892
Mouse (*Mus musculus*)	20,608	5,910	4,576	552	6,367	38,013	-	-	101,562
Mouse (*Mus musculus*)	20,858	5,974	4,680	558	6,777	38,847	-	-	103,134
Mouse (*Mus musculus musculus*)	20,179	5,816	4,414	546	5,739	36,694	-	-	101,394
Mouse (*Mus musculus domesticus*)	20,389	5,883	4,508	544	5,962	37,286	-	-	100,622
Rat (*Rattus norvegicus*)	22,250	3,288	5,122	524	1,668	32,852	-	-	41,078
Dog (*Canis lupus familiaris*)	20,567	6,485	3,437	22	610	31,121	-	-	55,335

Table C.1 continued from previous page

Species	Number of genes						Number of transcripts		
	Protein coding	Long non-coding	Small non-coding	Misc non-coding	Pseudo genes	Total	Protein coding	Long non-coding	Total
Opossum (*Monodelphis domestica*)	21,384	10,869	1,832	34	866	34,985	-	-	58,883
Chicken (Red jungle fowl) (*Gallus gallus*)	17,077	12,449	1,275	5	56	30,862	-	-	74,296

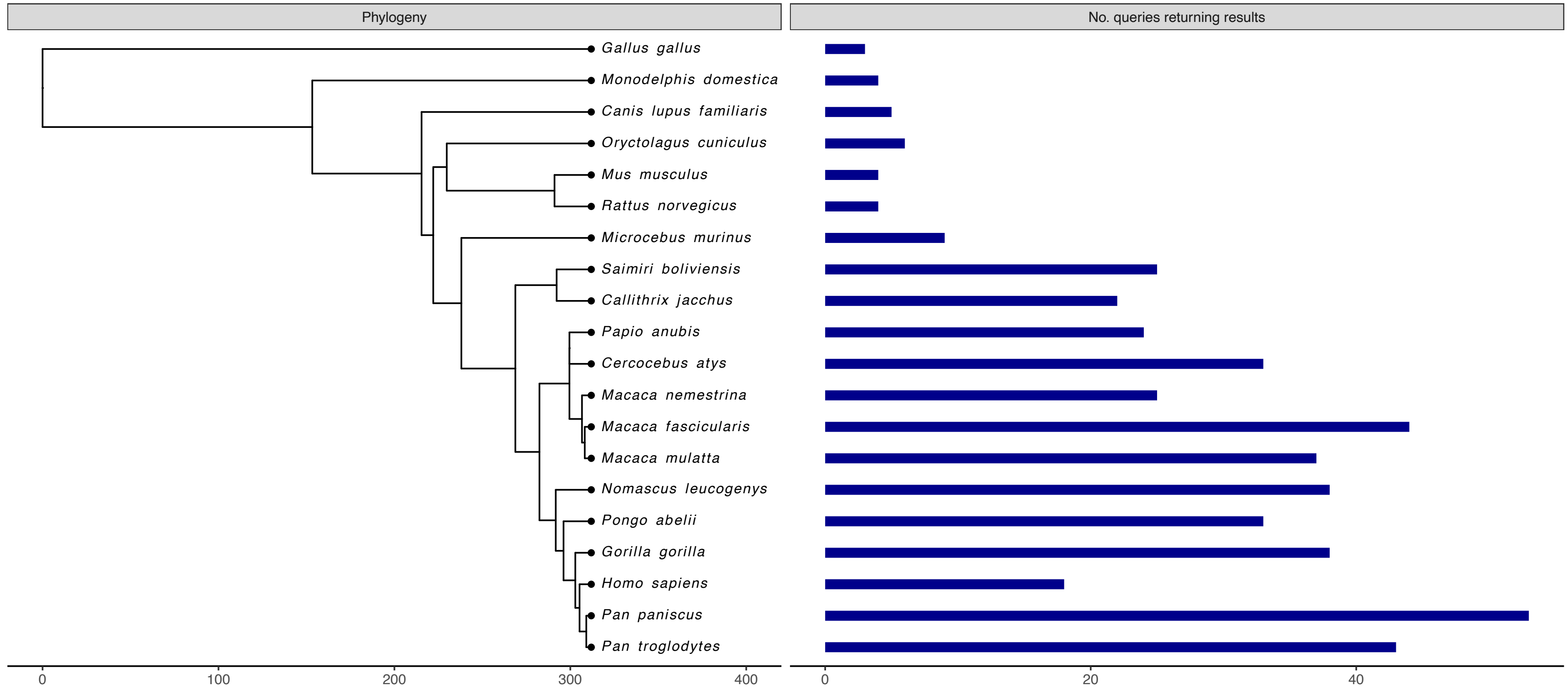


Figure C.1: **BLASTp hits were returned for lncRNA smORF peptides in all queried species.** Initial results from BLASTp v2.9.0+ (Altschul et al., 1990), as described in Figure 4.2.B. The phylogeny shows the species included in the conservation analysis, with estimated divergence times in millions of years ago (Kumar et al., 2017). The number of lncRNA smORF peptides which returned one or more potentially homologous sequences is shown for each queried species.

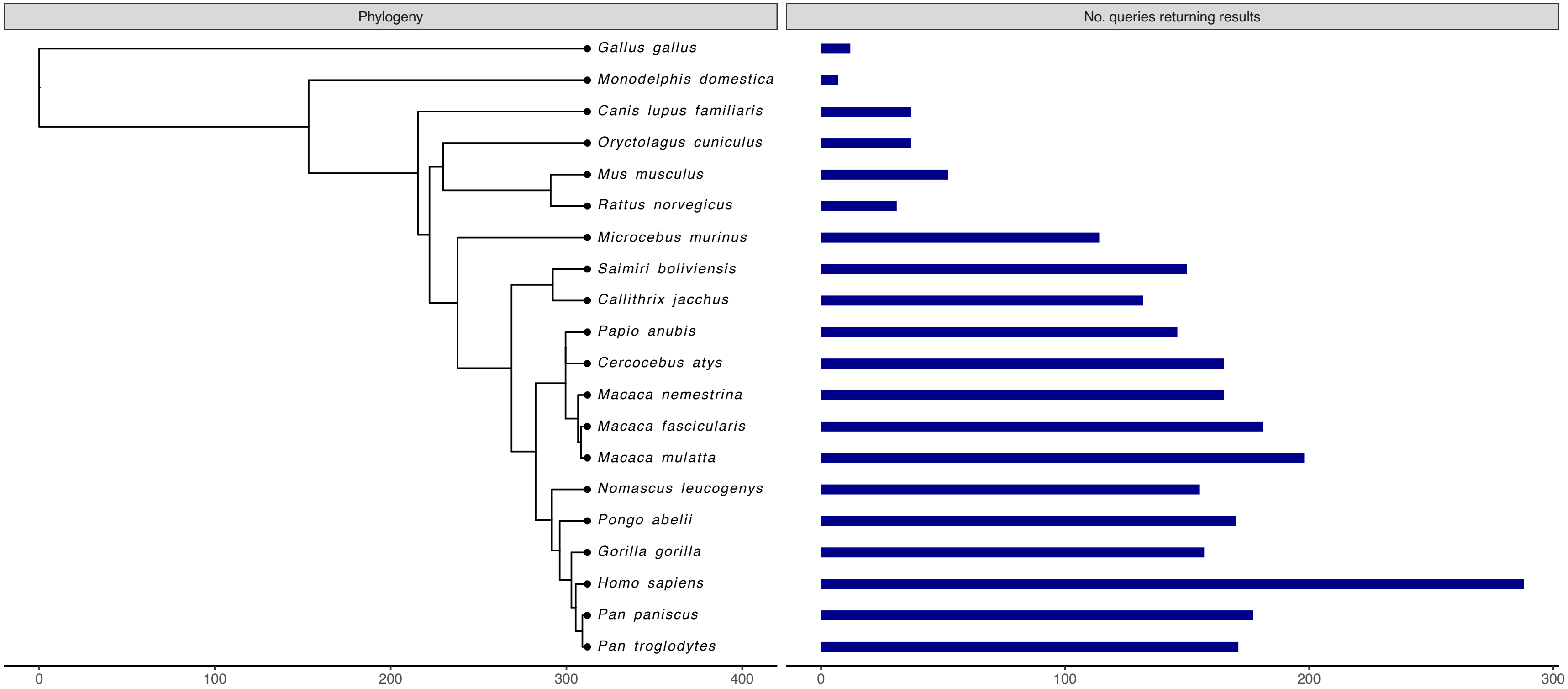


Figure C.2: **BLASTn hits were returned for translated lncRNA transcripts in all queried species.** Initial results from lncRNA transcript BLASTn v2.9.0+ (Altschul et al., 1990), as described in Figure 4.3B. The phylogeny shows the species included in the conservation analysis, with estimated divergence times in millions of years ago (Kumar et al., 2017). The number of lncRNA transcripts which returned one or more potentially homologous sequences is shown for each queried species.

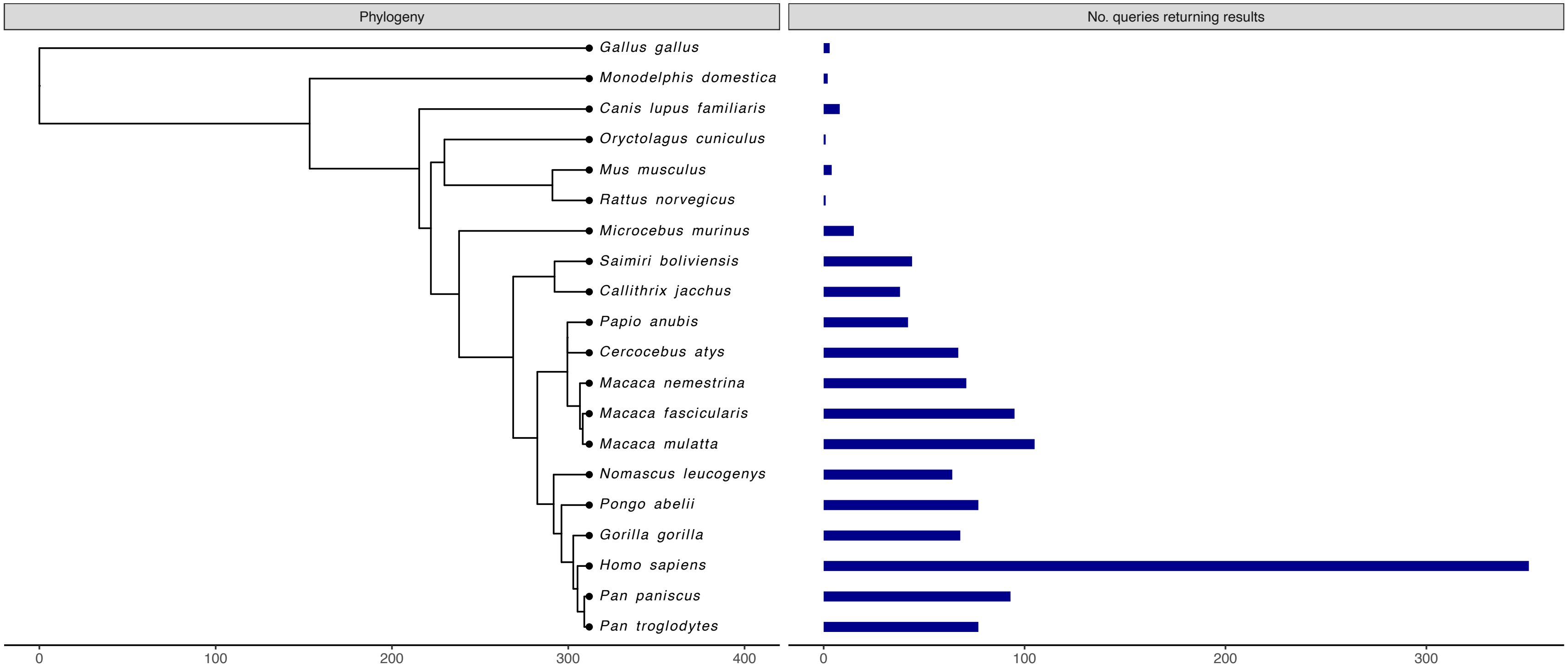


Figure C.3: **BLASTn hits were returned for translated lncRNA smORFs in all queried species.** Initial results from lncRNA smORF BLASTn v2.9.0+ (Altschul et al., 1990), as described in Figure 4.3D. The phylogeny shows the species included in the conservation analysis, with estimated divergence times in millions of years ago (Kumar et al., 2017). The number of lncRNA smORFs which returned one or more potentially homologous sequences is shown for each queried species.

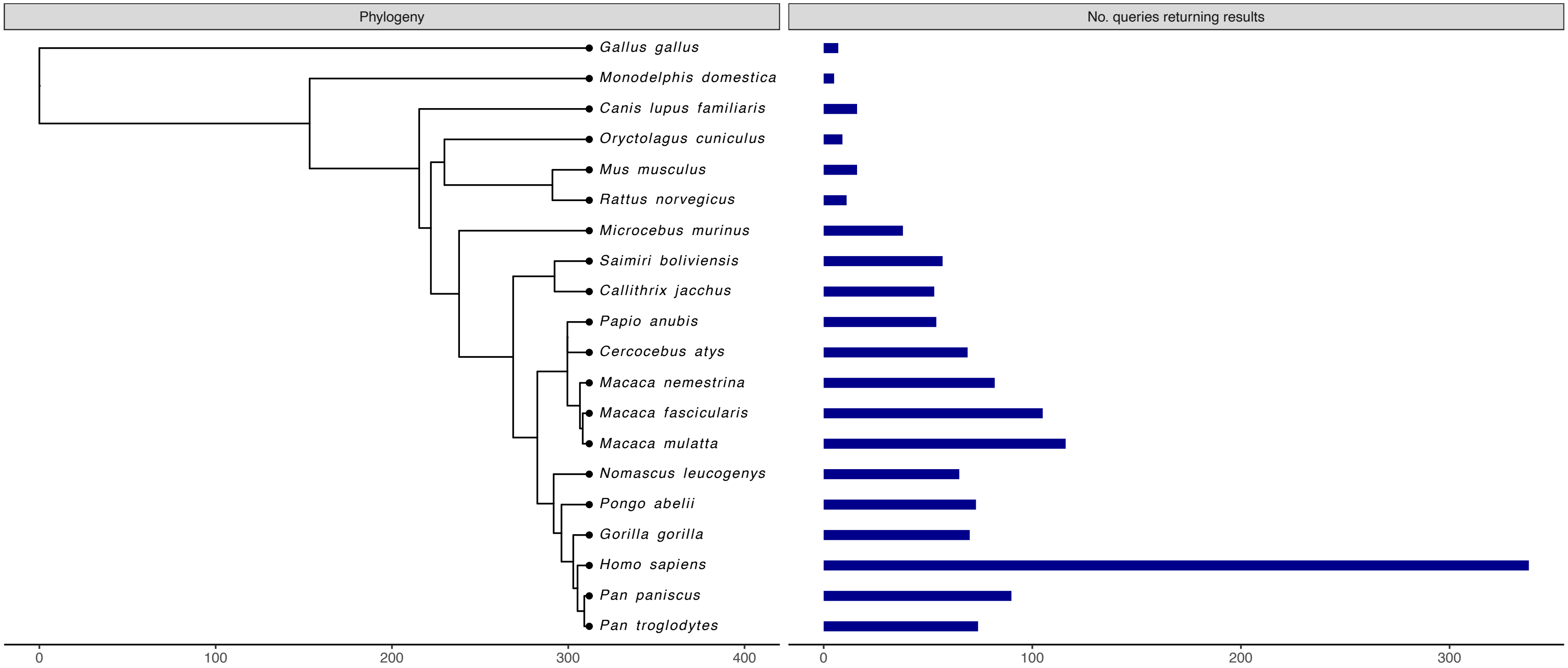


Figure C.4: **tBLASTn hits were returned for lncRNA smORF peptides in all queried species.** Initial results from tBLASTn v2.9.0+ (Altschul et al., 1990), as described in Figure 4.4B. The phylogeny shows the species included in the conservation analysis, with estimated divergence times in millions of years ago (Kumar et al., 2017). The number of lncRNA smORF peptides which returned one or more potentially homologous sequences is shown for each queried species.

Bibliography

Akbarian, Schahram, Chunyu Liu, James A. Knowles, Flora M. Vaccarino, Peggy J. Farnham, Gregory E. Crawford, Andrew E. Jaffe, Dalila Pinto, Stella Dracheva, Daniel H. Geschwind, Jonathan Mill, Angus C. Nairn, Alexej Abyzov, Sirisha Pochareddy, Shyam Prabhakar, Sherman Weissman, Patrick F. Sullivan, Matthew W. State, Zhiping Weng, Mette A. Peters, Kevin P. White, Mark B. Gerstein, Anahita Amiri, Chris Armoskus, Allison E. Ashley-Koch, Taejeong Bae, Andrea Beckel-Mitchener, Benjamin P. Berman, Gerhard A. Coetzee, Gianfilippo Coppola, Nancy Francoeur, Menachem Fromer, Robert Gao, Kay Grennan, Jennifer Herstein, David H. Kavanagh, Nikolay A. Ivanov, Yan Jiang, Robert R. Kitchen, Alexey Kozlenkov, Marija Kundakovic, Mingfeng Li, Zhen Li, Shuang Liu, Lara M. Mangravite, Eugenio Mattei, Eirene Markenscoff-Papadimitriou, Fábio C. P. Navarro, Nicole North, Larsson Omberg, David Panchision, Neelroop Parikshak, Jeremie Poschmann, Amanda J. Price, Michael Purcaro, Timothy E. Reddy, Panos Roussos, Shannon Schreiner, Soraya Scuderi, Robert Sebra, Mikihito Shibata, Annie W. Shieh, Mario Skarica, Wenjie Sun, Vivek Swarup, Amber Thomas, Junko Tsuji, Harm van Bakel, Daifeng Wang, Yongjun Wang, Kai Wang, Donna M. Werling, A. Jeremy Willsey, Heather Witt, Hyejung Won, Chloe C. Y. Wong, Gregory A. Wray, Emily Y. Wu, Xuming Xu, Lijing Yao, Geetha Senthil, Thomas Lehner, Pamela Sklar, and Nenad Sestan (Dec. 2015). “The PsychENCODE Project”. In: *Nature Neuroscience* 18.12, pp. 1707–1712. ISSN: 1546-1726. DOI: 10.1038/nn.4156.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter (2002). “DNA-Binding Motifs in Gene Regulatory Proteins”. In: *Molecular Biology of the Cell. 4th edition*.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (Oct. 1990). “Basic Local Alignment Search Tool”. In: *Journal of Molecular Biology* 215.3, pp. 403–410. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2.

Amr, Sami S. and Birgit Funke (Jan. 2015). “Chapter 16 - Targeted Hybrid Capture for Inherited Disease Panels”. In: *Clinical Genomics*. Ed. by Shashikant Kulkarni and John Pfeifer. Boston: Academic Press, pp. 251–269. ISBN: 978-0-12-404748-8. DOI: 10.1016/B978-0-12-404748-8.00016-2.

An, Ni A., Jie Zhang, Fan Mo, Xuke Luan, Lu Tian, Qing Sunny Shen, Xiangshang Li, Chunqiong Li, Fanqi Zhou, Boya Zhang, Mingjun Ji, Jianhuan Qi, Wei-Zhen Zhou, Wanqiu Ding, Jia-Yu Chen, Jia Yu, Li Zhang, Shaokun Shu, Baoyang Hu, and Chuan-Yun Li (Jan. 2023). “De Novo Genes with an lncRNA Origin Encode Unique Human Brain Developmental Functionality”. In: *Nature Ecology & Evolution*, pp. 1–15. ISSN: 2397-334X. DOI: 10.1038/s41559-022-01925-6.

Anders, Simon and Wolfgang Huber (Oct. 2010). “Differential Expression Analysis for Sequence Count Data”. In: *Genome Biology* 11.10, R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

Anderson, Douglas M., Kelly M. Anderson, Chi-Lun Chang, Catherine A. Makarewich, Benjamin R. Nelson, John R. McAnally, Prasad Kasaragod, John M. Shelton, Jen Liou, Rhonda Bassel-Duby, and Eric N. Olson (Feb. 2015). “A Micropeptide Encoded by a Putative Long

Noncoding RNA Regulates Muscle Performance". In: *Cell* 160.4, pp. 595–606. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.01.009.

Andreev, Dmitry E., Gary Loughran, Alla D. Fedorova, Maria S. Mikhaylova, Ivan N. Shatsky, and Pavel V. Baranov (May 2022). "Non-AUG Translation Initiation in Mammals". In: *Genome Biology* 23.1, p. 111. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02674-2.

Andrews, S (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data.*

Andrews, Shea J. and Joseph A. Rothnagel (Mar. 2014). "Emerging Evidence for Functional Peptides Encoded by Short Open Reading Frames". In: *Nature Reviews Genetics* 15.3, pp. 193–204. ISSN: 1471-0056. DOI: 10.1038/nrg3520.

Arun, Gayatri, Disha Aggarwal, and David L. Spector (June 2020). "MALAT1 Long Non-Coding RNA: Functional Implications". In: *Non-Coding RNA* 6.2, p. 22. ISSN: 2311-553X. DOI: 10.3390/ncrna6020022.

Aspden, Julie L, Ying Chen Eyre-Walker, Rose J Phillips, Unum Amin, Muhammad Ali S Mumtaz, Michele Brocard, and Juan-Pablo Couso (Aug. 2014). "Extensive Translation of Small Open Reading Frames Revealed by Poly-Ribo-Seq". In: *eLife* 3, e03528. ISSN: 2050-084X. DOI: 10.7554/eLife.03528.

Athey, John, Aikaterini Alexaki, Ekaterina Osipova, Alexandre Rostovtsev, Luis V. Santana-Quintero, Upendra Katneni, Vahan Simonyan, and Chava Kimchi-Sarfaty (Sept. 2017). "A New and Updated Resource for Codon Usage Tables". In: *BMC Bioinformatics* 18.1, p. 391. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1793-7.

Atkins, John F., Alan J. Herr, Christian Massire, Michael O'Connor, Ivaylo Ivanov, and Raymond F. Gesteland (2000). "Poking a Hole in the Sanctity of the Triplet Code: Inferences for Framing". In: *The Ribosome*. John Wiley & Sons, Ltd. Chap. 30, pp. 367–383. ISBN: 978-1-68367-252-4. DOI: 10.1128/9781555818142.ch30.

Attwood, T. K., M. E. Beck, A. J. Bleasby, and D. J. Parry-Smith (Sept. 1994). "PRINTS–a Database of Protein Motif Fingerprints". In: *Nucleic Acids Research* 22.17, pp. 3590–3596. ISSN: 0305-1048.

Banes, Graham L., Emily D. Fountain, Alyssa Karklus, Robert S. Fulton, Lucinda Antonacci-Fulton, and Joanne O. Nelson (Aug. 2022). "Nine out of Ten Samples Were Mistakenly Switched by The Orang-utan Genome Consortium". In: *Scientific Data* 9.1, p. 485. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01602-0.

Bao, Zhenyu, Zhen Yang, Zhou Huang, Yiran Zhou, Qinghua Cui, and Dong Dong (Jan. 2019). "LncRNADisease 2.0: An Updated Database of Long Non-Coding RNA-associated Diseases". In: *Nucleic Acids Research* 47.D1, pp. D1034–D1037. ISSN: 1362-4962. DOI: 10.1093/nar/gky905.

Barbosa, Cristina, Isabel Peixeiro, and Luísa Romão (Aug. 2013). "Gene Expression Regulation by Upstream Open Reading Frames and Human Disease". In: *PLoS Genetics* 9.8. Ed. by Elizabeth M. C. Fisher, e1003529. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003529.

Barczak, Wojciech, Simon M. Carr, Geng Liu, Shonagh Munro, Annalisa Nicastri, Lian Ni Lee, Claire Hutchings, Nicola Ternette, Paul Klenerman, Alexander Kanapin, Anastasia Samsonova, and Nicholas B. La Thangue (Feb. 2023). "Long Non-Coding RNA-derived Peptides Are Immunogenic and Drive a Potent Anti-Tumour Response". In: *Nature Communications* 14.1, p. 1078. ISSN: 2041-1723. DOI: 10.1038/s41467-023-36826-0.

Bazzini, Ariel A, Timothy G Johnstone, Romain Christiano, Sebastian D Mackowiak, Benedikt Obermayer, Elizabeth S Fleming, Charles E Vejnar, Miler T Lee, Nikolaus Rajewsky, Tobias C Walther, and Antonio J Giraldez (May 2014). "Identification of Small ORFs in Vertebrates Using Ribosome Footprinting and Evolutionary Conservation". In: *The EMBO Journal* 33.9, pp. 981–993. ISSN: 0261-4189. DOI: 10.1002/embj.201488411.

Begun, David J., Heather A. Lindfors, Melissa E. Thompson, and Alisha K. Holloway (Mar. 2006). "Recently Evolved Genes Identified From Drosophila Yakuba and D. Erecta Accessory Gland Expressed Sequence Tags". In: *Genetics* 172.3, pp. 1675–1681. ISSN: 0016-6731. DOI: 10.1534/genetics.105.050336.

Bengtsson, Henrik, Constantin Ahlmann-Eltze, Hector Corrada Bravo, Robert Gentleman, Jan Gleixner, Peter Hickey, Ola Hossjer, Harris Jaffee, Dongcan Jiang, Peter Langfelder, Brian Montgomery, Angelina Panagopoulou, Hugh Parsonage, and Jakob Peder Pettersen (2022). *matrixStats: Functions That Apply to Rows and Columns of Matrices (and to Vectors)*.

Benitez-Cantos, Maria S., Martina M. Yordanova, Patrick B.F. O'Connor, Alexander V. Zhdanov, Sergey I. Kovalchuk, Dmitri B. Papkovsky, Dmitry E. Andreev, and Pavel V. Baranov (July 2020). "Translation Initiation Downstream from Annotated Start Codons in Human mRNAs Coevolves with the Kozak Context". In: *Genome Research* 30.7, pp. 974–984. ISSN: 1088-9051. DOI: 10.1101/gr.257352.119.

Benoit Bouvrette, Louis Philip, Neal A.L. Cody, Julie Bergalet, Fabio Alexis Lefebvre, Cédric Diot, Xiaofeng Wang, Mathieu Blanchette, and Eric Lécuyer (Jan. 2018). "CeFra-seq Reveals Broad Asymmetric mRNA and Noncoding RNA Distribution Profiles in Drosophila and Human Cells". In: *RNA* 24.1, pp. 98–113. ISSN: 1355-8382. DOI: 10.1261/rna.063172.117.

Bhan, Arunoday and Subhrangsu S Mandal (Aug. 2015). "LncRNA HOTAIR: A Master Regulator of Chromatin Dynamics and Cancer." In: *Biochimica et biophysica acta* 1856.1, pp. 151–64. ISSN: 0006-3002. DOI: 10.1016/j.bbcan.2015.07.001.

Biedler, J. L., L. Helson, and B. A. Spengler (Nov. 1973). "Morphology and Growth, Tumorigenicity, and Cytogenetics of Human Neuroblastoma Cells in Continuous Culture". In: *Cancer Research* 33.11, pp. 2643–2652. ISSN: 0008-5472.

Biedler, June L., Suzanne Roffler-Tarlov, Melitta Schachner, and Lewis S. Freedman (Nov. 1978). "Multiple Neurotransmitter Synthesis by Human Neuroblastoma Cell Lines and Clones". In: *Cancer Research* 38.11_Part_1, pp. 3751–3757. ISSN: 0008-5472.

Birks, Diane K, Andrew M Donson, Purvi R Patel, Alexandra Sufit, Elizabeth M Algar, Christopher Dunham, B K Kleinschmidt-DeMasters, Michael H Handler, Rajeev Vibhakar, and Nicholas K Foreman (July 2013). "Pediatric Rhabdoid Tumors of Kidney and Brain Show Many Differences in Gene Expression but Share Dysregulation of Cell Cycle and Epigenetic Effector Genes". In: *Pediatric blood & cancer* 60.7, pp. 1095–1102. ISSN: 1545-5017. DOI: 10.1002/pbc.24481.

Blackburn, Elizabeth H. and Joseph G. Gall (Mar. 1978). "A Tandemly Repeated Sequence at the Termini of the Extrachromosomal Ribosomal RNA Genes in Tetrahymena". In: *Journal of Molecular Biology* 120.1, pp. 33–53. ISSN: 0022-2836. DOI: 10.1016/0022-2836(78)90294-2.

Blackburne, Benjamin P. and Simon Whelan (Feb. 2012). "Measuring the Distance between Multiple Sequence Alignments". In: *Bioinformatics* 28.4, pp. 495–502. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr701.

Blaxter, Mark (Dec. 2010). "Revealing the Dark Matter of the Genome". In: *Science* 330.6012, pp. 1758–1759. DOI: 10.1126/science.1200700.

Blevins, William R., Jorge Ruiz-Orera, Xavier Messeguer, Bernat Blasco-Moreno, José Luis Villanueva-Cañas, Lorena Espinar, Juana Díez, Lucas B. Carey, and M. Mar Albà (Jan. 2021). "Uncovering de Novo Gene Birth in Yeast Using Deep Transcriptomics". In: *Nature Communications* 12, p. 604. ISSN: 2041-1723. DOI: 10.1038/s41467-021-20911-3.

Blythe, Amanda J., Archa H. Fox, and Charles S. Bond (Jan. 2016). "The Ins and Outs of lncRNA Structure: How, Why and What Comes Next?" In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1859.1, pp. 46–58. ISSN: 1874-9399. DOI: 10.1016/J.BBAGRM.2015.08.009.

Borensztein, Maud, Laurène Syx, Katia Ancelin, Patricia Diabangouaya, Christel Picard, Tao Liu, Jun-Bin Liang, Ivaylo Vassilev, Rafael Galupa, Nicolas Servant, Emmanuel Barillot, Azim Surani, Chong-Jian Chen, and Edith Heard (Mar. 2017). "Xist-Dependent Imprinted X Inactivation and the Early Developmental Consequences of Its Failure". In: *Nature structural & molecular biology* 24.3, pp. 226–233. ISSN: 1545-9993. DOI: 10.1038/nsmb.3365.

Brenig, Katrin, Leonie Grube, Markus Schwarzländer, Karl Köhrer, Kai Stühler, and Gereon Poschmann (May 2020). "The Proteomic Landscape of Cysteine Oxidation That Underpins

Retinoic Acid-Induced Neuronal Differentiation". In: *Journal of Proteome Research* 19.5, pp. 1923–1940. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.9b00752.

Brockdorff, Neil (Dec. 2018). "Local Tandem Repeat Expansion in Xist RNA as a Model for the Functionalisation of ncRNA". In: *Non-Coding RNA* 4.4, p. 28. ISSN: 2311-553X. DOI: 10.3390/ncrna4040028.

Brockdorff, Neil, Alan Ashworth, Graham F. Kay, Penny Cooper, Sandy Smith, Veronica M. McCabe, Dominic P. Norris, Graeme D. Penny, Dipika Patel, and Sohaila Rastan (May 1991). "Conservation of Position and Exclusive Expression of Mouse Xist from the Inactive X Chromosome". In: *Nature* 351.6324, pp. 329–331. ISSN: 1476-4687. DOI: 10.1038/351329a0.

Brown, Carolyn J., Andrea Ballabio, James L. Rupert, Ronald G. Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F. Willard (Jan. 1991). "A Gene from the Region of the Human X Inactivation Centre Is Expressed Exclusively from the Inactive X Chromosome". In: *Nature* 349.6304, pp. 38–44. ISSN: 0028-0836. DOI: 10.1038/349038a0.

Brown, Carolyn J., Brian D. Hendrich, Jim L. Rupert, Ronald G. Lafrenière, Yigong Xing, Jeanne Lawrence, and Huntington F. Willard (Oct. 1992). "The Human XIST Gene: Analysis of a 17 Kb Inactive X-specific RNA That Contains Conserved Repeats and Is Highly Localized within the Nucleus". In: *Cell* 71.3, pp. 527–542. ISSN: 0092-8674. DOI: 10.1016/0092-8674(92)90520-M.

Brown, Jessica A., Max L. Valenstein, Therese A. Yario, Kazimierz T. Tycowski, and Joan A. Steitz (Nov. 2012). "Formation of Triple-Helical Structures by the 3′-End Sequences of MALAT1 and MENβ Noncoding RNAs". In: *Proceedings of the National Academy of Sciences* 109.47, pp. 19202–19207. DOI: 10.1073/pnas.1217338109.

Bryzghalov, Oleksii, Michał Wojciech Szcześniak, and Izabela Makałowska (Jan. 2020). "SyntDB: Defining Orthologues of Human Long Noncoding RNAs across Primates". In: *Nucleic Acids Research* 48.D1, pp. D238–D245. ISSN: 0305-1048. DOI: 10.1093/nar/gkz941.

Cai, Li-Jun, Li Tu, Xiao-Mo Huang, Jia Huang, Nan Qiu, Guang-Hong Xie, Jian-Xiong Liao, Wei Du, Ying-Yue Zhang, and Jin-Yong Tian (Sept. 2020). "LncRNA MALAT1 Facilitates Inflammasome Activation via Epigenetic Suppression of Nrf2 in Parkinson's Disease". In: *Molecular Brain* 13.1, p. 130. ISSN: 1756-6606. DOI: 10.1186/s13041-020-00656-8.

Calviello, Lorenzo, Antje Hirsekorn, and Uwe Ohler (Aug. 2020). "Quantification of Translation Uncovers the Functions of the Alternative Transcriptome". In: *Nature Structural & Molecular Biology* 27.8, pp. 717–725. ISSN: 1545-9985. DOI: 10.1038/s41594-020-0450-4.

Calviello, Lorenzo, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler (Feb. 2016). "Detecting Actively Translated Open Reading Frames in Ribosome Profiling Data". In: *Nature Methods* 13.2, pp. 165–170. ISSN: 1548-7105. DOI: 10.1038/nmeth.3688.

Carlevaro-Fita, Joana, Anisa Rahim, Roderic Guigó, Leah A Vardy, and Rory Johnson (2016). "Cytoplasmic Long Noncoding RNAs Are Frequently Bound to and Degraded at Ribosomes in Human Cells". In: DOI: 10.1261/rna.053561.115.

Carrieri, Claudia, Laura Cimatti, Marta Biagioli, Anne Beugnet, Silvia Zucchelli, Stefania Fedele, Elisa Pesce, Isidre Ferrer, Licio Collavin, Claudio Santoro, Alistair R. R. Forrest, Piero Carninci, Stefano Biffo, Elia Stupka, and Stefano Gustincich (Nov. 2012). "Long Non-Coding Antisense RNA Controls Uchl1 Translation through an Embedded SINEB2 Repeat". In: *Nature* 491.7424, pp. 454–457. ISSN: 1476-4687. DOI: 10.1038/nature11508.

Carrieri, Claudia, Alistair R. R. Forrest, Claudio Santoro, Francesca Persichetti, Piero Carninci, Silvia Zucchelli, and Stefano Gustincich (2015). "Expression Analysis of the Long Non-Coding RNA Antisense to Uchl1 (AS Uchl1) during Dopaminergic Cells' Differentiation in Vitro and in Neurochemical Models of Parkinson's Disease". In: *Frontiers in Cellular Neuroscience* 9, p. 114. ISSN: 1662-5102. DOI: 10.3389/fncel.2015.00114.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A Calderwood, Muhammed A Yildirim, Nicolas Simonis, Benoit Charloteaux, César A Hidalgo, Justin Barbette, Balaji Santhanam, Gloria A Brar, Jonathan S Weissman, Aviv Regev, Nicolas Thierry-Mieg, Michael E Cusick, and Marc Vidal (July 2012). “Proto-Genes and de Novo Gene Birth.” In: *Nature* 487.7407, pp. 370–4. ISSN: 1476-4687. DOI: 10.1038/nature11184.

Chai, Haina, Chao Sun, Jun Liu, Haihui Sheng, Renyan Zhao, and Zhiqiang Feng (May 2019). “The Relationship Between ZEB1-AS1 Expression and the Prognosis of Patients With Advanced Gastric Cancer Receiving Chemotherapy”. In: *Technology in Cancer Research & Treatment* 18, p. 1533033819849069. ISSN: 1533-0346. DOI: 10.1177/1533033819849069.

Charif, Delphine and Jean R. Lobry (2007). “SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis”. In: *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Ed. by Ugo Bastolla, Markus Porto, H. Eduardo Roman, and Michele Vendruscolo. Biological and Medical Physics, Biomedical Engineering. Berlin, Heidelberg: Springer, pp. 207–232. ISBN: 978-3-540-35306-5. DOI: 10.1007/978-3-540-35306-5_10.

Chen, Hanbo (2022). *VennDiagram: Generate High-Resolution Venn and Euler Plots.*

Chen, Jia-Yu, Qing Sunny Shen, Wei-Zhen Zhou, Jiguang Peng, Bin Z. He, Yumei Li, Chu-Jun Liu, Xuke Luan, Wanqiu Ding, Shuxian Li, Chunyan Chen, Bertrand Chin-Ming Tan, Yong E. Zhang, Aibin He, and Chuan-Yun Li (July 2015). “Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates”. In: *PLoS Genetics* 11.7, e1005391. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1005391.

Chen, Sidi, Benjamin H. Krinsky, and Manyuan Long (Sept. 2013). “New Genes as Drivers of Phenotypic Evolution”. In: *Nature Reviews Genetics* 14.9, pp. 645–660. ISSN: 1471-0064. DOI: 10.1038/nrg3521.

Chew, Guo-Liang, Andrea Pauli, John L. Rinn, Aviv Regev, Alexander F. Schier, and Eivind Valen (July 2013). “Ribosome Profiling Reveals Resemblance between Long Non-Coding RNAs and 5′ Leaders of Coding RNAs”. In: *Development* 140.13, pp. 2828–2834. ISSN: 0950-1991. DOI: 10.1242/dev.098343.

Choe, Sung E., Michael Boutros, Alan M. Michelson, George M. Church, and Marc S. Halfon (2005). “Preferred Analysis Methods for Affymetrix GeneChips Revealed by a Wholly Defined Control Dataset”. In: *Genome Biology* 6.2, R16. ISSN: 1474-760X. DOI: 10.1186/gb-2005-6-2-r16.

Choi, Seo-Won, Hyun-Woo Kim, and Jin-Wu Nam (June 2019). “The Small Peptide World in Long Noncoding RNAs”. In: *Briefings in Bioinformatics* 20.5, pp. 1853–1864. ISSN: 1467-5463. DOI: 10.1093/bib/bby055.

Chong, Chloe, Markus Müller, HuiSong Pak, Dermot Harnett, Florian Huber, Delphine Grun, Marion Leleu, Aymeric Auger, Marion Arnaud, Brian J. Stevenson, Justine Michaux, Ilija Bilic, Antje Hirsekorn, Lorenzo Calviello, Laia Simó-Riudalbas, Evarist Planet, Jan Lubiński, Marta Bryśkiewicz, Maciej Wiznerowicz, Ioannis Xenarios, Lin Zhang, Didier Trono, Alexandre Harari, Uwe Ohler, George Coukos, and Michal Bassani-Sternberg (Mar. 2020). “Integrated Proteogenomic Deep Sequencing and Analytics Accurately Identify Non-Canonical Peptides in Tumor Immunopeptidomes”. In: *Nature Communications* 11, p. 1293. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14968-9.

Chothani, Sonia, Eleonora Adami, John F. Ouyang, Sivakumar Viswanathan, Norbert Hubner, Stuart A. Cook, Sebastian Schafer, and Owen J. L. Rackham (Dec. 2019). “deltaTE: Detection of Translationally Regulated Genes by Integrative Analysis of Ribo-seq and RNA-seq Data”. In: *Current Protocols in Molecular Biology* 129.1, e108. ISSN: 1934-3639. DOI: 10.1002/cpmb.108.

Chothani, Sonia P., Eleonora Adami, Anissa A. Widjaja, Sarah R. Langley, Sivakumar Viswanathan, Chee Jian Pua, Nevin Tham Zhihao, Nathan Harmston, Giuseppe D’Agostino, Nicola Whiffin, Wang Mao, John F. Ouyang, Wei Wen Lim, Shiqi Lim, Cheryl Q. E. Lee, Alexandra Grubman, Joseph Chen, J. P. Kovalik, Karl Tryggvason, Jose M. Polo, Lena Ho, Stuart A. Cook, Owen J. L. Rackham, and Sebastian Schafer (Aug. 2022). “A High-Resolution Map of Human RNA Translation”. In: *Molecular Cell* 82.15, 2885–2899.e8. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2022.06.023.

Churbanov, Alexander, Igor B. Rogozin, Vladimir N. Babenko, Hesham Ali, and Eugene V. Koonin (2005). “Evolutionary Conservation Suggests a Regulatory Function of AUG Triplets in 5′-UTRs of Eukaryotic Genes”. In: *Nucleic Acids Research* 33.17, pp. 5512–5520. ISSN: 0305-1048. DOI: 10.1093/nar/gki847.

Clark, Michael B., Rebecca L. Johnston, Mario Inostroza-Ponta, Archa H. Fox, Ellen Fortini, Pablo Moscato, Marcel E. Dinger, and John S. Mattick (Jan. 2012). “Genome-Wide Analysis of Long Noncoding RNA Stability”. In: *Genome Research* 22.5, pp. 885–898. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.131037.111.

Conrad, Nicholas K and Joan A Steitz (May 2005). “A Kaposi’s Sarcoma Virus RNA Element That Increases the Nuclear Abundance of Intronless Transcripts”. In: *The EMBO Journal* 24.10, pp. 1831–1841. ISSN: 0261-4189. DOI: 10.1038/sj.emboj.7600662.

Cope, Alexander L, Felicity Anderson, John Favate, Michael Jackson, Amanda Mok, Anna Kurowska, Junchen Liu, Emma MacKenzie, Vikram Shivakumar, Peter Tilton, Sophie M Winterbourne, Siyin Xue, Kostas Kavoussanakis, Liana F Lareau, Premal Shah, and Edward W J Wallace (Apr. 2022). “Riboviz 2: A Flexible and Robust Ribosome Profiling Data Analysis and Visualization Workflow”. In: *Bioinformatics* 38.8, pp. 2358–2360. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac093.

Cordaux, Richard and Mark A. Batzer (Oct. 2009). “The Impact of Retrotransposons on Human Genome Evolution”. In: *Nature Reviews Genetics* 10.10, pp. 691–703. ISSN: 1471-0064. DOI: 10.1038/nrg2640.

Cotton, James A and Roderic D M Page (Feb. 2005). “Rates and Patterns of Gene Duplication and Loss in the Human Genome.” In: *Proceedings. Biological sciences* 272.1560, pp. 277–83. ISSN: 0962-8452. DOI: 10.1098/rspb.2004.2969.

Couso, Juan-Pablo and Pedro Patraquim (July 2017). “Classification and Function of Small Open Reading Frames”. In: *Nature Reviews Molecular Cell Biology* 18.9, pp. 575–589. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.58.

Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, Andrew Berry, Jyothish Bhai, Alexandra Bignell, Konstantinos Billis, Sanjay Boddu, Lucy Brooks, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Jose Gonzalez Martinez, Cristina Guijarro-Clarke, Arthur Gymer, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, José Carlos Marugán, Shamika Mohanan, Aleena Mushtaq, Marc Naven, Denye N Ogeh, Anne Parker, Andrew Parton, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, José G Pérez-Silva, William Stark, Emily Steed, Kyösti Sutinen, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Michal Szpak, Anja Thormann, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Natalie Willhoft, Andrea Winterbottom, Elizabeth Wass, Marc Chakiachvili, Bethany Flint, Adam Frankish, Stefano Giorgetti, Leanne Haggerty, Sarah E Hunt, Garth R IIsley, Jane E Loveland, Fergal J Martin, Benjamin Moore, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Sarah Dyer, Peter W Harrison, Kevin L Howe, Andrew D Yates, Daniel R Zerbino, and Paul Flicek (Jan. 2022). “Ensembl 2022”. In: *Nucleic Acids Research* 50.D1, pp. D988–D995. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1049.

Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, Julien Lagarde, Lavanya Veeravalli, Xiaoan Ruan, Yijun Ruan, Timo Lassmann, Piero Carninci, James B Brown, Leonard Lipovich, Jose M Gonzalez, Mark Thomas, Carrie A Davis, Ramin Shiekhattar, Thomas R Gingeras, Tim J Hubbard, Cedric Notredame, Jennifer Harrow, and Roderic Guigó (Sept. 2012). “The GENCODE v7 Catalog of Human Long Noncoding

RNAs: Analysis of Their Gene Structure, Evolution, and Expression.” In: *Genome research* 22.9, pp. 1775–89. ISSN: 1549-5469. DOI: 10.1101/gr.132159.111.

Devaux, Yvan, Jennifer Zangrando, Blanche Schroen, Esther E. Creemers, Thierry Pedrazzini, Ching-Pin Chang, Gerald W. Dorn II, Thomas Thum, and Stephane Heymans (July 2015). “Long Noncoding RNAs in Cardiac Development and Ageing”. In: *Nature Reviews Cardiology* 12.7, pp. 415–425. ISSN: 1759-5002. DOI: 10.1038/nrcardio.2015.55.

Diederichs, Sven (Apr. 2014). “The Four Dimensions of Noncoding RNA Conservation”. In: *Trends in Genetics* 30.4, pp. 121–123. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2014.01.004.

Dinman, Jonathan D. (2012). “Mechanisms and Implications of Programmed Translational Frameshifting”. In: *WIREs RNA* 3.5, pp. 661–673. ISSN: 1757-7012. DOI: 10.1002/wrna.1126.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras (Jan. 2013). “STAR: Ultrafast Universal RNA-seq Aligner”. In: *Bioinformatics* 29.1, pp. 15–21. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bts635.

Douka, Katerina, Michaela Agapiou, Isabel Birds, and Julie L. Aspden (2022). “Optimization of Ribosome Footprinting Conditions for Ribo-Seq in Human and Drosophila Melanogaster Tissue Culture Cells”. In: *Frontiers in Molecular Biosciences* 8. ISSN: 2296-889X.

Douka, Katerina, Isabel Birds, Dapeng Wang, Andreas Kosteletos, Sophie Clayton, Abigail Byford, Elton J. R. Vasconcelos, Mary J. O’Connell, Jim Deuchars, Adrian Whitehouse, and Julie L. Aspden (June 2021). “Cytoplasmic Long Non-Coding RNAs Are Differentially Regulated and Translated during Human Neuronal Differentiation”. In: *RNA*, rna.078782.121. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.078782.121.

Duret, Laurent, Corinne Chureau, Sylvie Samain, Jean Weissenbach, and Philip Avner (June 2006). “The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene”. In: *Science* 312.5780, pp. 1653–1655. DOI: 10.1126/science.1126316.

Edgar, Robert C. (2004). “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput”. In: *Nucleic Acids Research* 32.5, pp. 1792–1797. ISSN: 0305-1048. DOI: 10.1093/nar/gkh340.

Elisaphenko, Eugeny A., Nikolay N. Kolesnikov, Alexander I. Shevchenko, Igor B. Rogozin, Tatyana B. Nesterova, Neil Brockdorff, and Suren M. Zakian (June 2008). “A Dual Origin of the Xist Gene from a Protein-Coding Gene and a Set of Transposable Elements”. In: *PLOS ONE* 3.6, e2521. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0002521.

ENCODE Project Consortium (Sept. 2012). “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *Nature* 489.7414, pp. 57–74. ISSN: 1476-4687. DOI: 10.1038/nature11247.

Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann (2013). “Comet: An Open-Source MS/MS Sequence Database Search Tool”. In: *PROTEOMICS* 13.1, pp. 22–24. ISSN: 1615-9861. DOI: 10.1002/pmic.201200439.

Ensembl (Apr. 2018). *Ensembl Gene Annotation (e!92) Primate Clade*. https://mart.ensembl.org/info/genome/genebuild/Primate_clade_gene_annotation.pdf.

– (Sept. 2019). *Ensembl Gene Annotation (e!98) Rhesus Monkey (Macaca Mulatta)*. https://mart.ensembl.org/info/genome/genebuild/2019_09_macaca_mulatta_gene_ annotation.pdf.

– (Feb. 2021a). *Ensembl Gene Annotation (e!103) Crab-Eating Macaque - Macaca Fascicularis 6.0*. https://mart.ensembl.org/info/genome/genebuild/2020_12_macaca_fascicularis _gene_annotation.pd

– (Dec. 2021b). *Ensembl Gene Annotation (e!105) Primates Clade*. https://mart.ensembl.org/info/genome/genebuild/2021_02_primates_gene_annotation.pdf.

Erdmann, Volker A., Maciej Szymanski, Abraham Hochberg, Nathan De Groot, and Jan Barciszewski (Jan. 1999). *Collection of mRNA-like Non-Coding RNAs*. DOI: 10.1093/nar/27.1.192.

Erhard, Florian, Anne Halenius, Cosima Zimmermann, Anne L’Hernault, Daniel J. Kowalewski, Michael P. Weekes, Stefan Stevanovic, Ralf Zimmer, and Lars Dölken (May 2018). “Improved Ribo-seq Enables Identification of Cryptic Translation Events". In: *Nature Methods* 15.5, pp. 363–366. ISSN: 1548-7105. DOI: 10.1038/nmeth.4631.

Fabre, Bertrand, Jean-Philippe Combier, and Serge Plaza (Feb. 2021). "Recent Advances in Mass Spectrometry–Based Peptidomics Workflows to Identify Short-Open-Reading-Frame-Encoded Peptides and Explore Their Functions". In: *Current Opinion in Chemical Biology*. Omics 60, pp. 122–130. ISSN: 1367-5931. DOI: 10.1016/j.cbpa.2020.12.002.

Faghihi, Mohammad Ali, Farzaneh Modarresi, Ahmad M Khalil, Douglas E Wood, Barbara G Sahagan, Todd E Morgan, Caleb E Finch, Georges St Laurent, Paul J Kenny, Claes Wahlestedt, and Claes Wahlestedt (July 2008). "Expression of a Noncoding RNA Is Elevated in Alzheimer's Disease and Drives Rapid Feed-Forward Regulation of Beta-Secretase." In: *Nature medicine* 14.7, pp. 723–30. ISSN: 1546-170X. DOI: 10.1038/nm1784.

Faghihi, Mohammad Ali, Ming Zhang, Jia Huang, Farzaneh Modarresi, Marcel P. Van der Brug, Michael A. Nalls, Mark R. Cookson, Georges St-Laurent, and Claes Wahlestedt (May 2010). "Evidence for Natural Antisense Transcript-Mediated Inhibition of microRNA Function". In: *Genome Biology* 11.5, R56. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-5-r56.

Feng, Jianchi, Chunming Bi, Brian S. Clark, Rina Mady, Palak Shah, and Jhumku D. Kohtz (June 2006). "The Evf-2 Noncoding RNA Is Transcribed from the Dlx-5/6 Ultraconserved Region and Functions as a Dlx-2 Transcriptional Coactivator". In: *Genes & Development* 20.11, pp. 1470–1484. ISSN: 0890-9369. DOI: 10.1101/gad.1416106.

Fields, Alexander P., Edwin H. Rodriguez, Marko Jovanovic, Noam Stern-Ginossar, Brian J. Haas, Philipp Mertins, Raktima Raychowdhury, Nir Hacohen, Steven A. Carr, Nicholas T. Ingolia, Aviv Regev, and Jonathan S. Weissman (Dec. 2015). "A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation". In: *Molecular Cell* 60.5, pp. 816–827. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.11.013.

Flynn, Ryan A. and Howard Y. Chang (June 2014). "Long Noncoding RNAs in Cell-Fate Programming and Reprogramming". In: *Cell Stem Cell* 14.6, pp. 752–761. ISSN: 1934-5909, 1875-9777. DOI: 10.1016/j.stem.2014.05.014.

Forster, J. I., S. Köglsberger, C. Trefois, O. Boyd, A. S. Baumuratov, L. Buck, R. Balling, and P. M. A. Antony (June 2016). "Characterization of Differentiated SH-SY5Y as Neuronal Screening Model Reveals Increased Oxidative Vulnerability". In: *Journal of Biomolecular Screening* 21.5, pp. 496–509. ISSN: 1087-0571. DOI: 10.1177/1087057115625190.

Frankish, Adam, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L Howe, Toby Hunt, Osagie G Izuogu, Rory Johnson, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y Wolf, Jinuri Xu, Yucheng T Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Michael L Tress, and Paul Flicek (Jan. 2021). "GENCODE 2021". In: *Nucleic Acids Research* 49.D1, pp. D916–D923. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1087.

Freedman, Adam H. and Robert K. Wayne (2017). "Deciphering the Origin of Dogs: From Fossils to Genomes". In: DOI: 10.1146/annurev-animal-022114-110937.

Frith, Martin C., Alistair R. Forrest, Ehsan Nourbakhsh, Ken C. Pang, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, Timothy L. Bailey, and Sean M. Grimmond (Apr. 2006). "The Abundance of Short Proteins in the Mammalian Proteome". In: *PLOS Genetics* 2.4, e52. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0020052.

Fuke, Hiroyuki and Mutsuhito Ohno (Feb. 2008). "Role of Poly (A) Tail as an Identity Element for mRNA Nuclear Export." In: *Nucleic acids research* 36.3, pp. 1037–49. ISSN: 1362-4962. DOI: 10.1093/nar/gkm1120.

Gaertner, Bjoern, Sebastiaan van Heesch, Valentin Schneider-Lunitz, Jana Felicitas Schulz, Franziska Witte, Susanne Blachut, Steven Nguyen, Regina Wong, Ileana Matta, Norbert Hübner, and Maike Sander (Aug. 2020). "A Human ESC-based Screen Identifies a Role for the Translated lncRNA LINC00261 in Pancreatic Endocrine Differentiation". In: *eLife* 9, e58659. ISSN: 2050-084X. DOI: 10.7554/eLife.58659.

Gandhi, Minakshi, Maiwen Caudron-Herger, and Sven Diederichs (2018). "RNA Motifs and Combinatorial Prediction of Interactions, Stability and Localization of Noncoding RNAs". In: *Nature Structural & Molecular Biology* 25.12. ISSN: 1545-9993. DOI: 10.1038/s41594-018-0155-0.

Garnier, Simon, Noam Ross, Robert Rudis, Antonio P. Camargo, Marco Sciaini, and Cédric Scherer (2021). *Rvision - Colorblind-Friendly Color Maps for R*.

Ge, Qiwei, Dingjiacheng Jia, Dong Cen, Yadong Qi, Chengyu Shi, Junhong Li, Lingjie Sang, Luo-jia Yang, Jiamin He, Aifu Lin, Shujie Chen, and Liangjing Wang (Nov. 2021). "Micropeptide ASAP Encoded by LINC00467 Promotes Colorectal Cancer Progression by Directly Modulating ATP Synthase Activity". In: *The Journal of Clinical Investigation* 131.22, e152911. ISSN: 0021-9738. DOI: 10.1172/JCI152911.

Gelhausen, Rick, Teresa Müller, Sarah L Svensson, Omer S Alkhnbashi, Cynthia M Sharma, Florian Eggenhofer, and Rolf Backofen (Jan. 2022). "RiboReport - Benchmarking Tools for Ribosome Profiling-Based Identification of Open Reading Frames in Bacteria". In: *Briefings in Bioinformatics* 23.2, bbab549. ISSN: 1467-5463. DOI: 10.1093/bib/bbab549.

Girard, Angélique, Ravi Sachidanandam, Gregory J. Hannon, and Michelle A. Carmell (July 2006). "A Germline-Specific Class of Small RNAs Binds Mammalian Piwi Proteins". In: *Nature* 442.7099, pp. 199–202. ISSN: 1476-4687. DOI: 10.1038/nature04917.

Goudarzi, Mehdi, Kathryn Berg, Lindsey M Pieper, and Alexander F Schier (Jan. 2019). "Individual Long Non-Coding RNAs Have No Overt Functions in Zebrafish Embryogenesis, Viability and Fertility". In: *eLife* 8. ISSN: 2050-084X. DOI: 10.7554/eLife.40815.

Greco, Simona, Germana Zaccagnini, Paola Fuschi, Christine Voellenkle, Matteo Carrara, Iman Sadeghi, Claudia Bearzi, Biagina Maimone, Serenella Castelvecchio, Konstantinos Stellos, Carlo Gaetano, Lorenzo Menicanti, and Fabio Martelli (Apr. 2017). "Increased BACE1-AS Long Noncoding RNA and Beta-Amyloid Levels in Heart Failure". In: *Cardiovascular Research* 113.5, pp. 453–463. ISSN: 0008-6363. DOI: 10.1093/cvr/cvx013.

Gu, Zuguang, Roland Eils, and Matthias Schlesner (Sept. 2016). "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data". In: *Bioinformatics* 32.18, pp. 2847–2849. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw313.

Guo, Huili (July 2018). "Specialized Ribosomes and the Control of Translation". In: *Biochemical Society Transactions* 46.4, pp. 855–869. ISSN: 0300-5127. DOI: 10.1042/BST20160426.

Guttman, Mitchell and John L. Rinn (Feb. 2012). "Modular Regulatory Principles of Large Non-Coding RNAs". In: *Nature* 482.7385, pp. 339–346. ISSN: 1476-4687. DOI: 10.1038/nature10887.

Guttman, Mitchell, Pamela Russell, Nicholas T Ingolia, Jonathan S Weissman, and Eric S Lander (July 2013). "Ribosome Profiling Provides Evidence That Large Noncoding RNAs Do Not Encode Proteins." In: *Cell* 154.1, pp. 240–51. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.06.009.

Haerty, Wilfried and Chris P Ponting (May 2013). "Mutations within lncRNAs Are Effectively Selected against in Fruitfly but Not in Human". In: *Genome Biology* 14.5, R49. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-5-r49.

Hajjari, Mohammadreza and Abbas Salavaty (Mar. 2015). "HOTAIR: An Oncogenic Long Non-Coding RNA in Different Cancers." In: *Cancer biology & medicine* 12.1, pp. 1–9. ISSN: 2095-3941. DOI: 10.7497/j.issn.2095-3941.2015.0006.

Hannon, Gregory J. (2009). *FASTX-Toolkit*.

He, Sha, Shiping Liu, and Hao Zhu (Apr. 2011). "The Sequence, Structure and Evolutionary Features of HOTAIR in Mammals". In: *BMC Evolutionary Biology* 11.1, p. 102. ISSN: 1471-2148. DOI: 10.1186/1471-2148-11-102.

Hedges, S Blair, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar (2015). “Tree of Life Reveals Clock-Like Speciation and Diversification”. In: p. 11.

Heyer, Erin E. and Melissa J. Moore (Feb. 2016). “Redefining the Translational Status of 80S Monosomes”. In: *Cell* 164.4, pp. 757–769. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2016.01.003.

Hezroni, Hadas, Rotem Ben-Tov Perry, Zohar Meir, Gali Housman, Yoav Lubelsky, and Igor Ulitsky (Dec. 2017). “A Subset of Conserved Mammalian Long Non-Coding RNAs Are Fossils of Ancestral Protein-Coding Genes”. In: *Genome Biology* 18.1, p. 162. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1293-0.

Hezroni, Hadas, David Koppstein, Matthew G Schwartz, Alexandra Avrutin, David P Bartel, and Igor Ulitsky (May 2015). “Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species.” In: *Cell reports* 11.7, pp. 1110–22. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.04.023.

Hoagland, Mahlon B., Mary Louise Stephenson, Jesse F. Scott, Liselotte I. Hecht, and Paul C. Zamecnik (Mar. 1958). “A SOLUBLE RIBONUCLEIC ACID INTERMEDIATE IN PROTEIN SYNTHESIS”. In: *Journal of Biological Chemistry* 231.1, pp. 241–257. ISSN: 0021-9258. DOI: 10.1016/S0021-9258(19)77302-5.

Hodnett, J. L. and H. Busch (Dec. 1968). “Isolation and Characterization of Uridylic Acid-Rich 7 S Ribonucleic Acid of Rat Liver Nuclei”. In: *The Journal of Biological Chemistry* 243.24, pp. 6334–6342. ISSN: 0021-9258.

Hoerth, Katharina, Sonja Reitter, and Johanna Schott (Feb. 2022). “Normalized Ribo-Seq for Quantifying Absolute Global and Specific Changes in Translation”. In: *Bio-protocol* 12.4, e4323. ISSN: 2331-8325. DOI: 10.21769/BioProtoc.4323.

Hsu, Ming-Ta and Miguel Coca-Prados (July 1979). “Electron Microscopic Evidence for the Circular Form of RNA in the Cytoplasm of Eukaryotic Cells”. In: *Nature* 280.5720, pp. 339–340. ISSN: 1476-4687. DOI: 10.1038/280339a0.

Hulstaert, Niels, Jim Shofstahl, Timo Sachsenberg, Mathias Walzer, Harald Barsnes, Lennart Martens, and Yasset Perez-Riverol (Jan. 2020). “ThermoRawFileParser: Modular, Scalable and Cross-Platform RAW File Conversion”. In: *Journal of proteome research* 19.1, pp. 537–542. ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.9b00328.

Iacono, Michele, Flavio Mignone, and Graziano Pesole (Apr. 2005). “uAUG and uORFs in Human and Rodent 5′ Untranslated mRNAs”. In: *Gene* 349, pp. 97–105. ISSN: 0378-1119. DOI: 10.1016/j.gene.2004.11.041.

Ikemura, T. and H. Ozeki (1983). “Codon Usage and Transfer RNA Contents: Organism-Specific Codon-Choice Patterns in Reference to the Isoacceptor Contents”. In: *Cold Spring Harbor Symposia on Quantitative Biology* 47 Pt 2, pp. 1087–1097. ISSN: 0091-7451. DOI: 10.1101/sqb.1983.047.01.123.

Imam, Hasan, Aalia Shahr Bano, Paresh Patel, Prasida Holla, and Shahid Jameel (Mar. 2015). “The lncRNA NRON Modulates HIV-1 Replication in a NFAT-dependent Manner and Is Differentially Regulated by Early and Late Viral Proteins”. In: *Scientific Reports* 5.1, p. 8639. ISSN: 2045-2322. DOI: 10.1038/srep08639.

Ingolia, Nicholas T, Gloria A Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman (July 2013). “Genome-Wide Annotation and Quantitation of Translation by Ribosome Profiling.” In: *Current protocols in molecular biology* Chapter 4, Unit 4.18. ISSN: 1934-3647. DOI: 10.1002/0471142727.mb0418s103.

Ingolia, Nicholas T, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman (Apr. 2009). “Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling.” In: *Science (New York, N.Y.)* 324.5924, pp. 218–23. ISSN: 1095-9203. DOI: 10.1126/science.1168978.

Ingolia, Nicholas T., Liana F. Lareau, and Jonathan S. Weissman (Nov. 2011). “Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes”. In: *Cell* 147.4, pp. 789–802. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2011.10.002.

Jandura, Allison and Henry M. Krause (Oct. 2017). “The New RNA World: Growing Evidence for Long Noncoding RNA Functionality”. In: 33.10, pp. 665–676. ISSN: 0168-9525.

Ji, Ping, Sven Diederichs, Wenbing Wang, Sebastian Böing, Ralf Metzger, Paul M. Schneider, Nicola Tidow, Burkhard Brandt, Horst Buerger, Etmar Bulk, Michael Thomas, Wolfgang E. Berdel, Hubert Serve, and Carsten Müller-Tidow (Sept. 2003). “MALAT-1, a Novel Noncoding RNA, and Thymosin $B4$ Predict Metastasis and Survival in Early-Stage Non-Small Cell Lung Cancer”. In: *Oncogene* 22.39, pp. 8031–8041. ISSN: 1476-5594. DOI: 10.1038/sj.onc.1206928.

Ji, Zhe, Ruisheng Song, Aviv Regev, and Kevin Struhl (Dec. 2015). “Many lncRNAs, 5’UTRs, and Pseudogenes Are Translated and Some Are Likely to Express Functional Proteins”. In: *eLife* 4. DOI: 10.7554/eLife.08890.

Jiménez, J (2022). *Odseq: Outlier Detection in Multiple Sequence Alignments*.

Johnsson, Per, Leonard Lipovich, Dan Grandér, and Kevin V. Morris (Mar. 2014). “Evolutionary Conservation of Long Noncoding RNAs; Sequence, Structure, Function”. In: *Biochimica et biophysica acta* 1840.3, pp. 1063–1071. ISSN: 0006-3002. DOI: 10.1016/j.bbagen.2013.10.035.

Jørgensen, Flemming and Charles G. Kurland (Oct. 1990). “Processivity Errors of Gene Expression in Escherichia Coli”. In: *Journal of Molecular Biology* 215.4, pp. 511–521. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80164-0.

Kaessmann, Henrik (Oct. 2010). “Origins, Evolution, and Phenotypic Impact of New Genes.” In: *Genome research* 20.10, pp. 1313–26. ISSN: 1549-5469. DOI: 10.1101/gr.101386.109.

Kang, Jian, Natalie Brajanovski, Keefe T. Chan, Jiachen Xuan, Richard B. Pearson, and Elaine Sanij (Aug. 2021). “Ribosomal Proteins and Human Diseases: Molecular Mechanisms and Targeted Therapy”. In: *Signal Transduction and Targeted Therapy* 6.1, pp. 1–22. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00728-8.

Kapp, Lee D. and Jon R. Lorsch (2004). “The Molecular Mechanics of Eukaryotic Translation”. In: *Annual Review of Biochemistry* 73.1, pp. 657–704. DOI: 10.1146/annurev.biochem.73.030403.080419.

Kapranov, Philipp, Jill Cheng, Sujit Dike, David A. Nix, Radharani Duttagupta, Aarron T. Willingham, Peter F. Stadler, Jana Hertel, Jörg Hackermüller, Ivo L. Hofacker, Ian Bell, Evelyn Cheung, Jorg Drenkow, Erica Dumais, Sandeep Patel, Gregg Helt, Madhavan Ganesh, Srinka Ghosh, Antonio Piccolboni, Victor Sementchenko, Hari Tammana, and Thomas R. Gingeras (June 2007). “RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription”. In: *Science* 316.5830, pp. 1484–1488. DOI: 10.1126/science.1138341.

Kapusta, Aurélie, Zev Kronenberg, Vincent J. Lynch, Xiaoyu Zhuo, LeeAnn Ramsay, Guillaume Bourque, Mark Yandell, and Cédric Feschotte (Apr. 2013). “Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs”. In: *PLOS Genetics* 9.4, e1003470. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003470.

Karginov, Timofey A., Daniel Parviz Hejazi Pastor, Bert L. Semler, and Christopher M. Gomez (Feb. 2017). “Mammalian Polycistronic mRNAs and Disease”. In: *Trends in Genetics* 33.2, pp. 129–142. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2016.11.007.

Kassambara, Alboukadel (2023). *Ggpubr: ’ggplot2’ Based Publication Ready Plots*.

Katoh, K. and D. M. Standley (Apr. 2013). “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability”. In: *Molecular Biology and Evolution* 30.4, pp. 772–780. ISSN: 0737-4038. DOI: 10.1093/molbev/mst010.

Kedersha, N L and L H Rome (Sept. 1986). “Isolation and Characterization of a Novel Ribonucleoprotein Particle: Large Structures Contain a Single Species of Small RNA.” In: *Journal of Cell Biology* 103.3, pp. 699–709. ISSN: 0021-9525. DOI: 10.1083/jcb.103.3.699.

Khalil, Ahmad M., Mohammad Ali Faghihi, Farzaneh Modarresi, Shaun P. Brothers, and Claes Wahlestedt (Jan. 2008). “A Novel RNA Transcript with Antiapoptotic Function Is Silenced in Fragile X Syndrome". In: *PLoS ONE* 3.1, e1486. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0001486.

Khitun, Alexandra and Sarah A. Slavoff (2019). "Proteomic Detection and Validation of Translated Small Open Reading Frames". In: *Current Protocols in Chemical Biology* 11.4, e77. ISSN: 2160-4762. DOI: 10.1002/cpch.77.

Kim, Min-Sik, Jun Zhong, and Akhilesh Pandey (Mar. 2016). "Common Errors in Mass Spectrometry-Based Analysis of Post-Translational Modifications". In: *Proteomics* 16.5, pp. 700–714. ISSN: 1615-9853. DOI: 10.1002/pmic.201500355.

King, J L and T H Jukes (May 1969). "Non-Darwinian Evolution." In: *Science (New York, N.Y.)* 164.3881, pp. 788–98. ISSN: 0036-8075. DOI: 10.1126/science.164.3881.788. PMID: 5767777.

Kiniry, Stephen J, Audrey M Michel, and Pavel V Baranov (Nov. 2019). "Computational Methods for Ribosome Profiling Data Analysis." In: *Wiley interdisciplinary reviews. RNA*, e1577. ISSN: 1757-7012. DOI: 10.1002/wrna.1577. PMID: 31760685.

Kirk, Jessime M., Susan O. Kim, Kaoru Inoue, Matthew J. Smola, David M. Lee, Megan D. Schertzer, Joshua S. Wooten, Allison R. Baker, Daniel Sprague, David W. Collins, Christopher R. Horning, Shuo Wang, Qidi Chen, Kevin M. Weeks, Peter J. Mucha, and J. Mauro Calabrese (Oct. 2018). "Functional Classification of Long Non-Coding RNAs by k-Mer Content". In: *Nature Genetics* 50.10, pp. 1474–1482. ISSN: 1061-4036. DOI: 10.1038/s41588-018-0207-8.

Klinke, Sigbert (n.d.). *Plot.Matrix*.

Kochetov, Alex V. (2008). "Alternative Translation Start Sites and Hidden Coding Potential of Eukaryotic mRNAs". In: *BioEssays* 30.7, pp. 683–691. ISSN: 1521-1878. DOI: 10.1002/bies.20771.

Kolekar, Pandurang, Abhijeet Pataskar, Urmila Kulkarni-Kale, Jayanta Pal, and Abhijeet Kulkarni (July 2016). "IRESPred: Web Server for Prediction of Cellular and Viral Internal Ribosome Entry Site (IRES)". In: *Scientific Reports* 6.1, p. 27436. ISSN: 2045-2322. DOI: 10.1038/srep27436.

Komar, Anton A and Maria Hatzoglou (Jan. 2011). "Cellular IRES-mediated Translation: The War of ITAFs in Pathophysiological States." In: *Cell cycle (Georgetown, Tex.)* 10.2, pp. 229–40. ISSN: 1551-4005. DOI: 10.4161/cc.10.2.14472.

Koodli, Rohan V., Benjamin Keep, Katherine R. Coppess, Fernando Portela, Eterna Participants, and Rhiju Das (Apr. 2019). "EternaBrain: Automated RNA Design through Move Sets and Strategies from an Internet-scale RNA Videogame". In: *bioRxiv*, p. 326736. DOI: 10.1101/326736.

Kopp, Florian and Joshua T. Mendell (Jan. 2018). "Functional Classification and Experimental Dissection of Long Noncoding RNAs". In: *Cell* 172.3, pp. 393–407. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.01.011.

Kovalevich, Jane, Maryline Santerre, and Dianne Langford (2021). "Considerations for the Use of SH-SY5YSH-SY5Y NeuroblastomaNeuroblastoma Cells in Neurobiology". In: *Neuronal Cell Culture: Methods and Protocols*. Ed. by Shohreh Amini and Martyn K. White. Methods in Molecular Biology. New York, NY: Springer US, pp. 9–23. ISBN: 978-1-07-161437-2. DOI: 10.1007/978-1-0716-1437-2_2.

Kozak, M (Oct. 1987). "Effects of Intercistronic Length on the Efficiency of Reinitiation by Eucaryotic Ribosomes." In: *Molecular and Cellular Biology* 7.10, pp. 3438–3445. ISSN: 0270-7306.

– (Nov. 1989). "Context Effects and Inefficient Initiation at Non-AUG Codons in Eucaryotic Cell-Free Translation Systems." In: *Molecular and Cellular Biology* 9.11, pp. 5073–5080. ISSN: 0270-7306.

Kozak, Marilyn (Nov. 1980). "Evaluation of the "Scanning Model" for Initiation of Protein Synthesis in Eucaryotes". In: *Cell* 22.1, Part 1, pp. 7–8. ISSN: 0092-8674. DOI: 10.1016/0092-8674(80)90148-8.

– (Jan. 1986). “Point Mutations Define a Sequence Flanking the AUG Initiator Codon That Modulates Translation by Eukaryotic Ribosomes”. In: *Cell* 44.2, pp. 283–292. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/0092-8674(86)90762-2.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges (July 2017). “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times”. In: *Molecular Biology and Evolution* 34.7, pp. 1812–1819. ISSN: 0737-4038. DOI: 10.1093/molbev/msx116.

Ladoukakis, Emmanuel, Vini Pereira, Emile G Magny, Adam Eyre-Walker, and Juan Couso (Nov. 2011). “Hundreds of Putatively Functional Small Open Reading Frames in Drosophila”. In: *Genome Biology* 12.11, R118. ISSN: 1465-6906. DOI: 10.1186/gb-2011-12-11-r118.

Landan, Giddy and Dan Graur (Dec. 2007). “LOCAL RELIABILITY MEASURES FROM SETS OF CO-OPTIMAL MULTIPLE SEQUENCE ALIGNMENTS”. In: *Biocomputing 2008*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, pp. 15–24. DOI: 10.1142/9789812776136_0003.

Lander, Eric S. et al. (Feb. 2001). “Initial Sequencing and Analysis of the Human Genome”. In: *Nature* 409.6822, pp. 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062.

Langmead, Ben and Steven L Salzberg (Apr. 2012). “Fast Gapped-Read Alignment with Bowtie 2”. In: *Nature Methods* 9.4, pp. 357–359. ISSN: 1548-7091. DOI: 10.1038/nmeth.1923.

Lanz, R B, N J McKenna, S A Onate, U Albrecht, J Wong, S Y Tsai, M J Tsai, and B W O’Malley (Apr. 1999). “A Steroid Receptor Coactivator, SRA, Functions as an RNA and Is Present in an SRC-1 Complex.” In: *Cell* 97.1, pp. 17–27. ISSN: 0092-8674. PMID: 10199399.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins (Nov. 2007). “Clustal W and Clustal X Version 2.0”. In: *Bioinformatics* 23.21, pp. 2947–2948. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm404.

Latos, Paulina A, Florian M Pauler, Martha V Koerner, H Başak Şenergin, Quanah J Hudson, Roman R Stocsits, Wolfgang Allhoff, Stefan H Stricker, Ruth M Klement, Katarzyna E Warczok, Karin Aumayr, Pawel Pasierbek, and Denise P Barlow (Dec. 2012). “Airn Transcriptional Overlap, but Not Its lncRNA Products, Induces Imprinted Igf2r Silencing.” In: *Science (New York, N.Y.)* 338.6113, pp. 1469–72. ISSN: 1095-9203. DOI: 10.1126/science.1228110. PMID: 23239737.

Lawless, Craig, Richard D. Pearson, Julian N. Selley, Julia B. Smirnova, Christopher M. Grant, Mark P. Ashe, Graham D. Pavitt, and Simon J. Hubbard (Jan. 2009). “Upstream Sequence Elements Direct Post-Transcriptional Regulation of Gene Expression under Stress Conditions in Yeast”. In: *BMC Genomics* 10.1, p. 7. ISSN: 1471-2164. DOI: 10.1186/1471-2164-10-7.

Lee, Jeehyung, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Sungroh Yoon, Adrien Treuille, Rhiju Das, and EteRNA EteRNA Participants (Feb. 2014). “RNA Design Rules from a Massive Open Laboratory.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.6, pp. 2122–7. ISSN: 1091-6490. DOI: 10.1073/pnas.1313039111.

Lee, Sungyul, Florian Kopp, Tsung-Cheng Chang, Anupama Sataluri, Beibei Chen, Sushama Sivakumar, Hongtao Yu, Yang Xie, and Joshua T. Mendell (Jan. 2016). “Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins”. In: *Cell* 164.1, pp. 69–80. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2015.12.017.

Lefever, Steve, Jasper Anckaert, Pieter-Jan Volders, Manuel Luypaert, Jo Vandesompele, and Pieter Mestdagh (June 2017). “decodeRNA— Predicting Non-Coding RNA Functions Using Guilt-by-Association”. In: *Database: The Journal of Biological Databases and Curation* 2017, bax042. ISSN: 1758-0463. DOI: 10.1093/database/bax042.

Levine, Mia T, Corbin D Jones, Andrew D Kern, Heather A Lindfors, and David J Begun (June 2006). “Novel Genes Derived from Noncoding DNA in Drosophila Melanogaster Are Frequently X-linked and Exhibit Testis-Biased Expression.” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.26, pp. 9935–9. ISSN: 0027-8424. DOI: 10.1073/pnas.0509809103.

Li, Bo, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey (Feb. 2010). “RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty”. In: *Bioinformatics* 26.4, pp. 493–500. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp692.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup (Aug. 2009). “The Sequence Alignment/Map Format and SAMtools”. In: *Bioinformatics* 25.16, pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352.

Li, Jun, Cancan Chen, Xiancai Ma, Guannan Geng, Bingfeng Liu, Yijun Zhang, Shaoyang Zhang, Fudi Zhong, Chao Liu, Yue Yin, Weiping Cai, and Hui Zhang (June 2016). “Long Noncoding RNA NRON Contributes to HIV-1 Latency by Specifically Inducing Tat Protein Degradation”. In: *Nature Communications* 7.1, p. 11730. ISSN: 2041-1723. DOI: 10.1038/ncomms11730.

Li, Mingfeng, Gabriel Santpere, Yuka Imamura Kawasawa, Oleg V. Evgrafov, Forrest O. Gulden, Sirisha Pochareddy, Susan M. Sunkin, Zhen Li, Yurae Shin, Ying Zhu, André M. M. Sousa, Donna M. Werling, Robert R. Kitchen, Hyo Jung Kang, Mihovil Pletikos, Jinmyung Choi, Sydney Muchnik, Xuming Xu, Daifeng Wang, Belen Lorente-Galdos, Shuang Liu, Paola Giusti-Rodríguez, Hyejung Won, Christiaan A. de Leeuw, Antonio F. Pardiñas, Ming Hu, Fulai Jin, Yun Li, Michael J. Owen, Michael C. O’Donovan, James T. R. Walters, Danielle Posthuma, Mark A. Reimers, Pat Levitt, Daniel R. Weinberger, Thomas M. Hyde, Joel E. Kleinman, Daniel H. Geschwind, Michael J. Hawrylycz, Matthew W. State, Stephan J. Sanders, Patrick F. Sullivan, Mark B. Gerstein, Ed S. Lein, James A. Knowles, and Nenad Sestan (Dec. 2018). “Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risks”. In: *Science (New York, N.Y.)* 362.6420, eaat7615. ISSN: 0036-8075. DOI: 10.1126/science.aat7615.

Lilue, Jingtao, Anthony G. Doran, Ian T. Fiddes, Monica Abrudan, Joel Armstrong, Ruth Bennett, William Chow, Joanna Collins, Stephan Collins, Anne Czechanski, Petr Danecek, Mark Diekhans, Dirk-Dominik Dolle, Matt Dunn, Richard Durbin, Dent Earl, Anne Ferguson-Smith, Paul Flicek, Jonathan Flint, Adam Frankish, Beiyuan Fu, Mark Gerstein, James Gilbert, Leo Goodstadt, Jennifer Harrow, Kerstin Howe, Ximena Ibarra-Soria, Mikhail Kolmogorov, Chris J. Lelliott, Darren W. Logan, Jane Loveland, Clayton E. Mathews, Richard Mott, Paul Muir, Stefanie Nachtweide, Fabio C. P. Navarro, Duncan T. Odom, Naomi Park, Sarah Pelan, Son K. Pham, Mike Quail, Laura Reinholdt, Lars Romoth, Lesley Shirley, Cristina Sisu, Marcela Sjoberg-Herrera, Mario Stanke, Charles Steward, Mark Thomas, Glen Threadgold, David Thybert, James Torrance, Kim Wong, Jonathan Wood, Binnaz Yalcin, Fengtang Yang, David J. Adams, Benedict Paten, and Thomas M. Keane (Nov. 2018). “Sixteen Diverse Laboratory Mouse Reference Genomes Define Strain-Specific Haplotypes and Novel Functional Loci”. In: *Nature Genetics* 50.11, pp. 1574–1583. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0223-8.

Lin, Brian Y, Patricia P Chan, and Todd M Lowe (July 2019). “tRNAviz: Explore and Visualize tRNA Sequence Features”. In: *Nucleic Acids Research* 47.W1, W542–W547. ISSN: 0305-1048. DOI: 10.1093/nar/gkz438.

Lin, Nianwei, Kung-Yen Chang, Zhonghan Li, Keith Gates, Zacharia A. Rana, Jason Dang, Danhua Zhang, Tianxu Han, Chao-Shun Yang, Thomas J. Cunningham, Steven R. Head, Gregg Duester, P. Duc Si Dong, and Tariq M. Rana (Mar. 2014). “An Evolutionarily Conserved Long Noncoding RNA TUNA Controls Pluripotency and Neural Lineage Commitment”. In: *Molecular Cell* 53.6, pp. 1005–1019. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2014.01.021.

Liu, Chibo and Jianjun Lin (Oct. 2016). “Long Noncoding RNA ZEB1-AS1 Acts as an Oncogene in Osteosarcoma by Epigenetically Activating ZEB1”. In: *American Journal of Translational Research* 8.10, pp. 4095–4105. ISSN: 1943-8141.

Locke, Devin P. et al. (Jan. 2011). “Comparative and Demographic Analysis of Orang-Utan Genomes”. In: *Nature* 469.7331, pp. 529–533. ISSN: 1476-4687. DOI: 10.1038/nature09687.

Long, Manyuan, Nicholas W. VanKuren, Sidi Chen, and Maria D. Vibranovski (2013). “New Gene Evolution: Little Did We Know”. In: *Annual Review of Genetics* 47.1, pp. 307–333. DOI: 10.1146/annurev-genet-111212-133301.

Love, Michael I., Wolfgang Huber, and Simon Anders (Dec. 2014). “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2”. In: *Genome Biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Gabriele H. Marchler, James S. Song, Narmada Thanki, Roxanne A. Yamashita, Mingzhang Yang, Dachuan Zhang, Chanjuan Zheng, Christopher J. Lanczycki, and Aron Marchler-Bauer (Jan. 2020). “CDD/SPARCLE: The Conserved Domain Database in 2020”. In: *Nucleic Acids Research* 48.D1, pp. D265–D268. ISSN: 1362-4962. DOI: 10.1093/nar/gkz991.

Lyon, Mary F. (June 1962). “Sex Chromatin and Gene Action in the Mammalian X-Chromosome”. In: *American Journal of Human Genetics* 14.2, pp. 135–148. ISSN: 0002-9297.

Ma, Ming-Hui, Jia-Xiang An, Cheng Zhang, Jie Liu, Yu Liang, Chun-Dong Zhang, Zhen Zhang, and Dong-Qiu Dai (Feb. 2019). “ZEB1-AS1 Initiates a miRNA-mediated ceRNA Network to Facilitate Gastric Cancer Progression”. In: *Cancer Cell International* 19.1, p. 27. ISSN: 1475-2867. DOI: 10.1186/s12935-019-0742-0.

Magny, E. G., J. I. Pueyo, F. M. G. Pearl, M. A. Cespedes, J. E. Niven, S. A. Bishop, and J. P. Couso (Sept. 2013). “Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames”. In: *Science* 341.6150, pp. 1116–1120. ISSN: 0036-8075. DOI: 10.1126/science.1238802.

Márquez, Viter, Daniel N. Wilson, Warren P. Tate, Francisco Triana-Alonso, and Knud H. Nierhaus (July 2004). “Maintaining the Ribosomal Reading Frame: The Influence of the E Site during Translational Regulation of Release Factor 2”. In: *Cell* 118.1, pp. 45–55. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2004.06.012.

Martin, Marcel (May 2011). “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads”. In: *EMBnet.journal* 17.1, p. 10. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200.

Martinez, Thomas F., Qian Chu, Cynthia Donaldson, Dan Tan, Maxim N. Shokhirev, and Alan Saghatelian (Apr. 2020). “Accurate Annotation of Human Protein-Coding Small Open Reading Frames”. In: *Nature Chemical Biology* 16.4, pp. 458–468. ISSN: 1552-4469. DOI: 10.1038/s41589-019-0425-0.

Marx, Vivien (May 2014). “Cell-Line Authentication Demystified”. In: *Nature Methods* 11.5, pp. 483–488. ISSN: 1548-7105. DOI: 10.1038/nmeth.2932.

Matsumoto, Akinobu, John G. Clohessy, and Pier Paolo Pandolfi (May 2017). “SPAR, a lncRNA Encoded mTORC1 Inhibitor”. In: *Cell Cycle* 16.9, pp. 815–816. ISSN: 1538-4101, 1551-4005. DOI: 10.1080/15384101.2017.1304735.

Matsumoto, Akinobu, Alessandra Pasut, Masaki Matsumoto, Riu Yamashita, Jacqueline Fung, Emanuele Monteleone, Alan Saghatelian, Keiichi I. Nakayama, John G. Clohessy, and Pier Paolo Pandolfi (Jan. 2017). “mTORC1 and Muscle Regeneration Are Regulated by the LINC00961 Encoded SPAR Polypeptide”. In: *Nature* 541.7636, pp. 228–232. ISSN: 0028-0836. DOI: 10.1038/nature21034.

Mattick, John S. and John L. Rinn (Jan. 2015). “Discovery and Annotation of Long Noncoding RNAs”. In: *Nature Structural & Molecular Biology* 22.1, pp. 5–7. ISSN: 1545-9985. DOI: 10.1038/nsmb.2942.

McLysaght, Aoife and Laurence D. Hurst (Sept. 2016). “Open Questions in the Study of de Novo Genes: What, How and Why”. In: *Nature Reviews Genetics* 17.9, pp. 567–578. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.78.

Meijer, Hedda A. and Adri A.M. Thomas (Oct. 2002). “Control of Eukaryotic Protein Synthesis by Upstream Open Reading Frames in the 5′-Untranslated Region of an mRNA”. In: *Biochemical Journal* 367.1, pp. 1–11. ISSN: 0264-6021. DOI: 10.1042/bj20011706.

Menschaert, Gerben, Wim Van Criekinge, Tineke Notelaers, Alexander Koch, Jeroen Crappé, Kris Gevaert, and Petra Van Damme (July 2013). “Deep Proteome Coverage Based on

Ribosome Profiling Aids Mass Spectrometry-based Protein and Peptide Discovery and Provides Evidence of Alternative Translation Products and Near-cognate Translation Initiation Events*". In: *Molecular & Cellular Proteomics* 12.7, pp. 1780–1790. ISSN: 1535-9476. DOI: 10.1074/mcp.M113.027540.

Meredith, Robert W., Jan E. Janečka, John Gatesy, Oliver A. Ryder, Colleen A. Fisher, Emma C. Teeling, Alisha Goodbla, Eduardo Eizirik, Taiz L. L. Simão, Tanja Stadler, Daniel L. Rabosky, Rodney L. Honeycutt, John J. Flynn, Colleen M. Ingram, Cynthia Steiner, Tiffani L. Williams, Terence J. Robinson, Angela Burk-Herrick, Michael Westerman, Nadia A. Ayoub, Mark S. Springer, and William J. Murphy (Oct. 2011). "Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification". In: *Science* 334.6055, pp. 521–524. DOI: 10.1126/science.1211028.

Michel, Audrey M., James P. A. Mullan, Vimalkumar Velayudhan, Patrick B. F. O'Connor, Claire A. Donohue, and Pavel V. Baranov (Mar. 2016). "RiboGalaxy: A Browser Based Platform for the Alignment, Analysis and Visualization of Ribosome Profiling Data". In: *RNA Biology* 13.3, pp. 316–319. ISSN: 1547-6286. DOI: 10.1080/15476286.2016.1141862.

Mistry, Jaina, Alex Bateman, and Robert D Finn (Aug. 2007). "Predicting Active Site Residue Annotations in the Pfam Database". In: *BMC bioinformatics* 8, p. 298. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-298.

Mistry, Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta (July 2013). "Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions". In: *Nucleic Acids Research* 41.12, e121. ISSN: 0305-1048. DOI: 10.1093/nar/gkt263.

Moore, Andrew D. (Nov. 2022). *PfamScanner*.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold (July 2008). "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7105. DOI: 10.1038/nmeth.1226.

Mudge, Jonathan M., Jorge Ruiz-Orera, John R. Prensner, Marie A. Brunet, Ferriol Calvet, Irwin Jungreis, Jose Manuel Gonzalez, Michele Magrane, Thomas F. Martinez, Jana Felicitas Schulz, Yucheng T. Yang, M. Mar Albà, Julie L. Aspden, Pavel V. Baranov, Ariel A. Bazzini, Elspeth Bruford, Maria Jesus Martin, Lorenzo Calviello, Anne-Ruxandra Carvunis, Jin Chen, Juan Pablo Couso, Eric W. Deutsch, Paul Flicek, Adam Frankish, Mark Gerstein, Norbert Hubner, Nicholas T. Ingolia, Manolis Kellis, Gerben Menschaert, Robert L. Moritz, Uwe Ohler, Xavier Roucou, Alan Saghatelian, Jonathan S. Weissman, and Sebastiaan van Heesch (July 2022). "Standardized Annotation of Translated Open Reading Frames". In: *Nature Biotechnology* 40.7, pp. 994–999. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01369-0.

Müller, Kirill and Hadley Wickham (2021). *Tibble: Simple Data Frames*.

Murillo, Jimmy Rodriguez, Indira Pla, Livia Goto-Silva, Fábio C. S. Nogueira, Gilberto B. Domont, Yasset Perez-Riverol, Aniel Sánchez, and Magno Junqueira (Oct. 2018). "Mass Spectrometry Evaluation of a Neuroblastoma SH-SY5Y Cell Culture Protocol". In: *Analytical Biochemistry* 559, pp. 51–54. ISSN: 0003-2697. DOI: 10.1016/j.ab.2018.08.013.

Murphy, Daniel N. and Aoife McLysaght (Nov. 2012). "De Novo Origin of Protein-Coding Genes in Murine Rodents". In: *PLOS ONE* 7.11, e48650. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0048650.

Mus, El, Patrick R. Hof, and Henri Tiedge (June 2007). "Dendritic BC200 RNA in Aging and in Alzheimer's Disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.25, pp. 10679–10684. ISSN: 0027-8424. DOI: 10.1073/pnas.0701532104.

Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner, and Henrik Kaessmann (Jan. 2014). "The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods". In: *Nature* 505.7485, pp. 635–640. ISSN: 1476-4687. DOI: 10.1038/nature12943.

Nelson, Benjamin R, Catherine A Makarewich, Douglas M Anderson, Benjamin R Winders, Constantine D Troupes, Fenfen Wu, Austin L Reese, John R McAnally, Xiongwen Chen, Ege T Kavalali, Stephen C Cannon, Steven R Houser, Rhonda Bassel-Duby, and Eric N Olson (Jan. 2016). “A Peptide Encoded by a Transcript Annotated as Long Noncoding RNA Enhances SERCA Activity in Muscle.” In: *Science (New York, N.Y.)* 351.6270, pp. 271–5. ISSN: 1095-9203. DOI: 10.1126/science.aad4076.

Neme, Rafik and Diethard Tautz (Feb. 2013). “Phylogenetic Patterns of Emergence of New Genes Support a Model of Frequent de Novo Evolution”. In: *BMC Genomics* 14, p. 117. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-117.

Neuwirth, Erich (2022). *RColorBrewer: ColorBrewer Palettes*.

Noh, Ji Heon, Kyoung Mi Kim, Waverly G. McClusky, Kotb Abdelmohsen, and Myriam Gorospe (Mar. 2018). “Cytoplasmic Functions of Long Noncoding RNAs”. In: *Wiley Interdisciplinary Reviews: RNA*, e1471. ISSN: 17577004. DOI: 10.1002/wrna.1471.

Norris, Karl, Tayah Hopes, and Julie Louise Aspden (2021). “Ribosome Heterogeneity and Specialization in Development”. In: *Wiley Interdisciplinary Reviews. RNA* 12.4, e1644. ISSN: 1757-7004. DOI: 10.1002/wrna.1644.

Ohno, S (1972). “So Much "Junk" DNA in Our Genome.” In: *Brookhaven symposia in biology* 23, pp. 366–70. ISSN: 0068-2799. PMID: 5065367.

Olson, Wilma K, Mauricio Esguerra, Yurong Xin, and Xiang-Jun Lu (Mar. 2009). “New Information Content in RNA Base Pairing Deduced from Quantitative Analysis of High-Resolution Structures.” In: *Methods (San Diego, Calif.)* 47.3, pp. 177–86. ISSN: 1095-9130. DOI: 10.1016/j.ymeth.2008.12.003.

Oyama, Masaaki, Chiharu Itagaki, Hiroko Hata, Yutaka Suzuki, Tomonori Izumi, Tohru Natsume, Toshiaki Isobe, and Sumio Sugano (Oct. 2004). “Analysis of Small Human Proteins Reveals the Translation of Upstream Open Reading Frames of mRNAs”. In: *Genome Research* 14.10b, pp. 2048–2052. ISSN: 1088-9051. DOI: 10.1101/gr.2384604.

Ozadam, Hakan, Michael Geng, and Can Cenik (May 2020). “RiboFlow, RiboR and RiboPy: An Ecosystem for Analyzing Ribosome Profiling Data at Read Length Resolution”. In: *Bioinformatics* 36.9, pp. 2929–2931. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa028.

Palazzo, Alexander F and Eliza S Lee (2015). “Non-Coding RNA: What Is Functional and What Is Junk?” In: *Frontiers in genetics* 6, p. 2. ISSN: 1664-8021. DOI: 10.3389/fgene.2015.00002.

Palazzo, Alexander F. and Eugene V. Koonin (Nov. 2020). “Functional Long Non-coding RNAs Evolve from Junk Transcripts”. In: *Cell* 183.5, pp. 1151–1161. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.09.047.

Palmieri, Nicola, Carolin Kosiol, and Christian Schlötterer (Feb. 2014). “The Life Cycle of Drosophila Orphan Genes”. In: *eLife* 3. Ed. by Diethard Tautz, e01311. ISSN: 2050-084X. DOI: 10.7554/eLife.01311.

Pandey, Radha Raman, Tanmoy Mondal, Faizaan Mohammad, Stefan Enroth, Lisa Redrup, Jan Komorowski, Takashi Nagano, Debora Mancini-DiNardo, and Chandrasekhar Kanduri (Oct. 2008). “Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation”. In: *Molecular Cell* 32.2, pp. 232–246. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2008.08.022.

Patil, Indrajeet (May 2021). “Visualizations with Statistical Details: The ’ggstatsplot’ Approach”. In: *Journal of Open Source Software* 6.61, p. 3167. ISSN: 2475-9066. DOI: 10.21105/joss.03167.

Patraquim, Pedro, Emile G. Magny, José I. Pueyo, Ana Isabel Platero, and Juan Pablo Couso (Oct. 2022). “Translation and Natural Selection of Micropeptides from Long Non-Canonical RNAs”. In: *Nature Communications* 13.1, p. 6515. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34094-y.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford (Apr. 2017). “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression”. In: *Nature Methods* 14.4, pp. 417–419. ISSN: 1548-7091. DOI: 10.1038/nmeth.4197.

Patro, Rob, Stephen M. Mount, and Carl Kingsford (May 2014). “Sailfish Enables Alignment-Free Isoform Quantification from RNA-seq Reads Using Lightweight Algorithms”. In: *Nature biotechnology* 32.5, pp. 462–464. ISSN: 1087-0156. DOI: 10.1038/nbt.2862.

Patro, Rob, Avi Srivastava, and Hirak Sarkar (Oct. 2022). *SalmonTools*. COMBINE lab.

Pauli, Andrea, Megan L Norris, Eivind Valen, Guo-Liang Chew, James A Gagnon, Steven Zimmerman, Andrew Mitchell, Jiao Ma, Julien Dubrulle, Deepak Reyon, Shengdar Q Tsai, J Keith Joung, Alan Saghatelian, and Alexander F Schier (Feb. 2014). “Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors.” In: *Science (New York, N.Y.)* 343.6172, p. 1248636. ISSN: 1095-9203. DOI: 10.1126/science.1248636.

Peabody, D S (Mar. 1989). “Translation Initiation at Non-Aug Triplets in Mammalian Cells”. In: *Journal of Biological Chemistry* 264.9, pp. 5031–5035. ISSN: 00219258. DOI: 10.1016/S0021-9258(18)83694-8.

Perelman, Polina, Warren E. Johnson, Christian Roos, Hector N. Seuánez, Julie E. Horvath, Miguel A. M. Moreira, Bailey Kessing, Joan Pontius, Melody Roelke, Yves Rumpler, Maria Paula C. Schneider, Artur Silva, Stephen J. O’Brien, and Jill Pecon-Slattery (Mar. 2011). “A Molecular Phylogeny of Living Primates”. In: *PLOS Genetics* 7.3, e1001342. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001342.

Perez-Riverol, Yasset, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, Mathias Walzer, Shengbo Wang, Alvis Brazma, and Juan Antonio Vizcaíno (Jan. 2022). “The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences”. In: *Nucleic Acids Research* 50.D1, pp. D543–D552. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkab1038.

Petrzilek, Jan, Josef Pasulka, Radek Malik, Filip Horvat, Shubhangini Kataruka, Helena Fulka, and Petr Svoboda (Dec. 2022). “De Novo Emergence, Existence, and Demise of a Protein-Coding Gene in Murids”. In: *BMC Biology* 20.1, p. 272. ISSN: 1741-7007. DOI: 10.1186/s12915-022-01470-5.

Plenge, Robert M., Brian D. Hendrich, Charles Schwartz, J. Fernando Arena, Anna Naumova, Carmen Sapienza, Robin M. Winter, and Huntington F. Willard (Nov. 1997). “A Promoter Mutation in the XIST Gene in Two Unrelated Families with Skewed X-chromosome Inactivation”. In: *Nature Genetics* 17.3, pp. 353–356. ISSN: 1546-1718. DOI: 10.1038/ng1197-353.

Ponting, Chris P., Peter L. Oliver, and Wolf Reik (Feb. 2009). “Evolution and Functions of Long Noncoding RNAs”. In: *Cell* 136.4, pp. 629–641. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2009.02.006.

Prensner, John R., Jennifer G. Abelin, Leron W. Kok, Karl R. Clauser, Jonathan M. Mudge, Jorge Ruiz-Orera, Michal Bassani-Sternberg, Eric W. Deutsch, and Sebastiaan van Heesch (May 2023). *What Can Ribo-seq and Proteomics Tell Us about the Non-Canonical Proteome?* DOI: 10.1101/2023.05.16.541049.

Prensner, John R., Oana M. Enache, Victor Luria, Karsten Krug, Karl R. Clauser, Joshua M. Dempster, Amir Karger, Li Wang, Karolina Stumbraite, Vickie M. Wang, Ginevra Botta, Nicholas J. Lyons, Amy Goodale, Zohra Kalani, Briana Fritchman, Adam Brown, Douglas Alan, Thomas Green, Xiaoping Yang, Jacob D. Jaffe, Jennifer A. Roth, Federica Piccioni, Marc W. Kirschner, Zhe Ji, David E. Root, and Todd R. Golub (June 2021). “Non-Canonical Open Reading Frames Encode Functional Proteins Essential for Cancer Cell Survival”. In: *Nature biotechnology* 39.6, pp. 697–704. ISSN: 1087-0156. DOI: 10.1038/s41587-020-00806-2.

Pueyo, Jose I., Emile G. Magny, and Juan P. Couso (Aug. 2016). “New Peptides Under the s(ORF)Ace of the Genome”. In: *Trends in Biochemical Sciences* 41.8, pp. 665–678. ISSN: 0968-0004. DOI: 10.1016/J.TIBS.2016.05.003.

Qian, Weibin, Xinrui Cai, Qiuhai Qian, Wei Peng, Jie Yu, Xinying Zhang, Li Tian, and Can Wang (Feb. 2019). “lncRNA ZEB1-AS1 Promotes Pulmonary Fibrosis through ZEB1-mediated Epithelial–Mesenchymal Transition by Competitively Binding miR-141-3p”. In: *Cell Death & Disease* 10.2, pp. 1–12. ISSN: 2041-4889. DOI: 10.1038/s41419-019-1339-1.

Quek, Xiu Cheng, Daniel W Thomson, Jesper L V Maag, Nenad Bartonicek, Bethany Signal, Michael B Clark, Brian S Gloss, and Marcel E Dinger (Jan. 2015). “lncRNAdb v2.0: Expanding the Reference Database for Functional Long Noncoding RNAs.” In: *Nucleic acids research* 43.Database issue, pp. D168–73. ISSN: 1362-4962. DOI: 10.1093/nar/gku988.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Foundation for Statistical Computing, Vienna, Austria.

Raj, Anil, Sidney H Wang, Heejung Shim, Arbel Harpak, Yang I Li, Brett Engelmann, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard (May 2016). “Thousands of Novel Translated Open Reading Frames in Humans Inferred by Ribosome Footprint Profiling”. In: *eLife* 5. Ed. by Nicholas T Ingolia, e13328. ISSN: 2050-084X. DOI: 10.7554/eLife.13328.

Reinhardt, Josephine A, Betty M Wanjiru, Alicia T Brant, Perot Saelao, David J Begun, and Corbin D Jones (2013). “De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-Coding Sequences.” In: *PLoS genetics* 9.10, e1003860. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003860.

Reinhart, Brenda J., Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, H. Robert Horvitz, and Gary Ruvkun (Feb. 2000). “The 21-Nucleotide Let-7 RNA Regulates Developmental Timing in Caenorhabditis Elegans”. In: *Nature* 403.6772, pp. 901–906. ISSN: 1476-4687. DOI: 10.1038/35002607.

Renz, Peter F., Fabiola Valdivia-Francia, and Ataman Sendoel (Nov. 2020). “Some like It Translated: Small ORFs in the 5′UTR”. In: *Experimental Cell Research* 396.1, p. 112229. ISSN: 0014-4827. DOI: 10.1016/j.yexcr.2020.112229.

Ridings-Figueroa, Rebeca, Emma R. Stewart, Tatyana B. Nesterova, Heather Coker, Greta Pintacuda, Jonathan Godwin, Rose Wilson, Aidan Haslam, Fred Lilley, Renate Ruigrok, Sumia A. Bageghni, Ghadeer Albadrani, William Mansfield, Jo-An Roulson, Neil Brockdorff, Justin F. X. Ainscough, and Dawn Coverley (Jan. 2017). “The Nuclear Matrix Protein CIZ1 Facilitates Localization of Xist RNA to the Inactive X-chromosome Territory”. In: *Genes & Development* 31.9, pp. 876–888. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.295907.117.

Rinn, John L., Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha A. Brugmann, L. Henry Goodnough, Jill A. Helms, Peggy J. Farnham, Eran Segal, and Howard Y. Chang (June 2007). “Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs”. In: *Cell* 129.7, pp. 1311–1323. ISSN: 00928674. DOI: 10.1016/j.cell.2007.05.022.

Roberts, Adam and Lior Pachter (Jan. 2013). “Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments”. In: *Nature Methods* 10.1, pp. 71–73. ISSN: 1548-7105. DOI: 10.1038/nmeth.2251.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (Jan. 2010). “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data”. In: *Bioinformatics* 26.1, pp. 139–140. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616.

Robinson, Mark D and Alicia Oshlack (Mar. 2010). “A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data”. In: *Genome Biology* 11.3, R25. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-3-r25.

Rogozin, Igor B (2014). “Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing.” In: *Genetics research international* 2014, p. 516508. ISSN: 2090-3154. DOI: 10.1155/2014/516508.

Ross, R. A., B. A. Spengler, and J. L. Biedler (Oct. 1983). “Coordinate Morphological and Biochemical Interconversion of Human Neuroblastoma Cells”. In: *Journal of the National Cancer Institute* 71.4, pp. 741–747. ISSN: 0027-8874.

Roy, Siddhartha and Tapas K. Kundu (Jan. 2021). “V - Chemical Principles of DNA Sequence Recognition and Gene Regulation”. In: *Chemical Biology of the Genome*. Ed. by Siddhartha Roy and Tapas K. Kundu. Academic Press, pp. 171–223. ISBN: 978-0-12-817644-3. DOI: 10.1016/B978-0-12-817644-3.00005-2.

Ruiz-Orera, Jorge and M Mar Albà (July 2019). “Conserved Regions in Long Non-Coding RNAs Contain Abundant Translation and Protein–RNA Interaction Signatures”. In: *NAR Genomics and Bioinformatics* 1.1, e2. ISSN: 2631-9268. DOI: 10.1093/nargab/lqz002.

– (Dec. 2018). “Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation.” In: 0.0. ISSN: 0168-9525.

Ruiz-Orera, Jorge, Xavier Messeguer, Juan Antonio Subirana, and M Mar Alba (Sept. 2014). “Long Non-Coding RNAs as a Source of New Peptides”. In: *eLife* 3, e03523. ISSN: 2050-084X. DOI: 10.7554/eLife.03523.

Sahakyan, Anna, Yihao Yang, and Kathrin Plath (Dec. 2018). “The Role of Xist in X-Chromosome Dosage Compensation.” In: *Trends in cell biology* 28.12, pp. 999–1013. ISSN: 1879-3088. DOI: 10.1016/j.tcb.2018.05.005.

Sandmann, Clara-L., Jana F. Schulz, Jorge Ruiz-Orera, Marieluise Kirchner, Matthias Ziehm, Eleonora Adami, Maike Marczenke, Annabel Christ, Nina Liebe, Johannes Greiner, Aaron Schoenenberger, Michael B. Muecke, Ning Liang, Robert L. Moritz, Zhi Sun, Eric W. Deutsch, Michael Gotthardt, Jonathan M. Mudge, John R. Prensner, Thomas E. Willnow, Philipp Mertins, Sebastiaan van Heesch, and Norbert Hubner (Feb. 2023). “Evolutionary Origins and Interactomes of Human, Young Microproteins and Small Peptides Translated from Short Open Reading Frames”. In: *Molecular Cell*. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2023.01.023.

Santa, Francesca De, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, and Gioacchino Natoli (May 2010). “A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers”. In: *PLOS Biology* 8.5, e1000384. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000384.

Schmidt, E.E. and U. Schibler (1995). “High Accumulation of Components of the RNA Polymerase II Transcription Machinery in Rodent Spermatids”. In: *Development* 121.8.

Schmitz, Sandra U., Phillip Grote, and Bernhard G. Herrmann (July 2016). “Mechanisms of Long Noncoding RNA Function in Development and Disease”. In: *Cellular and Molecular Life Sciences* 73.13, pp. 2491–2509. ISSN: 1420-682X. DOI: 10.1007/s00018-016-2174-5.

Schoch, Conrad L., Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O’Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi (Jan. 2020). “NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools”. In: *Database: The Journal of Biological Databases and Curation* 2020, baaa062. ISSN: 1758-0463. DOI: 10.1093/database/baaa062.

Schorderet, Patrick and Denis Duboule (May 2011). “Structural and Functional Differences in the Long Non-Coding RNA Hotair in Mouse and Human”. In: *PLoS Genetics* 7.5, e1002071. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1002071.

Schuller, Anthony P. and Rachel Green (Aug. 2018). “Roadblocks and Resolutions in Eukaryotic Translation”. In: *Nature reviews. Molecular cell biology* 19.8, pp. 526–541. ISSN: 1471-0072. DOI: 10.1038/s41580-018-0011-4.

Schuster-Böckler, Benjamin, Jörg Schultz, and Sven Rahmann (Jan. 2004). “HMM Logos for Visualization of Protein Families”. In: *BMC Bioinformatics* 5.1, p. 7. ISSN: 1471-2105. DOI: 10.1186/1471-2105-5-7.

Seidl, Christine IM, Stefan H Stricker, and Denise P Barlow (Aug. 2006). “The Imprinted Air ncRNA Is an Atypical RNAPII Transcript That Evades Splicing and Escapes Nuclear Export”. In: *The EMBO Journal* 25.15, pp. 3565–3575. ISSN: 0261-4189. DOI: 10.1038/sj.emboj.7601245.

Sela, Itamar, Haim Ashkenazy, Kazutaka Katoh, and Tal Pupko (July 2015). “GUIDANCE2: Accurate Detection of Unreliable Alignment Regions Accounting for the Uncertainty of Multiple Parameters”. In: *Nucleic Acids Research* 43.W1, W7–W14. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv318.

Senís, Elena, Miriam Esgleas, Sonia Najas, Verónica Jiménez-Sábado, Camilla Bertani, Marta Giménez-Alejandre, Alba Escriche, Jorge Ruiz-Orera, Marta Hergueta-Redondo, Mireia Jiménez, Albert Giralt, Paolo Nuciforo, M. Mar Albà, Héctor Peinado, Daniel del Toro, Leif Hove-Madsen, Magdalena Götz, and María Abad (Dec. 2021). "TUNAR lncRNA Encodes a Microprotein That Regulates Neural Differentiation and Neurite Formation by Modulating Calcium Dynamics". In: *Frontiers in Cell and Developmental Biology* 9, p. 747667. ISSN: 2296-634X. DOI: 10.3389/fcell.2021.747667.

Sheng, Liang, Lan Ye, Dong Zhang, William P. Cawthorn, and Bin Xu (Sept. 2018). "New Insights Into the Long Non-coding RNA SRA: Physiological Functions and Mechanisms of Action". In: *Frontiers in Medicine* 5, p. 244. ISSN: 2296-858X. DOI: 10.3389/fmed.2018.00244.

Shukla, Chinmay J, Alexandra L McCorkindale, Chiara Gerhardinger, Keegan D Korthauer, Moran N Cabili, David M Shechner, Rafael A Irizarry, Philipp G Maass, and John L Rinn (Mar. 2018). "High-throughput Identification of RNA Nuclear Enrichment Sequences". In: *The EMBO Journal* 37.6, e98452. ISSN: 0261-4189. DOI: 10.15252/embj.201798452.

Silvera, Deborah, Rezina Arju, Farbod Darvishian, Paul H. Levine, Ladan Zolfaghari, Judith Goldberg, Tsivia Hochman, Silvia C. Formenti, and Robert J. Schneider (July 2009). "Essential Role for eIF4GI Overexpression in the Pathogenesis of Inflammatory Breast Cancer". In: *Nature Cell Biology* 11.7, pp. 903–908. ISSN: 1465-7392. DOI: 10.1038/ncb1900.

Slavoff, Sarah A., Andrew J. Mitchell, Adam G. Schwaid, Moran N. Cabili, Jiao Ma, Joshua Z. Levin, Amir D. Karger, Bogdan A. Budnik, John L. Rinn, and Alan Saghatelian (Jan. 2013). "Peptidomic Discovery of Short Open Reading Frame–Encoded Peptides in Human Cells". In: *Nature Chemical Biology* 9.1, pp. 59–64. ISSN: 1552-4469. DOI: 10.1038/nchembio.1120.

Sleutels, Frank, Ronald Zwart, and Denise P. Barlow (Feb. 2002). "The Non-Coding Air RNA Is Required for Silencing Autosomal Imprinted Genes". In: *Nature* 415.6873, pp. 810–813. ISSN: 1476-4687. DOI: 10.1038/415810a.

Spiroski, Ana-Mishel, Rachel Sanders, Marco Meloni, Ian R. McCracken, Adrian Thomson, Mairi Brittan, Gillian A. Gray, and Andrew H. Baker (Jan. 2021). "The Influence of the LINC00961/SPAAR Locus Loss on Murine Development, Myocardial Dynamics, and Cardiac Response to Myocardial Infarction". In: *International Journal of Molecular Sciences* 22.2, p. 969. ISSN: 1422-0067. DOI: 10.3390/ijms22020969.

Statello, Luisa, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte (Feb. 2021). "Gene Regulation by Long Non-Coding RNAs and Its Biological Functions". In: *Nature Reviews Molecular Cell Biology* 22.2, pp. 96–118. ISSN: 1471-0080. DOI: 10.1038/s41580-020-00315-9.

Stein, Colleen S., Pooja Jadiya, Xiaoming Zhang, Jared M. McLendon, Gabrielle M. Abouassaly, Nathan H. Witmer, Ethan J. Anderson, John W. Elrod, and Ryan L. Boudreau (June 2018). "Mitoregulin: A lncRNA-Encoded Microprotein That Supports Mitochondrial Supercomplexes and Respiratory Efficiency". In: *Cell Reports* 23.13, 3710–3720.e8. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2018.06.002.

Su, Wenjing, Miao Xu, Xueqin Chen, Ni Chen, Jing Gong, Ling Nie, Ling Li, Xinglan Li, Mengni Zhang, and Qiao Zhou (Aug. 2017). "Long Noncoding RNA ZEB1-AS1 Epigenetically Regulates the Expressions of ZEB1 and Downstream Molecules in Prostate Cancer". In: *Molecular Cancer* 16.1, p. 142. ISSN: 1476-4598. DOI: 10.1186/s12943-017-0711-y.

Subtelny, Alexander O., Stephen W. Eichhorn, Grace R. Chen, Hazel Sive, and David P. Bartel (Apr. 2014). "Poly(A)-Tail Profiling Reveals an Embryonic Switch in Translational Control". In: *Nature* 508.7494, pp. 66–71. ISSN: 0028-0836. DOI: 10.1038/nature13007.

Tao, Yunlong and Su-Chun Zhang (Nov. 2016). "Neural Subtype Specification From Human Pluripotent Stem Cells". In: *Cell stem cell* 19.5, pp. 573–586. ISSN: 1934-5909. DOI: 10.1016/j.stem.2016.10.015.

The UniProt Consortium (Jan. 2021). "UniProt: The Universal Protein Knowledgebase in 2021". In: *Nucleic Acids Research* 49.D1, pp. D480–D489. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1100.

Thompson, Julie D, Frédéric Plewniak, Raymond Ripp, Jean-Claude Thierry, and Olivier Poch (Dec. 2001). “Towards a Reliable Objective Function for Multiple Sequence alignments11Edited by J. Karn”. In: *Journal of Molecular Biology* 314.4, pp. 937–951. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.5187.

Toll-Riera, Macarena, Nina Bosch, Nicolás Bellora, Robert Castelo, Lluis Armengol, Xavier Estivill, and M. Mar Albà (Mar. 2009). “Origin of Primate Orphan Genes: A Comparative Genomics Approach”. In: *Molecular Biology and Evolution* 26.3, pp. 603–612. ISSN: 0737-4038. DOI: 10.1093/molbev/msn281.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter (Mar. 2012). “Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks”. In: *Nature Protocols* 7.3, pp. 562–578. ISSN: 1750-2799. DOI: 10.1038/nprot.2012.016.

Tripathi, Vidisha, Jonathan D. Ellis, Zhen Shen, David Y. Song, Qun Pan, Andrew T. Watt, Susan M. Freier, C. Frank Bennett, Alok Sharma, Paula A. Bubulya, Benjamin J. Blencowe, Supriya G. Prasanth, and Kannanganattu V. Prasanth (Sept. 2010). “The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation”. In: *Molecular Cell* 39.6, pp. 925–938. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2010.08.011.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén (Jan. 2015). “Tissue-Based Map of the Human Proteome”. In: *Science* 347.6220, p. 1260419. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.1260419.

Ulitsky, Igor (Oct. 2016). “Evolution to the Rescue: Using Comparative Genomics to Understand Long Non-Coding RNAs”. In: *Nature Reviews Genetics* 17.10, pp. 601–614. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.85.

Ulitsky, Igor and David P Bartel (July 2013). “lincRNAs: Genomics, Evolution, and Mechanisms.” In: *Cell* 154.1, pp. 26–46. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.06.020.

Ulitsky, Igor, Alena Shkumatava, Calvin H. Jan, Hazel Sive, and David P. Bartel (Dec. 2011). “Conserved Function of lincRNAs in Vertebrate Embryonic Development Despite Rapid Sequence Evolution”. In: *Cell* 147.7, pp. 1537–1550. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.11.055.

Ulveling, Damien, Claire Francastel, and Florent Hubé (Apr. 2011). “When One Is Better than Two: RNA with Dual Functions”. In: *Biochimie* 93.4, pp. 633–644. ISSN: 0300-9084. DOI: 10.1016/J.BIOCHI.2010.11.004.

Uszczynska-Ratajczak, Barbara, Julien Lagarde, Adam Frankish, Roderic Guigó, and Rory Johnson (Sept. 2018). “Towards a Complete Map of the Human Long Non-Coding RNA Transcriptome”. In: *Nature Reviews Genetics* 19.9, pp. 535–548. ISSN: 1471-0056. DOI: 10.1038/s41576-018-0017-y.

Vakirlis, Nikolaos, Omer Acar, Brian Hsu, Nelson Castilho Coelho, S. Branden Van Oss, Aaron Wacholder, Kate Medetgul-Ernar, Ray W. Bowman, Cameron P. Hines, John Iannotta, Saurin Bipin Parikh, Aoife McLysaght, Carlos J. Camacho, Allyson F. O’Donnell, Trey Ideker, and Anne-Ruxandra Carvunis (Feb. 2020). “De Novo Emergence of Adaptive Membrane Proteins from Thymine-Rich Genomic Sequences”. In: *Nature Communications* 11, p. 781. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14500-z.

Vakirlis, Nikolaos and Aoife McLysaght (2019). “Computational Prediction of De Novo Emerged Protein-Coding Genes”. In: *Computational Methods in Protein Evolution*. Ed. by Tobias Sikosek. Methods in Molecular Biology. New York, NY: Springer, pp. 63–81. ISBN: 978-1-4939-8736-8. DOI: 10.1007/978-1-4939-8736-8_4.

Vakirlis, Nikolaos, Zoe Vance, Kate M. Duggan, and Aoife McLysaght (Dec. 2022). “De Novo Birth of Functional Microproteins in the Human Lineage”. In: *Cell Reports* 41.12, p. 111808. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2022.111808.

van Heesch, Sebastiaan, Maarten van Iterson, Jetse Jacobi, Sander Boymans, Paul B. Essers, Ewart de Bruijn, Wensi Hao, Alyson W. MacInnes, Edwin Cuppen, and Marieke Simonis (Jan. 2014). “Extensive Localization of Long Noncoding RNAs to the Cytosol and Mono- and Polyribosomal Complexes”. In: *Genome Biology* 15.1, R6. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-1-r6.

van Heesch, Sebastiaan, Franziska Witte, Valentin Schneider-Lunitz, Jana F. Schulz, Eleonora Adami, Allison B. Faber, Marieluise Kirchner, Henrike Maatz, Susanne Blachut, Clara-Louisa Sandmann, Masatoshi Kanda, Catherine L. Worth, Sebastian Schafer, Lorenzo Calviello, Rhys Merriott, Giannino Patone, Oliver Hummel, Emanuel Wyler, Benedikt Obermayer, Michael B. Mücke, Eric L. Lindberg, Franziska Trnka, Sebastian Memczak, Marcel Schilling, Leanne E. Felkin, Paul J. R. Barton, Nicholas M. Quaife, Konstantinos Vanezis, Sebastian Diecke, Masaya Mukai, Nancy Mah, Su-Jun Oh, Andreas Kurtz, Christoph Schramm, Dorothee Schwinge, Marcial Sebode, Magdalena Harakalova, Folkert W. Asselbergs, Aryan Vink, Roel A. de Weger, Sivakumar Viswanathan, Anissa A. Widjaja, Anna Gärtner-Rommel, Hendrik Milting, Cris Dos Remedios, Christoph Knosalla, Philipp Mertins, Markus Landthaler, Martin Vingron, Wolfgang A. Linke, Jonathan G. Seidman, Christine E. Seidman, Nikolaus Rajewsky, Uwe Ohler, Stuart A. Cook, and Norbert Hubner (June 2019). “The Translational Landscape of the Human Heart”. In: *Cell* 178.1, 242–260.e29. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.05.010.

Wacholder, Aaron, Saurin Bipin Parikh, Nelson Castilho Coelho, Omer Acar, Carly Houghton, Lin Chou, and Anne-Ruxandra Carvunis (May 2023). “A Vast Evolutionarily Transient Translatome Contributes to Phenotype and Fitness”. In: *Cell Systems* 14.5, 363–381.e8. ISSN: 2405-4712. DOI: 10.1016/j.cels.2023.04.002.

Wagner, Günter P., Koryu Kin, and Vincent J. Lynch (Dec. 2012). “Measurement of mRNA Abundance Using RNA-seq Data: RPKM Measure Is Inconsistent among Samples”. In: *Theory in Biosciences* 131.4, pp. 281–285. ISSN: 1611-7530. DOI: 10.1007/s12064-012-0162-3.

Wang, Jiayi, Xiangfan Liu, Huacheng Wu, Peihua Ni, Zhidong Gu, Yongxia Qiao, Ning Chen, Fenyong Sun, and Qishi Fan (Sept. 2010). “CREB Up-Regulates Long Non-Coding RNA, HULC Expression through Interaction with microRNA-372 in Liver Cancer.” In: *Nucleic acids research* 38.16, pp. 5366–83. ISSN: 1362-4962. DOI: 10.1093/nar/gkq285.

Wang, Lantian, Jing Fan, Lili Han, Haonan Qi, Yimin Wang, Hongye Wang, Suli Chen, Lei Du, Sheng Li, Yunbin Zhang, Wei Tang, Gaoxiang Ge, Weijun Pan, Ping Hu, and Hong Cheng (May 2020). “The Micropeptide LEMP Plays an Evolutionarily Conserved Role in Myogenesis”. In: *Cell Death & Disease* 11.5, pp. 1–12. ISSN: 2041-4889. DOI: 10.1038/s41419-020-2570-5.

Wang, Yue, Zhenyu Xu, Junfeng Jiang, Chen Xu, Jiuhong Kang, Lei Xiao, Minjuan Wu, Jun Xiong, Xiaocan Guo, and Houqi Liu (Apr. 2013). “Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal”. In: *Developmental Cell* 25.1, pp. 69–80. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2013.03.002.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton (May 2009). “Jalview Version 2–a Multiple Sequence Alignment Editor and Analysis Workbench”. In: *Bioinformatics* 25.9, pp. 1189–1191. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp033.

Weiss, R. B., D. M. Dunn, A. E. Dahlberg, J. F. Atkins, and R. F. Gesteland (May 1988). “Reading Frame Switch Caused by Base-Pair Formation between the 3′ End of 16S rRNA and the mRNA during Elongation of Protein Synthesis in Escherichia Coli.” In: *The EMBO Journal* 7.5, pp. 1503–1507. ISSN: 0261-4189. DOI: 10.1002/j.1460-2075.1988.tb02969.x.

Wen, Zheng-Yang, Yu-Jian Kang, Lan Ke, De-Chang Yang, and Ge Gao (May 2023). “Genome-Wide Identification of Gene Loss Events Suggests Loss Relics as a Potential Source of Functional lncRNAs in Humans”. In: *Molecular Biology and Evolution* 40.5, msad103. ISSN: 1537-1719. DOI: 10.1093/molbev/msad103.

Wickham, Hadley (2007). “Reshaping Data with the Reshape Package.” In: *Journal of Statistical Software*.

– (2016). *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (Nov. 2019). “Welcome to the Tidyverse”. In: *Journal of Open Source Software* 4.43, p. 1686. ISSN: 2475-9066. DOI: 10.21105/joss.01686.

Wilke, Claus O and Brenton M. Wiernik (2022). *Ggtext: Improved Text Rendering Support for ’Ggplot2’.*

Wilkins, David (2023). *Gggenes: Draw Gene Arrow Maps in ’Ggplot2’.*

Wilusz, Jeremy E., Courtney K. JnBaptiste, Laura Y. Lu, Claus-D. Kuhn, Leemor Joshua-Tor, and Phillip A. Sharp (Nov. 2012). “A Triple Helix Stabilizes the 3’ Ends of Long Noncoding RNAs That Lack Poly(A) Tails”. In: *Genes & Development* 26.21, pp. 2392–2407. ISSN: 1549-5477. DOI: 10.1101/gad.204438.112.

Wohlgemuth, Ingo, Sibylle Brenner, Malte Beringer, and Marina V. Rodnina (Nov. 2008). “Modulation of the Rate of Peptidyl Transfer on the Ribosome by the Nature of Substrates *”. In: *Journal of Biological Chemistry* 283.47, pp. 32229–32235. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M805316200.

Wu, Dong-Dong, David M. Irwin, and Ya-Ping Zhang (Nov. 2011). “De Novo Origin of Human Protein-Coding Genes”. In: *PLOS Genetics* 7.11, e1002379. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002379.

Xiao, Nan, Dong-Sheng Cao, Min-Feng Zhu, and Qing-Song Xu (June 2015). “Protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences”. In: *Bioinformatics* 31.11, pp. 1857–1859. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv042.

Xie, Chen, Yong E. Zhang, Jia-Yu Chen, Chu-Jun Liu, Wei-Zhen Zhou, Ying Li, Mao Zhang, Rongli Zhang, Liping Wei, and Chuan-Yun Li (Sept. 2012). “Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs”. In: *PLoS Genetics* 8.9. Ed. by David J. Begun, e1002942. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002942.

Yang, Sumin, Key-Hwan Lim, Sung-Hyun Kim, and Jae-Yeol Joo (Apr. 2021). “Molecular Landscape of Long Noncoding RNAs in Brain Disorders”. In: *Molecular Psychiatry* 26.4, pp. 1060–1074. ISSN: 1476-5578. DOI: 10.1038/s41380-020-00947-5.

Yin, Qing-Fei, Li Yang, Yang Zhang, Jian-Feng Xiang, Yue-Wei Wu, Gordon G. Carmichael, and Ling-Ling Chen (Oct. 2012). “Long Noncoding RNAs with snoRNA Ends”. In: *Molecular Cell* 48.2, pp. 219–230. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.07.033.

Young, Robert S., Ana C. Marques, Charlotte Tibbit, Wilfried Haerty, Andrew R. Bassett, Ji-Long Liu, and Chris P. Ponting (2012). “Identification and Properties of 1,119 Candidate lincRNA Loci in the Drosophila Melanogaster Genome”. In: 4.4. ISSN: 1759-6653. DOI: 10.1093/gbe/evs020.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam (2017). “Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data”. In: *Methods in Ecology and Evolution* 8.1, pp. 28–36. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12628.

Yutani, Hiroaki (2022). *Gghighlight: Highlight Lines and Points in ’Ggplot2’.*

Zeng, Tao, Haitao Ni, Yue Yu, Mingke Zhang, Minjuan Wu, Qiaoling Wang, Liujun Wang, Sha Xu, Zhenyu Xu, Chen Xu, Jun Xiong, Junfeng Jiang, Yan Luo, Yue Wang, and Houqi Liu (July 2019). “BACE1-AS Prevents BACE1 mRNA Degradation through the Sequestration of BACE1-targeting miRNAs". In: *Journal of Chemical Neuroanatomy* 98, pp. 87–96. ISSN: 0891-0618. DOI: 10.1016/J.JCHEMNEU.2019.04.001.

Zerbino, Daniel R, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek (Jan. 2018). "Ensembl 2018". In: *Nucleic Acids Research* 46.D1, pp. D754–D761. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1098.

Zhang, Peng, Dandan He, Yi Xu, Jiakai Hou, Bih-Fang Pan, Yunfei Wang, Tao Liu, Christel M. Davis, Erik A. Ehli, Lin Tan, Feng Zhou, Jian Hu, Yonghao Yu, Xi Chen, Tuan M. Nguyen, Jeffrey M. Rosen, David H. Hawke, Zhe Ji, and Yiwen Chen (Nov. 2017). "Genome-Wide Identification and Differential Analysis of Translational Initiation". In: *Nature Communications* 8, p. 1749. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01981-8.

Zhang, Qi-Shun, Zhao-Hui Wang, Jian-Lei Zhang, Yan-Li Duan, Guo-Fei Li, and Dong-Lin Zheng (Oct. 2016). "Beta-Asarone Protects against MPTP-induced Parkinson's Disease via Regulating Long Non-Coding RNA MALAT1 and Inhibiting α-Synuclein Protein Expression". In: *Biomedicine & Pharmacotherapy* 83, pp. 153–159. ISSN: 0753-3322. DOI: 10.1016/j.biopha.2016.06.017.

Zhang, Yong E., Patrick Landback, Maria D. Vibranovski, and Manyuan Long (Oct. 2011). "Accelerated Recruitment of New Brain Development Genes into the Human Genome". In: *PLOS Biology* 9.10, e1001179. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001179.

Zhao, Li, Perot Saelao, Corbin D Jones, and David J Begun (Feb. 2014). "Origin and Spread of de Novo Genes in Drosophila Melanogaster Populations." In: *Science (New York, N.Y.)* 343.6172, pp. 769–72. ISSN: 1095-9203. DOI: 10.1126/science.1248286.

Zhou, Qi, Guojie Zhang, Yue Zhang, Shiyu Xu, Ruoping Zhao, Zubing Zhan, Xin Li, Yun Ding, Shuang Yang, and Wen Wang (Sept. 2008). "On the Origin of New Genes in Drosophila." In: *Genome research* 18.9, pp. 1446–55. ISSN: 1088-9051. DOI: 10.1101/gr.076588.108.

Zhu, Ying, Mingfeng Li, André MM Sousa, and Nenad Šestan (2014). "XSAnno: A Framework for Building Ortholog Models in Cross-Species Transcriptome Comparisons". In: *BMC Genomics* 15.1. DOI: 10.1186/1471-2164-15-343.

Zhu, Ying, André M. M. Sousa, Tianliuyun Gao, Mario Skarica, Mingfeng Li, Gabriel Santpere, Paula Esteller-Cucala, David Juan, Luis Ferrández-Peral, Mo Yang, Daniel J. Miller, Tomas Marques-Bonet, Yuka Imamura Kawasawa, Hongyu Zhao, and Nenad Sestan (Dec. 2018). "Spatio-Temporal Transcriptomic Divergence across Human and Macaque Brain Development". In: *Science (New York, N.Y.)* 362.6420, eaat8077. ISSN: 0036-8075. DOI: 10.1126/science.aat8077.